Exploring the Benefits of Visual Prompting in Differential Privacy

Yizhe Li[1], Yu-Lin Tsai[2], Chia-Mu Yu[2], Pin-Yu Chen[3], and Xuebin Ren[*†1]

[1]School of Computer Science and Technology, Xi'an Jiaotong University
[2]National Yang Ming Chiao Tung University
[3]IBM Research

Abstract

Visual Prompting (VP) is an emerging and powerful technique that allows sample-efficient adaptation to downstream tasks by engineering a well-trained frozen source model. In this work, we explore the benefits of VP in constructing compelling neural network classifiers with differential privacy (DP). We explore and integrate VP into canonical DP training methods and demonstrate its simplicity and efficiency. In particular, we discover that VP in tandem with PATE, a state-of-the-art DP training method that leverages the knowledge transfer from an ensemble of teachers, achieves the state-of-the-art privacy-utility trade-off with minimum expenditure of privacy budget. Moreover, we conduct additional experiments on cross-domain image classification with a sufficient domain gap to further unveil the advantage of VP in DP. Lastly, we also conduct extensive ablation studies to validate the effectiveness and contribution of VP under DP consideration. Our code is available at `https://github.com/EzzzLi/Prompt-PATE`*.*

1. Introduction

Originating from the domain of deep learning for natural language processing, prompt engineering has gained significant popularity as an emergent technique for efficient adoption and adaptation of pre-trained language models for solving different downstream tasks [24]. In recent years, the notion of prompting has been extended to other domains and data modalities, especially in computer vision and images [18, 3]. Specifically, the term *visual prompting* (VP) has been coined by [3], and the authors show competitive accuracy of VP on some downstream image classification tasks over linear probing (i.e., attaching a trainable linear head to a pre-trained model) when used with a large vision model such as CLIP [35] (only the image encoder). It is worth noting that VP in [3] can be viewed as a special case of *model reprogramming* (MR) [8] on a pre-trained model. MR inserts an input transformation layer and an output mapping layer into a pre-trained frozen model for fine-tuning downstream tasks. MR is equivalent to VP in [3] when the input transformation is a trainable input perturbation and the output mapping is a specified source-target label correspondence or a set of text prompts for label inference (e.g., "a photo of [predicted class]"). Throughout this paper, for ease of elucidation, we will use VP and MR interchangeably.

VP has been extensively studied for various use cases, ranging from image classification [3], enhancing adversarial robustness [6], image-inpainting [4], cross-domain adaptation [39, 31], to name a few. In this paper, we explore yet another benefit of VP with pre-trained models – deep learning with differential privacy (DP). In deep learning, scaling the training parameters of a neural network often leads to improved task performance (e.g., a classification model with higher accuracy) [19]. However, with a DP budget, training a larger neural network usually means more consumption of data privacy [27]. Motivated by this dilemma of the tradeoff between neural network capacity and DP, we aim to study the following fundamental question:

Can VP with a pre-trained model (trained on non-private data) improve the privacy-accuracy tradeoff in off-the-shelf DP-training mechanisms?

In this paper, we give an affirmative answer to this question, validated through a comprehensive analysis and empirical comparisons. We purposely focus on existing DP-training mechanisms, in order to study the benefit of improved performance contributed by VP. Our proposed approach applies VP (at data inputs) to off-the-shelf DP-training mechanisms, together with a pre-trained model trained on non-private data. Particularly, when VP is used in PATE (Private Aggregation of Teacher Ensembles) [33], a DP training

[*]Corresponding Author.

[†]This work was supported in part by the National Natural Science Foundation of China under Grants 61802298 and 62172329.

mechanism, we show that the classification accuracy under a privacy constraint can achieve the current state-of-the-art performance (SOTA) (over 97%) on the common benchmark of CIFAR-10 classification task. Furthermore, we also demonstrate that the performance increases with minimum expenditure of privacy budget. Consequently, our results uncovered new benefits of VP in DP and offer new use cases and insights into prompt engineering.

Contribution. We highlight our main contributions as follows. We are the first to explore the benefits of VP with pre-trained models in the design of DP classifiers. By leveraging VP, we present Prom-PATE as a training strategy for DP classifiers. While sophisticated backbones are usually difficult to be used in DP training, Prom-PATE has great flexibility in utilizing the high accuracy of the backbone without compromising privacy. Overall, Prom-PATE enjoys the following characteristics. Prom-PATE relies on VP to resolve the demand for huge data from PATE, improving practicality and accuracy. In the design, the public pre-trained models are utilized *twice*, significantly growing the accuracy. Through extensive experiments, we demonstrate that Prom-PATE outperforms current DP classifiers on CIFAR-10, showing an accuracy 97.07% under a privacy budget of $\epsilon = 1.019$. We also show significant accuracy gain of Prom-PATE in other datasets over existing methods.

2. Related Work and Background

Visual Prompting (VP) and Model Reprogramming (MR). Both VP and MR focus on the problem setup of reusing a pre-trained model to perform a new task (either in-domain or cross-domain) without changing the model weights during fine-tuning (i.g., the pre-trained model is "frozen"). MR was first studied through the lens of adversarial machine learning (ML). Elsayed et al. [14] showed that an attacker can "steal" an ML model's computation resource to perform another task without the model owner's consent. Later on, MR was shown to deliver competitive image classification results in data-limited and cross-domain settings [39, 31], wherein the authors demonstrated the possibility of reusing a pre-trained model from a source domain (e.g., general image classifiers or language models) to solve challenging image classification problems in a target domain (e.g., bio-medical measurements). We refer the readers to the survey paper of MR in [8] for more details. VP through a trainable (padded) universal input perturbation is revisited in [3], and the authors showed competitive results on some subset of 12 image classification tasks over linear probing and full fine-tuning on pre-trained image classifiers and the CLIP model [35]. Chen et al. [7] improved VP by introducing iterative label mapping during training. Beyond image classification, VP was extended to image inpainting tasks [4]. In this paper, we note that we limit the scope of VP to input-level prompt engineering as studied in [3, 6], and we leave the broader notion of VP via injecting trainable token embeddings (e.g., the visual prompt tuning as in [18]) to different layers of a pre-trained model as future work.

Differentially Private Classifiers. One of the most widely used techniques to achieve DP deep learning is DPSGD [1], where DP noise is added to the clipped gradient updates during the training process. The definition and properties of DP are provided in the Supplementary Material. DPSGD suffers from information loss due to the fact that the gradient clipping and the noise scale are proportional to the norm of clipped gradient. Recent research [10, 30] finds that we may overestimate the privacy loss for DPSGD because the attacker does not have access to the gradient in each training iteration. One of the current trends in training a DP classifier is to privately fine-tune large pre-trained models such as BERT variants and GPT-2 [45, 44, 23]. This private fine-tuning strategy can also be applied to the realm of images [16, 27, 21, 37, 11, 5]. For example, Tramèr and Boneh [37] improved the model utility by conducting private fine-tuning with SimCLR features [9]. De et al. [11] also pre-trained the model with the public data. After that, they apply many techniques including large batch size and weight standardization to improve accuracy. Bu et al. [5]'s DP classifier relies on the notion of ghost clipping to calculate the clipped gradient required by DPSGD.

PATE [33, 34] is another approach that trains a DP classifier. In PATE, the sensitive dataset is first partitioned into slices, with each *teacher model* trained on a different slice of the data (through SGD). Then, the non-sensitive samples labeled by the DP noisy votes from teacher models are used to train a *student model*, which turns out to be a DP classifier. Compared to DPSGD, fewer research efforts are put into the improvement of PATE. For example, Private-kNN [46] relies on the private release of k-nearest neighbor (kNN) queries to avoid splitting the training set in PATE.

Visual Prompting with DP. A recent work that combines VP and DP is Reprogrammable-FL [2]. Reprogrammable-FL is designed for DP federated learning (FL). More specifically, Reprogrammable-FL considers multiple clients, each with a common pre-trained model in each server-client interaction. The aim is to learn privatized visual prompts and label mappings for each client using DPSGD [1], enabling DPFL with more efficient use of the privacy budget. Reprogrammable-FL outperforms methods that rely on private fine-tuning from pre-trained models, currently considered the standard for achieving high accuracy in DPFL. However, in each training round of Reprogrammable-FL, the update of visual prompts and label mapping for each client is still subject to clipped noisy gradient updates to ensure privacy. As a result, the overall performance may still degrade compared to the non-private setting of visual prompting [3], as will be demonstrated in this paper.

3. Main Approach

In this section, we aim to investigate how VP can improve the privacy-utility trade-off of deep learning models.

Notations. As VP was originally proposed for model reutilization, we denote a source model $f_S(\theta_S; x)$ which is trained from a large, source (pubic) dataset $D_S := \{(x_S, y_S)\}$ with x_S, where x_S denotes the feature and y_S denotes the label, both from the source domain. We denote our target (private) dataset $D_T := \{(x_T, y_T)\}$ with x_T in which we re-utilize model $f_S(\theta_S; x)$ to accomplish the task in D_T via VP without modifying the weights θ_S.

3.1. Design Challenges for DP Classifiers

Though PATE outperforms DPSGD because of the reduced noise scale and no information loss from the gradient clipping, we identify three challenges for designing DP classifiers based on PATE.

- **(C1)** The performance of PATE is sensitive to data partitioning. In particular, the teacher models may perform badly when the sensitive data is limited in size. As also shown in [46], each teacher model has an accuracy under 50% due to only 200 images for each partition, given 250 teacher models for CIFAR-10. One might leverage transfer learning (TL), as suggested in [27], to train teacher models in PATE. Specifically, this involves using a public pre-trained model and fine-tuning it on the private dataset. However, Table 1 shows that this TL-based method leads to inefficient performance in PATE[1].

- **(C2)** A current trend in training a high-accuracy classifier in a DP manner is to take advantage of either public labeled data or a public pre-trained model. For example, De et al. [11] pre-train the model with ImageNet (seen as a public dataset) and then fine-tune the model with CIFAR-10 (seen as a private dataset) through DPSGD. De et al. achieve the predicting accuracy 94.7% under $\epsilon = 1$. While many pieces of evidence show that properly exploiting public datasets and models may significantly improve accuracy, a natural question that arises is whether exploiting public datasets and models more times in the design of DP classifiers benefits accuracy.

- **(C3)** Privately training a model pre-trained on the public dataset is a promising solution for DP-classifiers. However, take ImageNet and CIFAR-10 as examples. They may share a similar distribution and so make the above training strategy doubtful in DP guarantee [38].

CIFAR-10	**Prom-PATE (ours)**	**TL-based method**
ϵ	1.019	1.021
Accuracy ± Std(%)	**97.07 $\pm$ 0.50**	76.93 $\pm$ 0.81

Table 1. Comparison of Prom-PATE and TL-based method.

3.2. Prom-PATE

Here, we present a new approach, Prom-PATE, which leverages VP and PATE for private learning. The workflow of Prom-PATE is shown in Figure 1. Prom-PATE is a simple yet effective approach to training a classifier in a DP manner. Basically, Prom-PATE follows all of the steps in PATE [34, 33] except that each teacher model in PATE is reprogrammed from a pre-trained source model to a *re-teacher model*. The structure of re-teacher model is also shown in Figure 1. Such simplicity of Prom-PATE also enjoys the direct inheritance of DP guarantee from PATE.

Prom-PATE Procedures. Prom-PATE consists of three steps: (a) training re-teacher models, (b) executing private aggregation, and (c) training a student model. Step (a) considers a public pre-trained model as a *source model* and trains visual prompting and label mapping on sensitive data. In particular, we are aimed to train only the prompting parameter ω while the pre-trained source model is always fixed. The prompting parameter ω (including trainable parameters ω_1 and ω_2 in Eq. (1) and Eq. (2), respectively) and collectively called re-teacher model (see Figure 1). We note that the re-teacher model is trained on the sensitive dataset through SGD, and hence does not fulfill DP. The next step contributes to the DP guarantee of Prom-PATE. Step (b) uses PATE to aggregate the predictions of the re-teacher models; i.e., when a sample is fed into re-teacher models, all of them have votes and use the DP noisy top-1 outcome as the label. In step (c), a student model is trained using semi-supervised learning with a pre-trained classifier. In particular, certain unlabeled public samples with labels from the DP noisy votes are used to train the student model, which serves as the resulting DP classifier. One can easily prove that Prom-PATE satisfies DP; the proof can be found in the Supplementary Materials.

Training re-teacher Models. During the training of each re-teacher model, we keep the source model fixed while conducting SGD to update only the label mappings and vi-

[1] The poor accuracy of the TL-based method can be attributed to the over-partitioning of the sensitive data. In such a case, data are insufficient for the training of each teacher model.

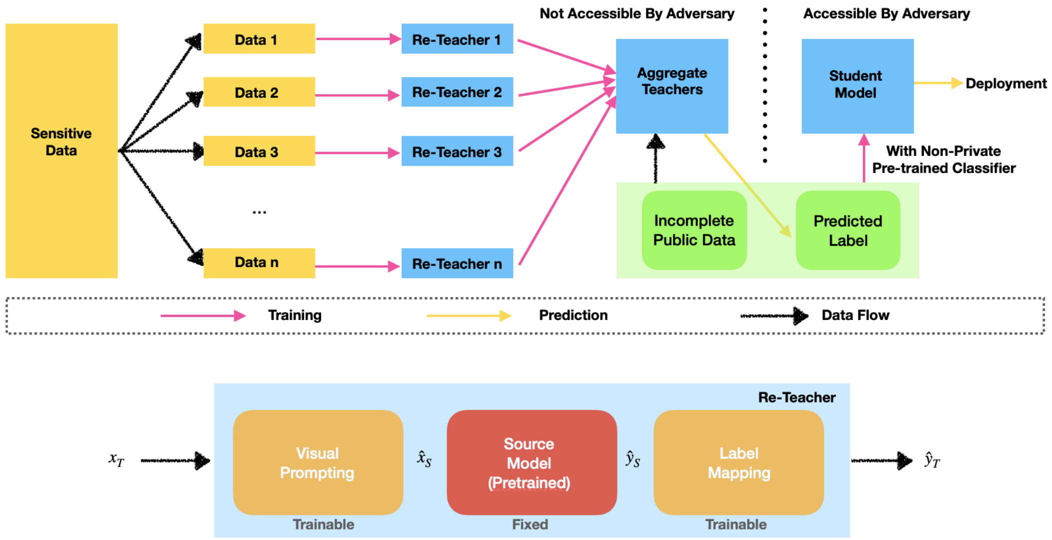

Figure 1. An overview of the proposed Prom-PATE framework.

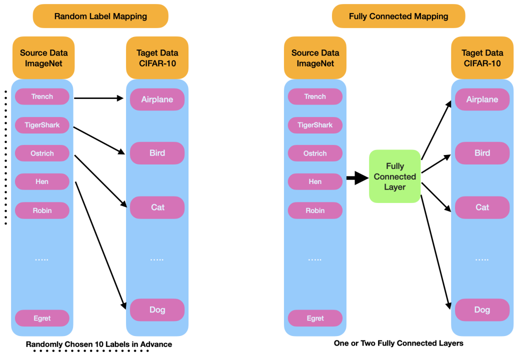

Figure 2. Illustration of different strategies for label mapping. *Left*: we follow the convention setting in VP [3] and apply randomly assigned label mapping that is pre-determined before training. *Right*: we simply apply a trainable fully-connected layer for the model to learn the appropriate mapping as proposed in [2]

sual prompts. The visual prompt $\hat{x}_S$ can be expressed as

$$\hat{x}_S = M \odot \omega_1 + (I - M) \odot \text{ZeroPad}(x_T), \qquad (1)$$

where $\odot$ stands for Hadamard product, ω_1 denotes the trainable noise parameter, and M denotes the binary mask of the same dimension with the source data x_S (i.e. $M \in \{0,1\}^{d_S}$, where d_S denotes the dimension of the source domain image). On the other hand, upon obtaining the pre-trained model output $\hat{y}_S := f_S(\theta_S; \hat{x}_S)$, we further render it through a label mapping function $f_\ell(\omega_2; \cdot)$ that maps the source labels to target labels and obtain the final prediction $\hat{y}_T$ which has the following form

$$\hat{y}_T = \text{softmax}(f_\ell(\omega_2; \hat{y}_S)). \qquad (2)$$

Algorithmic Details of Prom-PATE. Figure 2 illustrates different label mapping techniques used in Prom-PATE. To have a correspondence in label classes between the target and source domains, in the first approach, we conduct random label mapping [3, 39]. Particularly, before training, we establish a random mapping between the labels of two domains and train the model according to the predetermined label mapping (e.g., ImageNet label $i \rightarrow$ CIFAR-10 label j). In this case, ω_2 specifies the source-target label correspondence in VP. For the second approach, we consider using fully connected (FC) layers as part of the label mapping for greater expressiveness, as studied in [2]. This allows Prom-PATE to learn how to adapt labels from the source domain to the target domain. Overall, the re-teacher models in Prom-PATE only need to train the parameters $\omega := \{\omega_1, \omega_2\}$ on the private/sensitive dataset.

To enforce DP in Prom-PATE, we adopt the DP aggregation from PATE by considering Confident-GNMax [34, 33]. Specifically, given an unlabelled public data sample x, the aggregation mechanism would collect the response from every re-teacher model, establishing votes for each j-th class, $n_j(x)$. The aggregation then proceeds to determine whether the noisy votes are consent among re-teachers above a threshold T. Namely,

$$\max_j \{n_j(x)\} + \mathcal{N}(0, \sigma_1^2) \geq T. \qquad (3)$$

If the inequality is met, then the aggregation would proceed to offer noisy votes of re-teachers model as follows.

$$\arg\max_j \{n_j(x) + \mathcal{N}(0, \sigma_2^2)\}. \qquad (4)$$

Otherwise, the aggregation would output nothing.

To limit the privacy budget and further enhance performance in Prom-PATE, we use a subset of the public training data and label it using the private aggregation mechanism while conducting training for the rest of the training data in a semi-supervised fashion. Similar to PATE, this approach allows us to improve the privacy-utility trade-off by reducing the amount of data that needs to be labeled while still achieving high accuracy.

Since re-teacher models can adapt to the private domain under small sample complexity, we adopt the approach presented in [40] for our semi-supervised learning of the student model. We explain the details of this approach in Section 4.2, where we compare it to other baseline settings. Using this approach, we can achieve a better privacy-utility trade-off and improve overall performance of Prom-PATE.

3.3. Why Prom-PATE is Beneficial to DP?

This section provides an explanation as to why Prom-PATE, as a combination of VP and PATE, can attain an improved privacy-utility trade-off by overcoming the design challenges **(C1)**$\sim$**(C3)**.

- **(C1)** As mentioned in Section 3.1, though PATE is superior to DPSGD from the perspectives of noise scale and information loss, it can apply only to huge datasets because, otherwise, the teacher models fail to have decent accuracy, leading to poor student classifier accuracy. However, VP has proven to successfully transfer knowledge from large source domains to small target domains [39]. Thus, considering each partitioned slice of the sensitive dataset as a small target domain enables re-teacher models in Prom-PATE to avoid the problem of data insufficiency when increasing the number of re-teacher models, amplifying the benefits of ensemble learning in the ordinary PATE.
- **(C2)** Prom-PATE is featured by utilizing the public data *twice*; once in training re-teacher models and another one in training the student classifier. This can be attributed to our finding that PATE, in essence, can easily be modified to take advantage of pre-trained models (see the design of Prom-PATE in Section 3.2). Such efficient re-use of the public data can be highly beneficial to the resulting DP classifier accuracy, as shown in Table 2, where Prom-PATE, Prom-PATE w/o pre-trained classifier, and PATE means utilizing public data two, one, and zero times, respectively. Obviously, the accuracy grows with the increased number of times for utilizing public data.
- **(C3)** Due to cross-domain capability of VP/MR [39, 8, 42], even if the distribution of the dataset for the source model (used in training re-teacher models) is highly different from the distribution of the sensitive dataset, re-teacher models can still successfully attain high accuracies, which consequently improve the accuracy of the resulting DP classifier. Experiment evidence can be found in Section 4.4.

CIFAR-10	Prom-PATE	Prom-PATE w/o pre-trained classifier	PATE
ϵ	1.019	1.019	1.028
Accuracy ± Std(%)	**97.07 ± 0.50**	82.20 ± 1.14	32.53 ± 2.57

Table 2. Effect on the pre-trained classifier.

4. Experiments

In this section, we empirically evaluate the effectiveness of Prom-PATE on different datasets, with ImageNet serving as the public dataset for pre-training models. Additional experiments can be found in the Supplemental Materials.

4.1. Datasets and Implementation Details

We mainly use CIFAR-10 to benchmark image classification. However, we also report the results for CIFAR-100 in the Supplementary Material.

Cross Domain Dataset. To evaluate how Prom-PATE behaves in private domain adaptation with a large domain gap, we consider Blood-MNIST in our experiments. The Blood-MNIST dataset [43] contains images of blood cells sampled from uninfected patients, with an original shape of $3 \times 360 \times 363$. It contains 17,092 images of 8 different blood cells (11,959 of training and 3421 of testing) and has been processed to the size of $3 \times 28 \times 28$ [43]. We note that the sample distribution of Blood-MNIST is highly different from the sample distribution of ImageNet because the images in Blood-MNIST are taken under microscopic devices and planar in sight. Due to the large domain gap between Blood-MNIST and ImageNet, we consider Blood-MNIST in our experiments to resolve the concern of **(C3)**. Please see Section 4.4 for more details.

Implementation Details. All of the experiment results below are derived by averaging the results from three independent experiments. We use the official pre-trained models provided by PyTorch and set the parameters to default values for all pre-trained models. Regarding the training of each re-teacher model, since the source model is pre-trained on ImageNet, the visual prompt has a dimension of 224×224. When training the re-teacher model, we optimize the model with Adam whilst using a learning rate of 0.05 with a decay rate of 70%, batch size of 16, and training epoch of 10. In Section 4.7, we also investigate the effect of the binary mask M on visual prompt performance. For label mapping, we randomly select ten classes from the 1,000 source classes as a one-to-one mapping. We also use FC layers as the label mapping function in Section 4.8.

For the training of the student model, similar to the setting in PATE [34], in the case of CIFAR-10, the student has access to 9,000 samples that are partially labeled through the noisy aggregation mechanism (step (b) in Prom-PATE) in Section 3.2. The performance is evaluated on the remaining 1,000 samples in the testing set. Meanwhile, in the case of Blood-MNIST [43], the student has access to 2,421 samples that are as well partially labeled with privacy. The performance is evaluated on the remaining 1,000 samples in the testing set.

Privacy Parameter Setting. We use Rényi DP (RDP, see the definition in the Supplementary Materials) privacy accountant[2] to calculate the privacy budget ϵ. We adopt the $\delta \approx \frac{1}{n}$ convention and set $\delta = 10^{-5}$.

Evaluation Metrics. As the focus in this line of research mainly lies on image classification, we follow the convention and use the top-1 accuracy on CIFAR-10 as the metric.

4.2. Ablation Study of Prom-PATE

We conduct an ablation study on Prom-PATE for multiple baselines that can arise from our setting. In Prom-PATE, two key components for significant improvement of accuracy are re-teacher models and the use of a pre-trained classifier in student training. Thus, there are two dimensions for the ablation study: (i) VP-based re-teacher models, transfer learning-based teacher models, and train-from-scratch teacher models and (ii) using pre-trained or train-from-scratch classifiers in semi-supervised learning of the student model. Note that these pre-trained classifiers are all trained on ImageNet. The experiment results are shown in Table 3, where the setting A corresponds to Prom-PATE while the setting F corresponds to the ordinary PATE.

	Teacher	Student Training	ϵ	Accuracy $\pm$ Std(%)
A	VP-based re-teacher models	pre-trained	1.019	**97.07 $\pm$ 0.50**
B	VP-based re-teacher models	train-from-scratch	1.019	82.20 $\pm$ 1.14
C	transfer learning	pre-trained	1.021	96.10 $\pm$ 0.46
D	transfer learning	train-from-scratch	1.021	76.93 $\pm$ 0.81
E	train-from-scratch	pre-trained	1.028	49.00 $\pm$ 8.97
F	train-from-scratch	train-from-scratch	1.028	32.53 $\pm$ 2.57

Table 3. Ablation study of Prom-PATE.

From Table 3, we can observe that by comparing A with C and B with D, VP-based re-teacher models in Prom-PATE indeed hold an advantage over transfer learning-based teacher models when adapting the target domain of meager data, exceeding by a maximum of 5%. Secondly, suppose we compare A with B, C with D, and E with F, we can also see that utilizing a public pre-trained classifier in student training in Prom-PATE allows us to gain another performance improvement, ranging from 15% to 20 %. However, we particularly note that simply making use of a pre-trained classifier is not sufficient to have a great increase in accuracy, because the settings A and E, both containing a pre-trained classifier in the student training, have a difference of approximately 40% in terms of the predicting accuracy. The above results support the importance of re-teacher models in Prom-PATE. Lastly, we note that albeit holding a small difference against Prom-PATE and the transfer learning baseline, we note that under a sufficient domain gap, the re-teacher tends to perform much better at these diverse private domains. We refer the readers to Section 4.4 for more details.

	ϵ	sanitized ϵ	Accuracy on CIFAR-10
Arif et al. [2]	1.04	1.04	87.55%
Luo et al. [27]	1 1.5	1 1.5	76.64% 81.57%
Tramer et al. [37]	2	2	92.7%
Yu et al. [44]	1 2	1 2	94.3% 94.8%
De et al. [11]	1 2	1 2	94.7% 95.4%
Bu et al. [5]	1 2	1 2	96.7% 97.1%
Prom-PATE	1.019 1.505 1.943	1.209 1.670 2.250	**99.17%** **99.07%** **99.10%**

Table 4. Comparison between Prom-PATE and prior work.

4.3. Comparison with Existing DP Classifiers

We further compare Prom-PATE against the existing work including SOTA DP classifiers. Table 4 shows the comparison results, where the accuracies of the other methods are directly excerpted from the original papers except that Yu et al.'s experiment results are from [5]. Since Prom-PATE deploys a data-dependent bound in privacy calculation, we further follow [34] to sanitize our privacy budget using smooth sensitiy analysis, preventing data leakage. The smoothed budget is marked as *sanitized* ϵ in Table 4.

Table 4 shows that Prom-PATE achieves competitive performance over current existing works. In the low budget regime ($\epsilon \approx 1$), Prom-PATE outperforms all the other models and achieves the best accuracy of 99.17%. While the SOTA classification accuracy of CIFAR-10 (through ViT-H/14 [13]) in the non-private setting is 99.5%[3], Prom-PATE achieves a meaningful improvement in accuracy. The reason that Prom-PATE with $\epsilon = 1.019$ achieves 99.17% in Table 4 but achieves 97.07% in Tables 1$\sim$3 can be attributed to our choice of implementations. In particular, the pre-trained model for re-teachers, the pre-trained model for semi-supervised learning, and the algorithm for semi-supervised learning of Prom-PATE in Table 4 are Swin

[2] `https://github.com/tensorflow/privacy/tree/master/research/pate_2018`

[3] `https://paperswithcode.com/sota/image-classification-on-cifar-10` (last access: 2023/7)

Transformer [25], EVA [15], and FreeMatch [41], respectively, while those of Prom-PATE in Tables 1∼3 are Swin Transformer [25], ViT [13], and FixMatch [36]. In addition, unlike the other approaches [16, 27, 21, 37, 11, 5], Prom-PATE enjoys great flexibility in replacing source models (in re-teacher models) by the latest classifiers and up-to-date semi-supervised training method, so as to effortlessly improve the accuracy.

4.4. Cross-Domain Dataset Evaluation

We evaluate Prom-PATE under a cross-domain setting, where the re-teacher models with public pre-trained models are visually prompted toward a small private target domain. As mentioned in Section 4.1, we evaluate Prom-PATE on Blood-MNIST [43]. The experiment results are shown in Table 5, where Transfer-PATE is considered to use the same backbone source model of Prom-PATE and performs partial fine-tuning when training the teacher models.

Blood-MNIST	**Prom-PATE**	**Transfer-PATE**	**Arif et al. [2]**
ϵ	1.973	1.983	1.971
sanitized ϵ	2.521	2.508	1.971
Queries	1000	1000	-
Answered Queries	455	408	-
Answer Accuracy(%)	79.3	76.7	-
Threshold T	480	490	-
σ_1	150	150	-
σ_2	20	20	-
Accuracy(%)	**69.93**	61.33	63.45

Table 5. Effect on cross-domain datasets.

As one can see from Table 5, when adapting to a target domain with sufficient domain gap, Prom-PATE is able to manage the advantage of VP and maximize the accuracy gain given a fixed amount of privacy budget to vote for highly accurate labels that are beneficial for downstream student training, exceeding the Transfer-PATE by roughly 8%. On the other hand, Prom-PATE is also compared against Reprogrammable-FL [2], because the latter improves accuracy in the context of FL. Prom-PATE outperforms Reprogrammable-FL by approximately 2%. This can be attributed to much noisy perturbation of Reprogrammable-FL as stated in Section 2. Most importantly, due to the high discrepancy between ImageNet and Blood-MNIST, the high accuracy from such a train-on-ImageNet and test-on-Blood-MNIST setting also eliminates the suspicion **(C3)** from [38].

4.5. Numbers of Re-Teacher Models

In this section, we investigate the model performance under different numbers of re-teacher models. Table 6 reports the results, where Swin Transformer [25] is used as the source model for re-teacher models. As shown in Table 6, the best utility is achieved when using 1000 re-teacher models under a privacy budget of $\epsilon \approx 1$. We also note that the accuracy of all settings with 250, 500, and 1000 re-teacher models already exceed the performance of PATE [34] under a privacy budget of $\epsilon \approx 1$.

Number of re-teachers	**100**	**250**	**500**	**1000**
ϵ	1.095	1.095	1.04	1.019
Queries	1000	1000	1000	1000
Answered Queries	18	46	90	684
Threshold T	430	500	650	500
σ_1	150	150	150	200
σ_2	50	100	100	50
Accuracy%) ± Std	59.20 ± 0	85.87 ± 0.55	96.53 ± 0.74	**97.07 ± 0.50**

Table 6. Effect on different numbers of re-teacher models.

4.6. Different Pre-Trained Models

We study the effect of different pre-trained source models on Prom-PATE. Table 7 reports the results. In particular, using Swin Transformer [25] as the pre-trained source model results in the best performance of 99% on CIFAR-10. This is consistent with the theoretical relationship presented in [42], which states that the population risk on the target task of the reprogrammed model can be upper bounded by the source risk with an additional term in misalignment error. Therefore, as we can see from Table 8, which includes the accuracy of pre-trained models on the source domain (i.e., source risk), Swin Transformer has the least empirical risk and serves as a natural choice for the source model.

ImageNet	**Accuracy**
ResNet50	79.3
ResNet152	78.5
WideResNet	78.1
ViT	84.0
Swin Transformer	85.2

Table 8. Test accuracy of ImageNet source models.

4.7. Binary Mask in Visual Prompting

We further study how the different visual prompting techniques affect classification accuracy. Specifically, we consider two settings on whether to apply the binary mask M or not. Table 9 reports the results, where Swin transformer [25] as the source model with 1000 re-teacher models is considered.

Prompting Technique	**Without Mask M**	**With Mask M**
ϵ	1.017	1.019
Queries	1000	1000
Answered Queries	675	684
Answer Accuracy(%)	94.8	94.7
Threshold T	600	600
σ_1	200	200
σ_2	50	50
Accuracy ± Std(%)	96.53 ± 0.32	**97.07 ± 0.50**

Table 9. Effect on visual prompting technique

One can observe from Table 9 that using M could enhance performance. The rationale is that by utilizing M, we

	ϵ	Queries	Answered Queries	Answered Accuracy(%)	Threshold T	σ_1	σ_2	Accuracy ± Std(%)
ResNet50	1.081	1000	461	91.3	650	200	50	95.27 ± 0.80
ResNet152	1.009	1000	604	93.9	620	200	50	95.40 ± 0.40
WideResNet	1.068	1000	555	90.8	620	200	50	94.37 ± 0.25
ViT	1.007	1000	660	93.6	600	200	50	95.53 ± 0.51
Swin	1.019	1000	684	94.7	600	200	50	**97.07 ± 0.50**

Table 7. Effect on different pre-trained models.

can control the amount of noise placed in the visual prompt, hence controlling the ratio of target data x_T and noise parameter ω_1. This leads to a better trade-off between accuracy and the meager amount of private data each re-teacher model owns.

4.8. Label Mapping Techniques

Next, we proceed to investigate the effect of label mapping on Prom-PATE. Particularly, we consider the settings of using random label mapping (RLM), one fully-connected layer, and two fully-connected layers (see Figure 2). Table 10 shows the experiment results, where Swin transformer [25] as the source model with 1000 re-teacher models is considered. In particular, using one FC layer allows Prom-PATE to achieve the best performance. Furthermore, we note that randomly selecting ten classes for mapping would disrupt the behavior of the pre-trained model, as the mapping relations among source and target labels are randomly given but other remaining source classes might contain valuable information for the prediction. Such an explanation can be confirmed by the accuracy (i.e., noisy label accuracy) of RLM, which is only 22.9%, demonstrating that even with a high consensus of the re-teacher models, the ensemble prediction is likely to be wrong as well. On the other hand, while using two FC layers allows for more expressiveness, the number of training parameters is increased as well, leading to a slight degradation in accuracy with limited training data for each re-teacher model.

Mapping Technique	**RLM**	**1-Layer FC**	**2-Layer FC**
ϵ	1.042	1.019	1.026
Queries	1000	1000	1000
Answered Queries	109	684	336
Answer Accuracy(%)	22.9	94.7	92.6
Threshold T	650	600	670
σ_1	200	200	200
σ_2	50	50	50
Accuracy ± Std(%)	33.4 ± 0.66	**97.07 ± 0.50**	96.13 ± 0.41

Table 10. Effect on label mapping techniques.

4.9. Rescale Ratio in Visual Prompting

Usually, in VP/MR, the image from the target domain needs to be rescaled and surrounded by trainable noises, as shown in Eq. (1). The resulting $\hat{x}_S$ can then be fed into the source model. A higher rescale ratio generally leads to better performance. The rationale is that a higher rescale ratio provides more information from the target domain, which enables the re-teacher model to generate better visual prompts that can more effectively guide the source model in learning the relevant features of the target domain. However, a too-high rescale ratio could potentially result in overfitting to the target domain, leading to poor generalization performance. Hence, one strikes a balance between providing sufficient information from the target domain and avoiding overfitting. In our experiments, a rescale ratio of 0.6 achieves the best performance.

Rescale Size	ϵ	AQ	AA(%)	T	σ_1	σ_2	Accuracy ± Std(%)
64 × 64	1.028	408	86.3	650	200	50	93.03 ± 1.0
128 × 128	1.016	662	92.6	610	200	50	95.83 ± 0.1
160 × 160	1.016	655	93.7	610	200	50	95.07 ± 0.3
192 × 192	1.019	684	94.7	600	200	50	**97.07 ± 0.5**
210 × 210	1.016	655	93.7	610	200	50	95.30 ± 0.5

Table 11. Effect on the rescale ratio of target Data. The number of queries is 1,000. AQ, AA, and T denote answered queries, answered accuracy (%), and threshold, respectively.

As observed from Table 11, rescaling x_T to 192×192 for visual prompting achieved the highest utility. As explained in Section 4.7, the rescale size provides a ratio between the trainable parameter ω_1 and target data x_T. Too many noise parameters and a small target image might degrade performance due to the quality of the target image and insufficient data. Conversely, a larger target image and fewer parameters of ω_1 might cause sub-optimal input transformation from target to source, leading to a poor prompt.

5. Conclusion

In this paper, we conducted a comprehensive study and discovered a new benefit of VP in DP. In particular, we propose Prom-PATE, a new VP-empowered training method for constructing DP classifiers. Prom-PATE leverages VP to assist in the adaptation of pre-trained models in a more efficient way without losing privacy. Empirical evaluations show that Prom-PATE provides SOTA performance compared to several baselines and existing works. We also find that Prom-PATE achieves an even better accuracy gain when the target task has a sufficient domain gap against the pre-trained model (i.e., the ImageNet to Blood-MNIST setting), demonstrating the generality of Prom-PATE. Our findings suggest that VP is a promising approach to facilitating further research in building DP classifiers that improve or even extinguish the privacy-utility trade-off.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security (CCS)*, 2016.

[2] Huzaifa Arif, Alex Gittens, and Pin-Yu Chen. Reprogrammable-fl: Improving utility-privacy tradeoff in federated learning via model reprogramming. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022.

[4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A Efros. Visual prompting via image inpainting. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[5] Zhiqi Bu, Jialin Mao, and Shiyun Xu. Scalable and efficient training of large convolutional neural networks with differential privacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Aochuan Chen, Peter Lorenz, Yuguang Yao, Pin-Yu Chen, and Sijia Liu. Visual prompting for adversarial robustness. *arXiv preprint arXiv:2210.06284*, 2022.

[7] Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, and Sijia Liu. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023.

[8] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. *arXiv preprint arXiv:2202.10629*, 2022.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[10] R. Chourasia, Jiayuan Ye, and R. Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[11] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[12] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[14] Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. In *International Conference on Learning Representations*, 2019.

[15] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[16] Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022.

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022.

[19] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[21] Helena Klause, Alexander Ziller, Daniel Rueckert, Kerstin Hammernik, and Georgios Kaissis. Differentially private training of residual networks with scale normalisation. In *Theory and Practice of Differential Privacy (TPDP)*, 2022.

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

[23] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022.

[24] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[27] Zelun Luo, Daniel J Wu, Ehsan Adeli, and Li Fei-Fei. Scalable differential privacy with sparse network finetuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5059–5068, 2021.

[28] Ilya Mironov. Rényi differential privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

[29] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[30] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 2021.

[31] Paarth Neekhara, Shehzeen Hussain, Jinglong Du, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Cross-modal adversarial reprogramming. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2427–2435, 2022.

[32] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[33] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations (ICLR)*, 2017.

[34] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations (ICLR)*, 2018.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[37] Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations (ICLR)*, 2021.

[38] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *arXiv:2212.06470*, 2022.

[39] Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning (ICML)*, 2020.

[40] Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[41] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

[42] Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. Voice2series: Reprogramming acoustic models for time series classification. In *International Conference on Machine Learning*, pages 11808–11819. PMLR, 2021.

[43] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[44] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private finetuning of language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[45] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning (ICML)*, 2021.

[46] Yuqing Zhu, Xiang Yu, Manmohan Chandraker, and Yu-Xiang Wang. Private-knn: Practical differential privacy for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11854–11862, 2020.