In-distribution Public Data Synthesis with Diffusion Models for Differentially Private Image Classification

Jinseong Park[1†] Yujin Choi[1†] Jaewook Lee[1*]
[1]Seoul National University
{jinseong, uznhigh, jaewook} @snu.ac.kr

Abstract

To alleviate the utility degradation of deep learning image classification with differential privacy (DP), employing extra public data or pre-trained models has been widely explored. Recently, the use of in-distribution public data has been investigated, where tiny subsets of datasets are released publicly. In this paper, we investigate a framework that leverages recent diffusion models to amplify the information of public data. Subsequently, we identify data diversity and generalization gap between public and private data as critical factors addressing the limited public data. While assuming 4% of training data as public, our method achieves 85.48% on CIFAR-10 with a privacy budget of $\varepsilon = 2$, without employing extra public data for training.

1. Introduction

Differential privacy (DP) [18, 19] establishes a mathematical framework to ensure the privacy of training data. In deep learning, differentially private SGD (DP-SGD) [1] has become the de facto standard method to guarantee the models' privacy. However, differentially private training inevitably degrades performance compared to standard (non-DP) training [6, 45]. As a practical solution, leveraging *public pre-trained models* or *public data* has been explored to enhance utility [12, 39, 59, 65–67]. As using public data raises no privacy concerns, fine-tuning pre-trained models on private data can make use of learned features for free. Nevertheless, their effectiveness might be diminished when addressing *out-of-distribution (OOD) public data* of small shared characteristics with private data [59].

As an alternative, researchers have investigated the use of *in-distribution (ID) public data*, indicating that a small portion of in-distribution data is made public [38]. For example, some data owners decide to share their data publicly in exchange for economic incentives or public interest. This setup allows us to leverage public data with a distribution similar to private data. In contrast to previous approaches, such as utilizing extra OOD data or distribution shift, we need to assume a limited amount of ID public data since the public data might contain more sensitive information. However, the repeated use of limited-sized public data can raise critical issues related to memorization and overfitting [43]. To mitigate the aforementioned problems, previous studies mainly have focused on utilizing side information from the little public data to enhance the optimization of differentially private deep learning [2, 4, 38, 43].

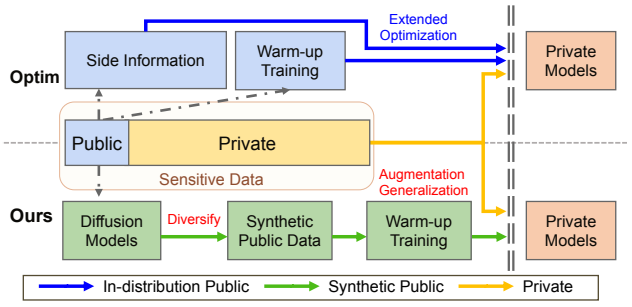


Figure 1. Illustration of (Top) existing optimization-based methods and (Bottom) proposed training procedure for amplifying in-distribution public data and utilizing the synthetic data.

Recently, Ganesh et al. [21] pointed out that warm-up training with public data plays a crucial in private optimization to find a good basin with a small loss, whereas DP-SGD from a random initial point faces a high loss and poor optimization. From this viewpoint, we suggest that enlarging the public dataset could improve the performance of private learning. By using the recent diffusion models [28, 30], we aim to enrich the contextual information of public data. We further identify that data diversity and generalization gap between public and private data act as critical factors, particularly addressing the limited size of public data.

The proposed framework is illustrated in Figure 1. [1] We summarize our approaches and their performance gains in

*Corresponding author. †Equal contribution.

[1]Our code is available at https://github.com/JinseongP/DPTrainer.

Table 1. Ablation study on the impact of various techniques, including synthesis, augmentation, and optimization, to enhance classification performance using in-distribution public data from CIFAR-10 under $(2, 10^{-5})$-DP. Refer to Section 3.1 for the details of setups and the relevant sections for the details of each method.

Setup	Training Settings	Test Acc
Baselines including previous SOTA methods (Sec. 3.1)		
Cold	Cold Baseline (WRN16-4) [12]	64.02%
Warm	Warm-up on public data	68.09%
WarmSyn	Warm-up on DDPM synthesis [43]	72.0%
WarmSE	WarmSyn (DDPM) + DOPE-SGD [43]	75.1%
In-distribution public data synthesis (Sec. 4.1) & Diversity (Sec. 4.2)		
WarmSyn	Warm-up on EDM synthesis	75.13 %
WarmSyn	EDM synthesis + generation diversity	77.66 %
WarmSyn	EDM synthesis + augmentation diversity	84.88%
Generalization for public-private datasets (Sec. 4.3)		
WarmSyn	Well-generalizing minimum during warm-up	85.48%

Table 1, assuming 4% of public data with a privacy budget of $\varepsilon = 2$ and $\delta = 10^{-5}$. The results outperform the existing state-of-the-art (SOTA) methods, i.e., boosting the accuracy from the previous 75.1% [43] to 85.48% on CIFAR-10. We emphasize that our approaches for dissecting warm-up training for DP are distinguishable from transfer tasks in non-DP setups, addressing data scarcity and improving underperforming DP-SGD optimization. Table 2 shows the gain and privacy cost of methods with public data usage compared to the training without public data, as in [21]. The results show higher gain and privacy costs with stronger privacy (lower ε), whereas ours shows the highest gain and smallest costs. Notably, we observe tiny improvements in non-private settings, highlighting the distinctiveness of DP scenarios and the unique strengths of our tailored approach.

Table 2. Gain (↑) in performance and the corresponding cost of privacy (↓, in parentheses) for each method compared to not exploiting public data. Refer to Section 3.1 for the details of the setups.

Methods	Non-priv.	$\varepsilon = 6$	$\varepsilon = 2$
Warm [43]	0.05% (-)	0.10% (18.80%)	3.20% (27.80%)
WarmSE [43]	0.13% (-)	3.00% (15.98%)	5.60% (25.48%)
Ours	0.13% (-)	**10.06%** (**8.92%**)	**20.58%** (**10.50%**)

2. Background and Related Work

2.1. Differentially Private Deep Learning

Differential privacy (DP) [19] can guarantee the privacy of training data as follows:

Definition 2.1 *(Differential privacy) For two adjacent inputs $d, d' \in \mathcal{D}$, a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ satisfies (ε, δ)-differential privacy for any set of possible outputs $\mathcal{S} \subseteq \mathcal{R}$ if*

$$Pr[\mathcal{M}(d) \in \mathcal{S}] \leq e^{\varepsilon} Pr[\mathcal{M}(d') \in \mathcal{S}] + \delta. \quad (1)$$

The privacy budget $\varepsilon \geq 0$ controls the level of privacy guarantee with the broken probability $\delta \geq 0$. In deep learning, differentially private SGD (DP-SGD) [1] is widely used with the following steps: (i) average the clipped per-sample gradient $\nabla \ell_i(\boldsymbol{w}) \coloneqq \nabla \ell(\boldsymbol{w}; \boldsymbol{x}_i)$ for weight $\boldsymbol{w}$ with respect to each individual private data sample $\boldsymbol{x}_i \in \boldsymbol{X}_t^{pr}$ and (ii) add Gaussian noise to the averaged gradient as follows:

$$\begin{aligned} \boldsymbol{g}_t^{pr} &= \frac{1}{|\boldsymbol{X}_t^{pr}|} \sum_{\boldsymbol{x}_i \in \boldsymbol{X}_t^{pr}} \texttt{clip}\left(\nabla \ell_i(\boldsymbol{w}_t), C\right), \\ \tilde{\boldsymbol{g}}_t^{pr} &= \boldsymbol{g}_t^{pr} + \mathcal{N}(\boldsymbol{0}, C^2 \sigma^2 \mathbf{I}). \end{aligned} \quad (2)$$

The weight is updated as $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \tilde{\boldsymbol{g}}_t^{pr}$ with a learning rate η and $\texttt{clip}(\boldsymbol{u}, C)$ projects $\boldsymbol{u}$ to the L_2-ball of radius C. The noise level σ is determined by the privacy budget (ε, δ), the number of training steps, and the sampling probability (refer to Appendix C for details).

To guarantee privacy with clipping and noise addition, DP-SGD inevitably degrades the performance during optimization. Therefore, various studies explored DP-friendly properties, including architecture [9, 59], loss and activation functions [44, 54], or smoothness [46, 55, 60]. Notably, De et al. [12] introduced a new era in utilizing large models for DP-SGD tasks by introducing the averaged loss of various augmentations called augmentation multiplicity, weight standardization [49], and optimization tricks. Note that adaptive clipping [3, 8] could improve our results.

2.2. Private Learning with Public Information

The benefits of massive OOD public data have been explored in the context of the pre-trained models in vision [7, 12] and language [39, 65–67] models. Additionally, the transfer learning with similar distribution, such as CIFAR-100 for CIFAR-10, was under investigated [58, 59].

ID public data $\boldsymbol{X}^{pub}$ requires different approaches from the massive OOD public data due to its limited size and similar distributional properties. Hence, prior studies [2, 4, 38, 43], categorized as *extended methods* in Section 3.1, have focused on mitigating errors in the optimization of DP-SGD by utilizing the gradients of public data as a proxy for true gradients. These methods update the weight with $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta f_g$, for $f_g(\boldsymbol{w}_t; \boldsymbol{X}_t^{pub}, \boldsymbol{X}_t^{pr})$ such as normalizing the private gradient with the public gradient norm [38], employing a linear combination of private and public gradients [2], or estimating low-rank approximation with public data (even unlabelled) to reduce noise levels [48, 65].

Nasr et al. [43] firstly introduced training a generative model as an augmentation technique. However, their primary focus is the optimization called DOPE-SGD, which updates towards the public gradient $\boldsymbol{g}_t^{pub}$ (without clipping and noise addition) and makes $\boldsymbol{g}_t^{pub} - \boldsymbol{g}_t^{pr}$ private as follows:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta(\boldsymbol{g}_t^{pub} + \texttt{clip}(\boldsymbol{g}_t^{pub} - \boldsymbol{g}_t^{pr}, C) + \mathcal{N}), \quad (3)$$

with the same noise level $\mathcal{N}$ in Equation (2). To the best of our knowledge, none of the prior studies deeply investigated the ID public data as a perspective of the data synthesis.

2.3. Diffusion Synthesis

The integration of generative models in classification tasks has been widely explored to enhance generalization performance without extra data samples [5, 22, 25]. In recent years, diffusion models [28, 56, 57] have shown promising results in image generation. Notably, Xiao et al. [62] argued that diffusion models can generate high-quality and diverse images. Subsequently, various studies have focused on enhancing the quality of generation [30, 32], conditioning sampling [15, 27], and text-to-image generation [50, 52].

Due to the significance of privacy concerns during image synthesis, differentially private diffusion synthesis is one of the actively investigated research topics [17, 41]. Given the effectiveness of private fine-tuning upon pre-trained diffusion models for generation and classification tasks [23], utilizing privacy-preserving diffusion models presents opportunities for our future works.

Measures To measure the quality of generated images, various measures are proposed: Inception Score (IS) [53] and Precision [36] to measure fidelity, Recall [36] to measure diversity, and Fréchet Inception Distance (FID) [26] to measure the distributional quality for mean and variance.

To assess the impact of synthetic data on classification tasks, Ravuri and Vinyals [51] proposed the Classification Accuracy Score (CAS), which evaluates the classification performance on a test set using a model trained on synthetic data. Thus, we use CAS to validate the performance of synthetic public data on test data. However, since our main focus is on improving classification performance including private learning, all of the previous measures cannot be perfectly aligned with our goals.

3. Framework

3.1. Training Procedure

Our objective is to investigate the potential of diffusion models with ID public data, departing from conventional approaches that focus on guiding optimization with the public data [2, 4, 38, 43, 48, 64]. Our work aligns closely with [43], but we specifically concentrate on synthesizing data from the public and examining the important properties of private learning. We summarize the training scenarios using ID public data as follows [43]:

- **Cold**: Conduct DP training (e.g., DP-SGD) on the private dataset without using public data.
- **Warm**: Train models through (i) non-DP *warm-up training* phase on the public dataset (e.g., SGD) and (ii) *private training* phase on the private dataset with DP methods.

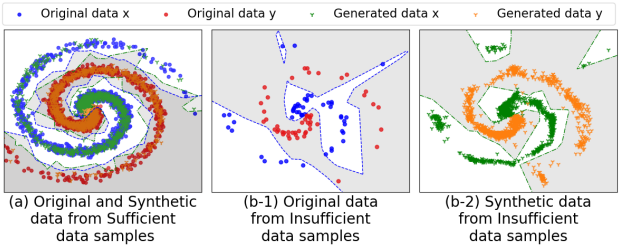


Figure 2. Toy experiment on diffusion synthesis and its effect on decision boundaries while varying the training sample size.

- **WarmExt**: During the private training (after warm-up training), leverage the public side information, such as gradients [2, 4, 38, 43], or low-rank approximation [48, 64], as *extended methods* of DP optimization.
- **WarmSyn** (**Ours**, [43]): Amplify the sparse information of public data using generative models or augmentation methods. Optionally, **WarmSE** indicates employing extended optimization with synthetic public data.

In terms of privacy, since utilizing public data does not pose any privacy concerns to the private data, we can leverage the public data multiple times, e.g., training warm-up classification models, diffusion models, or guiding private training. From now on, we denote in-distribution (ID) public data as public data unless otherwise specified.

3.2. Public Data Synthesis with Diffusion Models

For amplifying the sparse information in the WarmSyn setting, we first investigate amplifying the number of data samples using diffusion models. Recent studies [5, 22, 25, 61] have demonstrated the performance gain in complicated classification problems by training diffusion models to expand the size of training data samples within the data manifold, e.g., amplifying 50K CIFAR-10 training data into a 50M synthetic dataset improves adversarial training [61].

To analyze public data synthesis with limited data samples, we conduct an experiment using a spiral dataset as a toy example (detailed in Appendix C), as shown in Figure 2. We consider two settings: (a) with a sufficient and (b) with an insufficient number of original data samples. For each setting, we train a simple diffusion model to generate new data samples. Then, we train two-layer classifiers until convergence using the original and synthetic data, respectively.

With a large number of samples in (a), both diffusion and classification models demonstrate well-trained results. However, when the original data size is limited as in (b-1), the classification model tends to overfit individual data. Thus, despite correctly classifying almost all data samples, the model fails to discover the proper decision boundary. In contrast, in (b-2) where the diffusion model effectively approximates the data manifold, the classifier achieves a superior decision boundary compared to (b-1). Nevertheless, we should be aware that diffusion models can be apt to over-

Table 3. Quality comparison of synthetic data trained with 4% of public data on CIFAR-10.

Sampling	Fidelity		Diversity	Quality	CAS (↑)	Test Acc
	IS (↑)	Precision (↑)	Recall (↑)	FID (↓)	(%)	(%)
EDM (w_d=0)	**11.008**	**0.964**	0.157	7.799	62.82	75.13
EDM + DG (w_d=3)	10.815	**0.964**	0.153	**7.786**	62.47	75.31
EDM + DG (w_d=10)	10.796	0.946	0.170	8.274	64.61	75.98
EDM + DG (w_d=20)	10.157	0.873	0.191	11.497	**67.61**	**77.66**
EDM + DG (w_d=30)	9.113	0.785	**0.211**	19.748	66.53	77.22

Figure 3. Selected synthetic data samples from (Top) EDM and (Bottom) EDM + DG (w_d = 30) synthesis.

and underestimate the data distribution, as shown in (b-2).

To push further, we mathematically analyze how the rate of data samples affects synthesis. The ability of a generative model θ to approximate the 'entire' data can be determined by the ratio of 'seen' (observed during training) and the 'unseen' (not observed during training). Specifically, with a limited number of seen data, the model is required to estimate the unseen data to approximate the entire data.

Theorem 3.1 *For a finite number of 'entire' data samples $S_{data} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, split the data samples into 'seen' data $S_s = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, and 'unseen' data $S_u = \{\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N\}$, without loss of generality. Let p_{data}, p_s, and p_u be the probability distribution of each corresponding dataset. For any generative model θ trained on 'seen' data, p_θ represents the data distribution generated by θ. Then,*

$$\log \frac{1}{r_1} + \hat{p}_\theta^u \log \frac{\hat{p}_\theta^u}{r_2} \leq \tilde{D}_{KL}(p_\theta \| p_{data}) - \tilde{D}_{KL}(p_\theta \| p_s) \leq \log \frac{1}{r_1} + \hat{p}_\theta^u \log \frac{1}{r_2},$$

where $\hat{p}_\theta^u = \sum_{\boldsymbol{x} \in S_u} p_\theta(\boldsymbol{x}) \leq 1$ indicates the capacity of the model θ to generate 'unseen' data. For discrete probability distributions p and q, $\tilde{D}_{KL}(p\|q)$ indicates the relaxed Kullback–Leibler (KL) divergence, excluding the case when $q = 0$. $r_1 = \frac{n}{N}$ and $r_2 = \frac{N-n}{n}$ denote the ratios of 'seen' to 'entire', and 'unseen' to 'seen' data, respectively.

The detailed proof is presented in Appendix A. The above theorem indicates that the ratios r_1 and r_2 affect how the model θ can approximate the p_{data}. As $r_1 \to 1$ with sufficient seen data, then $r_2 \to 0$ and $S_u \to \varnothing$, so that $\hat{p}_\theta^u \to 0$ and the upper and lower bounds converge to zero. Thus, ensuring p_θ is close to p_s ensures proximity of p_θ to p_{data}.

However, when considering a small r_1, which is equivalent to a limited number of public data setups, the bounds depend on $\hat{p}_\theta^u$. For $r_1 \leq \frac{1}{e+1}$, the lower and upper bounds are decreasing functions with respect to $\hat{p}_\theta^u$. This implies that as the model's capacity to generate unseen data increases, the differences in divergences decrease. Thus, approximating the distribution of seen data is not enough to learn the whole data distribution for small r_1. Rather, it is important for the model to learn the data manifold effectively to better approximate the data distribution including the unseen data.

4. Generated Data for Private Learning

We now investigate how the aforementioned diffusion synthesis can be utilized in the context of public data and analyze how it can be further leveraged to enhance performance. For experiments, we primarily focus on the CIFAR-10 dataset with a privacy budget of $(2, 10^{-5})$, utilizing WRN-16-4[2] with 16 augmentation multiplicity and the techniques proposed in [12]. Within the training set, we randomly select 2K instances (4%) for the public samples, as suggested in [43]. These samples are uniformly drawn from each class. All the measures for evaluating the generated data in this section are calculated with the entire training set (50K) as a reference. Additional experimental details are provided in Section 5.1 and Appendix B.

4.1. Better Diffusion Synthesis with Public Data

Theorem 3.1 demonstrates that well-approximating the distribution of seen public data p_s enables the synthetic data distribution p_θ mimic the entire data distribution p_{data}, which is essential to obtain good decision boundary as in Figure 2. Thus, we initially replace the Denoising Diffusion Probabilistic Model (DDPM) [28], previously investigated by Nasr et al. [43] for public synthesis, with the Elucidating Diffusion Model (EDM) [30]. EDM is known to generate better images (e.g., lower FID scores) than DDPM by using a higher-order sampling process. Then, we train EDM on 2K public data samples without using external datasets and employ class-conditional sampling to match the original distribution. With the trained EDM, we calculate various measures in Table 3. The FID of EDM synthesis at 7.80 (correspondingly 7.89 on 40K) outperforms the reported FID of 12.8 of DDPM on 40K images [43], which also improves the classification performance as shown in Table 1.

However, the generation quality using public data is notably worse than the FID of 1.79 achieved with the entire set [30]. Within a limited public data, the model struggles to capture diversity, resulting in a recall of 0.16, even though precision remains high at 0.96. Note that repeating each public sample 25 times (thus 50K samples) results in a precision of 1.00, recall of 0.04, and FID of 13.64. The generated images and their memorization are illustrated in Ap-

[2]WRN-40-4 requires excessive GPU considering its performance [43].

Table 4. CAS (%), test accuracy (%), and their difference (%) with different augmentation methods during warm-up phase.

Augmentation	CAS(%)	Test Acc (%)	Diff (%)
No Aug (EDM)	62.82	75.13	13.31
Common	77.99	83.97	5.98
Common + Cutout	**80.72**	**84.88**	4.16
Common + Cutmix	65.73	82.20	**16.47**
AutoAugment	77.21	83.45	6.24

Table 5. Geometric measures of the models trained with SGD and SAM after warm-up phase.

Opt	Synthetic Data			Private Data		
	λ_{max}	λ_{max}/λ_5	$\mathrm{Tr}(\nabla^2\ell(\boldsymbol{w}))$	λ_{max}	λ_{max}/λ_5	$\mathrm{Tr}(\nabla^2\ell(\boldsymbol{w}))$
SGD	1.38	10.62	71.98	112.82	1.58	-2527.78
SAM	**0.44**	**2.63**	**53.21**	**58.32**	**1.50**	333.03

pendix E. No privacy concern occurs in diffusion synthesis since the diffusion model is solely trained with public data.

4.2. Data Diversity Matters for Public Data

With limited private data, the model's capacity to generate unseen data is required in Theorem 3.1. The lack of diversity in the generated data can lead to misclassification due to distorted decision boundaries, as described in Figure 2. Given the challenge of DP-SGD in identifying a good basin [21], the key to successful warm-up lies in finding a well-generalizing minimum with a high CAS. As observed in [25, 51], data diversity can play an important role in improving CAS. We hypothesize that the importance of diversity is more pronounced in the warm-up phase of private learning, where the generated images exhibit low recall values.

Diversity for generation To validate our hypothesis, we first investigate enforcing data diversity in the generation process. Obtaining diversity in generations solely through a diffusion model is a challenging task. Thus, to enhance the data diversity during generation, we can use variants of diffusion models for guidance [27, 32] or editing [10, 40, 42]. Among them, discriminator guidance (DG) [32] introduces a discriminator to judge whether the sampling is from the true data or synthesis, controlling the trade-off between fidelity and diversity of the generated images by adjusting the weight w_d of the discriminator. Higher w_d yields diversity. Refer to Appendix C for the details of DG.

As shown in Table 3, a higher weight of guidance ensures greater data diversity without any augmentation, but sacristies the fidelity. The best FID score is obtained with $w_d = 3$ while the best CAS is obtained at a bigger weight $w_d = 20$. Interestingly, we need a larger weight value to achieve a similar gain of diversity than standard training ($w_d = 1.5$ for best FID [32]). Figure 3 represents selected examples from two extreme cases, with $w_d = 0$ and $w_d = 30$ to visualize the difference. Despite the quality degradation of detailed features with $w_d = 30$, the CAS is higher due to the increased data diversity in features. Thus, we need to enhance diversity while maintaining quality for better classification. Similar to the EDM model trained with public data, no privacy concerns occur since the discriminator requires only original public and synthetic data from EDM.

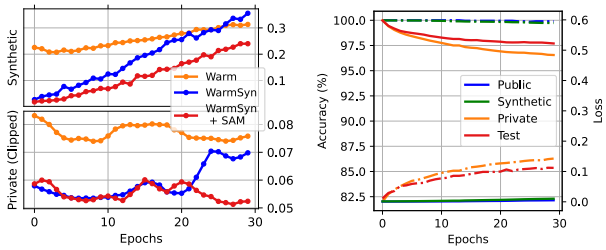
Diversity with augmentation Nonetheless, training an additional discriminator for DG can pose a bottleneck in terms of time and computational efficiency. Thus, to explicitly enhance the diversity of EDM synthetic data during the warm-up training, we employ various data augmentation techniques designed for classification tasks. Wang et al. [61] argued that utilizing appropriate augmentation in diffusion-based generated images can further improve the classification performance. Common augmentation [24] uses padding and random crop to the original size and horizontal flipping for the images. Cutmix [68] randomly replaces a part of the image with another and Cutout [14] randomly pad images. AutoAugment [11] chooses the best combination of augmentations such as color, rotation, or cutout. We set a baseline with no augmentation for synthetic images of EDM and compare these augmentations.

The results are summarized in Table 4. Adding cutout augmentation to common augmentation demonstrates the best performance in terms of CAS and test accuracy. However, the two measures are not always aligned. Rather, the best performance gain with private data, noted as *Diff (%) = Test Acc−CAS*, is obtained from Cutmix, which is known as one of the most diversified augmentations. This indicates that robust decision boundaries achieved through diversity facilitate easier private learning. Interestingly, the performance gain of augmentation surpasses the benefits of generation diversity, as illustrated in Appendix C.

4.3. Optimization with Synthetic Data

As the clipping of DP-SGD significantly lowers the performance, it is known that a smaller gradient norm ensures high performance in private training [44, 46, 54]. In Figure 4a, we illustrate the gradient norms during private training with different training setups. The norm of synthetic data is calculated without clipping, whereas the private data used for DP-SGD is summed over individually clipped gradients, each with a clipping value of $C = 1$. The results indicate that utilizing a synthetic dataset reduces the gradient norm, even in a private dataset where constraining the gradient norm is crucial to mitigate the effect of clipping.

However, even though training with diversified synthetic public data and augmentation, the models face the risk of being overfitted to public data. As public samples are selected from in-distribution data, the problem of the public-



(a) Gradient norm. (b) Accuracy (−·) and loss (−).

Figure 4. Learning dynamics of private training with different datasets and optimization methods.

private generalization gap is analogous to the training-test generalization gap in standard training. In Figure 4b, we illustrate the accuracy and loss for each public, synthetic, private, and test dataset during private training after the warm-up phase. The model consistently achieves near-zero loss and 100% accuracy on synthetic data (as well as on public data) but struggles with private data (as well as on test data).

To alleviate the generalization gap in standard training, various studies [16, 31, 33, 34, 37, 47] explored the geometric properties of the loss function in the weight space. The prominent optimization methods are designed to uncover flat minima, such as sharpness-aware minimization (SAM) [20] or stochastic weight averaging (SWA) [29]. Therefore, we apply SAM during the warm-up training and calculate geometric generalization measures, including metrics such as the top Hessian value λ_{max}, the ratio of Hessian λ_{max}/λ_5, and the trace of the Hessian matrix $\mathrm{Tr}(\nabla^2\ell(\boldsymbol{w}))$, for both synthetic and private data. Table 5 indicates that utilizing SAM results in lower (better) metric values, not only for synthetic data but also for private data. Note that SGD even demonstrates a negative trace value on private data.

Moreover, in Figure 4a, as the flat minimum of SAM leads to a decreased gradient norm on synthetic data, it also maintains similar results concerning private data. These results indicate that relieving overfitting to synthetic data can help the generalization of private data. Thus, the aforementioned approaches find suitable initial point of DP-SGD.

5. Experiments

5.1. Experimental setup

We assess the effectiveness of our proposed methods using public data primarily in two datasets: CIFAR-10 and CIFAR-100. For the public dataset, we randomly sample 4% of the training data (2K samples) uniformly drawn from each class, while the remaining data are used as private samples following [43]. We then train the EDM [30] models with the 2K public data samples and build 50K synthetic datasets with EDM sampling. To address the computational burden of training diffusion and sampling, we generate datasets once and reuse them for classifiers as in [43].

For classification models, we adopt WRN-16-4, following the techniques in [12] with 16 augmentations, and use pre-trained vision transformer models following [7]. Our experiments are conducted using PyTorch libraries [63] on eight NVIDIA GeForce RTX 3090 GPUs, partially on a cloud server with four NVIDIA A100 40GB GPUs. We report the mean and standard deviation for each experiment. For the detailed settings, refer to Appendix B.

5.2. Classification Performance with Public Data

Effects of individual techniques We first revisit Table 1, the ablation study of sequentially employing our approaches, with a privacy budget $(2, 10^{-5})$-DP on CIFAR-10. By only using better EDM synthesis without extra optimization techniques, we achieve the previous SOTA performance of 75.1% [43]. Additionally, recognizing the significance of diversity in data generation, we employ cutout augmentation techniques for diversity in classification, resulting in a performance of 84.88%. To mitigate potential overfitting to public information, we make use of generalization techniques. All these efforts collectively lead to an accuracy of 85.48% under $(2, 10^{-5})$-DP.

Despite achieving the best performance of 85.93% when combined with DG, we exclude DG in this section due to its marginal performance enhancement (0.45%p) considering its extended training time, as detailed in Appendix C.

CIFAR-10 with public data We report the performance comparison for our method with various previous approaches on a wide range of $\varepsilon \in \{1, 2, 3, 4, 6\}$ with $\delta = 10^{-5}$ in Table 6. Our approach, which incorporates EDM synthesis, augmentation, and optimization for the public data, exhibits superior classification performance when compared to existing methods including DDPM-based augmentation and extended optimizations. The accuracies of $\varepsilon = 0$ (after warm-up, without private data) and $\varepsilon = \infty$ (not private, with clipping) are 80.72% and 88.52%, respectively.

CIFAR-100 with public data We then explore the CIFAR-100 dataset for DP, which is not actively investigated without pre-trained models due to its complexity. We first train the EDM model with 2K images and generate 50K images, similar to CIFAR-10. Given the 100 classes in CIFAR-100, only 20 public samples are available per class, significantly fewer than in CIFAR-10. The FID on 50K images with EDM synthesis is 11.28, which reduces the original FID of 15.58 for replicating each public image 25 times. After the warm-up, the accuracy of synthetic images on $\varepsilon = 0$ is 38.04% (46.74% on $\varepsilon = \infty$), while the test accuracy of using 2K public samples without synthesis on $\varepsilon = 0$ is only 16.79%. In Table 7, we report the accuracies by sequentially adopting the aforementioned tech-

Table 6. Test accuracy (%) of private training on CIFAR-10 with privacy budgets of $\varepsilon \in \{1, 2, 3, 4, 6\}$. The public and synthesis columns denote the Warm and WarmSyn setups, respectively. Ours employs all the techniques in Table 1, i.e., synthesis, augmentation, and optimization. We highlight the best accuracy in **bold**.

Datasets	Architecture	Public	Synthesis	Method	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	$\varepsilon = 4$	$\varepsilon = 6$
CIFAR-10	CNN-Tanh (0.55M)	✗	✗	[44]	45.8	58.3	63.5	-	-
	ScatterNet (0.16M)	✗	✗	[59]	60.3	67.2	69.3	-	-
	DPNAS [9] (0.53M)	✗	✗	[46]	60.13±0.34	67.23±0.12	69.86±0.49	-	-
	WRN-16-4 (2.74M)	✗	✗	[12]	56.8±0.6	64.9±0.5	69.2±0.3	71.9±0.3	77.0±0.8
	WRN-40-4 (8.94M)	✗	✗	[12]	56.4±0.6	65.9±0.5	70.7±0.2	73.5±0.6	78.8±0.4
	WRN-16-4 (2.74M)	✓	✗	[2]†	-	68.7	-	73.1	77.2
		✓	✗	[38]†	-	68.7	-	73.5	77.9
		✓	✓	[2]†	-	70.5	-	74.5	78.2
		✓	✓	[38]†	-	69.1	-	74.1	78.1
		✓	✓	[43]†	-	75.1	-	77.9	80.0
		✓	✓	**Ours**	**84.30±0.11**	**85.48±0.12**	**86.03±0.09**	**86.49±0.13**	**87.06±0.24**

†We note the results reported in [43] to set the architecture same. All other baseline results are adopted from the original paper.

Table 7. Test accuracy (%) of private classification on CIFAR-100 on the privacy budget of $\varepsilon \in \{1, 2, 6, 10\}$. The public and synthesis columns indicate Warm and WarmSyn settings, respectively. We employ the techniques in Table 1 sequentially, i.e., synthesis, augmentation, and optimization. We highlight the best accuracy in **bold**.

Datasets	Architecture	Public	Synthesis	Methods	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 6$	$\varepsilon = 10$
CIFAR-100	Resnet-9	✗	✗	[35]†	18.1	24.9	-	40.8
	Resnet-9 (6.62M)	✗	✗	Cold†	8.35	14.42	29.89	35.11
	WRN-16-4 (2.74M)	✗	✗	Cold	9.28	18.19	33.61	39.09
	WRN-16-4 (2.74M)	✓	✗	Warm	20.84	25.15	33.47	38.89
		✓	✓	WarmSyn	26.13±0.20	31.53±0.04	35.53±0.11	40.82±0.10
		✓	✓	+Augmentation	40.96±0.30	44.52±0.01	50.47±0.04	54.29±0.05
		✓	✓	+Optimization	**45.93±0.06**	**48.61±0.59**	**54.36±0.52**	**56.56±0.11**

†Unfortunately, their DP-SGD results are not reproducible, even when using the same hyperparameters as in the original paper.

niques, where each of the approaches matters. As a result, we obtain 48.61% on $\varepsilon = 2$ without pre-trained models.

CIFAR-100 with pre-trained on ImageNet To push further, we demonstrate the effectiveness of our procedures when combined with pre-trained models with ID public data. We adopt the vision transformers following [7] pre-trained on ImageNet [13]. Upon the pre-trained models, we sequentially perform the warm-up training with ID public data and private learning. The results in Table 8 indicate that the synthesis can boost classification performance. Within a wide range of models and privacy budgets, our methods outperform the warm settings.

Distribution shift (CIFAR-100 → CIFAR-10) The distribution shifts from similar but not ID public data [59] can be combined with our methods. We consider 4% of CIFAR-100 data as public to enhance the performance of CIFAR-10 as a substitute for CIFAR-10 ID public data. In Table 9, we present the classification results for both the warm and our settings, where both models are trained on the public data and synthetic data from 4% of CIFAR-100, respectively. The experimental results show that the differences in accuracies exceed 20%, even though neither model has observed CIFAR-10 datasets before private training with DP-SGD. This indicates that amplifying public information is meaningful in distribution shifts. Additional sensitivity analysis and ablation studies can be found in Appendix D.

Pre-trained diffusion models Our main idea to use synthesis for public data is flexible in practical scenarios and diverse datasets. To mitigate the drawbacks of training diffusion on each dataset, we investigated using the pre-trained diffusion models, which are trained on OOD public data. Specifically, we make use of pre-trained Stable Diffusion (SD) [52] and Boomerang sampling [40]. We fed 4% of public data to Boomerang editing with SD backbone for generation diversity and generated 20 images per sample for classifier training. The similar experiments are shown in Appendix D.We maintain the other training procedures the same as 'Ours', which we call 'Ours-SD'. Table 10 indicates that our framework is straightforward and effective for real-world data, and generalizable without depending on specific diffusion models. We conclude that data diversity can enhance performance when the model is not trained on specific ID public data.

Table 8. Test accuracy (%) of private classification on CIFAR-100 using pre-trained models on the privacy budget of $\varepsilon \in \{0.5, 2\}$. The **bold** indicates results within the standard deviation of the top mean score.

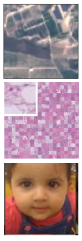
Datasets	Privacy budget	$\varepsilon = 0.5$			$\varepsilon = 2$		
	Architecture	Cold	Warm	Ours	Cold	Warm	Ours
CIFAR -100	CrossViT small 240 (26.3M)	60.50±1.70	73.17±0.24	**77.43±0.13**	71.36±0.52	77.00±0.27	**80.25±0.57**
	CrossViT 18 240 (42.6M)	67.03±0.46	77.19±0.44	**79.96±0.10**	74.28±0.22	80.30±0.02	**82.85±0.08**
	DeiT base patch16 224 (85.8M)	49.34±1.49	**79.56±0.74**	**79.28±0.76**	69.09±0.10	82.98±0.04	**83.06±0.01**
	CrossViT base 240 (103.9M)	66.24±0.47	75.59±0.21	**78.06±0.05**	75.15±0.03	79.25±0.05	**81.14±0.15**

Table 9. Comparison of distribution shift performance on CIFAR-10 with the warm and our models trained on 4% of CIFAR-100.

CIFAR-100 → CIFAR-10	Test Acc (%)		
	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 6$
Warm	50.65±0.37	56.54±0.20	59.49±1.39
Ours	**73.46±0.28**	**78.60±0.13**	**80.12±0.04**

Table 10. Classification results on real-world datasets.

Data	Method	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 6$
EuroSAT (Land use)	Warm (w/ Aug, Opt)	83.67	85.52	86.07
	WarmExt [43]	84.37	85.78	86.07
	Ours-SD	**90.37**	**91.74**	**92.52**
PathMNIST (Biomedical)	Warm (w/ Aug, Opt)	89.23	89.39	89.54
	WarmExt [43]	89.04	89.25	89.28
	Ours-SD	**91.64**	**91.94**	**92.10**
FairFace (w.r.t. Race)	Warm (w/ Aug, Opt)	53.00	54.55	54.83
	WarmExt [43]	53.33	54.74	54.84
	Ours-SD	**54.90**	**56.19**	**56.37**



5.3. Revisit Extended Optimization

On top of our method, we reevaluate the effectiveness of existing extended methods (thus WarmSE) in Table 11, particularly mirror GD [2] and DOPE-SGD [43]. These methods are trained after our synthesis, augmentation, and warm-up optimization on $\varepsilon \in \{2, 4, 6\}$. Remarkably, the performance of the extended methods falls within the standard deviation range of DP-SGD. This implies that, as the models have already extracted substantial side information from public data before private training, extended optimization does not further improve DP optimization. Note that the computational time of the public batch is marginal to DP-SGD.

However, extended methods sometimes exhibit poorer training stability than DP-SGD. Due to their reliance on public gradients, these methods are sensitive to the hyperparameter settings and easily encounter exploding gradients. In Figure 5, under the same setting as Figure 4a, we observe a gradual decrease in the private gradient norm, while the public gradient norm consistently increases. When the public norm increases to a certain level, the norm of both public and private data with extended methods may diverge after updating toward public gradients. For further insights about the learning dynamics in terms of the loss function and private gradient norm without clipping, please refer to Appendix C. However, we believe that undiscovered extended methods can be beneficial for the WarmSyn settings.

Table 11. Performance and computational time of different extended optimization methods.

Privacy budget ε ($\delta = 10^{-5}$)	Optimization				
	DP-SGD	Mirror GD [2]		DOPE-SGD [43]	
		Median	Min	Median	Min
$\varepsilon = 2$	85.48±0.17	85.52	10.00	85.53	10.00
$\varepsilon = 4$	86.49±0.13	86.31	10.00	86.69	10.00
$\varepsilon = 6$	87.06±0.24	86.73	10.00	86.84	10.00
Time (ms/image)	12.70	12.80		12.86	

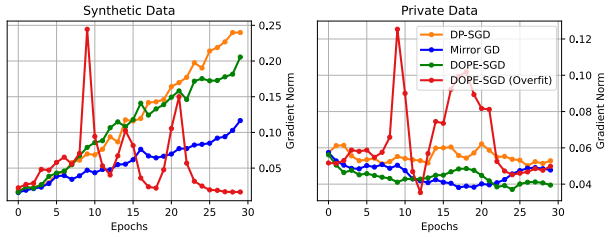


Figure 5. The norm of gradients with (Left) synthetic dataset and (Right) private dataset with clipping during private training with different optimization methods.

6. Conclusion

In this paper, we investigate the potential of diffusion models for in-distribution public data to enhance private classification performance. We demonstrate the importance of synthetic data diversity, augmentation techniques, and the importance of well-generalizing minima for private optimization. As a limitation, we leave the experiments on more extensive and sensitive datasets and the efficient implementation of our method. We hope that this work helps leverage diffusion models to address utility-privacy trade-offs.

Acknowledgements This work was supported by NRF (No. 2022R1A5A6000840, No. RS-2023-00272502) and IITP (No. 2022-0-00984, No. 2021-0-01343-004) grants funded by the Korea government (MSIT).

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 1, 2

[2] Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Vinith M Suriyakumar, Om Thakkar, and Abhradeep Thakurta. Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pages 517–535. PMLR, 2022. 1, 2, 3, 7, 8

[3] Galen Andrew, Om Thakkar, Hugh Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, 2021. 2

[4] Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 383–392. PMLR, 2021. 1, 2, 3

[5] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. gaosynthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 3

[6] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019. 1

[7] 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*, 2022. 2, 6, 7

[8] Zhiqi Bu, Hua Wang, Zongyu Dai, and Qi Long. On the convergence and calibration of deep learning with differential privacy. *Transactions on Machine Learning Research*, 2023. 2

[9] Anda Cheng, Jiaxing Wang, Xi Sheryl Zhang, Qiang Chen, Peisong Wang, and Jian Cheng. Dpnas: Neural architecture search for deep learning with differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6358–6366, 2022. 2, 7

[10] Yujin Choi, Jinseong Park, Hoki Kim, Jaewook Lee, and Saeroom Park. Fair sampling in diffusion models through switching mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 5

[11] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 5

[12] 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. 1, 2, 4, 6, 7

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 5

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[16] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017. 6

[17] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022. 3

[18] Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006. 1

[19] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. 1, 2

[20] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020. 6

[21] Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pages 10611–10627. PMLR, 2023. 1, 2, 5

[22] Cong Gao, Benjamin D Killeen, Yicheng Hu, Robert B Grupp, Russell H Taylor, Mehran Armand, and Mathias Unberath. Synthetic data accelerates the development of generalizable learning-based algorithms for x-ray image analysis. *Nature Machine Intelligence*, 5(3):294–308, 2023. 3

[23] Sahra Ghalebikesabi, Leonard Berrada, Sven Gowal, Ira Ktena, Robert Stanforth, Jamie Hayes, Soham De, Samuel L Smith, Olivia Wiles, and Borja Balle. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[25] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and XIAOJUAN QI. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 5

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 4

[29] P Izmailov, AG Wilson, D Podoprikhin, D Vetrov, and T Garipov. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885, 2018. 6

[30] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 1, 3, 4, 6

[31] Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. 6

[32] Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 16567–16598. PMLR, 2023. 3, 5

[33] Hoki Kim, Jinseong Park, Yujin Choi, and Jaewook Lee. Stability analysis of sharpness-aware minimization, 2023. 6

[34] Hoki Kim, Jinseong Park, Yujin Choi, Woojin Lee, and Jaewook Lee. Exploring the effect of multi-step ascent in sharpness-aware minimization. *arXiv preprint arXiv:2302.10181*, 2023. 6

[35] Moritz Knolle, Robert Dorfman, Alexander Ziller, Daniel Rueckert, and Georgios Kaissis. Bias-aware minimisation: Understanding and mitigating estimator bias in private sgd. *arXiv preprint arXiv:2308.12018*, 2023. 7

[36] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[37] Sungyoon Lee, Jinseong Park, and Jaewook Lee. Implicit Jacobian regularization weighted with impurity of probability output. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19141–19184. PMLR, 2023. 6

[38] Tian Li, Manzil Zaheer, Sashank Reddi, and Virginia Smith. Private adaptive optimization with side information. In *International Conference on Machine Learning*, pages 13086–13105. PMLR, 2022. 1, 2, 3, 7

[39] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022. 1, 2

[40] Lorenzo Luzi, Paul M Mayer, Josue Casco-Rodriguez, Ali Siahkoohi, and Richard Baraniuk. Boomerang: Local sampling on image manifolds using diffusion models. *Transactions on Machine Learning Research*, 2024. 5, 7

[41] Saiyue Lyu, Margarita Vinaroz, Michael F Liu, and Mijung Park. Differentially private latent diffusion models. *arXiv preprint arXiv:2305.15759*, 2023. 3

[42] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 5

[43] Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively using public data in privacy preserving machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 25718–25732. PMLR, 2023. 1, 2, 3, 4, 6, 7, 8

[44] Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9312–9321, 2021. 2, 5, 7

[45] Jinseong Park, Yujin Choi, Junyoung Byun, Jaewook Lee, and Saerom Park. Efficient differentially private kernel support vector classifier for multi-class classification. *Information Sciences*, 619:889–907, 2023. 1

[46] Jinseong Park, Hoki Kim, Yujin Choi, and Jaewook Lee. Differentially private sharpness-aware training. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27204–27224. PMLR, 2023. 2, 5, 7

[47] Jinseong Park, Hoki Kim, Yujin Choi, Woojin Lee, and Jaewook Lee. Fast sharpness-aware training for periodic time series classification and forecasting. *Applied Soft Computing*, page 110467, 2023. 6

[48] Francesco Pinto, Yaxi Hu, Fanny Yang, and Amartya Sanyal. Pillar: How to make semi-private learning more effective. *arXiv preprint arXiv:2306.03962*, 2023. 2, 3

[49] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019. 2

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[51] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019. 3, 5

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 7

[53] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 3

[54] Ali Shahin Shamsabadi and Nicolas Papernot. Losing less: A loss for differentially private deep learning. *Proceedings on Privacy Enhancing Technologies*, 3:307–320, 2023. 2, 5

[55] Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24552–24562, 2023. 2

[56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using

nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3

[58] Lichao Sun and Lingjuan Lyu. Federated model distillation with noise-free differential privacy. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021. 2

[59] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021. 1, 2, 7

[60] Wenxiao Wang, Tianhao Wang, Lun Wang, Nanqing Luo, Pan Zhou, Dawn Song, and Ruoxi Jia. Dplis: Boosting utility of differentially private deep learning via randomized smoothing. *Proceedings on Privacy Enhancing Technologies*, 4:163–183, 2021. 2

[61] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *Proceedings of the 40th International Conference on Machine Learning*, pages 36246–36263. PMLR, 2023. 3, 5

[62] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2021. 3

[63] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021. 6

[64] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2020. 3

[65] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021. 1, 2

[66] 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*, pages 12208–12218. PMLR, 2021.

[67] 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 fine-tuning of language models. In *International Conference on Learning Representations*, 2022. 1, 2

[68] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5