Differentially Private Prototypes for Imbalanced Transfer Learning

Dariush Wahdany[1,3*], **Matthew Jagielski**[2], **Adam Dziedzic**[3], **Franziska Boenisch**[3]
[1]Fraunhofer AISEC
[2]Google DeepMind
[3]CISPA Helmholtz Center for Information Security
dariush.wahdany@aisec.fraunhofer.de, jagielski@google.com, adam.dziedzic@cispa.de, boenisch@cispa.de

Abstract

Machine learning (ML) models have been shown to leak private information from their training datasets. Differential Privacy (DP), typically implemented through the differential private stochastic gradient descent algorithm (DP-SGD), has become the standard solution to bound leakage from the models. Despite recent improvements, DP-SGD-based approaches for private learning still usually struggle in the high privacy ($\varepsilon \leq 1$) and low data regimes, and when the private training datasets are imbalanced. To overcome these limitations, we propose Differentially Private Prototype Learning (DPPL) as a new paradigm for private transfer learning. DPPL leverages publicly pre-trained encoders to extract features from private data and generates DP prototypes that represent each private class in the embedding space and can be publicly released for inference. Since our DP prototypes can be obtained from only a few private training data points and without iterative noise addition, they offer high-utility predictions and strong privacy guarantees even under the notion of *pure DP*. We additionally show that privacy-utility trade-offs can be further improved when leveraging the public data beyond pre-training of the encoder: in particular, we can privately sample our DP prototypes from the publicly available data points used to train the encoder. Our experimental evaluation with four state-of-the-art encoders, four vision datasets, and under different data and imbalancedness regimes demonstrate DPPL's high performance under strong privacy guarantees in challenging private learning setups.

1 Introduction

Machine learning (ML) models are known to leak private information about their training datasets (Carlini et al. 2022; Fredrikson, Jha, and Ristenpart 2015; Shokri et al. 2017). As a solution to provably upper-bound privacy leakage, differential privacy (DP) (Dwork et al. 2006) has emerged as the de-facto standard for private training. It is usually implemented in ML through the differential private stochastic gradient descent (DP-SGD) algorithm which bounds the contribution of each data point during training and iteratively injects controlled amounts of noise (Abadi et al. 2016). Thereby, DP-SGD has been shown to increase training time and decrease the final model's utility.

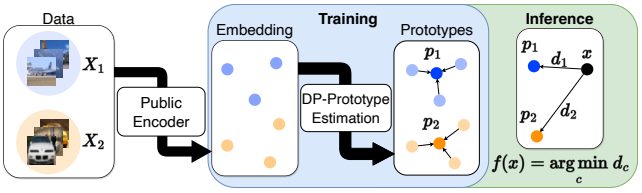


Figure 1: **Overview of DPPL.** We split the private data $\mathbf{X}$ per class c into $\mathbf{X}_c$'s, infer them through a publicly pre-trained encoder, and estimate per-class prototypes $\mathbf{p}_c$ in the embedding space with DP. Classification of samples is performed by returning the label of the closest prototype $\mathbf{p}_c$ in the embedding space according to some distance function d.

While, over the last years, there has been significant progress in improving both computational efficiency (Bu et al. 2021; Lee and Kifer 2021; Subramani, Vadivelu, and Kamath 2021) and privacy-utility trade-offs (Bu, Mao, and Xu 2022; De et al. 2022), there are a few relevant setups where DP training still yields unfavorable results. These include the *high privacy regime* (expressed in DP with small values of the privacy parameter ε, such as $\varepsilon \leq 1$), the low data regime, *i.e.,* when only a *few private data points* are available for training, and when the training dataset is *imbalanced*, *i.e.,* when some classes have significantly more data points than others (Buda, Maki, and Mazurowski 2018; Liu et al. 2019; Reed 2001).

There are various reasons why DP training is challenging in these setups (Feldman 2020; Esipova et al. 2023). One of these is that DP protects small sets of examples due to its "group privacy" property, providing a provable bound on how much a DP algorithm can learn from small data (Feldman 2020). Beyond this concern, the iterative noise addition weakens the signal from the training data, especially when only a few training data points are available. Moreover, standard approaches for learning in imbalanced setups, such as changing the sampling (Kubat, Matwin et al. 1997; Ling and Li 1998), generating synthetic data for the minority classes (Chawla et al. 2002), or weighing the training loss (Cao et al. 2019) are not directly compatible with DP or incur additional privacy costs. In a similar vein, each training iteration with DP training incurs additional privacy costs (Abadi et al. 2016), making it hard to keep ε low, *i.e.,* to stay in the high privacy regime.

*Part of the work was conducted as visiting researcher at CISPA

To address all of these challenges, we propose *Differential Private Prototype Learning* (DPPL), a novel approach for private learning that combines prototypical networks (Snell, Swersky, and Zemel 2017), a standard algorithm for non-private few shot learning, with recent advances in training high-performance private models with DP that leverage powerful encoder models pre-trained on public data (Caron et al. 2021; He et al. 2022; Radford et al. 2021) combined with private transfer learning (Yu et al. 2021; Gu, Kamath, and Wu 2022; Tramèr, Kamath, and Carlini 2022; Ganesh et al. 2023; Hu et al. 2021; Houlsby et al. 2019; Li et al. 2023; Mehta et al. 2023). The main idea of our DPPL is to use the encoder as a feature extractor for the private data and to generate DP prototypes in the embedding space for each private class. To classify new data points, we then simply have to infer these points through the encoder and to return the label of the closest prototype.

Relying on DP prototypes for private learning offers significant advantages over iterative private training or fine-tuning. First, our prototypes do not require iterative noise addition. This enables to obtain less noisy predictions at lower privacy costs and improves privacy-utility trade-offs in the high privacy regime. Second, the prototypes are inherently balanced, *i.e.,* it is possible to obtain good prototypes also at the low data regime or for underrepresented classes from imbalanced private training datasets. Third, DP prototypes are fast to obtain, enable fast inference, and, due to the DP post-processing guarantees—which express that no query to them will incur additional privacy costs—can be publicly released for performing predictions.

We propose multiple algorithms for obtaining DP prototypes, and find that prior approaches for training models with DP do not yet leverage the full capacity of the public data (Yu et al. 2021; Mehta et al. 2023): these prior approaches use the public data only for pre-training the encoder. Yet, we make the observation that we can leverage the public data additionally during the transfer learning step. By privately selecting per-class private prototypes from the public data, we can significantly decrease the privacy costs of our prototypes (even under the strong notion of *pure DP*, *i.e.,* ε-DP) and further improve privacy-utility trade-offs.

By performing thorough experimentation with four state-of-the-art encoders and four standard vision datasets, we show that DPPL provides strong utility in the high privacy regimes. Additionally, we highlight that DPPL is able to provide good privacy-utility trade-offs when only a few private training data points are available and that it yields state-of-the-art performance on imbalanced classification tasks. Thereby, DPPL represents a new powerful learning paradigm for private training with DP.

In summary, we make the following contributions:

- We propose DPPL, a novel alternative to private fine-tuning that combines recent advances in DP transfer learning with private few shot learning and can even yield pure DP guarantees.
- We perform extensive empirical evaluation which highlights that DPPL yields strong privacy-utility trade-offs, in particular in the high privacy regime and for imbalanced data.
- To further improve DP transfer learning, we show that we can leverage the public data beyond the pre-training step of the feature encoder by privately selecting public prototypes from it.

2 Background

Transfer Learning. We consider transfer learning where a publicly available pre-trained encoder $\hat{E}$ is used to extract features $\hat{\mathbf{X}}$ from a (private) dataset $D = (\mathbf{X}, \mathbf{y})$. Those features are then used to perform downstream classification by learning a function f maximizing $\Pr_{\mathbf{x},y\in D}[f(\hat{E}(\mathbf{x})) = y]$.

Differential Privacy. Differential privacy (DP) (Dwork et al. 2006) is a mathematical framework that provides privacy guarantees in ML by formalizing the intuition that a learning algorithm $\mathcal{A} : I \to S$, executed on two neighboring datasets D, D' that differ in only one data point,*i.e.,*, $D = D' \cup \{x\}$ (*add/remove DP*), will yield roughly the same output, *i.e.,* $\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(D') \in S] + \delta$ (*approximate DP*). In this inequality, ε is the privacy budget that specifies by how much the output is allowed to differ and δ is the probability of differing more. If $\delta = 0$, we refer to it as *pure DP*, a strictly stronger notion of privacy. We will also refer to zero-concentrated DP (zCDP) (Bun and Steinke 2016), which requires that $D_\alpha(\mathcal{A}(D)||\mathcal{A}(D')) \leq \xi + \rho\alpha \ \forall_{\alpha\in(1,\infty)}$, where D_α is the Rényi divergence of order α. zCDP is a relaxation of pure DP, but stricter than approximate DP. $(0, \rho)$-zCDP can also be expressed simply as ρ-zCDP. We provide more details on DP in Appendix A.1. The standard approach for learning ML models with DP guarantees is differentially private stochastic gradient descent (DPSGD) (Song, Chaudhuri, and Sarwate 2013; Abadi et al. 2016). DPSGD clips model gradients to a given norm to limit the impact of individual data points on the model updates and adds a controlled amount of Gaussian noise to implement formal privacy guarantees during training.

Exponential Mechanism. The exponential mechanism (McSherry and Talwar 2007) offers a way to implement pure DP guarantees. Given a set of possible outputs $\mathbf{X}'$, it samples an output $\mathbf{x}'$ according to some utility function u with probability $\Pr[\text{EM}_u(\mathbf{X}) = \hat{\mathbf{x}}] \propto \exp\left(\frac{\epsilon}{\Delta u} u(\mathbf{X}, \hat{\mathbf{x}})\right)$. This algorithm satisfies 2ϵ-DP. Appendix A.3 shows more details on the exponential mechanism and utility function.

Prototypical Networks. Prototypical networks (Snell, Swersky, and Zemel 2017) are used for few-shot classification, *i.e.,* they provide a way on adapting a classifier to new unseen classes with access only to a small number of data points from each new class. Their main components are a set of prototypes $\mathbf{p}_c \in \mathbb{R}^M$ and an embedding function $f_\phi : \mathbb{R}^D \to \mathcal{R}^M$. Each prototype for a class c is the mean of the embedded points belonging to that class, *i.e.,* $\mathbf{p}_c = \frac{1}{|\mathbf{X}_c|} \sum_{\mathbf{x}\in\mathbf{X}_c} f_\phi(\mathbf{x})$. Given a distance function $d : \mathbb{R}^M \times \mathbb{R}^M \to [0, \infty)$, the model classifies a point $\mathbf{x}$ based on its nearest prototype in the embedded space as $\hat{y}(\mathbf{x}) = \arg\min_c d(f_\phi(\mathbf{x}), \mathbf{p}_c)$.

DP Mean Estimation. Obtaining differentially private means $\mu = 1/n \sum_n \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^d$ is challenging in high

dimensions. A straightforward approach (Kamath and Ullman 2020) consists of clipping all samples to some ℓ_2 norm, adding noise scaled according to the clip norm and then reporting the noisy mean of the clipped samples. *FriendlyCore* (Tsfadia et al. 2022) is a framework for pre-processing the input data of private algorithms, such that the algorithms being executed on this pre-processed data need to be private only for relaxed conditions. It improves especially for the cases where the samples have a high ℓ_2 norm and high dimensionality d. The *CoinPress* algorithm (Biswas et al. 2020) estimates the mean iteratively, clipping the samples not w.r.t. the origin but to the estimated mean of the previous step. This approach is especially useful when the mean is far away from the origin and generally considered state-of-the-art for dimensionalities in the low thousands. Figure 10 shows that the straightforward approach outperforms all other methods given strong priors on the ℓ_2 norms of the samples. We provide more details in Appendix A.4.

3 Related Work

Private Transfer Learning. Standard approaches for DP transfer learning rely on the DPSGD algorithm to train a classifier on top of the representations output by a pre-trained encoder, and potentially also to privately update existing or added model parameters on the sensitive data (Yu et al. 2021; De et al. 2022; Mehta et al. 2022). Notably, there also exists an approach for transfer learning from few samples, DP-FiLM introduced by Tobaben et al. (2023). Such approaches have been shown effective, for loose DP guarantees (*i.e.,* large ε), yet suffer from severe utility drops in strong privacy regimes (*i.e.,* with small ε). This is because of the iterative nature of the DPSGD algorithm with multiple rounds of noise addition that negatively impact performance. To overcome these limitations, Mehta et al. (2023) proposed Differentially Private Least Squares (DP-LS), DP-Newton and DPSGD with Feature Covariance (DP-FC). DP-LS takes advantage of the closed form solution for least squares to avoid running many iterations of gradient descent. DP-Newton employs a second-order optimization to solve the smaller problem of transfer learning more efficiently. DP-FC integrates second order information by utilizing the covariance of the features without paying the composition cost of DP-Newton. All methods have three hyperparameters. In contrast to theirs, our method only has a single optional hyperparameter, does not rely on higher order optimization and utilizes parallel composition to solve each class independently in a single iteration, resulting in lower privacy costs especially for imbalanced datasets.

Leveraging Public Data for Private Training. Public data has, so far, been leveraged for privacy-preserving knowledge transfer to protect sensitive data (Papernot et al. 2017, 2018), to reduce the sample complexity within DP distribution learning (Bie, Kamath, and Singhal 2022; Ben-David et al. 2024), and for pre-training public encoders to then perform private transfer learning (Yu et al. 2021; Gu, Kamath, and Wu 2022; Tramèr, Kamath, and Carlini 2022; Ganesh et al. 2023; Hu et al. 2021; Houlsby et al. 2019; Li et al. 2023; Mehta et al. 2023). In a similar vein as previous work that determines the importance of public samples to private data (Ji and Elkan 2013), our approach goes beyond the latter and additionally leverages the public pre-training data of the encoder during the private transfer learning step by selecting public prototypes to represent our private classes.

Private Training on Unbalanced Datasets. DP has been shown to disproportionately harm utility for underrepresented sub-groups, *i.e.,* groups with fewer data points (Bagdasaryan, Poursaeed, and Shmatikov 2019; Suriyakumar et al. 2021). This is because the weak signal from these groups is more affected by the added noise. Additionally, the clipping operation in DPSGD changes the direction of the overall gradient, which adds a compounding bias over the runtime of the training, that disproportionally affects minority classes (Esipova et al. 2023). To mitigate this issue, Esipova et al. (2023) propose DPSGD-Global-Adapt, which clips only some gradients and instead scales most gradients, thus preserving the overall direction. The algorithm adaptively learns the clipping threshold, keeping the amount of clipped gradients low. Another approach is to add fairness through in- or post-processing (Jagielski et al. 2019), which trades off accuracy against fairness and requires additional privacy budget. In a non-private setting, solutions for improving utility of small subgroups include changing the sampling (Kubat, Matwin et al. 1997; Lewis and Catlett 1994; Ling and Li 1998), generating synthetic data for the minority classes (Chawla et al. 2002; Wang et al. 2018), or weighing the training loss (Cao et al. 2019). However, these approaches are not directly compatible with DP or incur additional privacy costs.

4 Differentially Private Prototyping

Setup and Assumptions. We aim at learning a private classifier based on a sensitive labeled dataset $D = (\mathbf{X}, \mathbf{y})$ with C different classes. We assume the availability of a standard public pre-trained vision encoder $\hat{M}$, such as DINO[1] or MAE[2] encoders, that return high-dimensional feature vectors for their input data points. Additionally, we assume the availability of a general purpose public dataset $\hat{D} = (\hat{\mathbf{X}}, \dots)$, such as ImageNet (Deng et al. 2009). Note that $\hat{D}$ can also be from a different distribution than D and $\hat{M}$'s pre-training data, as we show experimentally in Figure 7a, and does not require labels. In case of available labels for $\hat{D}$, we just discard them.

Overview. Our goal is to obtain private prototypes $\mathbf{p}_1, \dots \mathbf{p}_C$ that represent every class C from the private dataset D in the embedding space. To classify a new unseen data point $\mathbf{x}'$, we simply have to retrieve the most representative prototype and return its label. Concretely, we have to infer $\mathbf{x}'$ through the encoder $\hat{M}$, retrieve the prototype with the minimum distance in embedding space to $\mathbf{x}'$ and return its label as the prediction $y' = \min_{c \in C} d(\hat{M}(\mathbf{x}'), \mathbf{p}_c)$. We detail the general approach in Figure 1.

Note that if the private prototypes are obtained with DP

[1]https://github.com/facebookresearch/dinov2

[2]https://github.com/facebookresearch/mae

guarantees, using them for predictions will not incur additional privacy costs due to the DP post-processing guarantees. Hence, our DP prototypes can be publicly released, similar to privately trained ML models. We experimented with multiple ways for implementing DP prototypes and identified the two most promising approaches: `DPPL-Mean` generates a private prototype by calculating a DP mean on all data points of a given class in the embedding space. Our `DPPL-Public` takes advantage of the public dataset $\hat{D}$ and privately selects a data point from $\hat{D}$ to act as a prototype for each private class.

4.1 DPPL-Mean: Private Means

Intuition. Non-private prototypical networks (Snell, Swersky, and Zemel 2017) consist of two steps, namely the training of a projection layer at the output of the encoder and the estimation of the class prototypes. In the private setup, both these steps would depend on the private data and therefore each incur additional privacy costs. To keep privacy cost low, we forgo projection layer training, as we find it is unnecessary when given a strong pre-trained encoder (see Appendix C.6). Hence, for our private `DPPL-Mean`, we only implement the estimation of the prototypes without projection.

Non-Private Means. Given a training class c and corresponding samples $\mathbf{X}_c \in \mathbb{R}^{n_c \times d}$, the non-private prototype of each class is the mean of the embeddings $\frac{1}{n_c}\sum_{i=0}^{n_c} \hat{M}(\mathbf{x}_i)$

Our DPPL-Mean: Private Means. To privately estimate the means, we rely on the Gaussian Mechanism. We first clip each $\mathbf{x}_i \in \mathbb{R}^d$ to a ℓ_2 norm r. The estimate is then

$$\mathbf{p}_c = \mathcal{N}\left(\mathbf{0}, \frac{2r^2}{n_c^2 \rho}\right) + \frac{1}{n_c}\sum_{i=0}^{n_c} \hat{M}(clip_{\ell_2}(\mathbf{x}_i), r) \quad (1)$$

where ρ is the zCDP privacy budget. To improve the utility at strict privacy budgets, we include a single optional hyperparameter $k_{\text{pool}} \geq 1$, describing the kernel size of an average pooling layer before the mean estimation to reduce dimensionality, reducing the dimension from d to d/k_{pool}.

Privacy Analysis. The privacy analysis of `DPPL-Mean` follows the analysis of the Gaussian Mechanism. By clipping each sample to ℓ_2 norm of r we obtain $\Delta\mathbf{p}_c = 2r/n$, since the ℓ_2 distance between the previous mean and any new sample can be $2r$ at maximum and its influence diminishes with the number of samples n. We use parallel composition: each disjoint class computes a ρ-zCDP mean prototype, making the privacy cost ρ for the entire private dataset.

4.2 DPPL-Public: Privately Selecting Public Prototypes

Intuition. Our main idea for `DPPL-Public` is to leverage public data beyond the pre-training stage for learning a private classifier based on the sensitive data. Therefore, we privately select public prototypes for each training class, *i.e.,* a data point from the public dataset that represent the given class well.

Non-Private Selection. A good public prototype $\hat{\mathbf{x}}_c$ for a given training class c represents that class well in the embedding space of encoder $\hat{M}$. To select such a good prototype per class, we first calculate the embeddings $\mathbf{E} = \hat{M}(\mathbf{X})$ and $\hat{\mathbf{E}} = \hat{M}(\hat{\mathbf{X}})$ for the private and public data points, respectively. Then, based on the private labels $\mathbf{y}$, we split the embeddings of $\mathbf{X}$ in C subsets $\mathbf{E}_1, \ldots, \mathbf{E}_C$. Without any privacy considerations, a public prototype $\hat{\mathbf{x}}_c$ for class c could then be chosen as the data point that minimizes the average distance according to metric d, to all training data points $\mathbf{x}_i$ in class c as

$$\hat{\mathbf{x}}_c = \min_{\hat{\mathbf{x}} \in \hat{\mathbf{X}}} \frac{\sum_{i=0}^{|\mathbf{X}_c|} d(\hat{M}(\mathbf{x}_i), \hat{M}(\hat{\mathbf{x}}))}{|\mathbf{X}_c|}. \quad (2)$$

Our DPPL-Public: Private Selection. The previously described approach, however, does not take any privacy of the training data D into account. To perform public prototype selection with ε-DP guarantees, we rely on the exponential mechanism (McSherry and Talwar 2007). We use the cosine similarity and add $+1$ as our distance metric d, therefore obtaining a bounded and non-negative metric in $[0, 2]$. The utility function that indicates the goodness of each each public sample for a given class c is

$$u(\hat{\mathbf{x}}, c) = \sum_{i=0}^{|\mathbf{X}_c|} 1 + \frac{\hat{M}(\mathbf{x}_i) \cdot \hat{M}(\hat{\mathbf{x}})}{\|\hat{M}(\mathbf{x}_i)\| \|\hat{M}(\hat{\mathbf{x}})\|}. \quad (3)$$

To improve utility at strict privacy budgets, we include two optional hyperparameters $d_{\text{max}} \in (0, 2]$ and $d_{\text{min}} \in [0, d_{\text{max}})$, which clips the distances to $[d_{\text{min}}, d_{\text{max}}]$, reducing the sensitivity to $\Delta u = d_{\text{max}} - d_{\text{min}}$.

We detail the full algorithm for privately selecting public prototypes in Algorithm 1.

Privacy Analysis. Our proposed `DPPL-Public` fulfills ε-DP. We provide a sketch of the full proof from Appendix D.1 here. We first note that $\Delta u = d_{\text{max}} - d_{\text{min}}$. As mentioned above, we add +1 to each cosine similarity, which is therefore non-negative. Therefore, u is positively monotonic w.r.t. $\mathbf{X}$. The exponential mechanism with $\Pr[\text{EM}(\mathbf{X}) = \hat{\mathbf{x}}] \propto \exp \frac{\varepsilon u(\mathbf{X}, \hat{\mathbf{x}})}{\Delta u}$ is ε-DP if u is monotonic w.r.t. the private data $\mathbf{X}$ (McSherry and Talwar 2007). Therefore, executing `DPPL-Public` on a single class yields ε-DP. Additionally, since we calculate prototypes per-class and the classes are non-overlapping, parallel composition applies, *i.e.,*, the total privacy costs are also ε-DP.

Algorithm 1: Privately Select Public Prototypes

Input: Private dataset $D = (\mathbf{X}, \mathbf{y})$ with C classes, privacy budget ϵ, public pre-trained encoder $\hat{M}$, public dataset $\hat{D} = (\hat{\mathbf{X}}, \ldots)$, hyperparameters d_{max},d_{min}

Output: Prototypes $\mathbf{P} = \{\mathbf{p}_c \in \hat{D} | c \in \mathbf{y}\}$

Function `SelectPublicPrototypes()`**:**

$\mathbf{E} \leftarrow M(\mathbf{X})$ $\hat{\mathbf{E}} \leftarrow M(\hat{\mathbf{X}})$ **foreach** *class* $c \in C$ **do**

$\mathbf{E}_c \leftarrow \{\mathbf{e}_i \in \mathbf{E} | y_i = c\}$

$u_c(\hat{\mathbf{x}}_i) = \sum_{\mathbf{e} \in \mathbf{E}_c} \text{clip}(1 + \frac{\mathbf{e} \cdot \hat{\mathbf{e}}_i}{\mathbf{e}\hat{\mathbf{e}}_i}, d_{\text{max}}, d_{\text{min}})$;

$\mathbf{p}_c \propto \exp\left(\frac{\epsilon u_c}{d_{\text{max}} - d_{\text{min}}}\right)$

return $\{\mathbf{p}_c | c \in \mathbf{y}\}$;

5 Empirical Evaluation

Experiment Setup. We experiment with CIFAR10 (Krizhevsky 2009), CIFAR100 (Krizhevsky 2009), STL10 (Coates, Ng, and Lee 2011) and FOOD101 (Bossard, Guillaumin, and Van Gool 2014) as private datasets. From these datasets, we construct exponentially long-tailed imbalanced datasets with various imbalance ratios (IRs), the ratio between the number of samples in the largest and smallest class, following Cui et al. (2019) and Cao et al. (2019).[3] We compare our DP prototypes on the features obtained from three vision transformers based on the original architecture from Dosovitskiy et al. (2020) Vit-B-16 (Singh et al. 2022), namely ViT-L-16 (Oquab et al. 2023), ViT-H-14 (Singh et al. 2022) and a ResNet-50 (Caron et al. 2021; He et al. 2016). All models, except for ViT-L-16, which is trained on LVD-142M introduced by Oquab et al. (2023), are trained on ImageNet-1K (Deng et al. 2009). For `DPPL-Public`, we use the 64×64 downscaled version of ImageNet-1K upscaled to between 128×128 and 512×512 depending on the encoder. We evaluate our methods on the standard *balanced* test set. This corresponds to reporting a *balanced accuracy* for the imbalanced setups. A full description of our experimental setup is provided in Appendix B.

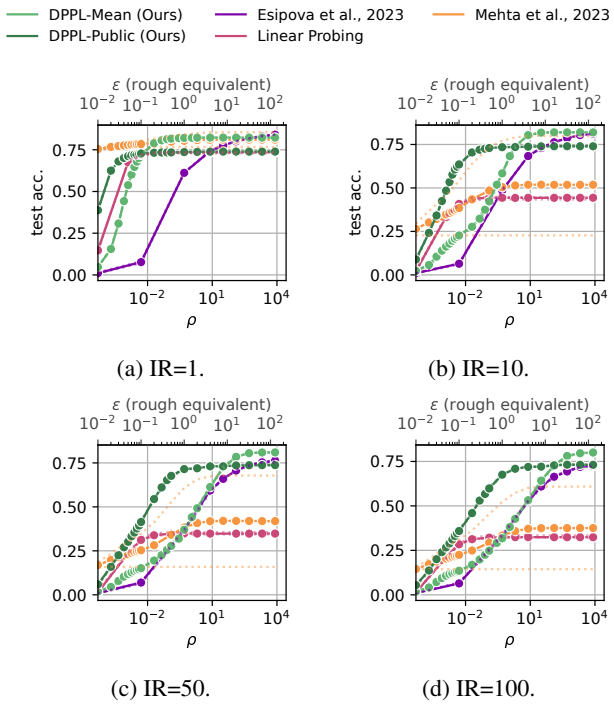
Baselines. For the baseline comparisons we compare to standard linear probing with DPSGD, as it's a common way of DP transfer-learning. Furthermore, we compare to DP-LS from Mehta et al. (2023) which is the current state-of-the-art for DP transfer learning across all privacy regimes and to DPSGD-Global-Adapt from Esipova et al. (2023) as it is specifically designed for for training on imbalanced datasets. We outline in Appendix E.4 the experimentally-supported reasons against including DP-FC and DP-FiLM.

Comparing Results over Different Notions of DP. Since we are comparing our new proposed methods that implement *pure DP* or *pure* ρ-zCDP guarantees against other baselines that also implement zCDP, we convert all privacy guarantees to ρ-zCDP. We also present a pure-DP ε equivalent by inverting the $\rho = \epsilon^2/2$ conversion from pure DP to zCDP. However, this does not imply that these algorithms fulfil pure DP. We detail all comparison implementations and conversion theorems used in Appendix D.2.

5.1 DP Prototypes: High Utility in High Privacy and Extreme Imbalance

We evaluate our DP prototypes at different privacy regimes in the range corresponding to standard approximate DP (Dwork et al. 2006) of $0.01 < \varepsilon < 100$ and under different IRs. In Figure 2, we benchmark our methods against DP-LS, the current state-of-the-art method for private transfer learning by Mehta et al. (2023), DPSGD-Global-Adapt by Esipova et al. (2023), a DP method deliberately designed to achieve high utility under imbalancedness of the private data, and standard DP linear probing on CIFAR100 with ViT-H-14. Our results highlight that over all levels of IRs larger than 1, our `DPPL-Public` significantly outperforms linear probing in all privacy regimes. Additionally `DPPL-Public` yields strong performance already at very low ε, such as $\varepsilon = 0.1$ for IR=1. As data becomes more imbalanced, all methods require larger privacy budgets to yield similarly high performance. Further, our results indicate that our `DPPL-Mean` method underperforms `DPPL-Public` and DP linear probing for low ε. We find that this results from the noise added during the mean calculation leading diverging estimations. We provide further detail on the cause in Appendix C.2. Yet, we observe that at higher epsilon, `DPPL-Mean` outperforms `DPPL-Public`. This suggests that the most beneficial way for leveraging DP prototypes might be an adaptive method where `DPPL-Public` is chosen for high performance in the high privacy regimes and `DPPL-Mean` can further boost performance for larger ε.

We further assess whether the observed trend holds over different datasets. Therefore, we depict the results of our methods vs. the baselines for different datasets and the ViT-H-14 encoder under the most challenging setup with IR$=$ 100 in Figure 3. The observed trends are indeed consistent between all datasets. We provide full results over all datasets and IRs in Appendix E.1.

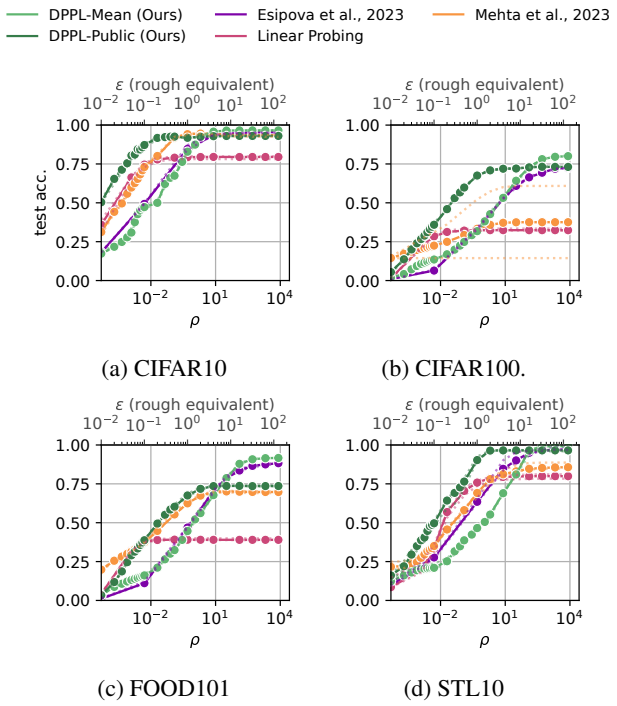


(a) IR=1. (b) IR=10.

(c) IR=50. (d) IR=100.

Figure 2: **Comparing against baselines on CIFAR100.** We present the results of our methods vs. state-of-the-art methods (DP-LS and DPSGD-Global-Adapt) on the CIFAR100 dataset using ViT-H-14 under different IRs. `DPPL-Public` uses ImageNet as public data. Dotted lines represent the upper/lower quantiles. Similar results for CIFAR10, Food101, and STL10 are presented in Appendix E.1.

[3] Concretely, the number of samples in each class decreases exponentially with a factor of $n(c) = \exp(-c\lambda)$, where $\lambda = \log(\text{IR})/C$. For the balanced case(IR $= 1$), $\lambda = \log(1)/C = 0$ and therefore $n(c) = 1\,\forall_c$. We detail the imbalancing process and depict the effect on the resulting absolute class sizes per dataset further in Appendix B.2.



(a) CIFAR10 (b) CIFAR100.

(c) FOOD101 (d) STL10

Figure 3: **DP Prototypes on various imbalanced datasets.** We present the balanced test accuracy for CIFAR10, CIFAR100, FOOD101 and STL10 at an imbalance ratio of 100 on ViT-H-14, using ImageNet as public data for `DPPL-Public`. We compare to standard Linear Probing with DP-SGD. We plot the mean test accuracy over multiple runs and represent the upper/lower quantiles by the dotted lines. Appendix E.1 shows more results.

5.2 DP Prototypes Improve over State-of-the-Art Baselines in Imbalanced Setups

Our results in Figures 2 and 3 highlight that while in the balanced setup, Mehta et al. (2023) outperforms the other methods, our DP prototypes outperform all other methods the more imbalanced the setup becomes. In particular `DPPL-Public` outperforms in high privacy regimes (*i.e.,* for small ε), while `DPPL-Mean` is better in lower privacy regimes, mostly outperforming even DPSGD-Global-Adapt.

The advantage of our methods against the baselines become even more obvious as we do not consider the accuracy over the entire balanced test set (equivalent to balanced accuracy), but look specifically at accuracy on minority classes, see Figure 4. Therefore, we take the smallest 25% of training classes in terms of number of their training data points and measure their accuracy on a balanced test set consisting of only those classes. Our results for CIFAR100 on ViT-H-14 under IR$=50$ in Figure 4 highlight that our DP prototypes significantly outperform all baselines. Full results for the minority classes are depicted in Appendix E.2.

5.3 Understanding the Success of DP Prototypes

To better understand the success of our DP prototypes, we perform various ablations. The full set of ablations and their results is presented in Appendix C.

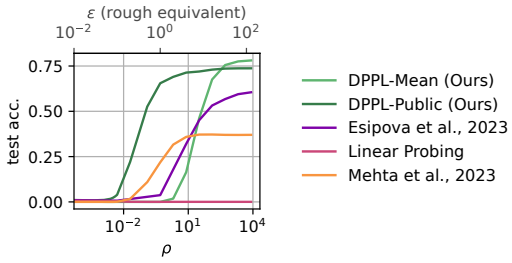
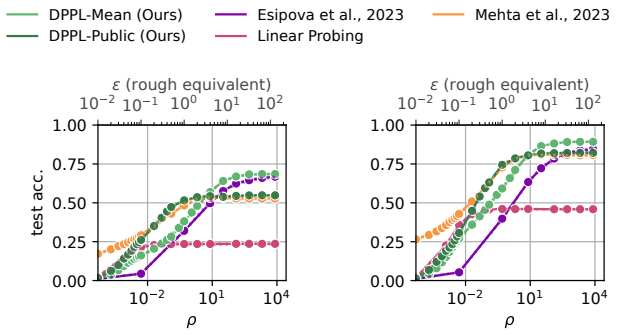


Figure 4: **Accuracies of the minority classes.** We depict the test accuracy on CIFAR100 with ViT-H-14 embeddings for the minority classes (smallest 25% of classes) at IR $= 50$.

Effect of the Publicly Pre-trained Encoder. We first assess the impact of the encoder used as a feature extractor. Therefore, we apply our method and the baselines with different encoder architectures. Our results in Figure 5 highlight that the encoder performance impacts all methods alike. In particular, we observe that all methods obtain better results with stronger encoders. For example, the ViT-H-14 yields to significantly higher private prediction accuracy that the much smaller ViT-B-16. Additionally, none of the methods yields satisfactory results using the ResNet50.

Impact of the Projection Layer. We evaluate whether a projection layer, usually part of a prototypical network, can increase the utility. We present our results in Figure 6 for CIFAR100 on ViT-H-14, using ImageNet as public for `DPPL-Public`. They highlight that `DPPL-Public` does not benefit from the projection. Even non-privately ($\epsilon = \infty$), the accuracy of `DPPL-Public` with projection is 72.6% and without projection is 74.0%, showing that it is not just the decreased privacy budget for the sampling that reduces the utility, but a fundamental misalignment between how the projection is trained and the public prototyping. We observe the same effect for `DPPL-Mean` and conclude that, although this limits adaptability (see Appendix F.2), with a strong enough pre-trained encoder, it is sufficient for `DPPL` to estimate prototypes without projection.

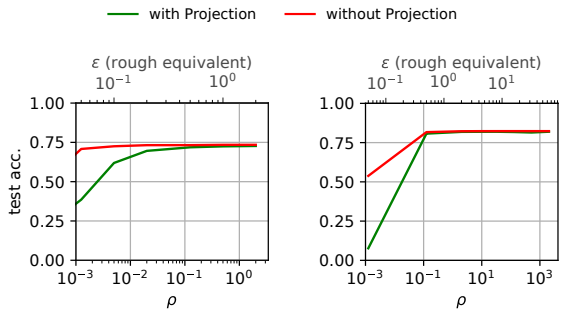
Improving through Multiple Per-Class Prototypes. We experiment with extending our `DPPL-Public` beyond a single per-class prototype—the common standard for prototypical networks. Therefore, we introduce the variation `DPPL-PublicK` which selects the top-K public prototypes per class. We extend the algorithm from Gillenwater et al. (2022) to sample multiple prototypes jointly using the exponential mechanism. Then, we classify based on the mean distance to each class's prototype. Our results in Figure 6 show that `DPPL-PublicK`'s privacy-utility trade-offs are between `DPPL-Mean` and `DPPL-Public`, indicating that the private means can be —to a certain degree— approximated by multiple public prototypes. `DPPL-PublicK` could therefore replace `DPPL-Mean` in cases of high dimensionality, where a mean estimation is not feasible. We provide more details, privacy proof, and a full set of results in Appendix C.3.



(a) ViT-B-16. (b) ViT-L-16.

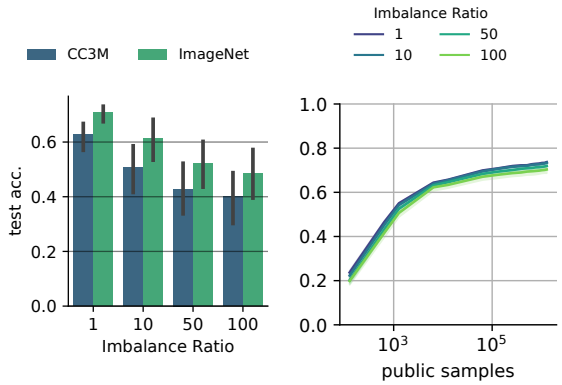
(c) ViT-H-14. (d) ResNet-50.

Figure 5: **Choice of Encoder.** We report the test accuracy for our methods and the baselines for CIFAR100, using ImageNet as public data for DPPL-Public. We observe that the success of all methods depends on the quality of the underlying encoder.



(a) DPPL-Public. (b) DPPL-Mean.

Figure 6: **Impact of the Projection Layer.**



(a) Public Dataset. (b) Public Data Size.

Figure 7: **Impact of the Public Data.**

Impact of the Public Data for Prototype Selection. To assess whether the public data for prototype estimation needs to be from the same distribution as the one used to pre-train the encoder, we conduct experiments with a different public dataset. We evaluate DPPL-Public using 2,298,112 samples from CC3M introduced by Sharma et al. (2018) as public data instead of ImageNet which is used to pre-train the encoder. We show the relation between the accuracy using ImageNet as public data and CC3M in Figure 7a for CIFAR100 on ViT-H-14. We find that DPPL-Public works well with both public datasets, highlighting the flexibility of our approach. Still, we observe that ImageNet yields better results which suggests that it is particularly beneficial to leverage the public data already available for pre-training also in the private transfer learning step.

Size of the Public Dataset for DPPL-Public. We also evaluate the success of DPPL-Public for different sizes of the public dataset that the public prototypes can be chosen from. Therefore, we randomly draw subsets of different sizes from ImageNet and apply DPPL-Public. Our results for CIFAR100 (100 classes) and ViT-H-14 in Figure 7b indicate that with growing public dataset size, our method's success increases. Figure 9 shows that tasks less similar to the pretraining data are more sensitive to dataset size. Note that even for public dataset of more than one million images, the selection of our public prototypes for 100 classes (*i.e.,* the "training") takes 34.3 seconds on a single GPU as we depict in Appendix E.3. For 10 classes (*e.g.,* CIFAR10), it takes 5 seconds. Hence, choosing a larger public dataset does not represent a practical limitation.

6 Conclusions and Future Work

We propose DPPL as a novel alternative to private fine-tuning with DP. DPPL builds DP prototypes on top of features extracted by a publicly pre-trained encoder, that can be later used as a classifier. The prototypes can be obtained without iterative noise addition and yield high utility even in high-privacy regimes, with few private training data points, and in unbalanced training setups. We show that we can further boost performance of our DP prototypes by leveraging the public data beyond training of the encoder and using them to draw the public prototypes from (DPPL-Public). Future work at improving utility of high-dimensional DP mean estimation will benefit our DPPL-Mean, which can, in the future, serve as an additional benchmark for private mean estimation algorithms.

Acknowledgements

This research project was supported by the Google Safety Engineering Center. The project was funded by the Initiative and Networking Fund of the Helmholtz Association in

the framework of the Helmholtz AI project call under the name "PAFMIM", funding number ZT-I-PF-5-227. Responsibility for the content of this publication lies with the authors. We thank the Helmholtz Information and Data Science Academy (HIDA) for supporting this research through the Helmholtz Visiting Researcher Grant.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. Vienna Austria: ACM. ISBN 978-1-4503-4139-4.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. *Advances in Neural Information Processing Systems*, 32.

Ben-David, S.; Bie, A.; Canonne, C. L.; Kamath, G.; and Singhal, V. 2024. Private distribution learning with public data: The view from sample compression. *Advances in Neural Information Processing Systems*, 36.

Bie, A.; Kamath, G.; and Singhal, V. 2022. Private estimation with public data. *Advances in neural information processing systems*, 35: 18653–18666.

Biswas, S.; Dong, Y.; Kamath, G.; and Ullman, J. 2020. CoinPress: Practical Private Mean and Covariance Estimation. In *Advances in Neural Information Processing Systems*, volume 33, 14475–14485. Curran Associates, Inc.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101 – Mining Discriminative Components with Random Forests. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 446–461. Cham: Springer International Publishing. ISBN 978-3-319-10599-4.

Bu, Z.; Gopi, S.; Kulkarni, J.; Lee, Y. T.; Shen, H.; and Tantipongpipat, U. 2021. Fast and memory efficient differentially private-sgd via jl projections. *Advances in Neural Information Processing Systems*, 34: 19680–19691.

Bu, Z.; Mao, J.; and Xu, S. 2022. Scalable and efficient training of large convolutional neural networks with differential privacy. *Advances in Neural Information Processing Systems*, 35: 38305–38318.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106: 249–259.

Bun, M.; and Steinke, T. 2016. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In Hirt, M.; and Smith, A., eds., *Theory of Cryptography*, Lecture Notes in Computer Science, 635–658. Berlin, Heidelberg: Springer. ISBN 978-3-662-53641-4.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022. Membership Inference Attacks from First Principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1519–1519. IEEE Computer Society.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Coates, A.; Ng, A.; and Lee, H. 2011. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 215–223. JMLR Workshop and Conference Proceedings. ISSN: 1938-7228.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. 9268–9277.

De, S.; Berrada, L.; Hayes, J.; Smith, S. L.; and Balle, B. 2022. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. ISSN: 1063-6919.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In Halevi, S.; and Rabin, T., eds., *Theory of Cryptography*, Lecture Notes in Computer Science, 265–284. Berlin, Heidelberg: Springer. ISBN 978-3-540-32732-5.

Esipova, M. S.; Ghomi, A. A.; Luo, Y.; and Cresswell, J. C. 2023. Disparate impact in differential privacy from gradient misalignment. In *The eleventh international conference on learning representations*.

Feldman, V. 2020. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 954–959. Chicago IL USA: ACM. ISBN 978-1-4503-6979-4.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1322–1333.

Ganesh, A.; Haghifam, M.; Nasr, M.; Oh, S.; Steinke, T.; Thakkar, O.; Thakurta, A. G.; and Wang, L. 2023. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, 10611–10627. PMLR.

Gillenwater, J.; Joseph, M.; Munoz, A.; and Diaz, M. R. 2022. A Joint Exponential Mechanism For Differentially Private Top-k. In *Proceedings of the 39th International Conference on Machine Learning*, 7570–7582. PMLR. ISSN: 2640-3498.

Gu, X.; Kamath, G.; and Wu, S. 2022. Choosing Public Datasets for Private Machine Learning via Gradient Subspace Distance. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. 770–778.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi Malvajerdi, S.; and Ullman, J. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning*, 3000–3008. PMLR. ISSN: 2640-3498.

Ji, Z.; and Elkan, C. 2013. Differential privacy based on importance weighting. *Machine Learning*, 93(1): 163–183.

Kamath, G.; and Ullman, J. 2020. A Primer on Private Statistics. ArXiv:2005.00010 [cs, math, stat].

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Kubat, M.; Matwin, S.; et al. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, 179. Citeseer.

Lee, J.; and Kifer, D. 2021. Scaling up differentially private deep learning with fast per-example gradient clipping. *Proceedings on Privacy Enhancing Technologies*.

Lewis, D. D.; and Catlett, J. 1994. Heterogeneous Uncertainty Sampling for Supervised Learning. In Cohen, W. W.; and Hirsh, H., eds., *Machine Learning Proceedings 1994*, 148–156. San Francisco (CA): Morgan Kaufmann. ISBN 978-1-55860-335-6.

Li, Y.; Tsai, Y.-L.; Yu, C.-M.; Chen, P.-Y.; and Ren, X. 2023. Exploring the benefits of visual prompting in differential privacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5158–5167.

Ling, C. X.; and Li, C. 1998. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, 73–79.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2537–2546.

McSherry, F.; and Talwar, K. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103. IEEE.

Mehta, H.; Krichene, W.; Thakurta, A. G.; Kurakin, A.; and Cutkosky, A. 2023. Differentially private image classification from features. *Transactions on Machine Learning Research*. ArXiv:2211.13403 [cs].

Mehta, H.; Thakurta, A.; Kurakin, A.; and Cutkosky, A. 2022. Large Scale Transfer Learning for Differentially Private Image Classification. ArXiv:2205.02973 [cs].

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. ArXiv:2304.07193 [cs].

Papernot, N.; Abadi, M.; Úlfar Erlingsson; Goodfellow, I.; and Talwar, K. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *Proceedings of the International Conference on Learning Representations*.

Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Úlfar Erlingsson. 2018. Scalable Private Learning with PATE. In *International Conference on Learning Representations (ICLR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reed, W. J. 2001. The Pareto, Zipf and other power laws. *Economics letters*, 74(1): 15–19.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565. Melbourne, Australia: Association for Computational Linguistics.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Singh, M.; Gustafson, L.; Adcock, A.; de Freitas Reis, V.; Gedik, B.; Kosaraju, R. P.; Mahajan, D.; Girshick, R.; Dollár, P.; and van der Maaten, L. 2022. Revisiting Weakly Supervised Pre-Training of Visual Perception Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 804–814.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Song, S.; Chaudhuri, K.; and Sarwate, A. D. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, 245–248. IEEE.

Subramani, P.; Vadivelu, N.; and Kamath, G. 2021. Enabling fast differentially private sgd via just-in-time compilation and vectorization. *Advances in Neural Information Processing Systems*, 34: 26409–26421.

Suriyakumar, V. M.; Papernot, N.; Goldenberg, A.; and Ghassemi, M. 2021. Chasing your long tails: Differentially private prediction in health care settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 723–734.

Tobaben, M.; Shysheya, A.; Bronskill, J. F.; Paverd, A.; Tople, S.; Zanella-Beguelin, S.; Turner, R. E.; and Honkela, A. 2023. On the Efficacy of Differentially Private Few-shot Image Classification. *Transactions on Machine Learning Research*.

Tramèr, F.; Kamath, G.; and Carlini, N. 2022. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*.

Tsfadia, E.; Cohen, E.; Kaplan, H.; Mansour, Y.; and Stemmer, U. 2022. FriendlyCore: Practical Differentially Private Aggregation. In *Proceedings of the 39th International Conference on Machine Learning*, 21828–21863. PMLR. ISSN: 2640-3498.

Wang, Y.-X.; Girshick, R.; Hebert, M.; and Hariharan, B. 2018. Low-Shot Learning from Imaginary Data. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7278–7286. ISSN: 2575-7075.

Yu, D.; Naik, S.; Backurs, A.; Gopi, S.; Inan, H. A.; Kamath, G.; Kulkarni, J.; Lee, Y. T.; Manoel, A.; Wutschitz, L.; et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.