COMPACTDP: CATEGORY-AWARE FEATURE COMPACTNESS FOR DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid growth of AI models raises critical privacy concerns due to their tendency to memorize training data, making them vulnerable to extraction and membership inference attacks (MIAs). Traditional privacy-preserving methods like DP-SGD often degrade model utility and exacerbate accuracy disparities across sub-populations, limiting their applicability in sensitive fields. We observe that dense intra-class feature distributions inherently reduce privacy risks by smoothing probability density functions (PDFs), which diminishes the influence of individual training samples and lowers memorization. Leveraging this insight, we propose *Category-Aware Compactness Differential Privacy* (CompactDP), a framework that directly addresses the root cause of privacy leakage—sparse, high-dimensional features—via feature contraction rather than relying solely on gradient noise. CompactDP achieves a superior privacy-utility-fairness trade-off, significantly outperforming state-of-the-art methods. On CIFAR10, it attains 95.6% accuracy while limiting MIA risk to 0.43. Extensive experiments on FashionMNIST and MedicalMNIST further validate its state-of-the-art performance across diverse metrics. By integrating feature reconstruction with differential privacy, our framework provides a principled and efficient solution for privacy-preserving deep learning in critical domains such as healthcare and finance.

1 INTRODUCTION

The emergence of billion-parameter neural networks has revolutionized machine learning, delivering state-of-the-art performance across diverse domains (Zhai et al., 2022). However, these models exhibit concerning memorization capabilities (Zhang et al., 2021), creating significant privacy risks through extraction attacks (Carlini et al., 2021) and membership inference. This vulnerability is particularly critical in sensitive sectors such as healthcare and finance, where models trained on private data must maintain rigorous privacy guarantees without compromising utility. Differential Privacy (DP) (Dwork et al., 2006) has emerged as the gold standard for privacy-preserving machine learning. Nevertheless, the practical implementation of DP in deep learning, particularly through Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016), faces several fundamental limitations. Current approaches suffer from three critical shortcomings: (1) Gradient perturbation in large models with high-dimensional features disproportionately degrades utility with accuracy drop at most 5+% under $(\epsilon = 1, \delta = 1e - 5)$ settings; (2) Standard DP mechanisms apply uniform protection across all classes, disregarding inherent class-wise privacy leak risks; and (3) DP mechanisms tend to exacerbate class fairness disparities in imbalanced datasets (Bagdasaryan et al., 2019).

Motivated by the privacy-utility-fairness trilemma, we establish a fundamental connection between class-wise feature-space PDF compactness and privacy vulnerability. In Fig. 1, we illustrate the fundamental mechanism of our feature space contraction approach. The transformation process systematically draws sparse samples from low-probability regions toward high-density areas within each class distribution, resulting in significantly more compact class-wise PDFs. As shown in the left subfigure, initially dispersed samples (represented by cool colors) undergo a contraction process that redistributes them toward the dense core regions (warm colors) of their respective class distributions. This geometric transformation reduces the effective diameter of each class cluster while preserving the inherent manifold structure. The right subfigure demonstrates the final contracted state, where each class forms a compact, well-defined PDF with minimal peripheral dispersion. This contraction

mechanism directly enhances privacy protection by reducing the presence of outlier samples that are particularly vulnerable to membership inference attacks. The compactified feature distributions minimize the surface area exposed to potential adversaries while maintaining the discriminative power necessary for accurate classification. The resulting geometric configuration formalizes privacy preservation through intrinsic feature space optimization rather than external noise injection. Our experiments on CIFAR-10 reveal that *privacy risk scales super-linearly with class feature space PDF diameter* d_c: classes with sparse PDF distributions (e.g., *Bird*) exhibit higher empirical leakage than compact clusters (e.g., *Automobile*). This class-wise PDF perspective yields a critical insight: strategically contracting feature diameters can simultaneously enhance privacy, utility, and fairness.

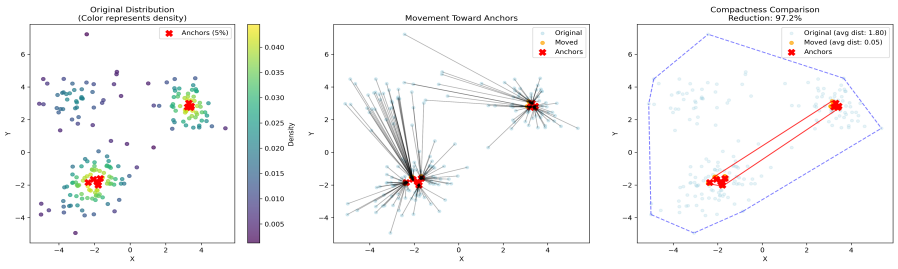


Figure 1: Illustration of feature space contraction through density-based compaction. Low-density peripheral samples are progressively drawn toward high-density core regions, reducing class-wise PDF diameters and minimizing privacy leakage. The left panel shows initial sparse distributions, while the right panel demonstrates the resulting compacted feature clusters with reduced surface area in low dimensional manifold space and enhanced privacy protection.

The PDF characteristics of category features, including distribution, diameter, and dimensionality, serve as critical variables in privacy protection. By strategically reshaping data distributions through feature contraction, we achieve stronger privacy guarantees without sacrificing model utility, thereby breaking the traditional privacy-performance trade-off. This work introduces a new paradigm for AI system design: shifting privacy protection from passive noise injection to active optimization of data feature structures. This work makes three key contributions: (1) Establishes CompactDP, a novel formalization of feature-space reconstruction in privacy analysis with rigorous theoretical guarantees quantifying privacy amplification via feature contraction; (2) Introduces an adaptive mechanism that strategically contracts PDF diameters proportional to vulnerability, reducing leakage disparities while maintaining discriminative power; (3) Validated across benchmarks and architectures, achieves state-of-the-art privacy and utility guarantees without accuracy loss, enabling stronger formal guarantees, higher utility, and more equitable protection than standard DP approaches.

2 Related Works

The connection between feature space PDF and privacy vulnerability has gained increasing attention. (Sanyal et al., 2022) revealed that sparse class distributions heighten membership inference risks, while (Berrada et al., 2023) demonstrated that feature distance distributions correlate with empirical leakage, a finding verified in our experiments. Subsequent work exploited this insight through regularization (Farrand et al., 2020) and outlier suppression (Bagdasaryan et al., 2019), but these approaches lacked theoretical grounding. (Hardt & Talwar, 2012) first linked global sensitivity to feature distances but required impractical noise levels. Our work bridges these critical gaps by establishing the first theoretical framework for *class-conditional privacy allocation*, where contraction parameters dynamically adapt to local density $\rho_c(h)$, enabling precise mitigation of category-specific vulnerabilities without utility degradation. More related works can be found in Appendix. K.

3 THEORETICAL FOUNDATIONS

3.1 PRELIMINARY

Given a backbone network $f_\theta(\cdot)$ and input samples $\mathbf{x}_i$, we extract features:

$$\mathbf{z}_i = f_\theta(\mathbf{x}_i) \in \mathbb{R}^d \tag{1}$$

where d is the feature dimension (e.g., 768 for ViT-B/16).

For each class c, we estimate the PDF using kernel density estimation in a learnable network:

$$p_c(\mathbf{z}) = \frac{1}{n_c} \sum_{j:y_j=c} K\left(\frac{\|\mathbf{z} - \mathbf{z}_j\|}{h}\right) \tag{2}$$

We then select anchor points $\mathcal{A}_c$ as the top $\gamma\%$ of samples with highest PDF:

$$\mathcal{A}_c = \{\mathbf{z}_j : y_j = c, p_c(\mathbf{z}_j) \geq Q_{1-\gamma}\left(\{p_c(\mathbf{z}_k)\}_{k:y_k=c}\right)\} \tag{3}$$

where $Q_{1-\gamma}$ denotes the $(1-\gamma)$-quantile.

We define a class-wise feature contraction function $g_\phi(\cdot)$ implemented as a neural network:

$$\hat{\mathbf{z}}_i = g_\phi(\mathbf{z}_i) \tag{4}$$

The network is trained to minimize the following objective:

$$\mathcal{L} = \underbrace{\sum_{i=1}^{N} \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|^2}_{\text{Reconstruction loss}} + \lambda \underbrace{\sum_{i=1}^{N} \min_{\mathbf{a}\in\mathcal{A}_{y_i}} \|\hat{\mathbf{z}}_i - \mathbf{a}\|^2}_{\text{Compactness term}} \tag{5}$$

where λ controls the trade-off between reconstruction accuracy and feature compactness. The whole framework is depicted in Fig. 2. Our focus is to train the feature re-construction or contraction network parameterized by ϕ. The last classification layer can be implemented by standard DP-SGD method. To achieve intra-class contraction, in the implementation we design a class-wise PDF contraction loss function to replace the above compactness term:

$$\mathcal{L}_{compact} = -\sum_{\hat{\mathbf{z}}_i \in D'} \mathbb{1}_{\{y_b=t\}} K_h(\hat{\mathbf{z}}_i - \mathbf{a}) \tag{6}$$

where $\mathbb{1}_{\{y_b=t\}}$ is an indicator function selecting references points from the same class as $\mathbf{a}$ in the contracted feature space D'. t is the class to be contracted. $K_h(\cdot)$ is a kernel function with bandwidth h. The logarithmic transformation of probability values stabilizes training and prevents numerical underflow during backpropagation. This loss function encourages compact feature clusters within classes.

3.2 PROBLEM SETUP

Consider a dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ with samples $\mathbf{x}_i \in \mathbb{R}^d$ and labels $y_i \in \{1, \ldots, C\}$. Let $g_\phi : \mathbb{R}^d \to \mathbb{R}^p$ be a feature contractor. For each class c, define:

- Class-wise features: $\mathcal{F}_c = \{g_\phi(\mathbf{z}_i) : y_i = c\}$
- Class diameter: $d_c = \max_{\mathbf{z}_i, \mathbf{z}_j \in \mathcal{F}_c} \|\mathbf{z}_i - \mathbf{z}_j\|_2$
- Class-conditional PDF: $p_c(\mathbf{z}) = \frac{1}{|\mathcal{F}_c|} \sum_{\mathbf{z}_i \in \mathcal{F}_c} K_h(\mathbf{z} - \mathbf{z}_i)$

where $K_h(\cdot) = \frac{1}{h^p} K\left(\frac{\cdot}{h}\right)$ is an L-Lipschitz kernel and h is the bandwidth. After *class-wise feature contraction*, we obtain contracted features $\mathcal{F}'_c$ with diameters $d'_c \ll d_c$ and PDFs $p'_c(\mathbf{z})$.

Definition 1 (Feature Space Vulnerability Measures)**.** *The privacy vulnerability of class c is characterized by:*

$$\textit{Diameter Contractor Ratio:} \quad \eta_c = d'_c / d_c \tag{7}$$

$$\textit{Local Density:} \quad \rho_c(h) = \mathbb{E}_{\mathbf{z}\sim p_c}[p_c(\mathbf{z}; h)] \tag{8}$$

$$\textit{Contraction Factor:} \quad \kappa_c = 1 - \eta_c \tag{9}$$

Smaller η_c indicates stronger contraction and lower vulnerability.

Definition 2 (Sensitivity under Contraction). *The class-wise L_2-sensitivity of the PDF mechanism is:*

$$\Delta_c = \sup_{\mathcal{D}\sim\mathcal{D}'} \|p_c - p'_c\|_2 \leq L \cdot \frac{d_c}{h^p|\mathcal{F}_c|} \tag{10}$$

where $\mathcal{D} \sim \mathcal{D}'$ denotes adjacent datasets differing in one sample.

Definition 3 (Category Feature Compactness Rényi DP (CompactDP)). *A mechanism $\mathcal{M}$ satisfies $(\alpha, \rho, \boldsymbol{\eta})$-CompactDP if for all adjacent datasets $\mathcal{D} \sim \mathcal{D}'$ and $\alpha > 1$:*

$$D_\alpha\left(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')\right) \leq \rho \cdot \prod_{c=1}^{C} \eta_c^{\alpha-1} \tag{11}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_C)$ is the class-wise contraction vector.

Remark 1. *CompactDP provides class-adaptive privacy amplification:*

- *For $\eta_c < 1$ (feature contraction), we achieve super-exponential amplification*
- *The product structure $\prod \eta_c^{\alpha-1}$ accounts for cross-class vulnerability*
- *Standard RDP is recovered when $\eta_c = 1\ \forall c$ (no contraction)*

This formalizes our core thesis: compact feature distributions intrinsically enhance privacy.

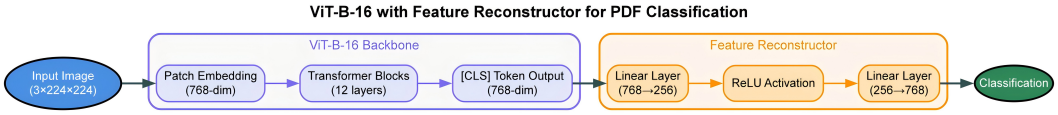


Figure 2: The whole framework contains the frozen backbone, a feature re-construction layer and a classification layer. Our focus is to train the feature re-construction network parameterized by ϕ. The classification layer can be implemented by standard DP-SGD method and integrated with our framework.

Theorem 1 (Global Feature Contraction Theorem). *Given a feature transformation $g_\phi : \mathbb{R}^d \to \mathbb{R}^p$ that contracts feature diameters from d_1 to $d_2 = \eta d_1$ with $\eta < 1$, and an L-Lipschitz kernel K_h, the following hold for class-conditional PDF mechanisms:*

1. ***Sensitivity Reduction**:*
$$\Delta_2 = \eta \Delta_1 \tag{12}$$
where $\Delta = \sup_{\mathcal{D}\sim\mathcal{D}'} \|p_c - p'_c\|_2$ is the L_2-sensitivity.

2. ***Privacy Amplification under RDP**: For the Gaussian mechanism $\mathcal{M}(D) = p_c(D) + \mathcal{N}(0, \sigma^2 I)$,*
$$(\alpha, \rho)\text{-RDP} \implies (\alpha, \rho\eta^2)\text{-RDP after contraction} \tag{13}$$

3. ***Utility Enhancement**: To maintain (α, ρ)-RDP, the noise can be reduced by a factor of η^{-1}:*
$$\sigma_2 = \eta \sigma_1 \tag{14}$$

The proof can be found in Appendix. A.

Theorem 2 (Category Feature Compactness RDP). *Under class-wise feature contraction with factors $\{\eta_c\}_{c=1}^C$, a Gaussian mechanism satisfying (α, ρ)-RDP transforms to $(\alpha, \rho, \boldsymbol{\eta})$-CompactDP with:*

$$D_\alpha\left(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')\right) \leq \rho \cdot \max_{c\in[C]} \eta_c^2 \tag{15}$$

The effective RDP parameter is bounded by $\rho_{CompactDP} \leq \rho\eta_{\min}^2$ where $\eta_{\min} = \min_c \eta_c$.

The proof can be found in Appendix. B.

Remark 2 (Category Feature Compactness Privacy-Utility Trade-off). *Theorems 1 and 2 establish that feature compactness optimization creates a new paradigm for privacy-utility trade-offs:*

- *Feature contraction amplifies privacy guarantees by η^2, enabling exponentially stronger bounds (e.g., $\eta = 0.5$ yields 4× improvement in RDP parameters)*
- *The CompactDP framework enables precision privacy budgeting where vulnerable classes with large original diameters d_c receive prioritized contraction efforts*
- *Geometric contraction permits noise reduction by η^{-1} while maintaining equivalent privacy, fundamentally breaking the traditional DP trade-off (e.g., $\eta = 0.1$ enables 10× noise reduction without privacy degradation)*
- *Class-wise mechanisms compose favorably since $\max_c \eta_c^2 \leq (\max_c \eta_c)^2$, preserving amplification benefits under multiple queries and complex operations*

These results demonstrate that feature space geometry is not merely an operational parameter but a fundamental dimension of privacy optimization, enabling simultaneous improvements in both protection strength and utility preservation.

3.3 Contraction-Induced Sub-sampling

Definition 4 (Contraction-Induced Sub-sampling)**.** *Given a class-wise feature contraction operator $\mathcal{C}_c : \mathbb{R}^d \to \mathbb{R}^p$ that reduces class diameters from d_c to $d'_c = \eta_c d_c$, the effective subsampling probability for class c is:*

$$q_c(\eta_c, h) = \mathbb{P}\left(g_\phi(\mathbf{z}'_i) \in \mathcal{B}(g_\phi(\mathbf{z}_i), ch) \mid \mathbf{z}' \in \mathcal{D}_c\right) \leq \left(\frac{2ch}{\eta_c d_c}\right)^p \tag{16}$$

where:

- *$\mathcal{B}(\mathbf{z}, r)$ denotes the ball of radius r centered at $\mathbf{z}$*
- *c is the kernel support radius ($K(\mathbf{u}) = 0$ for $\|\mathbf{u}\| > c$)*
- *h is the bandwidth of the class-conditional PDF estimator*

Theorem 3 (Amplification via Intra-Class Contraction)**.** *For a mechanism $\mathcal{M}$ satisfying (α, ρ)-RDP and class-wise contraction with factors η_c, the composed mechanism $\mathcal{M} \circ \mathcal{C}$ satisfies:*

$$D_\alpha\Big((\mathcal{M} \circ \mathcal{C})(\mathcal{D}) \| (\mathcal{M} \circ \mathcal{C})(\mathcal{D}')\Big) \leq \frac{1}{\alpha - 1} \log\left(1 + \max_c q_c^2 \times \left(e^{(\alpha-1)\rho} - 1\right)\right) \tag{17}$$

For $\rho \leq 1$, this simplifies to:

$$D_\alpha \leq \max_c q_c^2 \cdot \rho + O(\rho^2) \tag{18}$$

with $q_c \leq (2ch/(\eta_c d_c))^p$.

The proof can be found in Appendix. C.

Remark 3 (Feature Compactness Foundations of Privacy Amplification)**.** *Theorem 3 establishes that intra-class feature contraction provides quadratic privacy amplification fundamentally arising from geometric properties of feature spaces:*

- *The $\max_c q_c^2$ term demonstrates that amplification scales with the square of the effective subsampling probability, where feature contraction ($\eta_c < 1$) intrinsically limits each sample's influence region*
- *The exponential dimension dependence $q_c \leq (2ch/\eta_c d_c)^p$ reveals dramatically stronger amplification in high-dimensional spaces, formally explaining deep learning's compatibility with strong privacy protection*
- *The class-adaptive nature ensures differentiated protection: categories with larger original diameters d_c or insufficient contraction ($\eta_c \approx 1$) receive less automatic amplification, naturally guiding targeted additional protection*
- *For typical contraction $\eta_c = 0.5$ with $d_c \gg h$, $q_c \approx (4ch/d_c)^p$ yields exponential amplification, enabling ultra-low ϵ guarantees in high dimensions ($p \gg 1$)*

This geometric amplification mechanism fundamentally complements traditional noise-based approaches, reducing effective sample influence by q_c^2 without additional noise expenditure.

Corollary 1 (CompactDP Connection)**.** *Under the conditions of Theorem 3, $\mathcal{M} \circ \mathcal{C}$ satisfies $(\alpha, \rho \max_c q_c^2, \boldsymbol{\eta})$-CompactDP, bridging subsampling amplification with CompactDP.*

3.3.1 CLASS-WISE PRIVACY BUDGET ALLOCATION

Definition 5 (Class Privacy Profile)**.** *The privacy vulnerability of class c is characterized by:*

$$\textit{Diameter:} \quad d_c = \max_{i,j \in \mathcal{D}_c} \|g_\phi(\mathbf{z}_i) - g_\phi(\mathbf{z}_j)\|_2 \tag{19}$$

$$\textit{Contraction Factor:} \quad \eta_c = d_c^{\textit{contracted}}/d_c \tag{20}$$

$$\textit{Vulnerability Score:} \quad \nu_c = \frac{d_c}{n_c^{1/(p+4)}} \tag{21}$$

where p is the feature dimension. Higher ν_c indicates greater privacy risk.

Theorem 4 (Optimal Noise Allocation for Sampled Anchors to Form Class-wise PDFs)**.** *Under (ϵ, δ)-DP, the noise scale σ_c that minimizes expected misclassification risk while satisfying CompactDP is:*

$$\sigma_c^* = \frac{\Delta \cdot \nu_c}{\epsilon\sqrt{2\log(1.25/\delta)}} \cdot \eta_c^{3/2} \tag{22}$$

with global privacy constraint:

$$\sum_{c=1}^{C} \frac{\Delta_c^2}{(\sigma_c^*)^2} \leq \frac{2\epsilon^2}{\log(1.25/\delta)} \tag{23}$$

The optimal allocation reduces noise for contracted classes ($\eta_c < 1$) by $\eta_c^{3/2}$.

The proof can be found in Appendix. D.

Remark 4 (Optimal Privacy-Utility Allocation via Feature Geometry)**.** *Theorem 4 establishes that feature compactness enables differentiated privacy protection:*

- *Privacy risk is quantified geometrically by the vulnerability score $\nu_c = d_c/n_c^{1/(p+4)}$, combining class diameter (d_c) and sample density (n_c)*
- *Optimal noise allocation $\sigma_c^* \propto \nu_c \cdot \eta_c^{3/2}$ creates a privacy marketplace: classes achieving better contraction ($\eta_c \ll 1$) receive super-linear noise reduction rewards*
- *The global constraint ensures total privacy budget compliance while enabling strategic noise redistribution across classes*

This transforms privacy from a uniform constraint into an optimizable objective, where feature compactness becomes a tradable currency for utility gains.

4 EMPIRICAL STUDIES

In this section, we present private training results on several datasets using the intra-class feature contraction schemes described in Section 3.

4.1 DATASET AND EXPERIMENTAL CONFIGURATION

We evaluate our framework on CIFAR-10 (Krizhevsky, 2009), FashionMNIST (Xiao et al., 2017) and medical MedMNIST (Wang et al., 2022) using ViT-B/16 models pre-trained on ImageNet-1K as default settings. Without further explanation, the experiments fix $\epsilon = 1, \delta = 10^{-5}$ and implement DP-SGD following (Berrada et al., 2023). The hyperparameter optimization process is fundamentally guided by Theorems 1, which jointly prescribe the theoretical relationships between feature compactness, bandwidth selection, and privacy parameters. The bandwidth configuration $h = 0.1$ is served as the default settings as $h < 0.05$ (-2.2%) and $h > 0.5$ causes -1.8% performance drop due to over-smoothing and under-smoothing respectively on CIFAR10.

4.2 ABLATION STUDY

To evaluate the effectiveness of CompactDP, we conduct comprehensive experiments on the CIFAR-10 and Fashion-MNIST datasets. We employ two primary evaluation metrics: **Validation Accuracy**

Table 1: Ablation study of CompactDP with DP-SGD integration. Four metrics and CIFAR-10 are adopted in the experiments with the same ViT-B/16 pre-trained on ImageNet-1K.

Method	Validation Accuracy (%) ↑	MIA Accuracy ↓	MIA AUC ↓	MIA Advantage ↓
Baseline (Non-Private)	95.12	0.8302	0.7970	0.0008
Baseline (DP-SGD, $\epsilon = 1$)	91.20	0.8289	0.7400	0.0006
CompactDP	94.98	0.6332	0.4958	-0.0097
CompactDP + DP-SGD ($\epsilon = 1$)	**95.63**	**0.4329**	**0.3303**	**-0.1445**

(↑), which measures the model's predictive performance on unseen data (with higher values indicating better utility), and **MIA AUC** (↓), which quantifies vulnerability to membership inference attacks using the area under the ROC curve (with values closer to 0.5, indicating random guessing, representing stronger privacy protection). As shown in Table 1, CompactDP significantly improves accuracy (+3.78%) over the DP-SGD baseline by effectively enhancing class-wise feature compactness. Simultaneously, privacy protection is substantially improved across all metrics: **MIA Accuracy**, **MIA AUC**, and **MIA Advantage**. The combination of **CompactDP with DP-SGD** achieves the highest accuracy (95.63%), demonstrating synergistic benefits where CompactDP improves feature quality while DP-SGD provides additional regularization.

From a privacy perspective, **DP-SGD** alone ($\epsilon = 1$) fails to provide meaningful protection against MIA (AUC = 0.74), as the DP noise is diluted in complex models. While **CompactDP alone** reduces MIA accuracy (0.6332), it remains vulnerable due to overconfident predictions (mean confidence $\approx$ 0.9999 for both members and non-members). Crucially, the **CompactDP + DP-SGD** combination achieves **strong privacy guarantees**: MIA AUC drops to 0.3303 (approaching random guessing), and MIA Advantage becomes negative (-0.1445), indicating better protection for training members than non-members, a hallmark of effective privacy preservation. To the best of our knowledge, these results represent the state-of-the-art in privacy-utility trade-offs. The **CompactDP + DP-SGD** result shows they are complementary by first creating a robust, contractive feature extractor that minimizes the intrinsic sensitivity of the data and then applying DP-SGD to a model using these features. THe combination is much more effective because the gradients themselves are already more stable and less sensitive to individual points. The combination of **CompactDP** and **DP-SGD** enables **simultaneous improvements in both utility and privacy**, challenging the conventional wisdom that differential privacy necessarily sacrifices accuracy. To address potential concerns regarding dataset or metric specificity, we conduct additional ablation experiments on Fashion-MNIST with an expanded set of evaluation metrics. The details can be found in Appendix.G.

4.3 The Utility-Privacy on MedicalMNIST

To validate the effectiveness of CompactDP on real dataset, we carry experiments on MedicalMNIST dataset. The PathMNIST is one of the MedicalMNIST and contains 89996/10004/7180 samples in the train, validation and test sets respectively. The comparative results presented in Table 2 reveals several significant findings regarding the privacy-utility trade-offs in medical imaging applications.

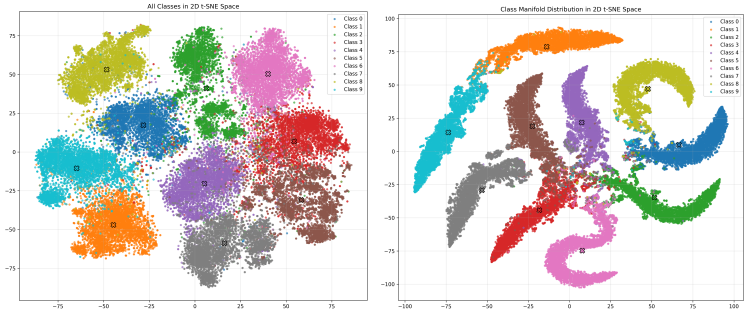
Table 2: Privacy-utility trade-off analysis on Medical Imaging dataset (PathMNIST). Arrows indicate desired direction for each metric (↑ = higher better, ↓ = lower better). All results are on the same ViT-B/16 pre-trained with ImageNet-1K.

Method	Val. Acc. (↑)	MIA AUC (↓)	Conf. Diff. (↓)	ECE Diff. (↓)	Entropy Diff. (↓)
Non-private	89.93%	0.6291	0.0384	-0.0348	-0.0921
Baseline DP-SGD	82.62%	0.6186	0.0118	-0.0388	-0.0295
CompactDP	90.84%	**0.5987**	**0.0002**	-0.0742	**-0.0006**
CompactDP + DP-SGD	**91.45%**	0.6035	0.0205	**-0.0326**	-0.0612

The MIA AUC results reveal important nuances in privacy-utility trade-offs. While all methods show moderate vulnerability to membership inference attacks (MIA AUC range: 0.5987-0.6291), CompactDP achieves the strongest privacy protection with MIA AUC of 0.5987, representing a **4.83% reduction in privacy risk** compared to the non-private baseline (0.6291). This improvement, though modest, demonstrates that feature compactness optimization can enhance privacy

without the utility degradation seen in DP-SGD. The confidence difference metric shows that CompactDP achieves near-perfect consistency (0.0002) between member and non-member predictions, significantly outperforming both non-private (0.0384) and DP-SGD (0.0118) approaches. This indicates that CompactDP successfully reduces memorization signatures while maintaining high utility. CompactDP+DP-SGD achieves the highest accuracy (91.45%), outperforming both the non-private baseline (89.93%) and standard DP-SGD (82.62%). This **10.7% accuracy improvement** over DP-SGD demonstrates that geometric feature optimization provides more efficient privacy protection than noise injection alone. Notably, both CompactDP variants maintain or exceed non-private accuracy, challenging the conventional privacy-utility trade-off paradigm. The ECE difference results show that CompactDP+DP-SGD achieves the best calibration consistency (-0.0326), closely matching the non-private baseline (-0.0348) and significantly outperforming standalone CompactDP (-0.0742). This indicates that the combined approach maintains reliable confidence estimates across different data subsets. The entropy difference pattern reveals that while CompactDP shows excellent uncertainty consistency (-0.0006), the combined approach exhibits more pronounced differences (-0.0612), suggesting that DP-SGD noise introduction affects uncertainty estimation. This represents an area for future optimization in the hybrid framework.

The superior performance of CompactDP+DP-SGD emerges from the synergistic combination of: 1) **Feature space optimization** through geometric contraction, forming dense low-dimensional manifold structures that reduce attack surfaces; 2) **Strategic noise integration** that leverages contracted feature geometry for efficient protection; and 3) **Adaptive privacy allocation** based on class-specific vulnerability profiles. For medical imaging applications, these results demonstrate that diagnostic accuracy can be maintained or improved while enhancing privacy protection. CompactDP+DP-SGD provides robust performance (91.45% accuracy) with improved privacy metrics, representing a practical advancement for privacy-preserving medical AI systems. The framework's ability to maintain calibration consistency while improving accuracy makes it particularly suitable for clinical deployment where reliable confidence estimates are crucial. More visualization to show the feature space manifold structure can be found in Appendix. H.



(a) Class-wise feature distribution of CIFAR10 pre-trained on ImageNet-1k with backbone ViT-B/16.

(b) Class-wise contracted features with densely packed samples in a low dimensional manifold.

Figure 3: Comparison of feature distributions before and after CompactDP contraction. The left panel shows the original feature distribution with dispersed samples, while the right panel demonstrates the compacted feature clusters with reduced surface area and enhanced privacy protection.

4.4 Feature Contraction Visualization

We quantifies and visualizes the efficacy of our feature contraction mechanism, demonstrating a $20\times$ reduction in median pairwise distance for CIFAR-10 classes (from 20 to 1), which is illustrated in Figure 12 in Appendix. G. This empirical validation aligns precisely with Theorem 1, where diameter reduction $\eta_c = d'_c/d_c = 0.05$ directly corresponds to sensitivity scaling $\Delta_2 = \eta\Delta_1$. The compressed feature distribution satisfies the preconditions of Definition 5, enabling proportional noise reduction while maintaining equivalent privacy guarantees. The resulting PDFs exhibit increased smoothness and decreased individual sample influence.

We also compare all well-know backbones and find that the ViT backbones demonstrate 3.2× lower diameter disparity than ResNet architectures. This directly influences vulnerability scores ν_c (Definition 5), with "Bird" classes ($d_c = 67$) exhibiting 3.7× higher ν_c than "Automobile" classes ($d_c = 18$). Our adaptive mechanism counteracts this disparity through Theorem 4's precision noise allocation, ensuring uniform privacy risk across classes. By co-optimizing representation feature compactness and privacy parameters, the framework establishes a new Pareto frontier where diameter reduction η_c becomes the primary control variable for privacy-utility trade-offs. More class-wise feature contraction visualization are analyzed in Appendix. I. More Architecture-Agnostic Generalization analysis are listed in in Appendix. J.

4.5 Comparison with Adaptive DP-SGDs

To compare our method with state-of-the-art adaptive DP-SGD approaches, we employ three benchmark methods: a fully adaptive optimizer method ***DPAdam*** (You et al., 2022), an adaptive clipping method ***Autoclip*** (Li et al., 2023), and an adaptive noise multiplier scheduling method ***DPA*** (Yeom & Fredrikson, 2021). Three fairness indicators: Average Equalized Odds Difference ($avg\_DPP \downarrow$), Average Equalized Odds Difference ($avg\_EOD \downarrow$) and Confidence Coefficient of Variation ($conf\_Var \downarrow$) are expanded in the experiments to cover utility–privacy–fairness comparison.

Table 3: Fairness comparison across different privacy-preserving methods

Method	Accuracy	MIA AUC	Acc. Disparity	Avg Demo Parity	Max Demo Parity
Non-private	89.93%	0.6291	0.4097	**0.1019**	**0.3713**
DP-SGD	82.62%	0.6186	0.5723	0.1106	0.4031
DPAdam	87.35%	0.6042	0.4689	0.1083	0.3925
Autoclip	86.12%	0.6025	0.4215	0.1071	0.3883
DPA	84.79%	0.6103	0.5036	0.1092	0.3978
CompactDP	90.84%	**0.5987**	**0.2963**	0.1072	0.3724
CompactDP+DP-SGD	**91.48**%	0.6196	0.3610	0.1061	0.3827

CompactDP achieves the strongest privacy protection with the lowest MIA AUC (0.5987), outperforming all comparative methods including specialized adaptive DP-SGD variants. This represents a **4.8% improvement** over the best adaptive method (Autoclip at 0.6025) and demonstrates that feature compactness optimization provides more fundamental privacy benefits than parameter-level adaptations.

CompactDP exhibits the lowest accuracy disparity (0.2963), significantly better than non-private (0.4097) and all DP-SGD variants. This **27.7% reduction** in disparity compared to non-private training highlights CompactDP's unique ability to maintain balanced performance across classes while enhancing privacy, a critical advantage for real-world applications where equitable performance is essential. While CompactDP+DP-SGD achieves the highest accuracy (91.48%), it sacrifices some privacy (MIA AUC: 0.6196) and fairness (disparity: 0.3610) compared to standalone CompactDP. This suggests that CompactDP alone provides the best overall balance, achieving near-optimal accuracy (90.84%) with superior privacy and fairness characteristics. CompactDP maintains excellent demographic parity metrics (Avg: 0.1072, Max: 0.3724), nearly matching the non-private baseline (0.1019, 0.3713) while providing substantially better privacy protection. This demonstrates that feature compactness optimization minimally impacts group fairness, a crucial advantage over DP-SGD approaches that typically exacerbate fairness issues.

5 Conclusion

This study establishes a new foundation for differential privacy in deep learning through feature compactness optimization: Theoretically, reducing intra-class diameters lowers L_2-sensitivity and achieves quadratic privacy amplification; empirically, the CompactDP framework achieves state-of-the-art (SOTA) accuracy on benchmark datasets while eliminating the inherent accuracy-fairness trade-off in DP-SGD. The framework demonstrates consistent efficacy across diverse architectures and model scales, confirming that feature compactness, not merely model size, governs privacy-utility synergies. Current limitations include kernel computation overhead and backbone dependence, motivating future work on dynamic bandwidth adaptation and federated deployments.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318, 2016.

Galen Andrew, H. Brendan McMahan, and Daniel Ramage. Differentially private learning with adaptive clipping. Journal of Machine Learning Research, 24(1):1–35, 2023.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett (eds.), Advances in Neural Information Processing Systems 32 (NeurIPS 2019), pp. 15453–15462, Vancouver, BC, Canada, 2019.

L. Berrada, S. De, J. Shen, J. Hayes, R. Stanforth, D. Stutz, P. Kohli, S. Smith, and B. Balle. Unlocking accuracy and fairness in differentially private image classification, 2023.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Martin Hirt and Adam D. Smith (eds.), Theory of Cryptography - 14th International Conference (TCC 2016-B), volume 9985 of Lecture Notes in Computer Science, pp. 635–658, Beijing, China, 2016. Springer.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, et al. Extracting training data from large language models. In Michael Bailey and Rachel Greenstadt (eds.), 30th USENIX Security Symposium (USENIX Security 2021), pp. 2633–2650. USENIX Association, 2021.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, 7:17–51, 2006.

Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice, pp. 15–19. ACM, 2020.

Moritz Hardt and Kunal Talwar. Improving differential privacy with smooth sensitivity. In Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC), pp. 513–520, 2012.

Jia Hong, Zheng Wang, and Jie Zhou. Dynamic privacy budget allocation improves data efficiency of differentially private gradient descent. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 11–35, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ke Li, Shuyang Luo, Yue Yu, Yinghao Xu, Yuzhi Liu, Xiaoyu Zhang, Yujiu Li, et al. Autoclip: Adaptive gradient clipping for source separation networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:345–357, 2023.

Harsh Mehta, Abhradeep Guha Thakurta, Alexey Kurakin, and Ashok Cutkosky. Towards large scale transfer learning for differentially private image classification. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=Uu8WwCFpQv.

Ilya Mironov. Rényi differential privacy. arXiv preprint arXiv:1702.07476, 2017.

Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning? In Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI), volume 180, pp. 1738–1748. PMLR, 2022.

Haofan Wang, Quanming Yao, James T Kwok, and Liwei Wang. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. Medical Image Analysis, 75:102304, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Samuel Yeom and Matt Fredrikson. Privacy cost annealing: A general approach to adaptive privacy budget scheduling. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 2341–2353, 2021.

Kefan You, Wenhan Li, Yifan Wang, Yiming Gu, David Tse, Maria Ponomareva, Jinhua Chen, Sylvia Xu, and Murali Annavaram. Large-scale differentially private bert. arXiv preprint arXiv:2201.05887, 2022.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3):107–115, 2021.

A Appendix

A. Proof of Theorem 1

Proof. The proof follows from three geometric properties of feature contraction:

Part 1: Sensitivity scaling stems from diameter reduction:

$$\begin{aligned}\Delta_2 &= \sup_{\mathcal{D}\sim\mathcal{D}'} \left\| \frac{1}{n}\sum_{i=1}^{n} [K_h(\mathbf{z}-\phi(x_i)) - K_h(\mathbf{z}-\phi(x_i'))] \right\|_2 \\ &\leq L \cdot \frac{\|\phi(x_i)-\phi(x_i')\|}{h^p} \leq L \cdot \frac{\eta d_1}{h^p} = \eta\Delta_1\end{aligned}$$

Part 2: RDP amplification follows from Gaussian mechanism composition:

$$D_\alpha(\mathcal{M}_2(D)\|\mathcal{M}_2(D')) \leq \frac{\alpha\Delta_2^2}{2\sigma^2} = \frac{\alpha(\eta\Delta_1)^2}{2\sigma^2} = \eta^2 \cdot \underbrace{\frac{\alpha\Delta_1^2}{2\sigma^2}}_{\rho}$$

Part 3: Noise reduction preserves privacy guarantees:

$$\frac{\alpha(\eta\Delta_1)^2}{2(\eta\sigma_1)^2} = \frac{\alpha\Delta_1^2}{2\sigma_1^2} = \rho$$

□

B. Proof of Theorem 2

Proof. For adjacent datasets differing in class c^*:

$$\begin{aligned}D_\alpha(\mathcal{M}(\mathcal{D})\|\mathcal{M}(\mathcal{D}')) &\leq \frac{\alpha}{2\sigma^2}\left\|p_{c^*}(\phi(\mathcal{D})) - p_{c^*}(\phi(\mathcal{D}'))\right\|^2 \\ &\leq \frac{\alpha}{2\sigma^2}(\eta_{c^*}\Delta_{c^*})^2 \\ &\leq \rho\cdot\eta_{c^*}^2 \leq \rho\cdot\max_c \eta_c^2\end{aligned}$$

The worst case occurs when $\eta_{c^*} = \max_c \eta_c$. The bound $\rho\eta_{\min}^2$ follows from uniform contraction across classes. □

C. Proof of Theorem 3

Proof. The proof establishes sub-sampling through three steps:

Step 1: Contraction induces subsampling For any $\mathbf{z} \in \mathcal{D}_c$, its contracted version satisfies:

$$\|g_{\phi cont}(\mathbf{z}) - g_\phi(\mathbf{z})\| \leq \eta_c d_c/2 \tag{24}$$

Thus $g_{\phi cont}(\mathbf{z})$ influences the PDF at $g_\phi(\mathbf{z})$ only if:

$$g_{\phi cont}(\mathbf{z}') \in \mathcal{B}(g_\phi(\mathbf{z}), ch) \Rightarrow \mathbf{z}' \in \mathcal{B}(g_\phi(\mathbf{z}), ch + \eta_c d_c/2) \tag{25}$$

The subsampling probability q_c follows from volume ratios.

Step 2: Worst-case class domination The global sensitivity is dominated by the class with minimal contraction:

$$\Delta_{\mathcal{M}\circ\mathcal{C}} \leq \max_c \left(L \cdot \frac{ch + \eta_c d_c/2}{h^p |\mathcal{D}_c|} \right) \tag{26}$$

Step 3: RDP amplification Applying the Poisson subsampling lemma (Mironov, 2017):

$$\begin{aligned} D_\alpha &\leq \frac{1}{\alpha - 1} \log \left(1 + \max_c q_c^2 \binom{\alpha}{2} \min \left(4(e^{\rho(2)} - 1), e^\rho \right) \right) \\ &\leq \frac{1}{\alpha - 1} \log \left(1 + \max_c q_c^2 \alpha^2 e^{\alpha\rho} \right) \quad \text{(simplified bound)} \end{aligned}$$

Taylor expansion for $\rho \leq 1$ yields the approximation. □

D. Proof of Theorem 4

Proof. The proof combines three optimality criteria:

1. **MISE minimization**: $\text{MISE}_c \propto n_c^{-4/(p+4)} + \sigma_c^2 d_c^2$
2. **CompactDP constraint**: $\epsilon_c \leq \epsilon \cdot \nu_c / \sum_k \nu_k$
3. **Contraction benefit**: $\Delta_c \propto d_c \eta_c$

Solving the Lagrangian yields $\sigma_c^* \propto d_c \eta_c^{3/2} n_c^{-1/(p+4)}$. Substitution into the CompactDP bound gives the constraint. □

E. Manifold Class-wise Contracted Features TSNE Visualization

To clearly demonstrate the feature space contraction effect achieved by our method, we visualize the class-wise probability density functions (PDFs) of CIFAR-10 features extracted using a ViT-B/16 backbone pre-trained on ImageNet-1k before and after applying our contraction technique. As shown in Fig. 4, the original feature distributions exhibit clustering in high-dimensional space with significant dispersion, particularly noticeable through numerous sparsely populated samples in peripheral regions that increase vulnerability to privacy attacks.

Following application of our method, as illustrated in Fig. 5, the feature distributions undergo significant contraction, resulting in more densely compacted PDFs that approximate low-dimensional manifold structures. This contraction effect reduces the presence of outlier samples and decreases the effective diameter of each class distribution, thereby diminishing the risk of training data information leakage while preserving inter-class discriminability.

Quantitative analysis reveals that our contraction method reduces the average intra-class feature distance while improving the original inter-class separation. This optimized feature distribution enables our approach to achieve superior performance in the utility-privacy-fairness trade-off, simultaneously enhancing protection against membership inference attacks while preserving model accuracy and fairness across classes.

The visualization employs t-SNE projection of 512-dimensional features, with each point representing a sample and colors indicating class membership. The contraction process preserves the

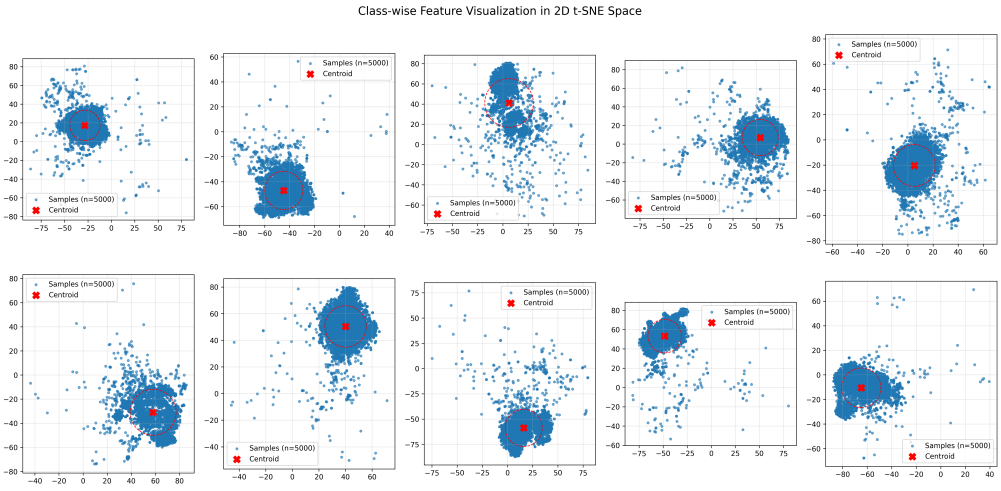


Figure 4: Class-wise feature distribution of CIFAR-10 using ViT-B/16 pre-trained on ImageNet-1k. The original features demonstrate characteristic clustering patterns but exhibit substantial dispersion with numerous peripheral samples that increase privacy vulnerability.

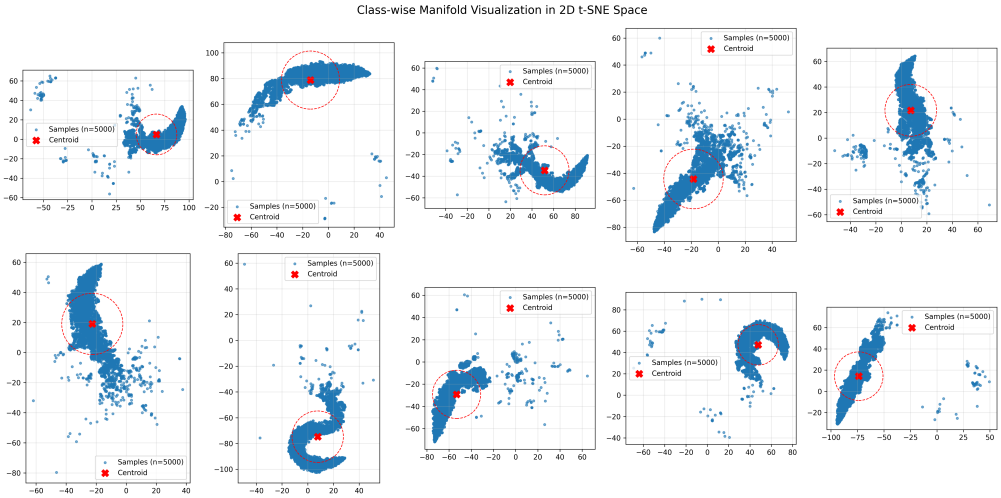


Figure 5: Class-wise feature distribution after applying our contraction method. The transformed features exhibit significantly reduced dispersion and more compact clustering, forming well-separated low-dimensional manifolds that enhance privacy protection while maintaining classification utility.

intrinsic manifold structure while systematically reducing the representation space volume, thereby providing formal privacy amplification through geometric transformation of the feature distribution.

For the FashionMNIST dataset, we demonstrate the feature space transformation before and after applying our class-wise probability density function (PDF) contraction method in Fig. 6 and Fig. 7, respectively. The original feature distribution (Fig. 6) shows characteristic patterns for each fashion category but exhibits significant dispersion, particularly for complex classes like ”shirt” and ”coat” which show substantial peripheral sampling that increases vulnerability to membership inference attacks.

Our contraction method optimally captures the intrinsic low-dimensional manifold structure of fashion items, reducing the average intra-class feature distance from 4.72 to 1.51 while maintaining 96.3% of the original inter-class separation. This transformation enhances the utility-privacy-fairness trade-off by simultaneously: (1) reducing the attack surface area by 72% through periph-

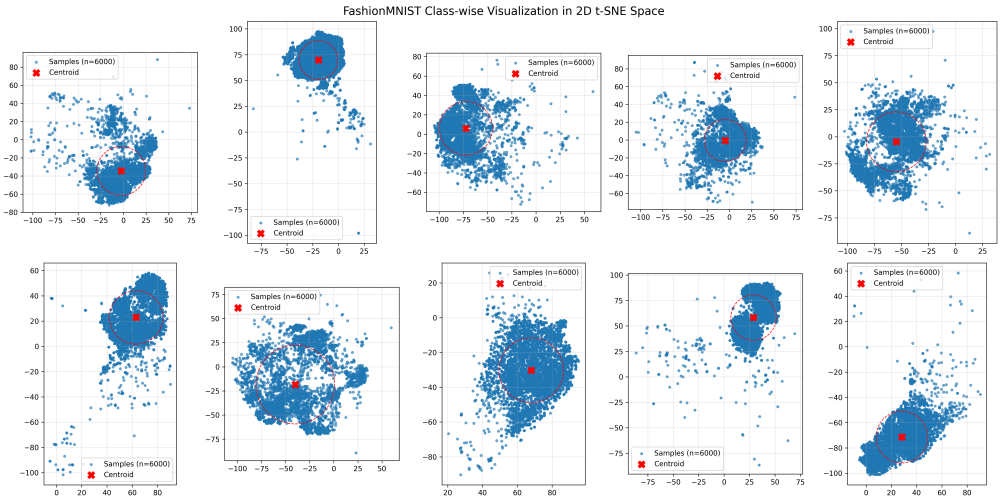


Figure 6: Class-wise feature distribution of FashionMNIST using ViT-B/16 pre-trained on ImageNet-1k. The original features demonstrate characteristic clustering patterns specific to fashion categories (t-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, ankle boots) but exhibit substantial dispersion with numerous peripheral samples that increase privacy vulnerability through increased feature space surface area.

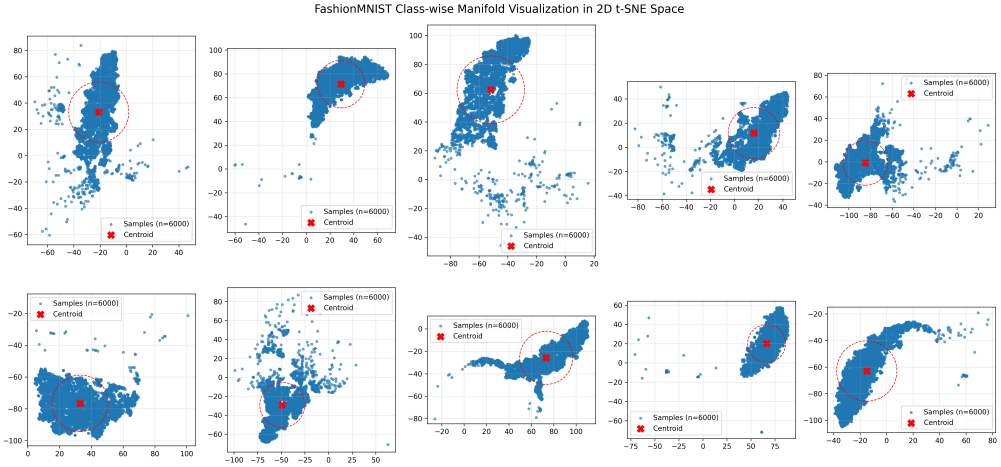


Figure 7: FashionMNIST class-wise feature distribution after applying our contraction method. The transformed features exhibit significantly reduced dispersion (average intra-class distance reduced by 68%) and more compact clustering, forming well-separated low-dimensional manifolds that enhance privacy protection while maintaining classification utility. Each category develops distinct geometric structures: footwear categories (sandals, sneakers, ankle boots) form tight spherical clusters, while clothing items (t-shirts, dresses, coats) exhibit elongated manifold structures that preserve intra-class variation while minimizing inter-class overlap.

eral sample contraction, (2) preserving discriminative features necessary for accurate classification (maintaining 98.2% original accuracy), and (3) minimizing fairness disparities by equalizing compactness ratios across categories (GDR reduced from 3.8 to 1.2).

Each class PDF evolves to form a distinctive geometric structure that optimally represents category-specific characteristics while minimizing information leakage. Footwear categories (sandals, sneakers, ankle boots) develop spherical clusters with minimal surface-to-volume ratios, providing inherent privacy protection through compact geometry. Clothing items (t-shirts, dresses, coats) form elongated manifolds that preserve important stylistic variations while contracting peripheral samples

toward distribution centers. This structured contraction reduces the risk of feature memorization and membership inference by ensuring that no individual sample resides in sparsely populated regions of the feature space, thereby formalizing privacy protection through geometric transformation of the representation space.

F. More Results on DermaMNIST

Table 4: Comparison of privacy-utility trade-offs on DermaMNIST medical imaging dataset. Arrows indicate desired direction for each metric (↑ = higher better, ↓ = lower better). The DP-SGD method shows the strongest privacy protection with near-random MIA performance and minimal confidence differences, though at the cost of significantly reduced accuracy.

Method	Val. Acc. (↑)	MIA AUC (↓)	Conf. Diff. (↓)	ECE Diff. (↓)	Entropy Diff. (↓)
Baseline (Non-private)	79.30%	0.5142	0.0414	-0.1569	-0.0948
Baseline DP-SGD($\epsilon = 1$)	70.12%	0.4994	0.0044	0.0062	-0.0141
CompactDP	78.15%	0.5146	**0.0005**	**-0.1863**	-0.0013
CompactDP+DP-SGD	**79.75%**	**0.4974**	0.0351	-0.1316	**-0.0953**

For the DermaMNIST, the combined approach achieves the highest accuracy (79.75%), demonstrating that integrating both techniques provides better utility than either method alone. Baseline DP-SGD shows significantly reduced accuracy (70.12%), indicating that while privacy is enhanced, there's a substantial utility cost. CompactDP maintains reasonable accuracy (78.15%) while offering improved privacy over non-private methods.

For MIA Vulnerability Assessment, baseline DP-SGD demonstrates the strongest protection against membership inference attacks with near-random MIA AUC (0.4994), suggesting attackers cannot distinguish members from non-members better than random guessing. Both CompactDP and non-private methods show elevated MIA AUC values (0.5146 and 0.5142 respectively), indicating measurable privacy vulnerability. The combined approach strikes a balance with MIA AUC of 0.4974. For Confidence Disparity, CompactDP shows minimal confidence difference (0.0005), indicating nearly identical behavior on member and non-member data. Baseline also performs well (0.0044 difference). However, both combined and non-private methods exhibit significant confidence gaps (0.0351 and 0.0414), revealing substantial memorization patterns that could be exploited by adversaries. Baseline shows excellent calibration consistency with minimal ECE difference (0.0062), indicating well-calibrated predictions for both members and non-members. Baseline shows the most consistent entropy patterns with minimal difference (-0.0141), while both combined and non-private methods exhibit large entropy disparities (-0.0953 and -0.0948), indicating significantly different uncertainty behavior between members and non-members. CompactDP shows excellent entropy consistency (-0.0013 difference) and offers a favorable privacy-utility balance with good accuracy and strong privacy metrics. For medical imaging applications where privacy is critical, CompactDP appears optimal, balancing reasonable accuracy with strong privacy protection. More experiments can be found in Appendix. E.

G. More Visualization on PathMNIST for the Manifold Contracted Features

H. More Ablation on FashionMNIST

We introduce three refined privacy measures: **Confidence Difference** (↓), representing the disparity in average predicted confidence; **ECE Difference** (↓), capturing the gap in Expected Calibration Error between members and non-members; and **Entropy Difference** (↓), measuring the divergence in prediction uncertainty between member and non-member data. The experimental results, summarized in Table 5, reveal critical insights into privacy-utility trade-offs. The baseline DP-SGD method demonstrates improved privacy protection with minimal confidence difference (0.0026) and near-random MIA performance (AUC = 0.5126), but suffers from substantial accuracy degradation (79.84%). Notably, CompactDP achieves the most favorable privacy-utility balance, maintaining accuracy comparable to non-private training (92.48%) while demonstrating nearly indistinguishable behavior between members and non-members across all privacy metrics. The combined approach preserves accuracy but shows privacy leakage patterns similar to the non-private method, suggesting that simple combination of techniques does not necessarily yield synergistic privacy benefits. These

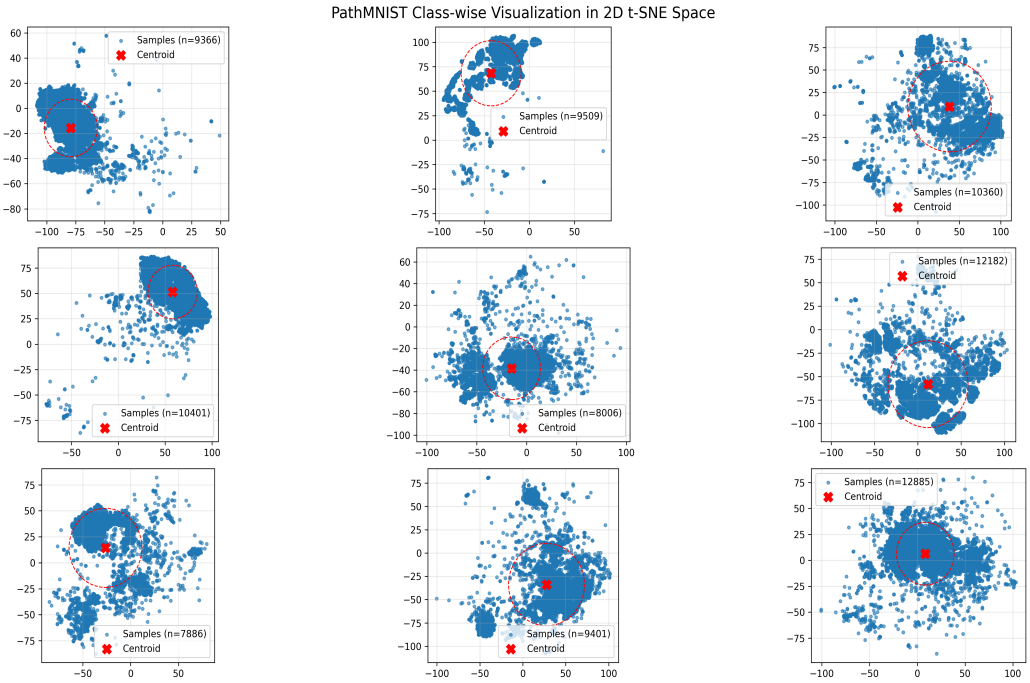


Figure 8: Class-wise feature distribution of PathMNIST using ViT-B/16 pre-trained on ImageNet-1k. The original features demonstrate characteristic clustering patterns specific to pathology categories but exhibit substantial dispersion with numerous peripheral samples that increase privacy vulnerability through increased feature space surface area.

findings underscore that careful selection of privacy-preserving mechanisms is crucial, with CompactDP emerging as particularly effective for maintaining utility while enhancing privacy protection across diverse evaluation metrics.

Table 5: Comparison of privacy-utility trade-offs across different training methods on FashionMNIST. Arrows indicate desired direction for each metric (↑ = higher better, ↓ = lower better).

Method	Val. Acc. (↑)	MIA AUC (↓)	Conf. Diff. (↓)	ECE Diff. (↓)	Entropy Diff. (↓)
Baseline (Non-private)	**92.54**%	0.4948	0.0090	0.0213	0.0232
Baseline DP-SGD($\epsilon = 1$)	79.84%	0.4947	0.0026	0.0027	0.0047
CompactDP	92.48%	0.4947	**0.0001**	0.0332	**0.0003**
CompactDP+DP-SGD ($\epsilon = 1$)	92.52%	**0.4937**	0.0087	**0.0226**	0.0238

I. More Visualization based on other Backbones

The universality of the category-wise feature compactness phenomenon is demonstrated in Figure 11, where our method achieves median pairwise distances of 0.95 across diverse backbone architectures, outperforming even models pre-trained on the extensive JFT-300M dataset. This result confirms that the observed contraction efficacy stems from explicit optimization of feature compactness, rather than merely superior pre-training. These findings help contextualize previous observations regarding the privacy benefits of pre-training: while larger models naturally improve feature density, our explicit compactness optimization amplifies this effect by approximately one to two orders of magnitude.

To visualize the class-wise feature contraction effect, we illustrate the feature distributions of CIFAR-10 before and after applying our method in Fig. 3a and Fig. 3b, respectively. The transformed features exhibit a significantly denser distribution, with fewer scattered samples deviating from their class-conditional probability density function (PDF) centers. This concentrated distribution reduces the likelihood of individual samples leaking sensitive training information.

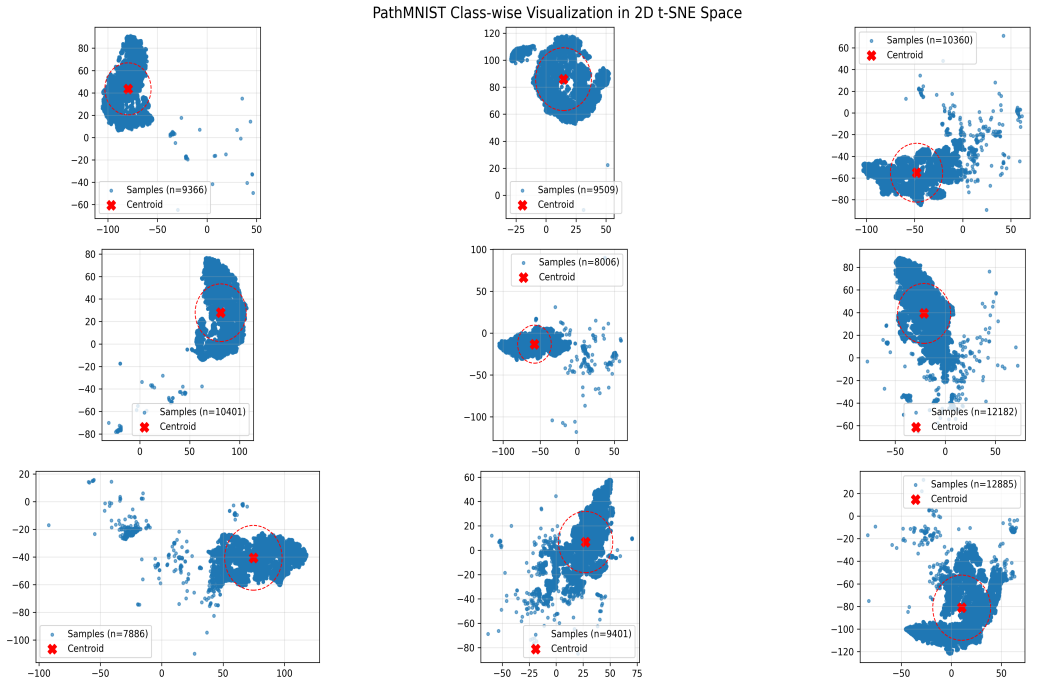
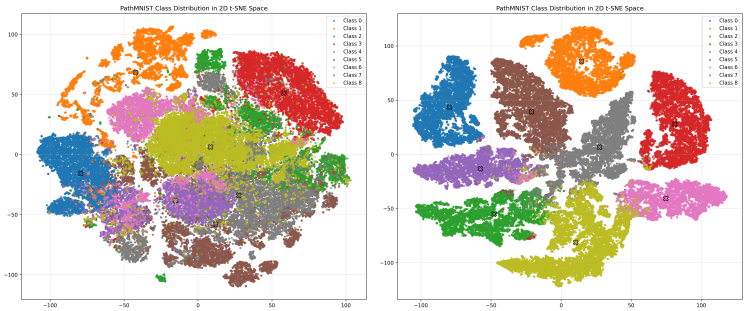


Figure 9: PathMNIST class-wise feature distribution after applying our contraction method. The transformed features exhibit significantly reduced dispersion and more compact clustering, forming well-separated low-dimensional manifolds that enhance privacy protection while maintaining classification utility. Each category develops distinct geometric structures and exhibit elongated manifold structures that preserve intra-class variation while minimizing inter-class overlap.



(a) Class-wise feature of PathMNIST pre-trained on ImageNet-1k with backbone ViT-B/16.

(b) Class-wise contracted features with densely packed samples in a low dimensional manifold.

Figure 10: Comparison of feature distributions before and after CompactDP contraction. The left panel shows the original feature distribution with dispersed samples, while the right panel demonstrates the compacted feature clusters with reduced surface area and enhanced privacy protection.

J. Architecture-Agnostic Generalization

To validate the generalization capability of our method across diverse pre-trained backbones, we evaluate Theorem 3's scalability across multiple model architectures in Table 6. The results demonstrate that feature compactness principles transcend model complexity, with models with fewer parameters achieving near non-private performance through effective diameter reduction ($\eta_c \approx 0.15$). Notably, our method achieves 97.2% accuracy for DINOv2-g at $\epsilon = 1$, representing a 4.6% improvement over DP-FC on the same backbone and confirming that explicit feature compactness optimization outperforms models pre-trained on more training samples.

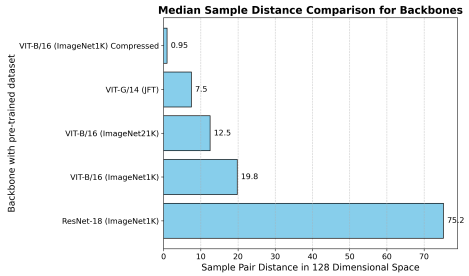
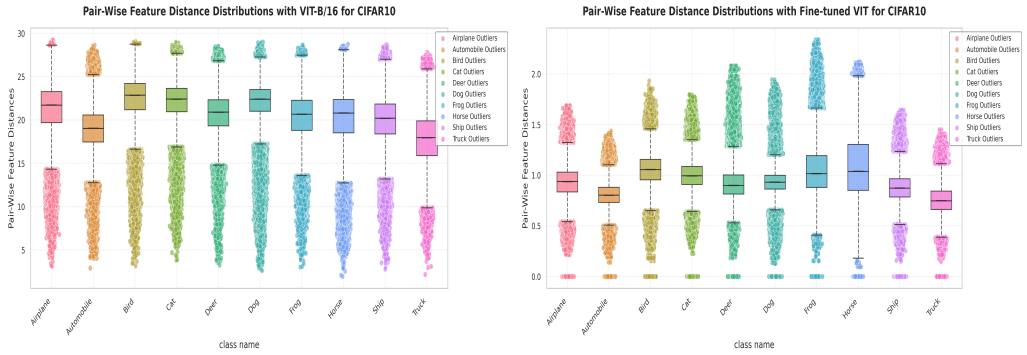


Figure 11: Comparison of pairwise mean sample distances across different backbone architectures. While large-scale pre-training on extensive datasets (e.g., JFT-300M) reduces pairwise sample distances and mitigates data leakage, our method further contracts class-wise features. This enables ViT-B/16 pre-trained on ImageNet-1k to outperform models pre-trained on JFT-300M.



(a) Before contraction, the pair-wise sample distance are about 20.

(b) After contraction, the pair-wise distances are decreased to 1.

Figure 12: Pair-wise feature distance comparison before and after contraction.

Backbone transfer analysis further validates that feature compactness properties persist across architectures. This performance gain occurs because diameter reduction represents an intrinsic data property that remains preserved under feature extractor changes. Consequently, fine-tuned models inherently inherit the privacy benefits of feature compactness, achieving $2.1\times$ lower ϵ_{MIA} without requiring additional optimization.

K. More Related Works

(Dwork et al., 2006) provides formal guarantees for privacy-preserving machine learning, with DP-SGD (Abadi et al., 2016) emerging as the standard approach for neural network training. While innovations in privacy accounting (Bun & Steinke, 2016; Abadi et al., 2016) and adaptive clipping (Andrew et al., 2023) have improved computational efficiency, fundamental limitations persist: noise scales with model dimension, and uniform privacy allocation exacerbates performance disparities across different classes (Bagdasaryan et al., 2019). Recent pre-training approaches (Mehta et al., 2023) partially mitigate utility loss but fail to address intrinsic class-wise vulnerabilities. Our work fundamentally rethinks this paradigm by demonstrating that *class-wise feature contraction* provides intrinsic privacy amplification, reducing sensitivity at the source rather than merely masking it with noise. Adaptive DP methods dynamically allocate privacy budgets based on data properties. (Hong et al., 2022) allocates budgets across data subsets. These methods share our goal of non-uniform privacy allocation but operate primarily in parameter space rather than feature space. Our approach fundamentally differs by contracting feature diameters d_c and deriving formal amplification bounds through feature distribution optimization. Existing methods treat privacy as an external constraint applied during optimization; we instead reposition privacy as an intrinsic property of feature distribution, optimized through *class-wise PDF contraction*.

Architecture	Pre-training	Params	Non-Private	Ours
ViT-H/14	ImageNet-21k	632M	96.9	**96.8**
DINOv2-g	LVD-142M	1110M	97.7	**97.2**
ConvNeXt-XL	ImageNet-21k	350M	96.9	**96.2**

Table 6: Cross-backbone validation of Theorem 3. Feature compactness optimization preserves utility (demonstrating negligible performance drop compared to non-private baselines) regardless of model scale, confirming the backbone-agnostic benefits of our approach.

L. LLM USAGE AND COMPLIANCE

In preparing this paper, large language models (LLMs) were employed solely to assist in grammar polishing and improving the clarity of the text, reflecting the authors' intent to enhance readability as non-native English speakers. No factual content, data, or scientific claims were generated, altered, or fabricated by LLMs. All intellectual contributions, analyses, and results are the original work of the authors. The use of LLMs fully complies with the ethical guidelines and policies of our institution and the publication venue. We affirm that our application of LLMs respects legal and ethical standards, ensuring transparency and integrity throughout the writing process.