**Discover** Computing

Research

# Towards auditing gradient privacy risks in image reconstruction attacks on deep learning models

**Tao Huang[1] · Xin Shi[1] · Qingyu Huang[1] · Ziyang Chen[1] · Liang Jiang[2] · Chenhuang Wu[2] · Guolong Zheng[1] · Xu Yang[1] · Wencheng Yang[3]**

Received: 10 October 2024 / Accepted: 11 April 2025
Published online: 30 April 2025

## Abstract

As artificial intelligence continues to drive advancements in computer vision, particularly in areas such as image analysis, object detection, and facial recognition, the ability to accurately recognize patterns in visual data has become a central focus of research. However, alongside these advances, concerns about the privacy risks associated with the training data used in AI models have also gained prominence. Deep learning models, frequently employed in computer vision tasks, can unintentionally expose sensitive information from the data they are trained on, raising the need for comprehensive research into privacy-preserving techniques. This paper explores the intersection of AI-driven pattern recognition and the privacy risks involved in training models on image data. Existing studies show that attackers can exploit the gradients from deep learning processes to reconstruct original image data, including personal and identifiable information, such as facial features. By iteratively adjusting input data, attackers can minimize the difference between the gradients of the random and stolen data, leading to the full reconstruction of private images. Current privacy protection methods fall short of explaining the relationship between an attacker's capacity to recover visual data and the structure of the targeted model. This paper introduces a novel privacy auditing framework that directly assesses the extent to which gradient-based attacks can reconstruct sensitive data. Unlike traditional methods, which mainly focus on mitigating privacy risks through model regularization or data obfuscation, our approach provides a systematic and quantitative evaluation of gradient leakage, filling a critical gap in existing privacy protection techniques. This paper investigates the relationships among reconstructed data, model gradients, and the original input data in the context of computer vision. By formalizing the connection between gradient similarity and data similarity, we propose a novel methodology that quantifies the vulnerability of deep learning models to data reconstruction attacks. Building on these insights, we propose a novel privacy auditing method aimed at evaluating the privacy risks associated with deep learning models used in pattern recognition for image data.

**Keywords** Pattern recognition · Differential privacy · Privacy risk auditing · Data mining

Xin Shi and Qingyu Huang contributed equally to this work.

✉ Guolong Zheng, gzheng@mju.edu.cn; ✉ Xu Yang, xu.yang@mju.edu.cn; Tao Huang, huang-tao@mju.edu.cn; Xin Shi, shixin@stu.mju.edu.cn; Qingyu Huang, qyhuang@stu.mju.edu.cn; Ziyang Chen, chenzy@stu.mju.edu.cn; Liang Jiang, ptjliang@163.com; Chenhuang Wu, ptuwch@163.com; Wencheng Yang, Wencheng.Yang@unisq.edu.au | [1]Fuzhou Institute of Oceanography, Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, School of Computer and Big Data, Minjiang University, Fuzhou 350108, China. [2]Fujian Key Laboratory of Financial Information Processing, Putian University, Putian 351100, China. [3]School of Mathematics, Physics and Computing, University of Southern Queensland, Brisbane 4350, Australia.

Discover Computing (2025) 28:53 | https://doi.org/10.1007/s10791-025-09551-z

## 1 Introduction

As deep learning models continue to drive advancements in computer vision [1, 2], their deployment in sensitive applications such as facial recognition [3, 4], medical diagnostics [5, 6], and autonomous systems [7, 8] has become widespread. However, these models introduce substantial privacy risks [9, 10], particularly due to potential leakage of sensitive data during the training process. One of the most concerning vulnerabilities arises from data reconstruction attacks [11, 12], where adversaries exploit gradient information to recover original training inputs. These attacks pose a serious privacy threat, as attackers can infer sensitive information solely based on gradients, even without direct access to the dataset.

Despite growing awareness of privacy risks in deep learning, existing privacy-preserving techniques primarily focus on mitigating leakage rather than systematically quantifying and auditing the extent of vulnerability posed by gradient-based attacks. A systematic auditing mechanism is essential for assessing the privacy risks associated with gradient leakage, as well as guiding the development of more effective mitigation strategies.

In this paper, we propose a novel privacy auditing framework that quantifies the degree of privacy vulnerability in deep learning models, particularly in the context of gradient-based data reconstruction attacks. Unlike prior works that focus on obfuscation techniques such as adversarial training and differential privacy, our approach systematically evaluates how effectively an attacker can leverage gradient similarity to approximate training data. By formalizing the relationship between gradient similarity and data similarity, our method provides a precise measure of privacy risk.

The proposed auditing framework serves as a practical tool for organizations and researchers, enabling them to identify and mitigate potential privacy vulnerabilities in their models. Our contributions are threefold:

- We introduce a new privacy auditing method specifically aimed at evaluating the privacy risks associated with deep learning models used in image pattern recognition. The framework assesses how adversaries can exploit gradient similarities to reconstruct private data, providing a systematic approach to understanding privacy vulnerabilities in deep learning models.
- We formalize the relationship between gradient similarity and data similarity, which helps in quantifying how gradient information can be used to reconstruct original data. They demonstrate how adversaries can iteratively refine their reconstructions using this relationship, offering deep insights into potential risks posed by stolen gradient data.
- We evaluate how differential privacy can mitigate the risks of gradient-based reconstruction attacks. Through experimental results, the study shows that applying differential privacy significantly increases the variance in gradient similarity, making it harder for attackers to reconstruct original data, thus demonstrating the effectiveness of differential privacy in safeguarding models from privacy breaches.

## 2 Related work

Existing research has explored various approaches to mitigating privacy risks in deep learning, primarily focusing on reducing gradient leakage rather than auditing and quantifying the risk posed by gradient similarity. In this section, we critically analyze these prior works in terms of the techniques employed and their technical shortcomings.

### 2.1 Gradient-based privacy attacks

Shokri et al. [13] introduced a model inversion attack that exploits gradients to reconstruct private training data, demonstrating that attackers can recover sensitive information without direct access to the original dataset. Their work provided foundational insights into gradient leakage but did not offer a systematic approach to quantifying privacy risks. Zhang et al. [14] extended this idea by analyzing the trade-offs between model accuracy and privacy protection, showing that simple obfuscation techniques are insufficient to fully prevent leakage. However, their study lacked an auditing framework that explicitly measures the risk of data reconstruction based on gradient similarity.

### 2.2 Mitigation techniques

Differential privacy (DP) has been widely adopted as a technique for mitigating gradient-based attacks. Hitaj et al. [15] demonstrated that adding noise to model gradients effectively reduces the information available to attackers. However, DP often results in a trade-off between privacy and model utility, leading to a loss in model performance. Papernot et al. [16] proposed adversarial training as a defense against model inversion attacks, aiming to obfuscate gradients and limit their effectiveness for reconstruction. While these techniques reduce privacy risks, they primarily function as defensive mechanisms rather than tools for systematically auditing vulnerabilities.

### 2.3 Limitations of existing methods

A major limitation of existing research is the lack of a structured approach to quantifying privacy risks. Most studies focus on proposing defense mechanisms, but few systematically measure the extent to which gradient similarity enables data reconstruction. Additionally, while prior works demonstrate that privacy risks exist, they do not establish a formal relationship between gradient similarity and data reconstruction accuracy. This gap in understanding limits the ability of organizations to assess their model's privacy vulnerabilities effectively.

Unlike prior research, our work does not propose another obfuscation technique; instead, we introduce a systematic auditing framework to quantify privacy risks. By formally defining the relationship between gradient similarity and data similarity, our approach provides a structured way to assess vulnerability. Moreover, while differential privacy and adversarial training are useful in mitigating attacks, our work highlights their effectiveness (or limitations) by evaluating their impact on gradient leakage in a measurable way and fills a critical gap by offering a formalized, quantifiable approach to assessing privacy risks in deep learning models. This structured auditing framework provides deeper insights into the risks posed by gradient leakage, enabling more informed decisions in deploying privacy-preserving machine learning models.

## 3 Preliminaries

For simplicity, we define key symbols in Table 1, which will be referenced throughout the following sections.

### 3.1 Gradient-based data reconstruction attacks

A data reconstruction attack aims to recover the original input $\mathbf{x}_0$ by exploiting gradient information. An adversary initializes a dummy input $\mathbf{x}$ and iteratively updates it to match the gradients of the target data $\mathbf{x}_0$. The optimization objective is:

$$\min_{\mathbf{x}} \|g_{\mathbf{x}_0} - g_{\mathbf{x}}\|, \tag{1}$$

where $\| \cdot \|$ represents a distance metric for gradient similarity.

This attack is particularly critical in privacy-sensitive domains such as healthcare or finance, where even partial reconstruction of data poses significant privacy risks.

### 3.2 Differential privacy as a defense mechanism

Differential privacy (DP) [17] provides mathematical privacy guarantees by ensuring the output of a function remains statistically indistinguishable when an individual's data is modified.

**Definition 1** (*Differential privacy* [17]) A randomized mechanism $\mathcal{M}$ satisfies $(\epsilon, \delta)$-differential privacy if for any two adjacent datasets $\mathbb{D}$ and $\mathbb{D}'$ differing by a single entry, and for all $S \subseteq \text{Range}(\mathcal{M})$,

$$\Pr[\mathcal{M}(\mathbb{D}) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(\mathbb{D}') \in S] + \delta. \tag{2}$$

To protect gradients during model training, DP is commonly implemented using Differentially Private Stochastic Gradient Descent (DP-SGD), which consists of four steps:

1. Compute the gradient $g$ of the loss function.
2. Clip $g$ to bound its sensitivity.
3. Add Gaussian noise to $g$: $g' = g + \mathcal{N}(0, \sigma^2 I)$.
4. Update model parameters using $g'$.

The noise scale $\sigma$ is selected based on the Gaussian mechanism:

**Lemma 1** (Gaussian Mechanism for Differential Privacy [17]) *Let* $\mathcal{M} : \mathbb{D} \to \mathbb{R}^k$ *have* $\ell_2$*-sensitivity*$\Delta_2\mathcal{M} = \|\mathcal{M}(\mathbb{D}) - \mathcal{M}(\mathbb{D}')\|$ *. The Gaussian Mechanism ensures* $(\epsilon, \delta)$*-DP if* $\sigma \geq \frac{\Delta_2\mathcal{M}\cdot\sqrt{2\log(1.25/\delta)}}{\epsilon}$.

By incorporating these privacy-preserving mechanisms, DP-SGD mitigates the risk of data reconstruction attacks.

# 4 Reconstruction privacy auditing method

## 4.1 Main ideas on reconstruction privacy auditing method

**Algorithm 1** Privacy Audit Algorithm Based on Gradient Similarity Analysis

```
Require: Public dataset $\mathbb{D} = \{\mathbf{x}_i \mid i = 1 \text{ to } N\}$
Require: Audited model $F(\mathbf{x}; W)$
Require: Similarity functions $Sim_{\text{data}}(\mathbf{x}, \mathbf{x}_i)$, $Sim_{\text{grad}}(g, g_i)$
Require: Number of reconstruction iterations $K$ return Mean and variance of
    coefficients $\{A_{\mathbf{x}_i}\}_{i=1}^N$ and $\{B_{\mathbf{x}_i}\}_{i=1}^N$
// Step 1: Initialization
Initialize empty lists: A_list $\leftarrow$ [ ], B_list $\leftarrow$ [ ]
for each data sample $\mathbf{x}_i$ in $D$ do
    // Step 2: Calculate gradient for the data sample
    $g_i \leftarrow \nabla F(\mathbf{x}_i; W)$
    if Audited Under Differential Privacy then
        $g_i \leftarrow g_i + \mathcal{N}(0, \sigma^2)$
    end if
    // Step 3: Conduct reconstruction attack
    Initialize dummy data $\mathbf{x}$ randomly
    Initialize empty lists: Sim_data_list $\leftarrow$ [ ], Sim_grad_list $\leftarrow$ [ ]
    for $k = 1$ to $K$ do
        $g \leftarrow \nabla F(\mathbf{x}; W)$
        sim_grad $\leftarrow$ Sim$_{\text{grad}}(g, g_i)$
        Append sim_grad to Sim_grad_list
        sim_data $\leftarrow$ Sim$_{\text{data}}(\mathbf{x}, \mathbf{x}_0)$
        Append sim_data to Sim_data_list
        // Update dummy data to minimize gradient difference
        $\mathbf{x} \leftarrow \mathbf{x} - \eta\nabla_{\mathbf{x}}\|g - g_i\|^2$
    end for
    // Step 4: Polynomial fitting for each $\mathbf{x}_0$
    Fit quadratic polynomial: sim_data $= A_{\mathbf{x}_i} \cdot (\text{sim\_grad})^2 + B_{\mathbf{x}_i} \cdot$ sim_grad
    Append coefficients $A_{\mathbf{x}_i}$ and $B_{\mathbf{x}_i}$ to A_list and B_list
end for
// Step 5: Calculate means and variances of coefficients
$mean_A \leftarrow$ Mean(A_list)
$var_A \leftarrow$ Variance(A_list)
$mean_B \leftarrow$ Mean(B_list)
$var_B \leftarrow$ Variance(B_list)
return $mean_A$, $var_A$, $mean_B$, and $var_B$
```

**Table 1** Notation and definitions

| Notation | Description |
|---|---|
| $\mathbf{x}_0$ | Original input data |
| $\mathbf{x}$ | Dummy input initialized randomly |
| $g_{\mathbf{x}_0}$ | Gradient of $\mathbf{x}_0$ |
| $g_{\mathbf{x}}$ | Gradient of $\mathbf{x}$ |
| $\|\cdot\|$ | Distance metric ($L_2$ norm) |
| $F(\mathbf{x};W)$ | Target model with parameters $W$ |
| $\mathbb{D}$ | Public dataset used by the auditor |
| $\mathcal{H}_F$ | Function mapping gradient similarity to data similarity |
| $\nabla\mathcal{H}_F$ | First-order derivative of $\mathcal{H}_F$ |
| $Sim(\mathbf{x},\mathbf{x}_0)$ | Data similarity measure |
| $Sim(g,g_0)$ | Gradient similarity measure |
| $\mathcal{N}(0,\sigma^2 I)$ | Gaussian noise for differential privacy |
| $\epsilon,\delta$ | Privacy budget parameters in differential privacy |
| $A_{\mathbf{x}_i}, B_{\mathbf{x}_i}$ | Coefficients of polynomial fit in privacy auditing |

Privacy risk auditing refers to the systematic evaluation of a machine learning model's susceptibility to privacy attacks by measuring the extent to which sensitive information can be inferred from model outputs, such as gradients. In the context of this work, privacy risk auditing assesses the vulnerability of a model to gradient-based data reconstruction attacks by quantifying the relationship between gradient similarity and data similarity. Our framework provides a structured approach to evaluate these risks and determine the effectiveness of privacy-preserving mechanisms such as differential privacy. We firstly give a formal definition of gradient similarity.

**Definition 2** (*Gradient Similarity*) Given two gradients $g = \nabla F(\mathbf{x};W)$ and $g_0 = \nabla F(\mathbf{x}_0;W)$ computed from two data points $\mathbf{x}$ and $\mathbf{x}_0$, the gradient similarity $Sim(g,g_0)$ is defined as the Mean Square Distance (MSE) between these gradient vectors:

$$Sim(g,g_0) = \|g - g_0\|_2 \tag{3}$$

where $\|\cdot\|_2$ represents the $\ell_2$-norm.

MSE is widely used in gradient-based analyses as it quantifies the alignment between gradients, which directly influences data reconstruction accuracy. A smaller $Sim(g,g_0)$ value indicates stronger alignment, implying greater privacy risk.

To audit the privacy risks of a machine learning model $F(\mathbf{x};W)$, the auditor analyzes the relationship between gradient similarity and data similarity using a public dataset $\mathbb{D}$. Given an input $\mathbf{x}$, the auditor computes gradients $g = \nabla F(\mathbf{x};W)$ and estimates a function $\mathcal{H}_F$:

$$Sim(\mathbf{x},\mathbf{x}_0) = \mathcal{H}_F(Sim(g,g_0)), \tag{4}$$

where $Sim(\mathbf{x},\mathbf{x}_0)$ measures data similarity and $Sim(g,g_0)$ measures gradient similarity. The attacker exploits $\mathcal{H}_F$ and its derivative $\nabla\mathcal{H}_F$ to refine dummy data $\mathbf{x}$ for improved reconstruction accuracy. The sensitivity of $\mathcal{H}_F$ (i.e., $\nabla\mathcal{H}_F$) determines how small changes in gradient similarity impact data similarity.

### 4.2 Conceptual framework

The overall auditing framework consists of key steps 1–6 involved in estimating $\mathcal{H}_F$ and assessing privacy risks:

1. Sampling data from a public dataset $\mathbb{D}$.
2. Computing gradients $g$ for each sample.
3. Simulating reconstruction attacks by iteratively adjusting $\mathbf{x}$.
4. Fitting a polynomial function to estimate $\mathcal{H}_F$.
5. Computing statistical properties of the fitted coefficients.
6. Generating a privacy audit report.

### 4.3 Steps to conduct privacy auditing

To conduct privacy audits, the auditor performs the following key operations:

**Step 1: Sampling data from $\mathbb{D}$**

The auditor selects samples from $\mathbb{D}$, which act as a surrogate for private training data. These samples provide reference points for evaluating privacy leakage.

**Step 2: Calculating gradients**

Each sample $\mathbf{x}_i \in \mathbb{D}$ is processed through $F(\mathbf{x};W)$ to compute gradients $g_i = \nabla F(\mathbf{x}_i;W)$. If differential privacy is applied, noise is added: $g_i \leftarrow g_i + \mathcal{N}(0, \sigma^2 I)$.

**Step 3: Simulating reconstruction attacks**

Dummy data $\mathbf{x}$ is iteratively adjusted to match $g_i$. During this process, $Sim(\mathbf{x}, \mathbf{x}_i)$ and $Sim(g, g_i)$ are recorded.

**Step 4: Polynomial fitting of $\mathcal{H}_F$**

The relationship between gradient similarity and data similarity is captured using a quadratic polynomial:

$$Sim(\mathbf{x}, \mathbf{x}_i) = A_{\mathbf{x}_i}(Sim(g, g_i))^2 + B_{\mathbf{x}_i}(Sim(g, g_i)). \tag{5}$$

The choice of a quadratic polynomial for fitting $\mathcal{H}_F$ is based on empirical observations from our experiments. Higher-order polynomials (e.g., cubic or quartic) were considered but did not provide significant improvements in goodness-of-fit while introducing unnecessary complexity and susceptibility to overfitting. Linear regression, on the other hand, failed to capture the non-linear relationship between gradient similarity and data similarity, leading to poorer approximation accuracy.

To validate this choice, we performed polynomial fitting with varying degrees and evaluated the residual sum of squares (RSS). Quadratic fitting consistently demonstrated a balance between accuracy and simplicity, with minimal error increase compared to higher-degree polynomials while avoiding overfitting artifacts. Furthermore, using a quadratic function aligns with prior works in adversarial robustness analysis, where second-order approximations are often sufficient to model local decision boundaries.

**Step 5: Statistical analysis of fitted coefficients**

The auditor computes the means and variances of $A_{\mathbf{x}_i}$ and $B_{\mathbf{x}_i}$ across all samples to assess the consistency of $\mathcal{H}_F$.

**Step 6: Privacy audit report generation**

The privacy audit report summarizes the strength and stability of $\mathcal{H}_F$, quantifying the privacy risks posed by data reconstruction attacks.

### 4.4 Interpreting the privacy report

The privacy report helps determine whether privacy risks can be mitigated by techniques like differential privacy or training adjustments.

Impact of $\mathcal{H}_F$ and $\nabla\mathcal{H}_F$: The means of $A_{\mathbf{x}_i}$ and $B_{\mathbf{x}_i}$ indicate the typical strength of the correlation between gradient and data similarity.

Attacker's confidence: A low variance in $A_{\mathbf{x}_i}$ and $B_{\mathbf{x}_i}$ suggests consistent privacy leakage, whereas high variance introduces uncertainty.

Reconstruction sensitivity: A higher mean of $\nabla\mathcal{H}_F$ implies that small changes in gradients lead to significant improvements in data reconstruction.

**Table 2** Model architectures and training configurations

| Model | Architecture | Training algorithm | Dataset |
|---|---|---|---|
| CNN | 1 Conv (12 filters, 5 x 5) 1 FC | Adam ($10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) | MNIST |
| LeNet | 2 Conv (6,16 filters, 5 x 5) 2 FC (120, 84) | Adam ($10^{-2}$) | CIFAR-10<br>CIFAR-100 |
| ResNet-18 | 18-layer residual network | SGD (momentum,$10^{-3}$) | FMNIST |

# 5 Experiments

## 5.1 Experiment setup

We evaluate our proposed privacy auditing method against gradient-based data reconstruction attacks using well-established deep learning models and benchmark datasets. The audited models include **CNN** [18], **LeNet** [19], and **ResNet-18** [20], trained on **MNIST**,[1] **CIFAR-100**,[2] **CIFAR-10**[3], and **FMNIST**[4]. These models are widely used in privacy and security-related research due to their simple architectures and publicly available datasets. We summarize the architectures and training configurations in Table 2.

For all models, parameters are randomly initialized and remain fixed during reconstruction attacks. Gradient-based data reconstruction is evaluated using gradients computed during training, and performance is measured based on reconstruction accuracy.

### 5.1.1 Machine configuration

Experiments were run on a machine with the following specifications: CPU: Intel Xeon Processor, GPU: NVIDIA Tesla V100 with 32 GB of VRAM, RAM: 128 GB, and OS: Ubuntu 20.04. All experiments were conducted using the PyTorch framework, with additional tools for logging and tracking results.

## 5.2 Experiment results

To evaluate the privacy auditing framework, we focus on the reconstruction process on public datasets which is presented in Algorithm 1. The core objective is to determine the coefficients of the function $\mathcal{H}_F$, which describes the relationship between the similarity of gradients and the underlying data. The models are audited under two different conditions: without differential privacy and with differential privacy applied during training.

For each condition, the privacy auditor selects a single class of images from the public dataset and attempts to reconstruct the images from their gradients. The coefficients of $\mathcal{H}_F$, which describe the relationship between gradient similarity and data similarity, are estimated for each image. We calculate both the mean and variance of these coefficients to evaluate the consistency and effectiveness of the gradient-based reconstruction process.

### 5.2.1 Without differential privacy

When gradients $g_i$ are not perturbed with differentially private noise, the gradient-based reconstruction process is highly effective. As shown in Table 3, the MSE between the original and reconstructed images decreases rapidly with each iteration, indicating quick convergence of the dummy data to the original images. The gradient similarity and the similarity of the reconstructed images are presented in Fig. 1. As observed, in the absence of differential privacy, as the gradient similarity gradually decreases, the similarity of the reconstructed images also diminishes, with both following a pattern that closely resembles a quadratic polynomial function. In contrast, when differential privacy is introduced, the image similarity tends toward zero when the gradient similarity becomes very small.

**5.2.1.1 Reconstruction accuracy** The low MSE values signify that the reconstructed images are nearly identical to the original images. For instance, in the case of CIFAR-10, the MSE drops from 1.4347 at iteration 0 to 0.0001 at iteration 150. Similar trends are observed for CIFAR-100, MNIST, and FMNIST. This rapid decrease in MSE demonstrates that an attacker can successfully reconstruct the original data with high accuracy using gradient information alone. A low MSE value indicates that the reconstructed images are almost identical to the original images, demonstrating a high privacy risk.

**5.2.1.2 Coefficient behavior** As shown in Table 4, the coefficients of $\mathcal{H}_F$ are very stable across all images. For example, the mean value of $A_{\mathbf{x}_i}$'s on LeNet trained with CIFAR10 is 0.00550, while the mean value of $B_{\mathbf{x}_i}$'s is 0.0050. More impor-

[1] https://yann.lecun.com/exdb/mnist/.
[2] https://www.cs.toronto.edu/kriz/cifar.html.
[3] https://www.cs.toronto.edu/k̃riz/cifar.html.
[4] https://github.com/zalandoresearch/fashion-mnist.

tantly, the variance of these coefficients is extremely small. For example, the variance value of $A_{\mathbf{x}_i}$'s on LeNet trained with CIFAR10 is 9e-6, while the variance value of $B_{\mathbf{x}_i}$'s is 8e-6. This low variance indicates a strong and consistent relationship between gradient similarity and data similarity, meaning the auditor could reliably reconstruct the original images with high accuracy based on the gradients alone.

**5.2.1.3 Parameters and resultant output** We substitute the means of $A_{\mathbf{x}}$ and $B_{\mathbf{x}}$ into Eq. (5) and calculate the MSE estimate using the similarity of the gradients at the termination of the reconstruction attack process. The results are recorded in Table 5. As observed, the estimated MSE closely matches the actual MSE, which can be attributed to the small variance of $A_{\mathbf{x}}$ and $B_{\mathbf{x}}$ in the absence of differential privacy (as shown in Table 4). In this case, the reconstruction error can be effectively estimated.

**5.2.1.4 Implication** The small variance in the coefficients suggests that the reconstruction process is highly predictable and effective. The coefficients $A_{\mathbf{x}_i}$ and $B_{\mathbf{x}_i}$ provide a measure of how easily an attacker can correlate gradients with the original data. Lower variance in these coefficients indicates that reconstruction is consistently successful across different data points. The attacker can use the estimated $\mathcal{H}_F$ function to consistently reconstruct images from gradient information, posing a significant privacy risk.

#### 5.2.2 With differential privacy

In contrast, when the gradients are perturbed with differentially private noise, namely the reconstruction process is based on the noisy gradients $g_i \leftarrow g_i + \mathcal{N}(0, \sigma^2)$. We fix $\sigma^2 = 0.00001$and we find the gradient-based reconstruction process is significantly less effective under this situation. As shown in Fig. 1, under the influence of differential privacy, the similarity of the reconstructed images does not drop to zero. Instead, all models exhibit a noticeable lower bound for the image similarity of the reconstructed images. The key findings are as follows:

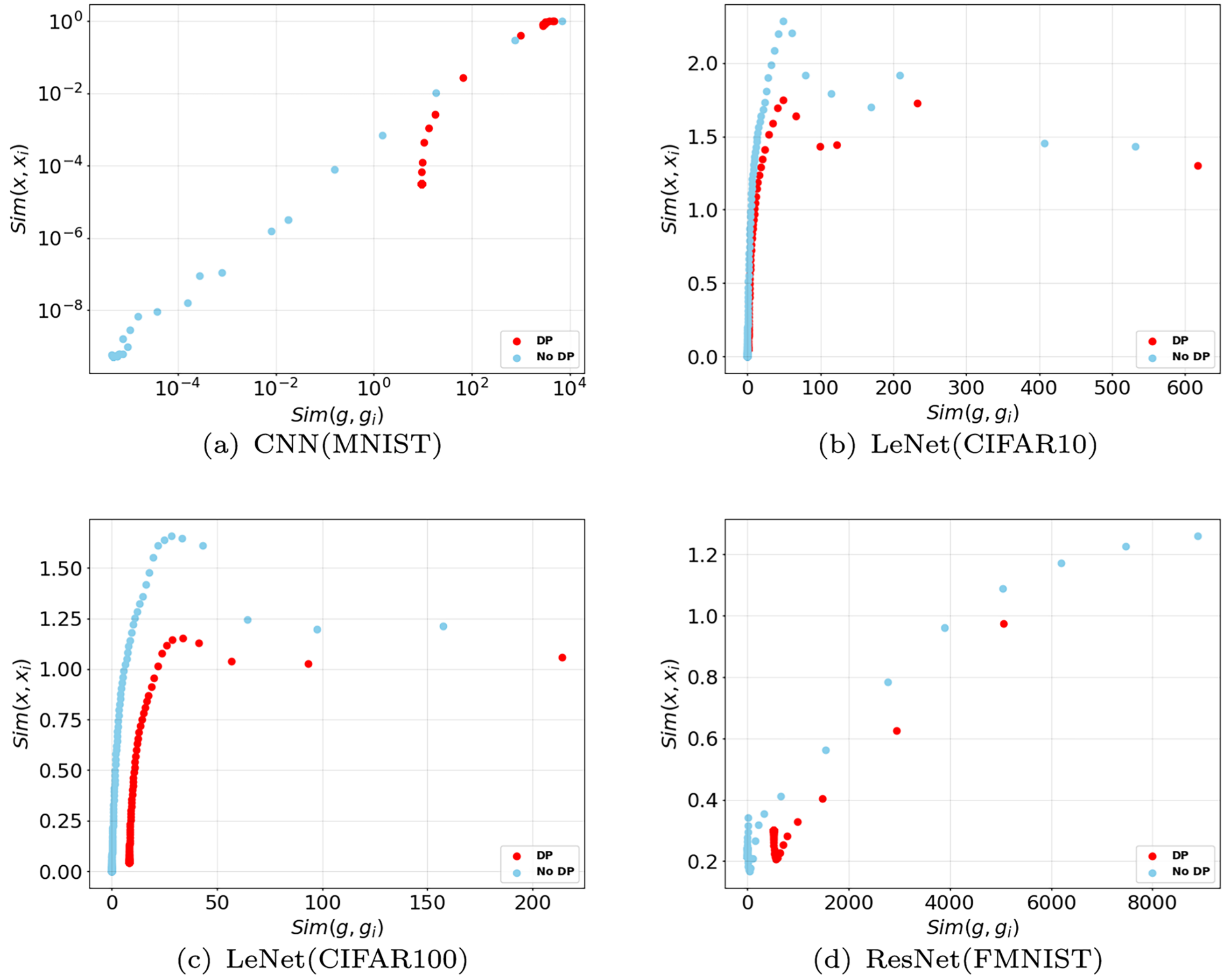
**5.2.2.1 Reconstruction accuracy** As illustrated in Table 3, the MSE remains high even after many iterations when differential privacy is applied. For example, in FMNIST with DP, the MSE decreases from 1.4347 at iteration 0 to 0.5785 at iteration 30. In CIFAR-10 with DP, the MSE decreases only marginally from 1.3089 at iteration 0 to 0.0537 at iteration 150, which is 500 times compared with the situation without differential privacy. This indicates that the reconstructed images remain significantly different from the original images, and the attacker is unable to recover meaningful data. Overall, differential privacy introduces significant noise into the gradients, rendering reconstruction attacks less effective across all datasets and models.

**5.2.2.2 Coefficient behavior** As shown in Table 4, the coefficients of $\mathcal{H}_F$ exhibit much higher variance when the gradients are perturbed with differentially private noise. The variance values of $A_{\mathbf{x}_i}$'s and $B_{\mathbf{x}_i}$'s increase by 10 times or even 100 times, respectively. This large variance suggests that the relationship between gradient similarity and data similarity became unstable, and the attacker could no longer rely on gradient similarity to accurately reconstruct the original data.

**5.2.2.3 Parameters and resultant output** When differential privacy is introduced, we similarly substitute the means of $A_{\mathbf{x}}$ and $B_{\mathbf{x}}$ into Eq. (5) and use the similarity of the gradients at the termination of the reconstruction attack process to compute the MSE estimate. The results are recorded in Table 6. It can be observed that the estimated MSE differs significantly from the actual MSE. This discrepancy is due to the larger variance of $A_{\mathbf{x}}$ and $B_{\mathbf{x}}$ under differential privacy (Table 4), making the reconstruction error difficult to estimate accurately.

**Table 3** Mean squared error (MSE) of data reconstruction over iterations

| Iteration | CIFAR-10 (No DP / DP) | CIFAR-100 (No DP / DP) | MNIST (No DP / DP) | FMNIST (No DP / DP) |
|---|---|---|---|---|
| 0 | 1.4347/1.3089 | 1.2130/1.15880 | 1.1604/1.0220 | 1.5997/1.4347 |
| 30 | 0.6886/0.6860 | 0.5496/0.3414 | 5.1062e-10/3.0911e-5 | 0.2424/0.5785 |
| 60 | 0.0409/0.0774 | 0.0713/0.9833 | –/– | –/– |
| 90 | 0.0018/0.0462 | 0.0063/0.0672 | –/– | –/– |
| 120 | 0.0001/0.0516 | 0.0007/0.0477 | –/– | –/– |
| 150 | 0.0001/0.0537 | 0.0001/0.0489 | –/– | –/– |



(a) CNN(MNIST)

(b) LeNet(CIFAR10)

(c) LeNet(CIFAR100)

(d) ResNet(FMNIST)

**Fig. 1** $Sim(\mathbf{x}, \mathbf{x}_i)$ v.s. $Sim(g, g_i)$ (DP v.s. No DP)

**Table 4** Privacy auditing results for various models and datasets

| Dataset | Model | DP | $mean_A$ | $var_A$ | $mean_B$ | $var_B$ | MSE |
|---|---|---|---|---|---|---|---|
| CIFAR-10 | LeNet | No | 0.00650 | 9e–6 | 0.0050 | 8e–6 | 0.0001 |
| | | Yes | 0.00020 | 5e–4 | 0.0030 | 0.0004 | 0.5000 |
| CIFAR-100 | LeNet | No | 0.00500 | 0.00007 | 0.0028 | 0.0006 | 0.0001 |
| | | Yes | 0.00007 | 0.00100 | 0.0002 | 0.0090 | 0.6000 |
| MNIST | CNN | No | 0.00077 | 0.00005 | 0.0008 | 0.0002 | 5e–6 |
| | | Yes | 0.00008 | 0.0016 | 0.0020 | 0.0054 | 0.1000 |
| FashionMNIST | ResNet9 | No | 0.01000 | 0.1130 | 0.0180 | 0.0100 | 0.2424 |
| | | Yes | 0.00030 | 0.9200 | 0.0025 | 0.3200 | 0.5477 |

**5.2.2.4 Implication** The large variance in the coefficients of $\mathcal{H}_F$ indicates that differential privacy introduces significant uncertainty into the reconstruction process. By adding noise to the gradients, differential privacy effectively hampers the attacker's ability to find a consistent mapping between gradient and data similarity, thereby protecting the model

**Table 5** Parameters and resultant output (Without DP)

| Model (dataset) | $A_{\mathbf{x}}$ | $B_{\mathbf{x}}$ | Estimated MSE | Actual MSE | Classification performance (%) |
|---|---|---|---|---|---|
| CNN(MNIST) | 0.00077 | 0.0008 | 6.8e−6 | 5e−6 | 98 |
| LeNet(CIFAR10) | 0.00650 | 0.0050 | 0.00014 | 0.0001 | 87 |
| LeNet(CIFAR100) | 0.00500 | 0.0028 | 0.00019 | 0.0001 | 70 |
| ResNet9(MNIST) | 0.01000 | 0.0180 | 0.21170 | 0.2424 | 93 |

from gradient-based attacks. This demonstrates the effectiveness of differential privacy in reducing privacy risks, even at the cost of reduced reconstruction accuracy.

## 5.3 Discussions

### 5.3.1 Model complexity

Model complexity plays a crucial role in the vulnerability to gradient-based reconstruction attacks. The results show that even without differential privacy, models trained on datasets with higher complexity (CIFAR-10, CIFAR-100) are slightly harder to reconstruct compared to simpler datasets (MNIST). This is reflected in the variance of the coefficients of $\mathcal{H}_F$ and the reconstruction accuracy:

*Simpler models and datasets* (like MNIST, which involves grayscale digits) have a lower variance in the coefficients and achieve nearly perfect reconstruction. The consistent structure of the images allows the auditor to accurately match gradients, making it easier to recover the original data.

*More complex datasets* (like CIFAR-100, which includes diverse color images from 100 classes) introduce slightly more variability in the reconstruction process. The MSE is higher, and the variance of the coefficients increases, although it remains small without differential privacy. This shows that while complex models and datasets are somewhat more resistant to attacks, they are still vulnerable to gradient-based reconstruction. These findings indicate that the complexity of the model and the dataset can influence the attacker's success but is not sufficient on its own to protect against privacy risks. Even with complex data, the absence of privacy-preserving techniques leaves models exposed to reconstruction attacks.

### 5.3.2 Importance of differential privacy

The experiments clearly show that differential privacy is an essential defense mechanism in mitigating gradient-based attacks. When models are trained with differential privacy, the reconstruction quality is drastically reduced, and the variance of the coefficients of $\mathcal{H}_F$ increases significantly. This highlights several important points:

**5.3.2.1 Noise introduction** Differential privacy works by introducing noise into the gradients during the training process. This noise breaks the direct correlation between gradient similarity and data similarity, making it much harder for an attacker to reconstruct the original images. The increase in variance of the coefficients shows how differential privacy effectively disrupts the attacker's ability to find consistent patterns.

**Table 6** Parameters and resultant output (with DP)

| Model (dataset) | $A_{\mathbf{x}}$ | $B_{\mathbf{x}}$ | Estimated MSE | Actual MSE | Classification performance (%) |
|---|---|---|---|---|---|
| CNN(MNIST) | 0.00008 | 0.00200 | 0.45800 | 0.1000 | 91 |
| LeNet(CIFAR10) | 0.00020 | 0.00300 | 0.37882 | 0.5000 | 79 |
| LeNet(CIFAR100) | 0.00007 | 0.00020 | 0.8732 | 0.6000 | 64 |
| ResNet9(MNIST) | 0.00030 | 0.00250 | 0.9211 | 0.5477 | 85 |

**5.3.2.2 Effectiveness across datasets** The impact of differential privacy is consistent across different datasets, indicating that the technique can be applied broadly to safeguard models, regardless of the dataset's complexity. However, more complex datasets (like CIFAR-100) still exhibit higher variance in reconstruction, suggesting that the protective effects of differential privacy are even more pronounced for complex data.

**5.3.2.3 Trade-offs** While differential privacy reduces privacy risks, it does come at the cost of reducing the model's performance. In sensitive applications where privacy is paramount, this trade-off is justified. However, for other scenarios where high accuracy is critical, balancing privacy and performance may require fine-tuning the noise level in the differential privacy mechanism. These findings suggest that while differential privacy offers strong protection against gradient-based attacks, it comes at the cost of reduced model accuracy. In sensitive applications, such as healthcare or finance, these trade-offs must be carefully considered to balance privacy and performance.

#### 5.3.3 The need for privacy auditing

The results underline the importance of privacy auditing in understanding the privacy risks posed by machine learning models. Even without active attacks, auditing gradients can reveal the degree to which sensitive data can be reconstructed:

**5.3.3.1 Small variance without differential privacy** The small variance in $\mathcal{H}_F$ coefficients when models are trained without privacy-preserving techniques underscores how predictable and effective the reconstruction process is. This predictability is a major concern because it suggests that attackers could reliably exploit gradients to extract sensitive data.

**5.3.3.2 Large variance with differential privacy** Conversely, the large variance observed when differential privacy is applied demonstrates how auditing can help organizations gauge the effectiveness of their privacy defenses. If the variance in $\mathcal{H}_F$ is large, as shown in our experiments, it suggests that the model is more resistant to gradient-based attacks.

**5.3.3.3 Proactive defense** Regular privacy auditing, especially in sensitive applications like healthcare or finance, can help organizations proactively identify potential privacy leaks. It provides a framework for evaluating how well privacy-preserving techniques are working and whether additional measures, such as stronger differential privacy or adversarial training, are necessary.

### 5.4 Limitations

Despite promising results, our study has several limitations:

- Benchmark datasets: We primarily evaluate our framework using standard datasets, which may not fully represent real-world complexities in sensitive domains.
- Effectiveness of defenses: The performance of differential privacy and other defenses may vary depending on model architectures and specific applications.
- Adversarial adaptation: Attackers may develop more sophisticated gradient-based attacks that circumvent existing defenses, necessitating continuous updates to auditing methodologies.

Future work should focus on mitigating these limitations by extending experiments to diverse real-world datasets and continuously improving defense mechanisms.

## 6 Conclusion and future work

In this paper, we introduced a privacy auditing framework to assess the vulnerability of deep learning models to gradient-based data reconstruction attacks. Our findings highlight that attackers can effectively exploit gradient similarity to reconstruct sensitive data, posing significant privacy risks.

Through extensive experiments, we demonstrated that deep learning models such as CNN, LeNet, and ResNet remain susceptible to these attacks in the absence of privacy-preserving mechanisms. The application of differential privacy effectively mitigates these risks by introducing noise, though at the cost of reduced model accuracy. This highlights the inherent trade-off between privacy and utility in machine learning systems.

Looking ahead, future research should explore additional privacy-enhancing techniques such as homomorphic encryption, secure multi-party computation (SMPC), and federated learning to further strengthen model security. Expanding our privacy auditing framework to evaluate broader attack vectors, including model inversion and membership inference attacks, would provide a more comprehensive assessment of privacy risks.

In summary, our work contributes a systematic approach for evaluating privacy vulnerabilities in deep learning models and serves as a foundation for future advancements in privacy-preserving machine learning.

**Author contributions** All authors contributed to the study conception and design. Tao Huang was responsible for the conceptualization, methodology, and initial drafting of the manuscript. Qingyu Huang, who contributed equally to Xin Shi, was involved in the experimental design, data collection, and analysis, as well as manuscript writing and revisions. Data collection, processing, and visualization were performed by Ziyang Chen. Liang Jiang and Chenhuang Wu provided technical support and resources for the experiments. They contributed to the interpretation of the results and reviewed the manuscript. Guolong Zheng and Xu Yang are the corresponding authors. They supervised the research project, provided critical feedback on the study's design, and were involved in manuscript writing and revisions. Wencheng Yang provided expertise in the theoretical framework and overall direction of the study. Wencheng also contributed to data interpretation and critical review of the final manuscript.

**Funding** This study was funded by Natural Science Foundation of Fujian Province (Grant No. [2024J08276] and [2024J8278]), National Natural Science Foundation of China (Grant No. 62402203), Key Research Project for Young and Middle-aged Researchers by the Fujian Provincial Department of Education (Grant No. [JZ230044]) and Open Research Project of Fujian Province Key Laboratory of Financial Information Processing (Grant No. [JXC202402]).

**Data availability** Data is provided within the manuscript or supplementary information files.

## Declarations

**Ethics approval and consent to participate** We confirmed that the experiments follow the criteria of ethics approval and consent to participate. No human subjects were harmed in this research. The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

**Consent for publication** Not applicable.

**Clinical trial registration** Not applicable.

**Permission to reproduce material from other sources** Not applicable.

**Competing Interests** The authors declare no competing interests.

## References

1. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience. 2018;2018(1):7068349.
2. Chai J, Zeng H, Li A, Ngai EW. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications. 2021;6: 100134.
3. Kaur P, Krishan K, Sharma SK, Kanchan T. Facial-recognition algorithms: A literature review. Medicine, Science and the Law. 2020;60(2):131–9.
4. Li L, Mu X, Li S, Peng H. A review of face recognition technology. IEEE access. 2020;8:139110–20.
5. Aggarwal R, Sounderajah V, Martin G, Ting DS, Karthikesalingam A, King D, Ashrafian H, Darzi A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine. 2021;4(1):65.
6. Shehab M, Abualigah L, Shambour Q, Abu-Hashem MA, Shambour MKY, Alsalibi AI, Gandomi AH. Machine learning in medical applications: a review of state-of-the-art methods. Computers in Biology and Medicine. 2022;145: 105458.

7. Zhang T, Li Q, Zhang C-S, Liang H-W, Li P, Wang T-M, Li S, Zhu Y-L, Wu C. Current trends in the development of intelligent unmanned autonomous systems. Frontiers of information technology & electronic engineering. 2017;18:68–85.
8. Kuutti S, Bowden R, Jin Y, Barber P, Fallah S. A survey of deep learning applications to autonomous vehicle control. IEEE Transactions on Intelligent Transportation Systems. 2020;22(2):712–33.
9. Hesamifard E, Takabi H, Ghasemi M, Wright R.N. Privacy-preserving machine learning as a service. Proceedings on Privacy Enhancing Technologies (2018)
10. Jin X, Chen P-Y, Hsu C-Y, Yu C-M, Chen T. Cafe: Catastrophic data leakage in vertical federated learning. Advances in Neural Information Processing Systems. 2021;34:994–1006.
11. Salem A, Bhattacharya A, Backes M, Fritz M, Zhang Y. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In: 29th USENIX Security Symposium (USENIX Security 20), 2020:1291–1308
12. Balle B, Cherubin G, Hayes J. Reconstructing training data with informed adversaries. In: 2022 IEEE Symposium on Security and Privacy (SP), 2022:1138–1156 . IEEE
13. Hayes J, Melis L, Danezis G, De Cristofaro E. Logan: Membership inference attacks against generative models. arXiv preprint arXiv:1705.07663 (2017)
14. Ha T, Dang T.K, Dang T.T, Truong T.A, Nguyen M.T. Differential privacy in deep learning: an overview. In: 2019 International Conference on Advanced Computing and Applications (ACOMP), 2019:97–102 . IEEE
15. Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the gan: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017:603–618
16. Papernot ., McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP), 2016:582–597 . IEEE
17. Dwork C, Roth A, et al. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science 2014;9(3–4):211–407
18. Ketkar N, Moolayil J, Ketkar N, Moolayil J. Convolutional neural networks. Deep learning with Python: learn best practices of deep learning models with PyTorch, 2021:197–242
19. Al-Jawfi R. Handwriting arabic character recognition lenet using neural network. Int Arab J Inf Technol. 2009;6(3):304–9.
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770–778

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.