Cybersecurity

RESEARCH

Open Access

# Iterative and mixed-spaces image gradient inversion attack in federated learning

Linwei Fang[1,2*], Liming Wang[1] and Hongjia Li[1]

**Abstract**

As a distributed learning paradigm, federated learning is supposed to protect data privacy without exchanging users' local data. Even so, the *gradient inversion attack*, in which the adversary can reconstruct the original data from shared training gradients, has been widely deemed as a severe threat. Nevertheless, most existing researches are confined to impractical assumptions and narrow range of applications. To mitigate these shortcomings, we propose a comprehensive framework for gradient inversion attack, with well-designed algorithms for image and label reconstruction. For image reconstruction, we fully utilize the generative image prior, which derives from wide-used generative models, to improve the reconstructed results, by additional means of iterative optimization on mixed spaces and gradient-free optimizer. For label reconstruction, we design an adaptive recovery algorithm regarding real data distribution, which can adjust previous attacks to more complex scenarios. Moreover, we incorporate a gradient approximation method to efficiently fit our attack for FedAvg scenario. We empirically verify our attack framework using benchmark datasets and ablation studies, considering loose assumptions and complicated circumstances. We hope this work can greatly reveal the necessity of privacy protection in federated learning, while urge more effective and robust defense mechanisms.

**Keywords** Privacy preserving, Federated learning, Gradient inversion

## Introduction

In the era of mobile networks, rapidly growing smart devices have become scattered data resources for knowledge discovery and data mining. Nowadays, traditional central machine learning approaches, which have to collect training data in a central server prior to the training phase, cannot fit that situation in terms of data collecting and sharing, because of increasing data privacy legislations and growing public privacy concerns (Lim et al. 2020). Meanwhile, federated learning (FL) (McMahan et al. 2017; Yang et al. 2019), a novel distributed machine learning paradigm, which can collaboratively train a shared model in distributed system, has been widely studied and applied in privacy-sensitive learning tasks, such as next-word prediction (Hard et al. 2018), medical data analysis (Brisimi et al. 2018) or vision object (image) classification.

At the same time, the privacy threats in federated learning are not negligible (Lim et al. 2020; Lyu et al. 2020; Bouacida and Mohapatra 2021). Despite its implicit data protection guarantee that the private data will not leave their local devices in training rounds, a line of studies have recently revealed many possible privacy leakages in FL, from sensitive information inference (Shokri et al. 2017) to original data reconstruction (Zhu et al. 2019).

As one severe privacy attack, *Gradient Inversion* mainly aims to restoring the local training data directly from shared gradients. Previous works have demonstrated various reconstruction approaches, while most of them can merely work well under particular conditions. Specifically, to cope with the highly ill-posed inversion problem,

*Correspondence:
Linwei Fang
fanglinwei@iie.ac.cn
[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

existing researches always introduce different domain knowledge, also known as *prior* knowledge, to simplify the optimization process, while in return restrict their applications in narrow range. In cases of image reconstruction, many approaches choose to impose some natural image priors in form of regularization terms on image space, such as direct constraints like total variation (Geiping et al. 2020) or clip-scale constraint (Geng et al. 2021), and indirect constraints from BN statistics (Yin et al. 2021) or intermediate representations (Jin et al. 2021).

In this work, we propose a comprehensive gradient attack framework concerning the image classification task in FL. We separately design two algorithms for image and label recovery, with an approximation method for gradient estimation in FedAvg scenario. We experiment under real and complex circumstances to validate the effectiveness of our approaches. Our framework shows fairly good optimization efficiency and adaptation to different gradient inversion cases. We hope our attack framework can serve as a supplemental evaluation for privacy leakage, and facilitate the research of robust defense mechanism in federated learning.

The specific contributions of our proposed attack framework are as follows:

- For image reconstruction, we propose an iterative and mixed-spaces gradient inversion algorithm. We not only fully utilize the implicit image prior in generative models, but also carefully design an iterative mixed optimization strategy and adopt the gradient-free optimizer, obtaining fairly natural and realistic reconstructed results. We also verify its strengths in optimization efficiency and restoration accuracy, particularly compared to previous related works.
- For label recovery, we propose a high-accuracy analysis-based recovery algorithm. Through experimental comparison, we show its better adaptation and effectiveness to handle more realistic label distributions than existing methods. It successfully supports existing attack approaches to work normally when faced with troublesome situations, e.g. target batch with repeated-classes samples.
- To mitigate the difficulties in computation and inversion under FedAvg scenarios, we additionally introduce a gradient approximation method. We empirically validate its benefit in cases of weight update, balancing between reconstruction quality and computation workload.
- We demonstrate our framework on different datasets with various image resolutions (i.e. CIFAR, FFHQ and ImageNet), which shows the superiority of our work to other state-of-the-art ones, especially in scenarios of high-resolution images, large batches and repeated labels. Our research also reveals the necessity of stronger defense mechanisms in federated learning.

## Related work

### Privacy threats in FL

The early privacy leakage researches related to FL include member inference attack (Shokri et al. 2017; Nasr et al. 2019) and model inversion attack (Fredrikson et al. 2015). Later researches propose a GAN-based method in FL to construct representations of class-level (Hitaj et al. 2017) and user-level (Wang et al. 2019) private data. Further property inference attack (Melis et al. 2019), shows that more sensitive and detailed information can be extracted from training gradients. Since then, gradient inversion attack that can directly invert training data from shared gradients, has become a hot spot.

### Analysis-based gradient inversion

The basic analytical gradient inversion study originates from Aono et al. (2017). The authors notice that the input of one biased full connection layer can be derived directly from its parameter gradients. Then, Zhu and Blaschko (2020) studies how to recursively reconstruct the input of each layer in network by solving a sequence of linear systems. Moreover, Qian et al. (2020) extensively studies the feasibility of analytical inversion in wider range, such as complex target model and large restored batch, and provides the according minimum success conditions. Nevertheless, even newly analytical methods cannot well handle the case when the batch size is larger than one.

### Optimization-based gradient inversion

Due to their good scalability, optimization-based inversion approaches, which use iterative optimizers such as Stochastic Gradient Descent, are supposed to adapt in wider circumstances. At first, Zhu et al. (2019) formulates the inversion optimization problem using gradient matching loss and L-BFGS (Liu and Nocedal 1989), where the images and labels have to be jointly optimized. Following work (Zhao et al. 2020) designs an analytical label recovery for single sample. After that, many researches (Geiping et al. 2020; Yin et al. 2021; Geng et al. 2021) try to use various image-domain knowledge to guide the space search, by means of adding various regularizers to the cost function, and obtain pretty good results. Besides, inspired by GAN inversion techniques, Jeon et al. (2021) and Li et al. (2022) incorporate the generator into the original inversion optimization, and utilize such implicit image prior to improve the reconstruction quality.

However, nearly all existing studies have problems of too strong assumptions and poor adaptation to some

realistic scenarios, such as repeated labels and large batches. Our work is expected to give the corresponding solutions.

## Preliminaries

### Federated learning

As a distributed learning paradigm, federated learning can inherently protect the privacy of training data held by local devices (users) to some extent. According to different applications, FL has spawned diverse variants and deployments (Yang et al. 2019). To simplify discussion, in this work, we focus on one brief and canonical form. Some necessary descriptions are as follows.

In each FL training round, the central server first broadcasts the global model to users. After that, each user executes local training on its private data and uploads model update (shared gradients) back to the server. The server then aggregates all shared updates to complete this round of training. Typical update and aggregation implementations include FedSGD and FedAvg (McMahan et al. 2017). The main difference between them is whether users perform multi-steps local update in each training round.

For FedSGD:

$$\theta^{t+1} = \theta^t - \sum_{i=1}^{N} \frac{n_i}{n} \nabla\theta_i^t, \tag{1}$$

where $\theta_i^t$ and $\nabla\theta_i^t$ denote the trained parameters and gradients from one-step training on node $i$, in the $t$-th global round, while $\frac{n_i}{n}$ denotes the aggregation weight for node $i$.

For FedAvg:

$$\theta^{t+1} = \sum_{i=1}^{N} \frac{n_i}{n} \theta_i^t, \tag{2}$$

where $\theta_i^t$ denotes the local model weights updated over multi-steps training on node $i$, in the $t$-th global round.

### Attack background

Like most existing researches, we assume the attack occurs in *semi-honest* scenario. The adversary is an honest-but-curious central server in FL, who has access to the global model and FL hyperparameters, as well as the shared data provided by users: for FedSGD, it is the gradients from one-step training on local batch; for FedAvg, it is the local model after multi-steps update. The objective of the adversary is to recover local training data of users from their shared information.

In previous studies, researchers tend to make many strong assumptions and need additional requirement of much useful domain knowledge, often contrary to reality. Compared to them, we assume the attacker can barely obtain extra information except the shared gradients in extreme cases, leading these previous approaches not to work well in our settings. For instance, fixing the BN layers in target model makes it impossible to use BN statistics (Yin et al. 2021) as a regularization.

It is also noteworthy that some researchers (Fowl et al. 2021; Wen et al. 2022; Boenisch et al. 2023) recently study the gradient attack in *malicious* scenario. Under their assumption, the attacker can make arbitrary modification in target model to achieve high-performance inversion. Our study does not consider and compare with such studies, because they only work in more stringent and restricted circumstances.

### Problem formulation

In this section, we describe the formulation of gradient inversion. Specifically, in our work, we focus on the image classification task in FL. For each user, the local training objective is as follows:

$$\min_{\theta} \frac{1}{B} \sum_{(\boldsymbol{x},y)\in\mathbb{D}^B} \ell(f_\theta(\boldsymbol{x}), y). \tag{3}$$

The local batch $\mathbb{D}^B$ is $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_i, y_i) | i \in B\}$, where the batch size is $B$, and the target model $f$ is a neural network parametrized by $\theta$. The $\ell$ is a sample-wise loss function.

To better understand the inversion problem in brief, we herein consider the FedSGD scenario. In this way, the update gradients trained on the local batch is:

$$\nabla\theta = \frac{1}{B} \sum_{i=1}^{B} \nabla\ell(f_\theta(\boldsymbol{x_i}), y_i). \tag{4}$$

The attacker then collects the shared gradients and try to achieve the reconstructed images and labels $(\boldsymbol{X}^*, \boldsymbol{y}^*)$. In most studies, the restoration of label and image can be decoupled. Therefore, if the recovered labels $\boldsymbol{y}^*$ are known, the detailed objective function of image reconstruction is as follows, which takes one gradient matching loss as the core component:

$$\begin{aligned} \arg\min_{\{\boldsymbol{x}_i^* i\in B\}} \; & \mathcal{L}_{\mathrm{GM}}\left(\frac{1}{B}\sum_{i=1}^{B} \nabla\ell\left(f_\theta\left(\boldsymbol{x}_i^*\right), y_i^*\right); \nabla\theta\right) \\ & + \mathcal{R}_{\mathrm{AUX}}\left(\boldsymbol{X}^*\right), \end{aligned} \tag{5}$$

where the gradient discrepancy measure has many options, e.g. $\ell_2$ distance or cosine similarity. Another auxiliary component is a set of regularization terms.

One practical solution for this ill-posed problem is to directly optimize the objective Function (5) on image space, which is generally adopted by most existing approaches. To make the inversion results closer to real images, some state-of-the-art works additionally adopt some regularizers, which essentially implies the guide of various domain knowledge. For instance, $\mathcal{R}_{\text{TV}}$ (Geiping et al. 2020), $\mathcal{R}_{\text{BN}}$ (Yin et al. 2021), $\mathcal{R}_{\text{Clip}}$ (Geng et al. 2021), or $\mathcal{R}_{\text{Feat}}$ (Jin et al. 2021).

## Methodology

To solve the gradient inversion problem, we propose a comprehensive attack framework, the overview is provided in Fig. 1. For image reconstruction, we design an optimization-based algorithm, named as **I**terative **G**radient **I**nversion on **M**ixed **S**paces (IGIMS); while for label recovery, we design an analysis-based algorithm, named as **A**daptive **L**abel **R**ecovery (ALR). We additionally introduce the **A**verage **G**radient **A**pproximation (AGA) method to facilitate our attack in FedAvg cases. We describe all components of our framework in the following sections.

### Iterative gradient inversion on mixed spaces

In this section, we describe our IGIMS algorithm for image reconstruction. We first introduce the *generative image prior*, which supports our gradient inversion as a crucial domain knowledge, and subsequently elaborate on the overall optimization procedure. The profile of this algorithm is in Fig. 1. To simplify writing, in remain of this section, we focus on the case of *single* sample reconstruction. For batch reconstruction, we as well present the complete pseudocode in Algorithm 1.

#### *Generative image prior*

Obviously, due to the non-linearity of deep classification model, the problem of inverting gradients back to original images is ill-posed. Direct search on image space easily falls into local optima, which leads to unnatural spatial structures and artifacts in reconstructed images. To obtain more natural and realistic results, it is feasible to simply impose some natural regularizers on image space, which is the subject of many researches, but also a difficult heuristic work. Different from them, we introduce the *generative image prior*, a novel domain knowledge which is adopted in quite different form as to previous handcrafted regularizers.

Here, we consider a generator $G$ in GAN or any other generative model theories. It is generally believed that, a well-trained generator can memorize a natural image manifold and generate realistic images from inputs of random noises. For the gradient inversion problem, if the reconstructed image space can be constrained on that natural manifold, the final results should have sufficient real image characteristics. Moreover, Dmitry et al. (2020) finds that even given a $G$ with randomly-initialized weights, its outputs still hold some low-level natural image characteristics. Such implicit regularization is called deep image prior, which somewhat represents the pixel-level self-similarity derived from the shared convolution operators in $G$. All

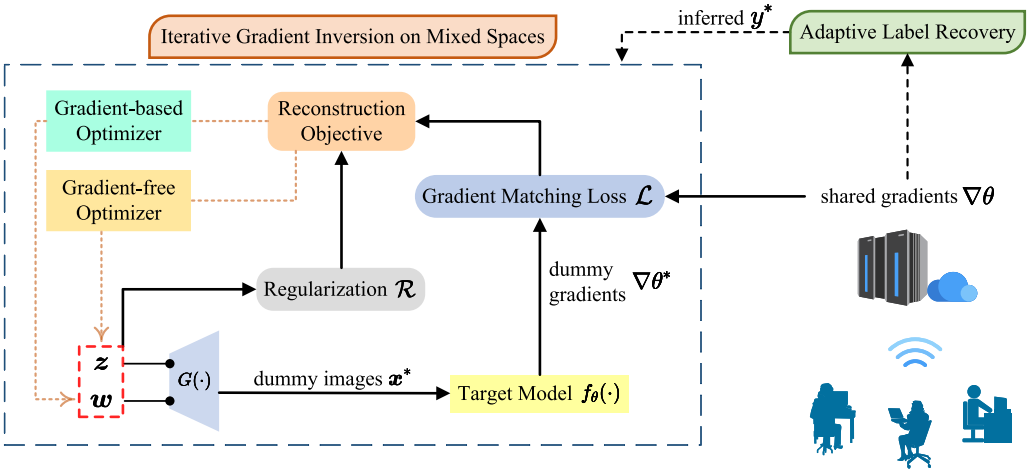



**Fig. 1** An overview of our attack framework, including two recovery algorithms, separately for image (Left, blue dashed box) and label (right, black dashed lines). For image reconstruction, black solid lines denote the forwarding process, light dashed lines denote the back propagation, and the red dashed box contains the real optimized variables. In each attack round, the attacker first collects shared gradients from the server side, then recovers the label, and last inverts the image through optimization

in all, Dmitry et al. (2020) summarizes that, whether $G$ is trained or not, its network architecture can have a major effect on its outputs with a high noise impedance. In other words, the outputs of $G$ often bias low noises and apparent regular spatial structures in visual. To sum up, in our work, these above discussed properties are all referred to the *generative image prior*. Next, we will describe how to apply that prior knowledge to our gradient inversion optimization.

#### *Objective loss function*

We denote a generator $G(\cdot) : \boldsymbol{z} \mapsto \boldsymbol{x}$, which maps a latent code $\boldsymbol{z} \in \mathbb{R}^k$ (commonly sampled from one multivariate Gaussian distribution) to an image $\boldsymbol{x} \in \mathbb{R}^m$, and is parametrized by $\boldsymbol{w} \in \mathbb{R}^n$. As discussed above, to utilize the generative image prior, we need to incorporate the generator into the image search process. Like existing GAN Inversion studies (Zhu et al. 2016; Bau et al. 2019a, b; Huh et al. 2020), we parametrize the reconstructed image $\boldsymbol{x}$ by $\boldsymbol{x} = G(\boldsymbol{z}, \boldsymbol{w})$, and modify the origin objective Function (5) as follows:

$$\arg\min_{\boldsymbol{z},\boldsymbol{w}} \mathcal{L}_{\mathrm{GM}}\left(\nabla \ell\left(f_\theta\left(G(\boldsymbol{z},\boldsymbol{w})\right)\right); \nabla\theta\right) + \mathcal{R}_{\mathrm{AUX}}(\boldsymbol{z},\boldsymbol{w}). \tag{6}$$

Note that we not only try to find the optimal $\boldsymbol{z}$, but also fine-tune the $\boldsymbol{w}$ to obtain the best reconstructed result [similar to Image-Specific Adaptation in Bau et al. (2019a)]. We hereinafter describe the whole optimization procedure, which comprises an external loop with two internal search stages.

#### *Search on mixed spaces*

We firstly define the internal two-stages search procedure in our algorithm. For each stage, the optimization is conducted on one variable space. Note that in the following two formulas (7, 8), we use superscript '$*$' to denote the truly optimized variable in different stages.

In the first stage, the optimizer searches the optimal latent code $\boldsymbol{z}^*$ with fixed $\boldsymbol{w}$. In other words, we try to reproduce the best match image on fixed image manifold. If $G$ already holds enough domain knowledge to reconstruct accurately, this search procedure will be efficient. Except for that, the latent space $k$ is normally much smaller than the image space $m$, which means the difficulty of inversion should be reduced to some extent. The optimization objective at this stage is as follows:

$$\arg\min_{\boldsymbol{z}^*} \mathcal{L}_{\mathrm{GM}}\left(\nabla \ell\left(f_\theta\left(G(\boldsymbol{z}^*,w)\right)\right); \nabla\theta\right) + \mathcal{R}_{\mathrm{AUX}}(\boldsymbol{z}^*), \tag{7}$$

note that $\mathcal{R}_{\mathrm{AUX}}(\boldsymbol{z}^*)$ is a KL-based regularization, which penalizes $\boldsymbol{z}^*$ against deviating the prior distribution (commonly set to a normal distribution).

In the second stage, after fixing the updated $\boldsymbol{z}$, the optimizer further fine-tunes the $\boldsymbol{w}^*$ to better fit the target image. In real situation, the generator may not well-trained, which causes imperfect reconstruction. Therefore, such instance-level model adaptation can encourage $G$ to match better in image details. The optimization objective at this stage is as follows:

$$\arg\min_{\boldsymbol{w}^*} \mathcal{L}_{\mathrm{GM}}\left(\nabla \ell\left(f_\theta\left(G(z,\boldsymbol{w}^*)\right)\right); \nabla\theta\right) + \mathcal{R}_{\mathrm{AUX}}(\boldsymbol{w}^*), \tag{8}$$

note that $\mathcal{R}_{\mathrm{AUX}}(\boldsymbol{w}^*)$ is a $\ell_2$-based regularization $\|\boldsymbol{w}^* - \boldsymbol{w_0}\|_2$, which penalizes $\boldsymbol{w}^*$ against moving too far from original $\boldsymbol{w_0}$, otherwise $G$ may overfit easily and output more artifacts.

#### *Multi-rounds external loop*

In the complete Algorithm 1, we extra set an external loop with multiple rounds, and execute the above two-stages search in each round, since we believe such incremental update over iterations can help to improve the final result. Particularly, in order to avoid the unexpected deviation in overall optimization procedure, we limit the update step in each round, by reducing the learning rates and epochs of internal search. In that way, the fine-tuning of generator may barely overfit and keep stable.

#### *Gradient-free optimization strategy*

Almost previous studies choose gradient-based optimizer in their attacks, such as L-BFGS (Liu and Nocedal 1989) and Adam (Kingma and Ba 2014). However, due to the high ill-posedness of gradient inversion, especially when involved with generator, these gradient-based optimizers are easily trapped in local minima, and extremely sensitive to initialization. Herein, we adopt one gradient-free optimizer in latent space search, i.e. CMA (Hansen 2016), which combines covariance adaptation with evolution strategy. We empirically verify its strength in solving inversion problem involving generator.

**Algorithm 1** Iterative Gradient Inversion on Mixed Spaces

**Input:** target model $f_\theta$, gradients $\nabla\theta$ from local batch $\mathbb{D}^B$ $\{(\boldsymbol{x}_i, y_i)|i \in B\}$, reconstructed label $\boldsymbol{y}^*$, generator $G$, internal iterations $(M, N)$, external iterations $T$.
**Output:** reconstructed images $\{\boldsymbol{x}_i^*|i \in B\}$.
**CMA Configuration:**
  Initial Parameters $(\mu_{\boldsymbol{z}}, \Sigma_{\boldsymbol{z}})_i \leftarrow (0, I) \quad i \in B$;
  Budget $K$.

**for** $T$ *iterations* **do** `// External Loops`
  **for** $M$ *iterations* **do** `// Latent Space Search`
    **foreach** $i \in B$ **do**
      $\{\boldsymbol{z}_i\}_{1:K} \leftarrow$ `SampleCMA` $((\mu_{\boldsymbol{z}}, \Sigma_{\boldsymbol{z}})_i)$;
    $D_{\boldsymbol{z}} \Leftarrow \mathcal{L}(\frac{1}{B}\sum\nabla\ell(f_\theta(G(\boldsymbol{z}_i^*))); \nabla\theta) + \mathcal{R}_{\boldsymbol{Z}^*}$;
    **foreach** $i \in B$ **do**
      `UpdateCMA` $(\{\boldsymbol{z}_i\}_{1:K}, \{D_{\boldsymbol{z}}\}_{1:K})$;
      $\boldsymbol{z}_i^* \leftarrow$ `SampleCMA` $((\mu_{\boldsymbol{z}}, \Sigma_{\boldsymbol{z}})_i)$;
  **for** $N$ *iterations* **do** `// Parameter Space Search`
    $D_{\boldsymbol{w}} \Leftarrow \mathcal{L}(\frac{1}{B}\sum\nabla\ell(f_\theta(G(\boldsymbol{w}_i^*))); \nabla\theta) + \mathcal{R}_{\boldsymbol{W}^*}$;
    **foreach** $i \in B$ **do**
      $\boldsymbol{w}_i^* \leftarrow$ `UpdateWithAdam` $(D_{\boldsymbol{w}})$;
**foreach** $i \in B$ **do** $\boldsymbol{x}_i^* \leftarrow G(\boldsymbol{z}_i^*, \boldsymbol{w}_i^*)$;
**return** $\{\boldsymbol{x}_i^*|i \in B\}$.

#### *Comparison of IGIMS with other SoAs*

To justify our design, here we theoretically compare our algorithm with some other state-of-the-arts. GI in Yin et al. (2021), sets some stronger assumptions than ours, such as available BN statistics, which easily takes no effect if the BN layers are fixed in FL. Besides, we utilize the novel generative image prior to improve the inversion results. GGL in Li et al. (2022), lacks the fine-tuning procedure, which makes it difficult to carry out any attack in common cases if not having a well pretrained generator. GIAS in Jeon et al. (2021), establishes a similar two-stages sequential space search as ours, while ours additionally conduct extra global iteration with smaller update step, which performs better in optimization efficiency. For the above analyses, we provide some experimental illustrations in Sect. 5.

### Adaptive label recovery

Almost all existing researches base on one default assumption, i.e. there is no repeated label in the training batch, which is excessively strong in reality. Besides, for existing analytical recovery algorithms, e.g. iDLG (Zhao et al. 2020) and zero-shot label restoration (Yin et al. 2021), their essential idea is to simply analyse the signs of weight gradients in classification layer, which cannot deal with more complex data distributions and varying model structures. Compared to them, we propose an adaptive and effective label recovery algorithm, particularly fit for realistic label distributions. Hereinafter, we demonstrate our analysis procedure, in which the objective loss of the target model is cross-entropy loss $\mathcal{L}$.

First, we start from one single sample in the $K$-size training batch $\mathbb{D}^K$. Note that $(\boldsymbol{x}, \boldsymbol{y})$ is a single sample, and its label is one-hot encoded, which has $\boldsymbol{y} \in \{0, 1\}^N$, where $N$ is the total number of classes. The output logits $\boldsymbol{l}$ of the final full connection layer, i.e. classification layer, is denoted by $\boldsymbol{l} = \boldsymbol{W}^{(FC)^T}\boldsymbol{a} + \boldsymbol{b}^{(FC)}$, where $\boldsymbol{W}^{(FC)}$ is the weight matrix, $\boldsymbol{b}^{(FC)}$ is the bias vector, and $\boldsymbol{a}$ is the input vector of classification layer. Then we can observe the relation between $\frac{\partial\mathcal{L}}{\partial\boldsymbol{l}}$ and $\frac{\partial\mathcal{L}}{\partial\boldsymbol{b}^{(FC)}}$ through the following deduction:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial l_i} &= \frac{e^{l_i}}{\sum_1^N e^{l_i}} - y_i \\ &= p_i - y_i \\ &= \frac{\partial \mathcal{L}}{\partial b_i^{(FC)}},\end{aligned} \tag{9}$$

note that the subscript $i \in N$ refers to the $i$-th element of each vector, and $\boldsymbol{p}$ denotes the probability vector after Softmax layer.

When considering the batch size is $K$, the gradient of cross-entropy loss $\mathcal{L}$ w.r.t. $\boldsymbol{b}_i^{(FC)}$ is:

$$\frac{\partial \mathcal{L}}{\partial b_i^{(FC)}} = \frac{1}{K}\sum_{k=1}^{K}(p_{i,k} - y_{i,k}), \tag{10}$$

where the subscript $k$ refers to the $k$-th sample in batch.

Premise that the target model is not trained, which means the network parameters of each layer are merely randomly-initialized. In that case, no matter what input data is fed into the network, the input vector $\boldsymbol{a}$ of classification layer may empirically conform to a certain uniform distribution. Therefore, the attacker can firstly estimate the mean $\overline{\boldsymbol{a}}$ by feeding dummy inputs into the model, and further infer the mean $\overline{p_i}$ for each $i \in N$. Based on the above analyses, the appearances of each class in batch $\mathbb{D}^K$ can be inferred from the following equation:

$$\sum_{k=1}^{K} y_{i,k} = \sum_{k=1}^{K} p_{i,k} - K\frac{\partial \mathcal{L}}{\partial b_i^{(FC)}}. \tag{11}$$

Due to the variety of model parameters and structures, the $\overline{p_i}$ may be inconsistent between different models. Our algorithm can mitigate this adverse impact by adapting to corresponding discrepancies, hence able to handle more complex label distributions.

#### Average gradient approximation

For FedAvg, the user's local training has multiple epochs. While previous studies rarely consider this common scenario, only Geiping et al. (2020) defines the gradient matching loss regarding (multi-steps) weight update, which we refer to as the *Federated Averaging inversion* operator. Nevertheless, due to the high complexity of this operator, the process of a standard FedAvg inversion has the problem of massive computation and difficult optimization. In this section, we propose a simple but effective approximate method, averaging the cumulative update to each local step, therefore the FedAvg inversion problem can directly adopt the same optimization algorithm described in "Iterative gradient inversion on mixed spaces" section, as in FedSGD.

It can be assumed that the attacker, i.e. the central server, knows the local learning rate $\eta$ and the total local epochs $T$. Then, given the parameters $\theta^0$ and $\theta^T$, which denote the local model states at the 0-th and $T$-th step, the average approximate gradient for each step is inferred as:

$$\overline{\nabla\theta} = \frac{\theta^0 - \theta^T}{\eta\ T}. \tag{12}$$

Therefore, similar to previous Function (6), the corresponding gradient matching loss function under FedAvg is:

$$\mathcal{L}_{\text{GM}}\left(\frac{1}{B}\sum_{i=1}^{B}\nabla\ell(f_\theta(\boldsymbol{x}_i), y_i); \overline{\nabla\theta}\right). \tag{13}$$

Considering the numerical instability of the average approximation, the most fit gradient discrepancy measure is reasonably the cosine similarity. In Sect. 5, We empirically justify the effectiveness of this method, especially in reducing the computation workload and inversion complexity.

## Experiments

### Experimental setup

Towards the image classification task in FL, we comprehensively verify our attack framework in various experiments. Some important settings are as follows.

*Datasets.* We evaluate on supervised datasets with different resolutions and classes, i.e. CIFAR-10/-100 (Krizhevsky et al. 2009) (10/100 classes, 32×32), FFHQ (Karras et al. 2019) (10 classes for age attribute, resized down to 32×32), and ImageNet (Russakovsky et al. 2015) (1000 classes, cropped to 224×224). These datasets cover common objects, human faces and animals, which can show the applicability of our approach to different image styles and scenarios.

*Implementations.* For CIFAR-10 and FFHQ, we consider the minimum prior condition, i.e. no access to any pretrained generator, and use the GIML method in Jeon et al. (2021) to train a DCGAN (Radford et al. 2015) solely from shared gradients. For ImageNet, we use a pretrained BigGAN (Brock et al. 2018). We mainly evaluate on the ResNet-18 (He et al. 2016) target model with randomly-initialized weights, which is a practical choice in line with reality and complex enough for inversion study. We also accomplish some additional experiments on ResNet-50 in Appendix 6, and some other detailed settings are in Appendix 6.

*Approaches.* We mainly compare our design with other two SoA approaches, i.e. GIAS (Jeon et al. 2021) and IG

(Geiping et al. 2020). *Inversion Gradients* (IG) uses cosine matching loss and total variation regularizer to execute gradient inversion on image space. *GradInversion* (GI) (Yin et al. 2021) is more advanced with two extra strong regularizers. Given these new regularizations may not work under weaker assumptions, we regard IG as a more stable baseline. For GIAS, it utilizes the generative image prior as well and optimizes in alternate spaces of generator.

*Evaluation Metrics.* We use a lot of measures to quantitatively evaluate the similarity between the reconstructed image and the target image, including: **MSE** (pixel-wise Mean Square Error), **PSNR** (Peak Signal-to-Noise Ratio), **SSIM** (Structural Similarity), and **LPIPS** (Zhang et al. 2018) (Learned Perceptual Image Patch Similarity).

### Comprehensive results

Above all, we comprehensively evaluate our approach with other two SoAs, i.e. GIAS and IG, in two base scenarios.

Firstly, we consider the *Single* sample scenario in FedSGD. It is the most primary that there is only one sample and one-step update during each local training. Table 1 shows the quantitative performance of our approach with other counterparts on all three datasets. In addition, Figs. 2, 4 show some visualization samples. From quantitative views, our approach performs best on all evaluation metrics, while GIAS is consistently better than IG, reflecting the significant improvement brought by generative image prior. From qualitative views, for small-scale cropped samples of CIFAR-10 and FFHQ, our results have less noise and better clarity in visual; for larger-scale ImageNet, our approach can recover more accurate details, in terms of object posture, texture and color.

Secondly, for the *Batch* samples scenario with non-repeated label in FedSGD, we respectively conduct experiments on CIFAR-100 and ImageNet (results shown in Fig. 3). We find that our approach maintains high performance over all batch sizes, superior to GIAS and IG. Especially on ImageNet, the two generator-based approaches have significantly improvement over IG in reconstruction quality. It could be somewhat attributed to the fact that, when facing the large batch inversion, compared to generator-based methods, IG has to carry out much more complex and difficult optimization, due to its far larger joint search space on batch images.

### Ablation experiments

In different scenarios, we correspondingly set up ablation studies to justify the correctness and effectiveness of each component in our framework.

#### *Iterative gradient inversion on mixed spaces*

For our image reconstruction algorithm, we further study the following problems:

1. If the gains from *gradient-free optimizer* and *parameter fine-tuning* are obvious?
2. If the *multi-rounds small update* over two-spaces search in IGIMS is necessary?
3. Can the gradient inversion be carried out with untrained generator?

We consider the *Single* scenario and respectively investigate on CIFAR-10 and ImageNet. For CIFAR-10, we configure two variants of our approach: IGIMS*, using gradient-based optimizer Adam in both search spaces; and IGIMS*(untrained), the former using an untrained generator. The experimental results are shown in Figs. 5a, 6. For ImageNet, we also compare with two other approaches in Li et al. (2022): GGL, which can be regarded as a portion of IGIMS, using gradient-free optimizer while only searching on latent space; and its variant GGL*, the former using gradient-based optimizer. The experimental results are shown in Fig. 5b.

**Table 1** Quantitative comparison of IGIMS and other two SoA algorithms in *Single* scenario

| Metric | CIFAR-10 | | | FFHQ | | | ImageNet | | |
|---|---|---|---|---|---|---|---|---|---|
| | **IGIMS** | **GIAS** | **IG** | **IGIMS** | **GIAS** | **IG** | **IGIMS** | **GIAS** | **IG** |
| MSE ↓ | **0.0030** | 0.0043 | 0.0065 | **0.0041** | 0.0056 | 0.0073 | **0.0217** | 0.0246 | 0.0420 |
| PSNR ↑ | **26.33** | 24.41 | 22.88 | **25.00** | 23.27 | 21.98 | **16.68** | 15.86 | 14.06 |
| SSIM ↑ | **0.8924** | 0.8602 | 0.8176 | **0.8737** | 0.8521 | 0.8001 | **0.4727** | 0.4274 | 0.2313 |
| LPIPS ↓ | **0.0069** | 0.0092 | 0.0134 | **0.0073** | 0.0129 | 0.0166 | **0.4665** | 0.5510 | 0.7974 |

We highlight the best performances in bold

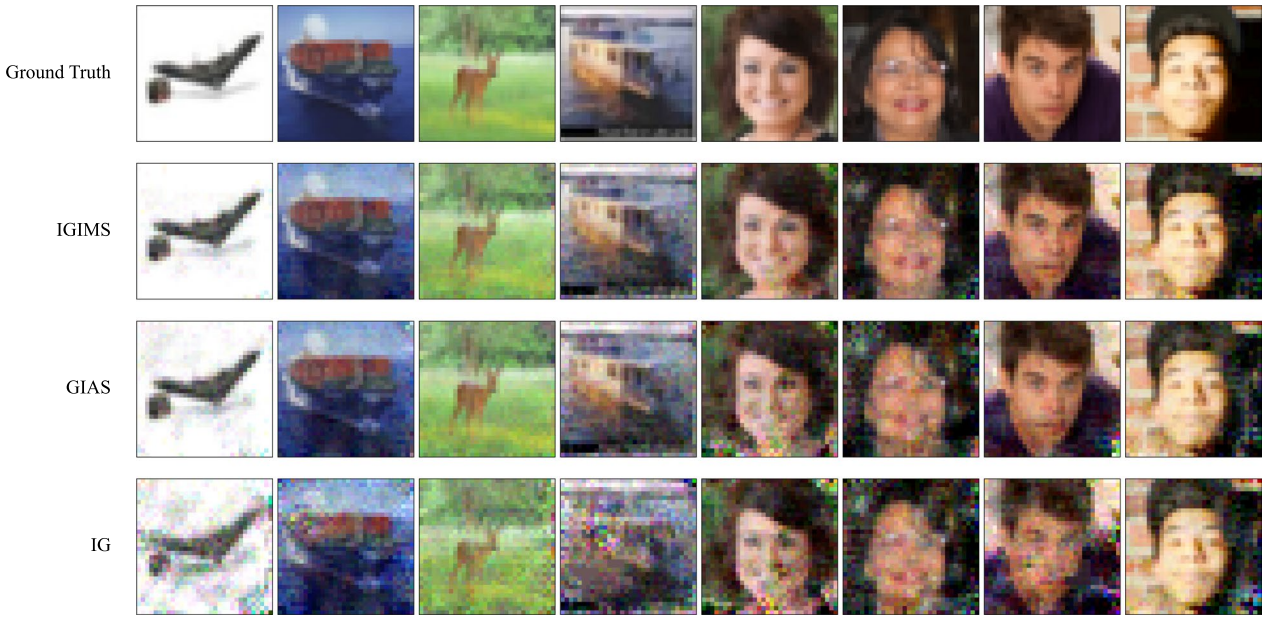
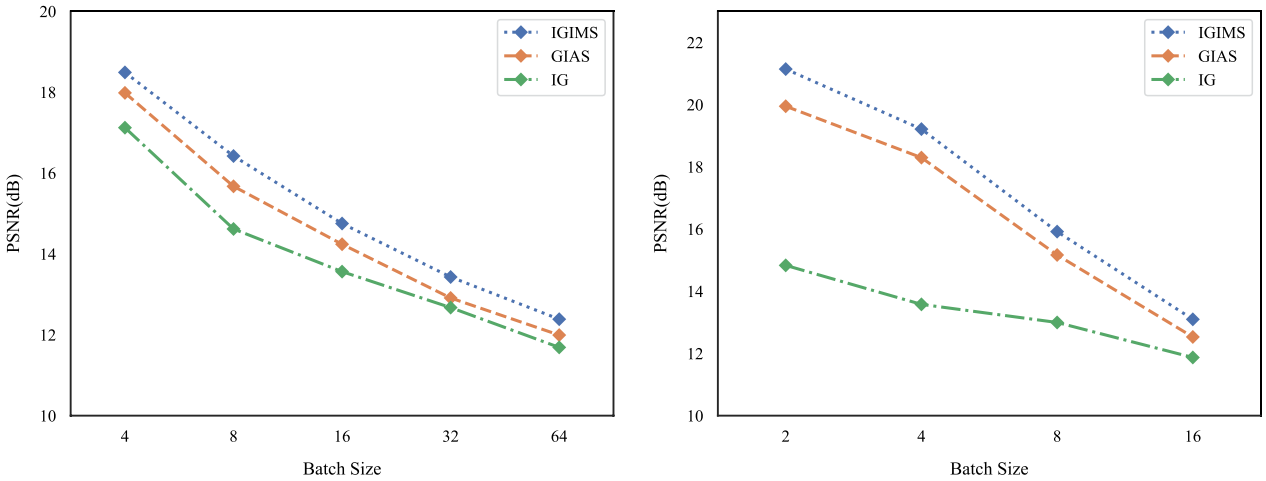



**Fig. 2** Visualization samples in *Single* scenario of all 3 algorithms, on CIFAR-10 (left 4 cols) and FFHQ (right 4 cols)




(a) CIFAR-100 (b) ImageNet

**Fig. 3** The *Batch* scenario results of all 3 algorithms over different batch sizes

We have the following observations: first, using *gradient-free optimizer* in space search certainly improves the reconstruction quality, as IGIMS and GGL perform better than IGIMS* and GGL*, which is consistent with the theoretical analyses in "Iterative gradient inversion on mixed spaces" section. Therefore, when meeting the intractable gradient inversion problem involved with a deep generator, it is advisable to incorporate a gradient-free strategy in optimization procedure. Second, the model adaptation by *fine-tuning parameters* of generator is empirically proved to be necessary, since it can indeed improve the inversion results, as IGIMS and GIAS have better performances over GGL as a whole. Third, the loss convergence curve in Fig. 6 along with the quantitative results in Fig. 5a validate the superior convergence efficiency of our *multi-rounds small update*, which shows a quick loss drop process. Moreover, despite using untrained generator leads to slightly loss in image quality,

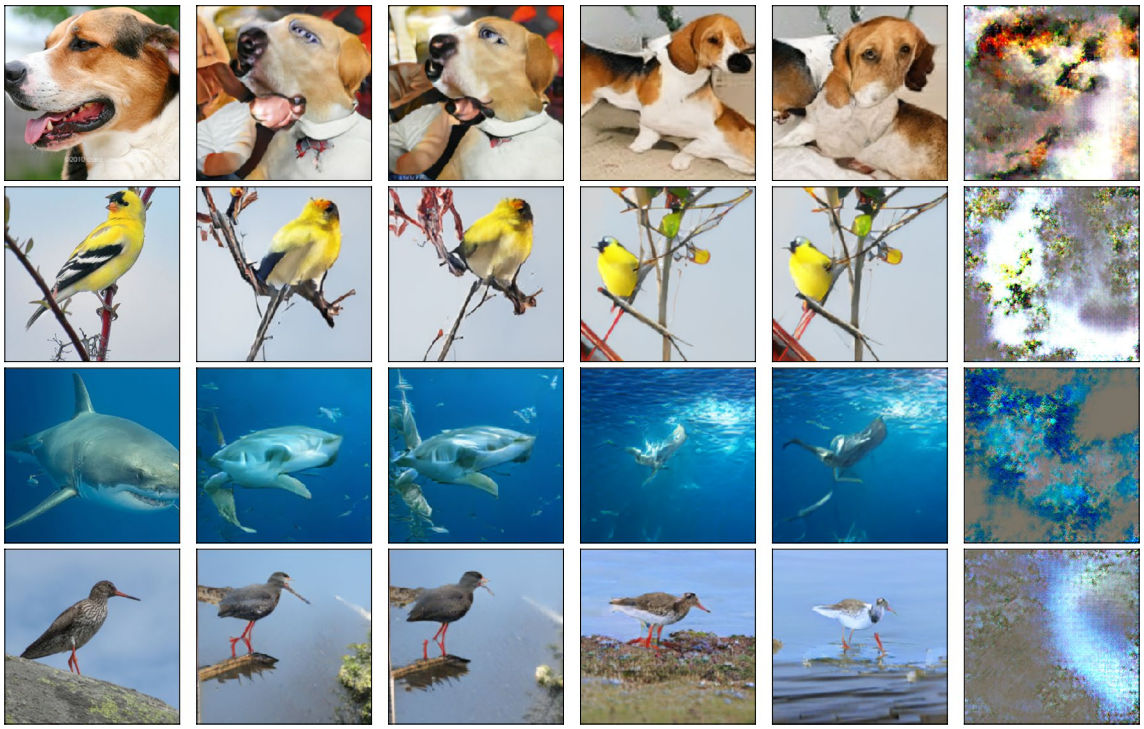



**Fig. 4** Visualization samples for *Single* scenario on ImageNet, also the ablation study for image reconstruction algorithm

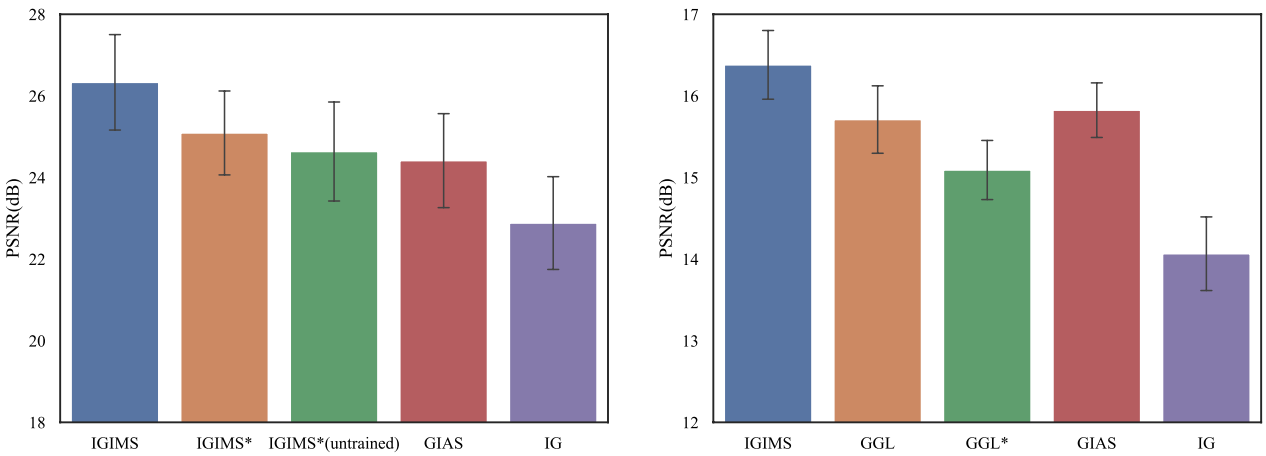



**Fig. 5** The ablation experiment for our image reconstruction algorithm on different datasets

our gradient inversion is still feasible and can reconstruct satisfactory results superior to other methods. It illustrates the deep image prior described in "Iterative gradient inversion on mixed spaces" section and Dmitry et al. (2020), which can implicitly impose a strong regularization on generated images. On the other hand, it also reflects the robustness of our approach to limited prior condition.

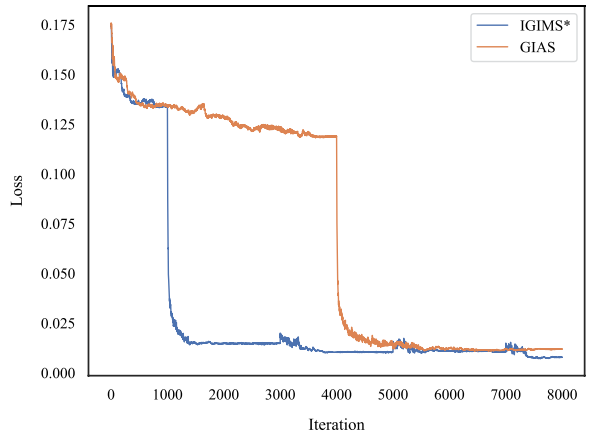



**Fig. 6** The loss convergence curve of IGIMS* and GIAS, note that the variant IGIMS* adopts the same gradient-based optimizer as GIAS. The target dataset is CIFAR-10

### *Average gradient approximation*

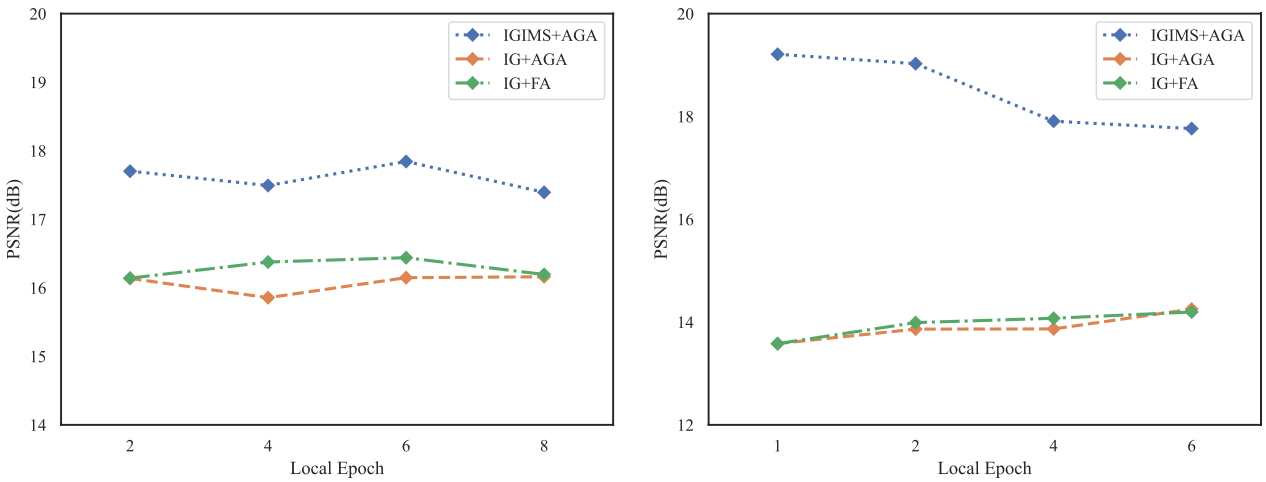
In order to verify our average gradient approximation, we intend to study the influence of our proposed method for reconstruction task in FedAvg. We consider the FedAvg scenario with non-repeated label and separately investigate on CIFAR-10 and ImageNet, where the local batch size is set to 8/4, and the local step is up to 8/6.

In this experiment, we set three integrated approaches. Among them, We denote the *federated averaging inversion* operator in Geiping et al. (2020) by FA, and our gradient approximation method by AGA. The ablation results are shown in Fig. 7. We find that, whether IG combined with our approximation method or not, IGIMS+AGA can obtain better results over all local epochs, and maintain stable performance as local epoch increases. Apart from that, though approximation strategy may theoretically result in certain loss in accuracy, the IG+AGA can still maintain similar performance compared to IG+FA, yet at a lower computation cost. On our experimental platform, the reduction of computational time can reach up to 4×.

### *Adaptive label restoration*

In order to verify our label recovery algorithm, we consider the scenario of complex data distribution with repeated labels, which is fairly common in reality while lacking in sufficient studies in previous researches. Here, we fix the experiment on one selected unbalanced distribution, in which 50% samples belong to one class, 25% belong to another class, and the remaining 25% are individually assigned to different classes.

First, we conduct a label recovery experiment, comparing our adaptive label recovery (ALR) with the existing zero-shot label restoration (ZLR) in Yin et al. (2021). We randomly select batches of different sizes from the above unbalanced distribution. The results are shown in Table 2. Owing to the robust adaptation for varying background information including model parameters and structures, our algorithm can gain better accuracy for arbitrary batch sizes, especially larger ones.



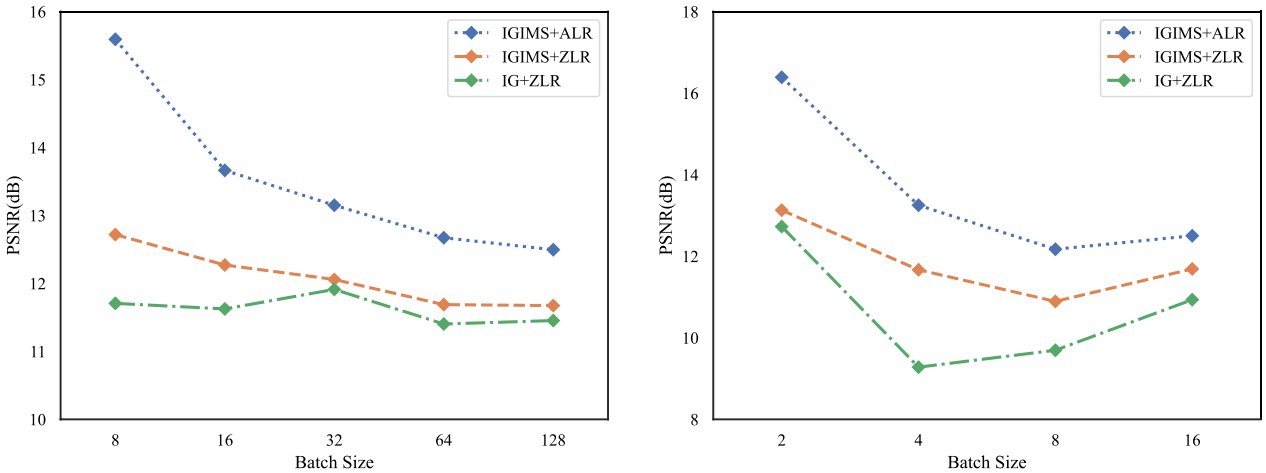

(a) CIFAR-10 (b) ImageNet

**Fig. 7** The ablation experiment for our gradient approximation method on different local epochs. The batch size for CIFAR-10 is 8, while for ImageNet is 4

**Table 2** Label reconstruction accuracy of our adaptive label recovery and zero-shot label restoration in Yin et al. (2021)

| Dataset | Method | Batch size | | | | |
|---|---|---|---|---|---|---|
| | | **16** | **32** | **64** | **128** | **256** |
| **CIFAR-100** | ALR | **1.000** | **1.000** | **0.984** | **0.977** | **0.926** |
| | ZLR | 0.375 | 0.313 | 0.266 | 0.242 | 0.254 |
| **Dataset** | **Method** | **Batch size** | | | | |
| | | **4** | **8** | **16** | **32** | **64** |
| **ImageNet** | ALR | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| | ZLR | 0.750 | 0.500 | 0.375 | 0.313 | 0.281 |

We highlight the best results in bold




(a) CIFAR-100 (b) ImageNet

**Fig. 8** The ablation experiment for our label recovery algorithm on different batch sizes

Then, we illustrate the integral reconstruction results for different unbalanced batches respectively on CIFAR-100 and ImageNet. We also configure three integrated approaches. The results are shown in Fig. 8. We observe that our integrated approach achieves better results over all batch sizes. Besides, even though adopting the inaccurate zero-shot restoration, our IGIMS+ZLR still performs better compared with IG+ZLR, proving the robustness of our image reconstruction algorithm utilizing the generative image prior.

## Conclusion

In this work, we propose an image gradient inversion framework for federated learning. We first introduce our image reconstruction approach with generative image prior, a novel domain knowledge, in form of using a generator to parametrize the image data. From theoretical and empirical views, we verify its facilitation to gradient inversion. Moreover, we design a high-accuracy adaptive label recovery method, which can expend our attack to more complicated and realistic circumstances. In addition, a gradient approximation strategy is adopted in the final integrated approach, in order to alleviate the computation workload in practice. We experiment in a variety of scenarios, and compare with some other state-of-the-art approaches, which reveals the superiority of our attack framework, particularly under much more relaxed assumptions. Our study reflects a severe privacy threat in federated learning, and our future work includes two aspects: First, we try to expend the proposed gradient inversion framework to other FL tasks, such as text analysis with language models; Second, we hope to

design some corresponding defense mechanisms through some security techniques, e.g. homomorphic encryption or differential privacy, to better protect the vulnerable shared gradients and related privacy information.

## Appendix A Experiment settings

Unless stated otherwise, we consider an image classification task on the validation set of all datasets, i.e. CIFAR-10/-100, FFHQ, and ImageNet. We choose a randomly-initialized ResNet-18 (He et al. 2016) as the target classification model, which is an appropriate choice for two reasons: first, previous research (Geiping et al. 2020) finds that shallower and wider models may make the gradient inversion easier; second, we empirically observe that, our counterpart IG, can only output fairly degraded images if given untrained models. Therefore, we assert that an untrained ResNet-18 is enough challenging for our inversion study. We fix the negative cosine similarity as the gradient matching loss. For IGIMS, the internal iteration $M$ on latent space is 300(CMA)/1000(Adam), and iteration $N$ on parameter space is 1000, while the external loop $T$ is adjusted according to GIAS in specific experiment. For IG, it optimizes over 8000/20,000 iterations respectively for CIFAR(FFHQ)/ImageNet. For all approaches, we use total variation regularizer $\mathcal{R}_{\text{TV}}$ with weight $\lambda_{TV} = 10^{-6}$ for CIFAR(FFHQ), and $\lambda_{TV} = 10^{-4}$ for ImageNet. For CIFAR(FFHQ), we use DCGAN (Radford et al. 2015) for IGIMS and GIAS. For ImageNet, we use pretrained BigGAN (Brock et al. 2018) for IGIMS, GIAS and GGL, with regularizer $\mathcal{R}(\boldsymbol{w})$ weight $\lambda_{\boldsymbol{w}} = 1$, and given the modifiability of this large model, only one external iteration is carried out in practice. About Adam optimizer, with fixed momentum coefficients ($\beta_1 = 0.9, \beta_2 = 0.999$) and decayed factor of 0.1 at 3/8, 5/8, 7/8 of total iterations, separately we set the initial learning rate $lr = 0.1$ for IG, while initial learning rates $\eta_{\boldsymbol{z}} = 3 \times 10^{-2}$, $\eta_{\boldsymbol{w}} = 10^{-3}$ for GIAS on CIFAR(FFHQ), and initial learning rates $\eta_{\boldsymbol{z}} = 3 \times 10^{-2}$, $\eta_{\boldsymbol{w}} = 10^{-5}$ for GIAS on ImageNet. About CMA optimizer, we set the initial distribution parameters $\mu, \Sigma = (0, I)$, and the budget is 50. All experiments are performed on our experimental platform equipped with NVIDIA RTX 4080 GPU.

Note that apart from $\mathcal{R}_{\text{TV}}$, we also consider some other advanced fidelity regularizers, such as regularizer of intermediate representation ($\mathcal{R}_{\text{Feat}}$) (Jin et al. 2021). However, we do not adopt $\mathcal{R}_{\text{Feat}}$ in our experiments finally, because we empirically find it degrades the original approaches to output much more blurred images, while it cannot fit for batch reconstruction.

In the single scenarios, we randomly select 30 samples from different datasets for each experiment, while in the batch scenarios, we randomly select 3 batches from different datasets for each experiment. To show the performances in the text, we take the average results of each group of reconstruction.

## Appendix B Additional experiments on deeper target model

In this section, we show our additional evaluation on one deeper target model: ResNet-50 (He et al. 2016). Here, we consider the *single* scenario, and the detailed settings are the same as the former experiments in Sect. 5.2 and Appendix 6. The quantitative results are listed in the following Table 3.

From the quantitative results, we observe overall degradations for all reconstruction approaches compared to results in Sect. 5.2, which is consistent with the theoretical inference that deeper models may lead to harder optimization process. Besides, our approach still shows good performance superior to other two approaches, reflecting the robustness of our design when facing models of larger scale.

**Table 3** Quantitative results on ResNet-50 in *Single* scenario

| Metric | CIFAR-10 | | | ImageNet | | |
|---|---|---|---|---|---|---|
| | IGIMS | GIAS | IG | IGIMS | GIAS | IG |
| MSE ↓ | **0.0199** | 0.0272 | 0.0319 | **0.0433** | 0.0490 | 0.0569 |
| PSNR ↑ | **17.19** | 15.85 | 14.99 | **13.70** | 13.14 | 12.62 |
| SSIM ↑ | **0.5627** | 0.4722 | 0.5008 | 0.2946 | **0.3058** | 0.2799 |
| LPIPS ↓ | **0.0539** | 0.0646 | 0.0596 | **0.5270** | 0.5703 | 0.8811 |

We highlight the best performances in bold

**Acknowledgements**
The authors would like to thank the editor and anonymous referees.

**Author contributions**
The design of the proposed approaches, the experiment deployment and the draft of the manuscript: LF. Providing critical guidance and suggestions for revision: LW and HL. All authors read and approved the final manuscript.

**Funding**
This work was supported by the National Key R &D Program of China under Grant 2019YFB1005200.

**Availability of data and materials**
The datasets used in this article are freely available on Internet. Detailed experimental results can be obtained from the author if necessary.

## Declarations

**Competing interests**
The authors declare that they have no competing interests.

Received: 24 November 2023 Accepted: 26 February 2024
Published online: 05 April 2024

**References**

Aono Y, Hayashi T, Wang L et al (2017) Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans Inf Forensics Secur 13(5):1333–1345
Bau D, Strobelt H, Peebles W et al (2019) Semantic photo manipulation with a generative image prior. ACM Trans Graph (TOG) 38(4):1–11
Bau D, Zhu JY, Wulff J, et al (2019b) Seeing what a gan cannot generate. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4502–4511
Boenisch F, Dziedzic A, Schuster R, et al (2023) When the curious abandon honesty: Federated learning is not private. In: 2023 IEEE 8th European symposium on security and privacy (EuroS &P), IEEE, pp 175–199
Bouacida N, Mohapatra P (2021) Vulnerabilities in federated learning. IEEE. Access 9:63229–63249
Brisimi TS, Chen R, Mela T et al (2018) Federated learning of predictive models from federated electronic health records. Int J Med Inform 112:59–67
Brock A, Donahue J, Simonyan K (2018) Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096
Dmitry U, Vedaldi A, Victor L (2020) Deep image prior. Int J Comput Vis 128(7):1867–1888
Fowl L, Geiping J, Czaja W, et al (2021) Robbing the fed: Directly obtaining private data in federated learning with modified models. arXiv preprint arXiv:2110.13057
Fredrikson M, Jha S, Ristenpart T (2015) Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pp 1322–1333
Geiping J, Bauermeister H, Dröge H et al (2020) Inverting gradients-how easy is it to break privacy in federated learning? Adv Neural Inf Process Syst 33:16937–16947
Geng J, Mou Y, Li F, et al (2021) Towards general deep leakage in federated learning. arXiv preprint arXiv:2110.09074
Hansen N (2016) The cma evolution strategy: a tutorial. arXiv preprint arXiv:1604.00772
Hard A, Rao K, Mathews R, et al (2018) Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604
He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
Hitaj B, Ateniese G, Perez-Cruz F (2017) Deep models under the gan: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp 603–618
Huh M, Zhang R, Zhu JY, et al (2020) Transforming and projecting images into class-conditional generative networks. In: European conference on computer vision, pp 17–34
Jeon J, Lee K, Oh S et al (2021) Gradient inversion with generative image prior. Adv Neural Inf Process Syst 34:29898–29908
Jin X, Chen PY, Hsu CY et al (2021) Cafe: Catastrophic data leakage in vertical federated learning. Adv Neural Inf Process Syst 34:994–1006
Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4401–4410
Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
Krizhevsky A, Hinton G, et al (2009) Learning multiple layers of features from tiny images. Technical report
Li Z, Zhang J, Liu L, et al (2022) Auditing privacy defenses in federated learning via generative gradient leakage. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10132–10142
Lim WYB, Luong NC, Hoang DT et al (2020) Federated learning in mobile edge networks: a comprehensive survey. IEEE Commun Surv Tutor 22(3):2031–2063
Liu DC, Nocedal J (1989) On the limited memory bfgs method for large scale optimization. Math Program 45(1–3):503–528
Lyu L, Yu H, Yang Q (2020) Threats to federated learning: a survey. arXiv preprint arXiv:2003.02133
McMahan B, Moore E, Ramage D, et al (2017) Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics, PMLR, pp 1273–1282
Melis L, Song C, De Cristofaro E, et al (2019) Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE symposium on security and privacy (SP), IEEE, pp 691–706
Nasr M, Shokri R, Houmansadr A (2019) Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE symposium on security and privacy (SP), IEEE, pp 739–753
Qian J, Nassar H, Hansen LK (2020) Minimal model structure analysis for input reconstruction in federated learning. arXiv preprint arXiv:2010.15718
Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434
Russakovsky O, Deng J, Su H et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252
Shokri R, Stronati M, Song C, et al (2017) Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP), IEEE, pp 3–18
Wang Z, Song M, Zhang Z, et al (2019) Beyond inferring class representatives: user-level privacy leakage from federated learning. In: IEEE INFOCOM 2019-IEEE conference on computer communications, IEEE, pp 2512–2520
Wen Y, Geiping J, Fowl L, et al (2022) Fishing for user data in large-batch federated learning via gradient magnification. arXiv preprint arXiv:2202.00580
Yang Q, Liu Y, Chen T et al (2019) Federated machine learning: concept and applications. ACM Trans Intell Syst Technol (TIST) 10(2):1–19
Yin H, Mallya A, Vahdat A, et al (2021) See through gradients: image batch recovery via gradinversion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16337–16346
Zhang R, Isola P, Efros AA, et al (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 586–595
Zhao B, Mopuri KR, Bilen H (2020) idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610
Zhu J, Blaschko MB (2020) R-gap: Recursive gradient attack on privacy. In: International conference on learning representations
Zhu JY, Krähenbühl P, Shechtman E, et al (2016) Generative visual manipulation on the natural image manifold. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14, Springer, pp 597–613
Zhu L, Liu Z, Han S (2019) Deep leakage from gradients. Adv Neural Inf Process Syst 32

## Publisher's Note