Breaching FedMD: Image Recovery via Paired-Logits Inversion Attack

Hideaki Takahashi[*] Jingjing Liu[†] Yang Liu[‡]

Abstract

Federated Learning with Model Distillation (FedMD) is a nascent collaborative learning paradigm, where only output logits of public datasets are transmitted as distilled knowledge, instead of passing on private model parameters that are susceptible to gradient inversion attacks, a known privacy risk in federated learning. In this paper, we found that even though sharing output logits of public datasets is safer than directly sharing gradients, there still exists a substantial risk of data exposure caused by carefully designed malicious attacks. Our study shows that a malicious server can inject a PLI (Paired-Logits Inversion) attack against FedMD and its variants by training an inversion neural network that exploits the confidence gap between the server and client models. Experiments on multiple facial recognition datasets validate that under FedMD-like schemes, by using paired server-client logits of public datasets only, the malicious server is able to reconstruct private images on all tested benchmarks with a high success rate.

1 Introduction

Federated Learning (FL) [28] is a distributed learning paradigm, where each party sends the gradients or parameters of its locally trained model to a centralized server that learns a global model with the aggregated gradients/parameters. While this process allows clients to hide their private datasets, a malicious server can still manage to reconstruct private data (only visible to individual client) from the shared gradients/parameter, exposing serious privacy risks [30, 43, 49].

One effective solution against such attacks is Federated Learning with Model Distillation (FedMD) [21], where each party uses both its *private data* and a public dataset for local training and sends its predicted logits on the public dataset (termed *public knowledge*) to the server, who then performs aggregation and sends the aggregated logits back to each party for next iteration. Therefore, in FedMD, the server can only access the predicted logits on public dataset, thus preventing leakage from gradients or parameters and retaining data and model heterogeneity with low communication cost [26]. Prior work [4, 14, 18, 26, 29] suggests that FedMD is a viable defense to the original FL framework, and many studies [4, 6, 14, 21, 24] consider FedMD as a suitable solution to solving real-life problems. This has catalyzed a broad application of FedMD-like schemes [5, 6] in industrial scenarios [16, 18, 25, 29, 37, 46, 47].

Despite its popularity, few studies have investigated the safety of sharing *public knowledge* in FedMD. Contrary to the common belief that FedMD is reliably safe, we found that it is actually vulnerable to serious privacy threats, where an adversary can reconstruct a party's *private data* via shared *public knowledge* only. Such a reconstruction attack is nontrivial, since to guarantee that the adversary recovers its *private data* strictly from *public knowledge* only, the attack needs to follow two inherent principles: *Knowledge-decoupling* and *Gradient-free*. 'Gradient-free' means that the adversary can recover private data without access to gradients. 'Knowledge-decoupling' means that the attack 'decouples' private information from learned knowledge on public data only, and directly recovers private data. However, existing inversion attacks [11, 42, 49] fail to meet the 'knowledge-decoupling' requirement as they do not consider the case where the target model is trained on both private and public datasets. Recent TBI [42] proposed a gradient-free method with inversion network, but still relied on availability of private data (as shown in Tab. 1).

In this paper, we propose a novel *Paired-Logits Inversion* (PLI) attack that is both gradient-free and fully decouples private data from public knowledge. We observe that a local model trained on *private data* will produce more confident predictions than a server model that has not seen the *private data*, thus creating a "confidence gap" between a client's logits and the server's logits. Motivated by this, we design a logit-based inversion neural network that exploits this confidence gap between the predicted logits from an auxiliary server model and those from the client model (*paired-logits*). Specifically, we first train an inversion neural network model based on public dataset. The input of the

[*] The University of Tokyo
takahashi-hideaki567@g.ecc.u-tokyo.ac.jp
[†] Institute for AI Industry Research, Tsinghua University, jjliu@air.tsinghua.edu.cn
[‡] Corresponding Author, Institute for AI Industry Research, Tsinghua University, Shanghai Artificial Intelligence Laboratory, liuy03@air.tsinghua.edu.cn

inversion network is the predicted logits of server-side and client-side models on the public data. The output is the original public data (e.g., raw image pixels in an image classification task). Then, through confidence gap optimization via paired-logits, we can learn an estimation of server and client logits for the target private data. To ensure the image quality of reconstructed data, we also propose a prior estimation algorithm to regulate the inversion network. Lastly, we feed those estimated logits to the trained inversion model to generate original private data. To the best of our knowledge, this is the first study to investigate the privacy weakness in FedMD-like schemes and to successfully recover private images from shared *public knowledge* with high accuracy.

We evaluate the proposed PLI attack on facial recognition task, where privacy risks are imperative. Extensive experiments show that despite the fact that logit inversion attacks are more difficult to accomplish than gradient inversion attacks (logits inherently contain less information than gradients), PLI attack successfully recovers original images in the private datasets across all tested benchmarks. Our contributions are three-fold: 1) we reveal a previously unknown privacy vulnerability of FedMD; 2) we design a novel paired-logits inversion attack that reconstructs private data using only the output logits of public dataset; 3) we provide a thorough evaluation with quantitative and qualitative analysis to demonstrate successful PLI attacks to FedMD and its variants.

Table 1. Comparison between PLI attack and existing inversion attacks. GF: *Gradient-free*. KD: *Knowledge-decoupling*.

Method	Leak	GF	KD
MI-FACE [11]	logit/gradient	✗	✗
TBI [42]	logit	✓	✗
DLG [49], GS [12] iDLG [45], CPL [40]	gradient	✗	✗
mGAN-AI [39], Secret Revealer [44], GAN Attack [13]	parameter	✗	✗
PLI (Ours)	logit	✓	✓

2 Problem Definition

In this study, we investigate the vulnerability of FedMD-like schemes and design successful attacks that can breach FedMD to recover private confidential data. As a case study, we choose image classification task, specifically face recognition, as our focus scenario. This is because in real applications (e.g., financial services, social networks), the leakage of personal images poses severe privacy risks. In this section, we first provide a brief overview of FedMD framework with notations. Then, we give a formal definition of image classification task under the federated learning setting.

FedMD [21] is a federated learning setting where each of K clients has a small private dataset $D_k := \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$, which is used to collaboratively train a classification model of J classes. There is also a public dataset $D_p := \{(x_i^p, y_i^p)\}_{i=1}^{N_p}$ shared by all parties. Each client k trains a local model f_k with parameters W_k on either D_k, D_p or both, and sends its predicted logits $\boldsymbol{l}^k := \{\boldsymbol{l}_i^k\}_{i=1}^{N_p}$ on D_p to the server, where $\boldsymbol{l}_i^k = f_k(W_k; x_i^p)$. The server aggregates the logits received from all clients to form consensus logits $\boldsymbol{l}^p := \{\boldsymbol{l}_i^p\}_{i=1}^{N_p}$. The consensus logits are then sent back to the clients for further local training (See the complete algorithm in Appendix B).

Several schemes follow FedMD with slight variations. For example, FedGEMS [6] proposes a larger server model f_0 with parameters w_0 to further train on the consensus logits and transfer knowledge back to clients. DS-FL [15] uses unlabeled public dataset and entropy reduction aggregation (see Appendix B for details). These frameworks only use public logits to communicate and transfer knowledge, which is the targeted setting of our attack.

Image Classification task can be defined as follows. Let $(x_i^k,\ y_i^k)$ denote image pixels and class label ID, respectively, and L denote the set of target labels $\bigcup_{k=1}^{K}\{y_i^k\}_{i=1}^{N_k}$ .We aim to use public logits $\boldsymbol{l}^k$ obtained for D_p to reconstruct private class representation x_j for any target label $j \in L$. We further assume that D_p is made up of two disjoint subsets $D_0 := \{(x_i^0, y_i^0)\}$ and $D_a := \{(x_i^a, y_i^a)\}$, where D_0 consists of public data of non-target labels, while D_a contains images from a different domain for all target and non-target labels. For example, in face recognition tasks, the feature space of D_a might include an individual's insensitive images, such as masked or blurred faces (see Fig. 1). We also suppose the server is honest-but-curious, meaning it cannot observe or manipulate gradients, parameters, or architectures of local models.

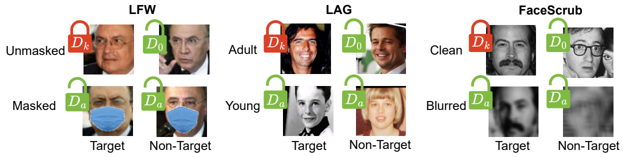

Figure 1. Examples from three datasets containing private and public data. Public dataset consists of two subsets: $D_p = D_0 \cup D_a$. D_0 and D_k consist of images of non-target and target labels(*Unmasked* (LFW), *Adult* (LAG) and *Clean* (FaceScrub)), while D_a contains images from a different domain (*Masked* (LFW), *Young* (LAG), *Blurred* (FaceScrub)).

3 Paired Logits Inversion Attack

In this section, we first describe how to train the inversion neural network for private data recovery (Sec. 3.1), then explain how to estimate private logits by optimizing the confidence gap between the logits predictions of server and clients (Sec. 3.2). To prevent too much deviation from real data, we also introduce a prior estimation to regulate the inversion network training (Sec. 3.3). Lastly, we will explain how to recover private data using the trained inversion network, the estimated logits and the learned prior (Sec. 3.4).

3.1 Inversion Neural Network

Logit-based inversion is a model inversion attack that recovers the representations of original training data by maximizing the output logits *w.r.t.* the targeted label class of the trained model. This inversion typically requires access to model parameters, which is not accessible in FedMD. In order to perform a gradient-free inversion attack, [42] proposes a training-based inversion method (TBI) that learns a separate inversion model with an auxiliary dataset, by taking in output logits and returning the original training data. Inspired by this, we insert an inversion attack on the server side of FedMD. However, we show that logit-based inversion is more challenging than gradient-based inversion since logits inherently contain less information about the original data (See Appendix. E). To tackle the challenge, we train a server-side model f_0 on the public dataset with parameters W_0. The server-side logits are then updated as :

$$\boldsymbol{l}_i^0 = f_0(W_0; x_i^0) \tag{1}$$

Next, we train an inversion model using client-server paired logits, $\boldsymbol{l}^k$ and $\boldsymbol{l}^0$ on the pubic data subset D_0 only, denoted as G_θ:

$$\begin{aligned} &\min_\theta \sum_i ||G_\theta(p_{i,\tau}^0, p_{i,\tau}^k) - x_i^0||_2 \\ &p_{i,\tau}^0 = \text{softmax}(\boldsymbol{l}_i^0, \tau), \quad p_{i,\tau}^k = \text{softmax}(\boldsymbol{l}_i^k, \tau) \end{aligned} \tag{2}$$

where softmax$(, \tau)$ is the softmax function with temperature τ. Note the distribution gap between $\boldsymbol{l}^0$ and $\boldsymbol{l}^k$ is the key driver for successfully designing such a paired-logits inversion attack, as will be demonstrated in Sec 3.2.

To enhance the quality of reconstruction, we leverage the auxiliary domain features D_a to obtain a prior estimation for each data sample $\bar{x}_i$ via a translation algorithm (detailed in Sec. 3.3). The reconstruction objective is therefore summarized as:

$$\min_\theta \sum_i ||G_\theta(p_{i,\tau}^0, p_{i,\tau}^k) - x_i^0||_2 + \gamma||G_\theta(p_{i,\tau}^0, p_{i,\tau}^k) - \bar{x}_i||_2 \tag{3}$$

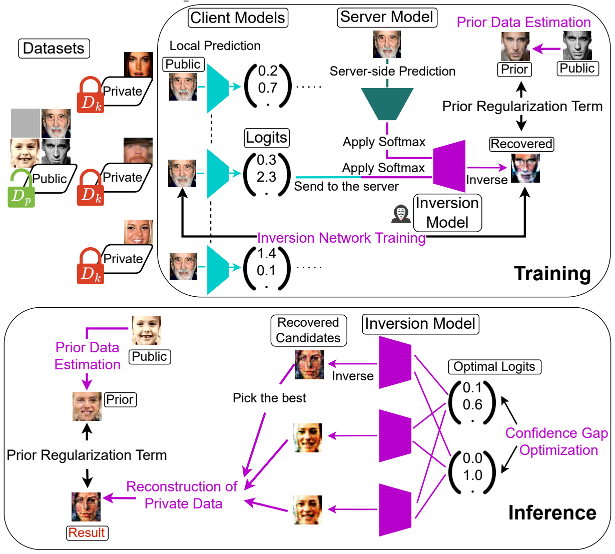

Figure 2. Overview of PLI. After receiving the output logits $\boldsymbol{l}^k$ on the public data D_p from k-th client, the server applies softmax to the logits corresponding D_0 of both client and server models and forwards them to the inversion model G_θ^k. The inversion model reconstructs the original input. We first optimize the inversion model with Eq. 3, while utilizing the prior data synthesized from D_a by transformation model A_ϕ. Then, the server recovers the private date of target labels with optimal logits and trains the inversion model with Eq. 10. The final reconstructed class representation will be picked from the set of reconstructed data with Eq. 11.

The second term enforces the recovered data to be close to the prior estimation.

Since the server already has access to the pubic dataset and its corresponding output logits, it can train G_θ^k to minimize Eq. 3 via backpropagation (line 6 in Algo 1). The architecture of the inversion model typically consists of transposed convolutional layers [8, 42]. Notice that the inversion model is trained with public data only and did not observe any private data at training time. Once the inversion model is trained, it is able to reconstruct data in the public dataset.

However, the model cannot be used directly to reconstruct private data with sensitive labels yet, since it has never seen any true logits of the private data, the distribution of which is different from that of public data. Most existing inversion attacks [11–13, 39, 40, 44, 45, 49] require either gradients or parameters, thus not *gradient-free*. To reconstruct private data without gradients, we propose a new path for estimating the input logits of private data, by exploiting the *confidence gap* between server and clients, as below.

3.2 Confidence Gap Optimization

We observe that the server and client models exhibit a confidence gap on their predictions of the public data, mani-

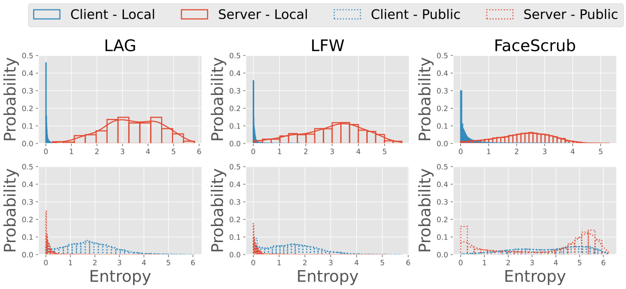

Figure 3. Confidence gap between the server and the client under FedMD setting on public and private data. This figure represents the normalized histogram of entropy on public and local datasets and estimated distribution. Lower entropy means that the model is more confident. Client consistently has higher confidence on private dataset than server, indicating a significant confidence gap.

Algorithm 1 PLI Attack

```
Input: The number of communication $T$, clients $K$, the set of target labels $L$, the inversion model $G_\theta^k$, the translation model $A_\phi$, the global model $f_0$, softmax temperature $\tau$, and public dataset $D_p$.
Output: Class representations in the private dataset of $L$
Generate prior $\bar{x}_j$ for each $j \in L$ with $A_\phi$.    ▷ Eq. 9
for $t = 1 \leftarrow T$ do
    Receive $\{\boldsymbol{l}_i^k\}$ from $k = 1 ... K$
    for $k = 1 \leftarrow K$ do
        Train $G_\theta^k$ with Eq. 3 on $D_0$.
        Train $G_\theta^k$ with Eq. 10 for each $j \in L$.
    Update global logits by aggregating $\{\boldsymbol{l}_i^k\}$
    Train $f_0$ and send the global logits to each client.
for $j \in L$ do
    Reconstruct $\{\hat{x}_j^k \leftarrow G_\theta^k(\hat{p}_{j,\tau}^0, \hat{p}_{j,\tau}^k)\}_{k=1}^K$
    $\hat{x}_j \leftarrow$ Pick the best from $\{\hat{x}_j^k\}$.    ▷ Eq. 11
return $\{\hat{x}_j\}_{j \in L}$
```

fested by their predicted logits (Fig. 3). Specifically, the client model is more confident on the private dataset, resulting in a higher entropy of server logits. To exploit this, we propose a new metric for optimizing private logits, then analytically obtain the optimal solution, which is used as the input logits for the inversion model to generate private data.

Specifically, we adopt the following metric to measure the quality of the reconstructed class representation x_j^k for arbitrary target label j owned by the k-th client:

$$Q(x_j^k) \coloneqq p_{\underline{j},\tau}^k + p_{\underline{j},\tau}^0 + \alpha H(p_{j,\tau}^0) \tag{4}$$

where $p_{\underline{j},\tau}^k$ and $p_{\underline{j},\tau}^0$ denote the j-th elements of $p_{j,\tau}^k$ and $p_{j,\tau}^0$ respectively. $H(\cdot)$ is the entropy function, and α is a weighting factor. The first and second terms grow bigger when the client and server are more confident that the recovered data belongs to class j, while the third term penalizes the strong confidence of server's prediction, which plays an important role in differentiating public and private feature spaces. With Eq. 4, finding the optimal input logits for our inversion attack against FedMD is equivalent to the following maximization problem:

$$\underset{p_{j,\tau}^k, p_{j,\tau}^0}{\arg\max}\, Q(x_j^k) \tag{5}$$

Recall that the private and public data domains for the target label j are different. Since the server-side model is not directly trained on the private dataset, we can assume that f_0 returns less confident outputs on private data compared to f_k, in this sense Q reinforces the *knowledge-decoupling* principle since Q increases when the reconstructed data of j are similar to the private data, and decreases when too close to the public data. For instance, Q takes a lower value when the reconstructed data is masked, even if its label seems j.

Analytical Solution We can analytically solve Eq. 5 *w.r.t.* $p_{j,\tau}^k$ and $p_{j,\tau}^0$, and the optimal values for target label j that maximizes Q, as follows:

$$\hat{p}_{\underline{u},\tau}^k = \begin{cases} 1 & (u = j) \\ 0 & (u \neq j) \end{cases}, \quad \hat{p}_{\underline{u},\tau}^0 = \begin{cases} \frac{\sqrt[\alpha]{e}}{J-1+\sqrt[\alpha]{e}} & (u = j) \\ \frac{1}{J-1+\sqrt[\alpha]{e}} & (u \neq j) \end{cases} \tag{6}$$

Detailed derivation can be found in Appendix A. Given an inversion model G_θ^k that takes paired logits and returns the original input x, we have the following equation:

$$\underset{x_j^k}{\arg\max}\, Q(x_j^k) = G_\theta^k(\hat{p}_{j,\tau}^0, \hat{p}_{j,\tau}^k) \tag{7}$$

The empirical impact of feature space gap between public and private data is examined in Appendix E.

3.3 Prior Data Estimation

To prevent the reconstructed image from being unrealistic, we design a prior estimation component to regulate the inversion network. A naive approach is using the average of D_0 (public data with the same domain) as the prior data for any target label i:

$$\bar{x}_i = \frac{1}{|D_0|} \sum_{i \in D_0} x_i^0 \tag{8}$$

Here the server uses the same prior for all the target labels.

When the public dataset is labeled, such as for FedMD and FedGEMS, the adversary can generate tuned prior for each label by converting D_a with the state-of-the-art translation method such as autoencoder [7, 23, 34] and GAN [17, 19, 48]. Specifically, the server can estimate the prior data $\bar{x}_i$ for the target label i with translation model A_ϕ as follows:

$$\bar{x}_j = \frac{1}{|D_a|} \sum_{i \in D_a} A_\phi(x_i^a) \tag{9}$$

3.4 Reconstruction of Private Data

Once the inversion model is trained with Eq. 3, the attacker uses the optimal logits (obtained from Sec. 3.2) to estimate the private data with G_θ. To make the final prediction more realistic, the attacker further fine-tunes the inversion model by restricting the distance between the reconstructed image and their prior data (obtained from Sec. 3.3) in the same way as in Eq. 3:

$$\min_\theta \sum_{j \in L} \gamma || G_\theta^k(\hat{p}_{j,\tau}^0, \hat{p}_{j,\tau}^k) - \bar{x}_j ||_2 \tag{10}$$

Finally, given a reconstructed image from each client, the attacker needs to determine which of the k images $\{\hat{x}_j^k\}_{k=1}^K$ from k clients is the best estimation for target label j. We design the attacker to pick the most distinct and cleanest image as the best-reconstructed data based on: 1) similarity to groundtruth image via SSIM metric [38]; 2) data readability measured by Total Variation (TV) [35], which is commonly used as regularizer [12,43]. Specifically, the attacker chooses reconstructed private data $\hat{x}_j$ by:

$$\arg\min_{\hat{x}_j^k} \sum_{k=1, k \neq k'}^{K} \text{SSIM}(\hat{x}_j^k, \hat{x}_j^{k'}) + \beta \text{TV}(\hat{x}_j^k) \tag{11}$$

where $\hat{x}_j^k = G_\theta^k(\hat{p}_{j,\tau}^0, \hat{p}_{j,\tau}^k)$. Lower SSIM indicates that the image is not similar to any other reconstructed privat pictures, and smaller TV means the image has better visual quality.

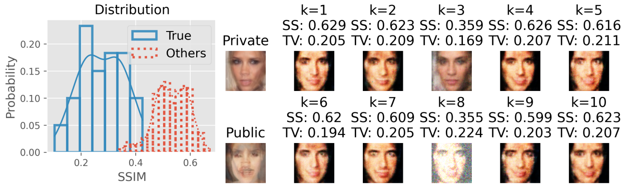

Figure 4. SSIM distribution (left) and Reconstructed Images (right) for LAG.

The reasoning for choosing Eq. 11 as the criterion is demonstrated in Fig. 4, where the left figure shows the distribution of the first term of Eq. 12 for clients who own the private data for the target label (blue) and clients who don't (red). Fig. 4 also provides reconstructed examples from all clients, where $k = 3$ is the ground-truth client, whose recovered image has the (almost) lowest SSIM. This figure also demonstrates the importance of β (the trade-off weight in Eq. 11, where a highly-noisy image (k=8) results in an even lower SSIM but is penalized by high TV score via a balanced weight ($\beta \geq 0.1$ in this case).

The complete algorithm is shown in Algo 1.

4 Experiments

We evaluate our proposed attack on face recognition task, which inherits high privacy risks in practice. We attack three schemes: FedMD [21], DS-FL [15], and FedGEMS [6] (see pseudo code in Appendix B). First, we describe detailed comparison between our approach and the baseline method TBI on various attacks. Then, we provide a thorough analysis on the impact of the learned prior, as well as the trade-off between image quality and attack accuracy in different attacks. Ablation studies further provide thorough analysis and insights on the effect of each PLI component and its comparison with gradient-based attack.

4.1 Experimental Setup

Datasets and Metrics Three standard datasets are used in our experiments: Large-Age-Gap (LAG) [3], Masked LFW [33], and FaceScrub [31]. LAG [3] contains both youth and adult images of celebrities. Masked LFW [33] is created by automatically adding masks to the faces of images in the LFW dataset [20]. For LAG and LFW, we treat adult and non-masked images as sensitive features, and young and masked images as insensitive features. We use the images of 1000 subjects who have the highest numbers of photos, so the number of classes is 1000. Among these, we randomly pick 200 classes as private labels, and treat the remaining classes as public labels. For FaceScrub [31], we randomly select 330 persons and treat all their pictures as the public dataset. Then, we blur half of the images of the remaining 200 individuals with box kernel and merge them to the public dataset as auxiliary domain D_a. The private datasets consist of the remaining clean photos. All pictures of each dataset are resized to 64x64 and normalized to [-1, 1].

An attack is considered successful when: $\text{SSIM}(\hat{x}_j, x_j) > max(\text{SSIM}(\hat{x}_j, x_{-j}), \text{SSIM}(\hat{x}_j, x_p))$. x_j: averaged private image of label j. x_{-j}: averaged private image of each label except j. x_p: averaged public image of any label. Assume *M* is the number of target labels with successful attack, and *N* is the number of all target labels. *AttackAccuracy* is defined by M/N.

Attack Targets We target three types of schemes: FedMD [21], DS-FL [15], and FedGEMS [6]. For simplicity, the server and clients use the same neural network with a convolutional layer, a max-pooling layer, and a linear layer with ReLU (Code. 1 in Appendix B). DS-FL assumes that the public dataset is unlabeled, so the total number of classes in DS-FL equals the number of classes in the local dataset, 200. Classification loss is cross entropy for all schemes, and distillation loss is L1-loss for FedMD, KL divergence for FedGEMS, and cross-entropy for DS-FL (as in original papers). See Appendix C for the detailed hyper parameters.

For the translation model A_ϕ for learned prior, the attacker trains CycleGAN [48] for LAG and LFW, and DeblurGAN-v2 [19] for FaceScrub in advance on the public dataset when it is labeled (Eq. 9). CycleGan is used for facial aging and mask removing in some work [10, 32, 36], and DeblurGAN-v2 is one of the state-of-the-art methods to deblur images. For DS-FL, the server uses the average of public sensitive features as the prior (Eq. 8). For the inversion model, we adopt the same architecture used in TBI (see Code. 2 in Appendix B). Detailed hyper-parameters can be found in Appendix C.

4.2 Comparison with Baseline

Tab. 2 reports the attack accuracy of PLI in comparison with TBI. We use τ=3.0 and γ=0.03 here as the empirically optimal setting across schemes and datasets. Results show that PLI outperforms TBI in all settings. Fig. 5 shows some reconstructed images from FedMD. The malicious party tries to recover private images. As the communication goes (higher t), the quality of reconstructed images improves (higher similarity to the grounthuth of private images). These images successfully capture the distinct sensitive features of the original private data, demonstrating that public logits can leak private information. On the other hand, the reconstructed images from TBI tend to be closer to the insensitive public features. More recovered examples from FedGEMS and DS-FL can be found in Appendix D.

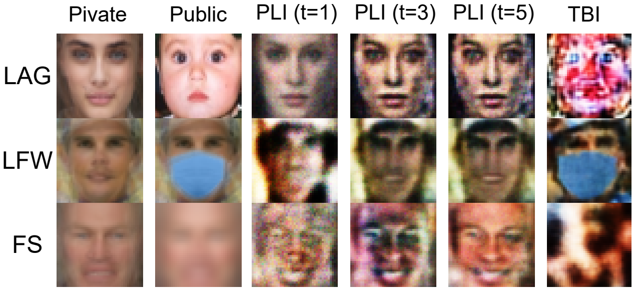

Figure 5. Examples of reconstructed images during training by our method where t represents the number of communications (FS: FaceScrub). ”TBI” represents the reconstructed results from TBI. Although TBI generates insensitive images, our PLI successfully reconstructs sensitive images as the learning progresses.

In general, attack accuracy on FedGEMS is lower than that on the other two schemes. We hypothesize that this is because FedGEMS trains iteratively on both private and hard-labeled public data, compared to FedMD that trains each model on labeled public dataset only. DS-FL does not even use labels on public data. Thus, local models in FedGEMS are more affected by the public dataset compared to FedMD and DS-FL, and more difficult to accurately recover private data.

4.3 Analysis on Impact of Prior

We also discuss how the quality of prior data is related to optimal γ, the tunable weight for prior.

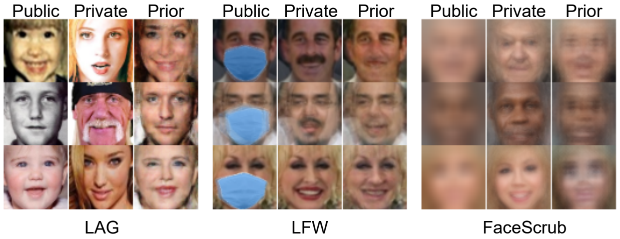

Figure 6. Examples of Prior for the labeld public dataset. While we can obtain relatively good prior images for LFW, the prior pictures for LAG fail to estimate the change of the shape of faces. For FaceScrub, the prior images are still not sharp enough.

Fig. 6 shows some qualitative examples of prior synthesized from the generator A_ϕ with the labeled public data. For LFW, CycleGAN successfully replaces the masked region with unmasked mouth. However, some distinct features under the mask, such as beard or lip color, are still not recovered. For LAG, faces in prior data look older than the public images, but are still more similar to private ones. Specifically, CycleGAN renders public images in adult appearances while keeping the original face outline. In addition, the recovered image is often too young or too old compared to the private image, since the attacker does not know the true age of the target person. For FaceScrub, DeblurGAN-v2 can deblur public images to some extent, but estimated prior data is still closer to public images.

Fig. 7 reports attack accuracy and SSIM value with various weights of prior γ with fixed τ=3.0. Note that the attacker cannot access GAN-based prior for DSFL, thus uses the average of all sensitive public features as prior (Eq. 8). While PLI can reconstruct private class representations without prior(γ=0.0), using prior as the regularization improves the attack performance in some cases. On the one hand, when the prior is relatively reliable (such as for LFW), higher γ increases both attack accuracy and SSIM. On the other hand, for LAG and FaceScrub, lower γ increases attack accuracy but damages quality. This effect of

Dataset		FaceScrub			LAG			LFW	
Scheme	DS-FL	FedGEMS	FedMD	DS-FL	FedGEMS	FedMD	DS-FL	FedGEMS	FedMD
TBI (K = 1)	87.0	1.0	92.5	70.0	0.5	16.5	73.5	2.5	2.0
PLI (K = 1)	**91.5**	**29.0**	**94.0**	**71.0**	**17.0**	**60.0**	**99.5**	**91.0**	**99.5**
TBI (K = 10)	2.0	0.5	7.0	6.5	0.0	0.0	17.5	9.5	10.0
PLI (K = 10)	**62.5**	**20.0**	**74.5**	**15.0**	**26.5**	**63.5**	**15.5**	**71.5**	**79.0**

Table 2. Results on attack accuracy (%).

γ is reasonable as increasing γ makes the reconstructed images closer to prior images, which generally decreases the noise but also eliminates some distinctive features (such as beards) that the prior image does not have (see Fig. 8).

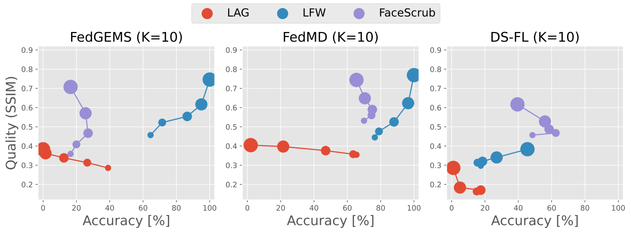

Figure 7. Trade-off between quality and accuracy with various γ. Dot size corresponds to the magnitude of γ. Larger γ leads to high quality but ruins attack accuracy for LAG and FaceScrub, where the prior images are not close enough to the ground-truth private images. For LFW, the prior sufficiently resemble the private images, and increasing γ improves both accuracy and quality.

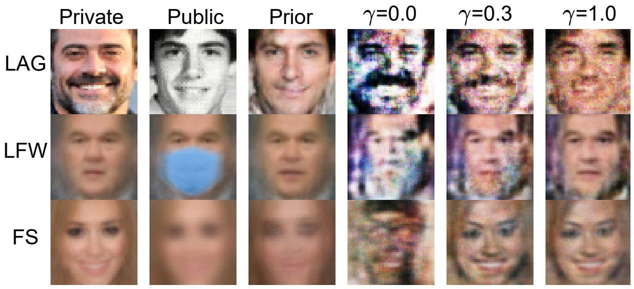

Figure 8. Effect of γ in FedMD (FS: FaceScrub) with fixed τ. Higher γ makes the reconstructed images noiseless, but sometimes drops distinctive elements (e.g., beard in the first row).

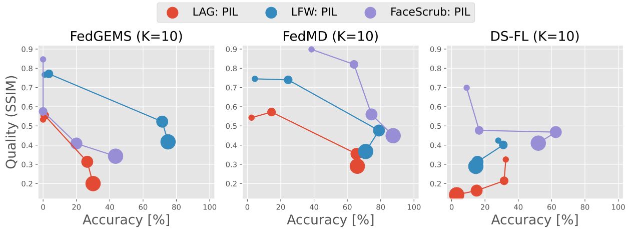

Figure 9. Trade-off between quality and accuracy with various τ. Dot size is proportional to the magnitude of τ. Larger τ leads to high accuracy but ruins quality for FedMD and FedGEMS, where the local models can use the labeled public dataset. The appropriate τ is lower for DSFL, where the public data is unlabeled and the predicted logits on the public data are more ambiguous.

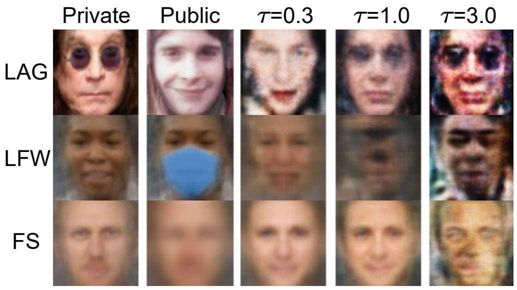

Figure 10. Effect of τ in FedMD with fixed γ (FS: FaceScrub). Higher τ preserves the characteristic features but damages readability (e.g., skin color in the second row).

4.4 Attack Accuracy vs. Image Quality

Experiments show that temperature τ controls the trade-off between attack accuracy and image quality. Fig. 9 shows the attack performance on various temperature τ with fixed $\gamma = 0.1$ (see Appendix D for the detailed numerical values). We observe that higher τ improves accuracy but damages quality. The influence of τ is consistent with previous work [15], suggesting that a higher τ amplifies non-highest scores of predicted logits while losing class information. We also find that the best temperature for DS-FL is lower in some cases. It is natural because the predicted logit on public data is more vague when the trained model has not observed actual public labels. Fig. 10 shows some qualitative examples, and we can observe that increasing τ leads to reconstructed images being created by compositing multiple classes, which preserves distinctive elements but reduces the quality of image. The same pattern occurs with TBI (more results can be found in Appendix D).

In summary, these experiments exhibit a trade-off between accuracy and quality, as optimizing attack accuracy does not guarantee improved quality. By finding a sweet spot (via γ and τ), we can empirically learn an optimal setting with high accuracy and acceptable quality.

4.5 Ablation Study

	FaceScrub	LAG	LFW
p_j^k	55.5	57.0	71.5
$+p_j^0 + \alpha H(p^0)$	**70.0**	**65.5**	**76.5**

Table 3. Ablation studies on Q (FedMD with K=10) in terms of attack accuracy. PLI (second row) outperforms using only client-side logits p_j^k in terms of accuracy.

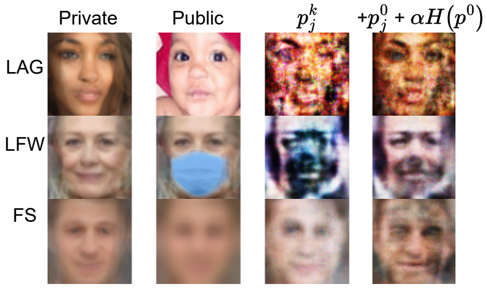

Figure 11. Qualitative analysis: eliminating $\alpha H(p^0)$ and p_j^0 from Q leads to reconstructed images with featureless faces or higher noise. (FS: FaceScrub)

We further investigate the influence of each term of Q. Tab. 3 shows the results of ablation studies on FedMD of $K = 10, \tau = 3.0$ and $\gamma = 0$, by gradually adding server and local logits of Eq. 4. Note that the second row is PLI, and the optimal logits for the first row is $(\hat{p}_j^k, \hat{p}_j^0) = (1, -)$. The results show that it is effective to add both p_j^0 and $\alpha H(p^0)$. This phenomenon validates the importance of taking into consideration the confidence gap between the server and client logits. The effects of public dataset size and magnitude of feature domain gap are reported in Appendix D. Appendix E and F also show that gradients lead to more severe privacy violations than public logits.

Limitations One limitation of our method is that it generates representative data of each class, not pixel-wise accurate image within the private dataset. Since the attacker cannot obtain anything directly calculated from the private dataset, reconstructing the exact image is still challenging.

5 Related Work

Federated Learning [28] is one of the representative algorithms for distributed learning. To overcome data and model heterogeneity, many studies focus on knowledge distillation [26]. In addition to the three schemes targeted in this work, there are several methods [22, 41] that communicate both distilled knowledge and gradients. While these schemes are the combination of gradient-based and knowledge-distillation-based federated learning, they are subject to existing gradient-based attacks, which are more severe attacks (as shown in Appendix E and F).

Model inversion attack aims to recover training data via machine learning models, which takes in information about the model (output logits, gradients, or parameters) and generates training data. Gradient inversion attack reconstructs the original training data by minimizing the distance between the received gradient that each client calculates on its local dataset and model and the gradient of synthesized data on the loss function, with model parameters used on the client side [12, 40, 45, 49]. Parameter-based inversion attacks utilize parameters of the target model. For example, [13, 39, 44] build GAN with built-in local models to recover the private data. Such attacks may not apply to FedMD scenarios when the server cannot access the parameters of client models. MI-FACE [11] reconstructs training data by directly optimizing synthesized data, so the learned model generates a high probability for the target class on that data. However, it requires the gradient of the model w.r.t synthesized data, thus not satisfying the gradient-free principle. In this sense, MI-FACE can be categorized as a combination of gradient-based and logit-based attacks.

6 Conclusion

This paper introduces an attack against FedMD and its variants that reconstructs private training data from the output logits of the model trained on both private and public datasets. Our proposed attack does not rely on gradients, parameters or model structures. Experiments on face recognition benchmarks demonstrate that confidential information can be recovered from output logits of public datasets. By demonstrating FedMD-like methods are not as safe as previously believed, we hope to inspire future design of more secure FL algorithms for diverse tasks. Potential future work includes adaptively choosing γ based on the quality of the prior, and investigating protection mechanisms such as differential privacy [1,9] and homomorphic encryption [2,27].

7 Acknowledgement

This work was supported by the National Key R&D Program of China under Grant No.2022ZD0160504, Tsinghua-Toyota Joint Research Institute inter-disciplinary Program, and Xiaomi AI Innovation Research under grant No.202-422-002.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308–318, 2016. 8

[2] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 13(5):1333–1345, 2017. 8

[3] Simone Bianco. Large age-gap face verification by feature injection in deep networks. Pattern Recognition Letters, 90:36–42, 2017. 5

[4] Nader Bouacida and Prasant Mohapatra. Vulnerabilities in federated learning. IEEE Access, 9:63229–63249, 2021. 1

[5] Hongyan Chang, Virat Shejwalkar, Reza Shokri, and Amir Houmansadr. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. arXiv preprint arXiv:1912.11279, 2019. 1

[6] Sijie Cheng, Jingwen Wu, Yanghua Xiao, and Yang Liu. Fedgems: Federated learning of larger server models via selective knowledge fusion. arXiv preprint arXiv:2110.11027, 2021. 1, 2, 5

[7] Tai-Yin Chiu and Danna Gurari. Photowct2: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2868–2877, 2022. 5

[8] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285, 2016. 3

[9] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014. 8

[10] Farnaz Farahanipad, Mohammad Rezaei, Mohammadsadegh Nasr, Farhad Kamangar, and Vassilis Athitsos. Gan-based face reconstruction for masked-face. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, pages 583–587, 2022. 6

[11] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, pages 1322–1333, 2015. 1, 2, 3, 8

[12] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems, 33:16937–16947, 2020. 2, 3, 5, 8

[13] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pages 603–618, 2017. 2, 3, 8

[14] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. ACM Comput. Surv., jan 2022. Just Accepted. 1

[15] Sohei Itahara, Takayuki Nishio, Yusuke Koda, Masahiro Morikura, and Koji Yamamoto. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. IEEE Transactions on Mobile Computing, pages 1–1, 2021. 2, 5, 7

[16] Meirui Jiang, Hongzheng Yang, Chen Cheng, and Qi Dou. Iop-fl: Inside-outside personalization for federated medical image segmentation. arXiv preprint arXiv:2204.08467, 2022. 1

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019. 5

[18] Abhinav Kumar, Vishal Purohit, Vandana Bharti, Rishav Singh, and Sanjay Kumar Singh. Medisecfed: Private and secure medical image classification in the presence of malicious clients. IEEE Transactions on Industrial Informatics, 18(8):5648–5657, 2021. 1

[19] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In The IEEE International Conference on Computer Vision (ICCV), Oct 2019. 5, 6

[20] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014. 5

[21] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581, 2019. 1, 2, 5

[22] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. Advances in Neural Information Processing Systems, 33:2351–2363, 2020. 8

[23] Zhi-Song Liu, Vicky Kalogeiton, and Marie-Paule Cani. Multiple style transfer via variational autoencoder. In 2021 IEEE International Conference on Image Processing (ICIP), pages 2413–2417. IEEE, 2021. 5

[24] Guodong Long, Tao Shen, Yue Tan, Leah Gerrard, Allison Clarke, and Jing Jiang. Federated learning for privacy-preserving open innovation future on digital health. In Humanity Driven AI, pages 113–133. Springer, 2022. 1

[25] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In Federated learning, pages 240–254. Springer, 2020. 1

[26] Lingjuan Lyu, Han Yu, Xingjun Ma, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S Yu. Privacy and robustness in federated learning: Attacks and defenses. arXiv preprint arXiv:2012.06337, 2020. 1, 8

[27] Abbass Madi, Oana Stan, Aurélien Mayoue, Arnaud Grivet-Sébert, Cédric Gouy-Pailler, and Renaud Sirdey. A secure federated learning framework using homomorphic encryption and verifiable computing. In 2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge (RDAAPS), pages 1–8, 2021. 8

[28] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017. 1, 8

[29] Plamena Tsankovaand Galina Momcheva. Sentiment detection with fedmd: Federated learning via model distillation. Information Systems and Grid Technologies, 2020. 1

[30] Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. Future Generation Computer Systems, 115:619–640, 2021. 1

[31] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In 2014 IEEE International Conference on Image Processing (ICIP), pages 343–347, 2014. 5

[32] Sveinn Palsson, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Generative adversarial style transfer networks for face aging. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2018. 6

[33] Hai Phan and Anh Nguyen. Deepface-emd: Reranking using patch-wise earth mover's distance improves out-of-distribution face identification. arXiv preprint arXiv:2112.04016, 2021. 5

[34] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In International Conference on Machine Learning, pages 5210–5219. PMLR, 2019. 5

[35] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena, 60(1-4):259–268, 1992. 5

[36] Neha Sharma, Reecha Sharma, and Neeru Jindal. Comparative analysis of cyclegan and attentiongan on face aging application. Sādhanā, 47(1):1–20, 2022. 6

[37] Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. Feded: Federated learning via ensemble distillation for medical relation extraction. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pages 2118–2128, 2020. 1

[38] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004. 5

[39] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In IEEE INFOCOM 2019-IEEE Conference on Computer Communications, pages 2512–2520. IEEE, 2019. 2, 3, 8

[40] Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. A framework for evaluating client privacy leakages in federated learning. In European Symposium on Research in Computer Security. Springer, 2020. 2, 3, 8

[41] Chuhan Wu, Fangzhao Wu, Ruixuan Liu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Fedkd: Communication efficient federated learning via knowledge distillation. arXiv preprint arXiv:2108.13323, 2021. 8

[42] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19, page 225–240, New York, NY, USA, 2019. Association for Computing Machinery. 1, 2, 3

[43] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16337–16346, 2021. 1, 5

[44] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 253–261, 2020. 2, 3, 8

[45] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610, 2020. 2, 3, 8

[46] Shuai Zhao, Roshani Bharati, Cristian Borcea, and Yi Chen. Privacy-aware federated learning for page recommendation. In 2020 IEEE International Conference on Big Data (Big Data), pages 1071–1080. IEEE, 2020. 1

[47] Yiqing Zhou, Jian Wang, and Zeru Wang. Bearing faulty prediction method based on federated transfer learning and knowledge distillation. Machines, 10(5), 2022. 1

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017. 5, 6

[49] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. Advances in Neural Information Processing Systems, 32, 2019. 1, 2, 3, 8