# Federated Morozov Regularization for Shortcut Learning in Privacy Preserving Learning with Watermarked Image Data

Tao Ling
Hong Kong Polytechnic University
Hong Kong SAR, China
cstling@comp.polyu.edu.hk

Siping Shi
Hong Kong Polytechnic University
Hong Kong SAR, China
cssshi@comp.polyu.edu.hk

Hao Wang
Stevens Institute of Technology
Hoboken, USA
hwang9@stevens.edu

Chuang Hu
Wuhan University
Wuhan, Hubei, China
handc@whu.edu.cn

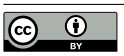
Dan Wang*
Hong Kong Polytechnic University
Hong Kong SAR, China
csdwang@comp.polyu.edu.hk

## Abstract

Federated learning is a promising privacy-preserving learning paradigm in which multiple clients can collaboratively learn a model with their image data kept local. For protecting data ownership, personalized watermarks are usually added to the image data by each client. However, the introduced watermarks can lead to a shortcut learning problem, where the learned model performs predictions over-rely on the simple watermark-related features and represents a low accuracy on real-world data. Existing works assume the central server can directly access the predefined shortcut features during the training process. However, these may fail in the federated learning setting as the shortcut features of the heterogeneous watermarked data are difficult to obtain.

In this paper, we propose a federated Morozov regularization technique, where the regularization parameter can be adaptively determined based on the watermark knowledge of all the clients in a privacy-preserving way, to eliminate the shortcut learning problem caused by the watermarked data. Specifically, federated Morozov regularization firstly performs lightweight local watermark mask estimation in each client to obtain the locations and intensities knowledge of local watermarks. Then, it aggregates the estimated local watermark masks to generate the global watermark knowledge with a weighted averaging. Finally, federated Morozov regularization determines the regularization parameter for each client by combining the local and global watermark knowledge. With the regularization parameter determined, the model is trained as normal federated learning. We implement and evaluate federated Morozov regularization based on a real-world deployment of federated learning on 40 Jetson devices with real-world datasets. The results show that federated Morozov regularization improves model accuracy by 11.22% compared to existing baselines.

*Corresponding author.




*MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia*

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681480

## CCS Concepts

• **Computing methodologies** → **Distributed computing methodologies**.

## Keywords

Federated Learning, Watermark, Shortcut Learning

**ACM Reference Format:**
Tao Ling, Siping Shi, Hao Wang, Chuang Hu, and Dan Wang. 2024. Federated Morozov Regularization for Shortcut Learning in Privacy Preserving Learning with Watermarked Image Data. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681480

## 1 Introduction

With the growth of applying advanced multimedia technology to commercial applications, concerns about user data privacy have greatly increased [21], and research on privacy-preserving learning has come into being. Federated learning [12, 31, 52] emerges as a promising privacy-preserving learning paradigm, where multiple clients can collaboratively learn a model without exposing their private data to the central server. Federated learning has been widely adopted in many multimedia applications such as medical image classification [28], anomaly detection in public safety surveillance [61], and sentiment analysis in social media content [59].

For data ownership identification and copyright protection, digital watermarking technologies are developed and applied in many multimedia applications [13, 53], through adding the well-designed digital watermark into the image data by the data owner [5, 20]. Training models with the watermarked data may lead to the shortcut learning problem, that is the learned model makes predictions based on the simple shortcut features in the training data, rather than learning the underlying complex core features of the target domain, and presents a good performance on the training dataset but decreased model accuracy on the unseen data [4, 22, 29, 55, 58]. For example, in medical image classification, a trained model detects pneumonia in chest X-rays (CXRs) relying on watermarks that represent which hospital the patient was seen instead of lung pathophysiology used by a radiologist [8, 55].

There are many works proposed to overcome the shortcut learning problem. According to where the shortcut feature is processed, existing works can be divided into data preprocesing-based [32,

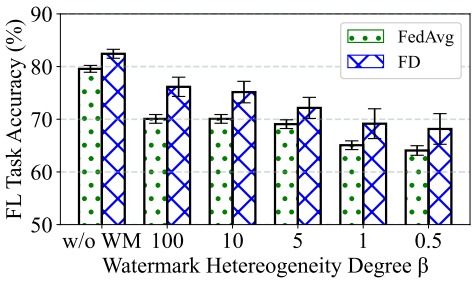



**Figure 1: The task accuracy of a model learned with various watermark heterogeneity.**

36, 38, 44] and regularization-based [18, 34] methods. The data preprocessing-based methods assume the shortcut features of data are useless, and they eliminate the shortcut learning problem by detecting and removing the shortcut features from the training dataset. These methods may fail in learning with the watermarked data as the shortcut features (i.e., the watermark-related features) are important for data ownership identification, and cannot be directly removed in practice. For regularization-based methods, shortcut features are regularized based on certain prior knowledge during each training iteration. For example, FD [18] assumes the shortcut features are represented in specific frequency, and designs a feature-level regularization technique where a randomized filtering layer is applied after each convolution layer to prevent CNNs from learning frequency-specific imaging features. wMMD-T [34] assumes the causal Directed Acyclic Graph (DAG) indicating the relationship between the input image and output label is known and designs a regularizer that leverages the causal DAG to efficiently learn a classifier. These regularization-based methods can work well in the centralized learning setting where the central server can directly obtain certain characteristics of the shortcut features. However, they may fail in a federated learning setting with watermarked data, as the characteristics of each client's watermark features are uncertain and unknown to the server for privacy protection.

Moreover, different clients may apply various digital watermarking techniques on the local data, including explicit watermarking (such as logos for copyright in media data, patient information text or physical markers in CT images) and implicit watermarking (such as frequency domain embedding with Discrete Cosine Transform (DCT) and spatial domain embedding with least significant bits (LSB)). This results in watermark heterogeneity, which further degrades the accuracy of the learned model. Our initial experiments show the impact of watermark heterogeneity under different regularization-based methods. As shown in Fig. 1, with the watermark heterogeneity degree $\beta$ (detail setting can be seen in Sec. 4) increasing from 100 to 0.5, the accuracy of the learned model decreases up to 15.5% under all baselines.

In this paper, we propose a federated Morozov regularization method to solve the shortcut learning problem of learning with watermarked data in a privacy-preserving way. Specifically, we first perform the local watermark mask estimation with the maximum a posteriori (MAP) method to generate the watermark mask, a matrix that can represent the characteristics of the watermarks. We observe that the embedded watermarks with various digital watermarking technologies can all be presented by the *location* and *intensity* map. Therefore, we estimate the watermark mask based on the distinct statistical distributions of natural images and artificial watermarks, capturing the divergence in their spatial and frequency domain characteristics. Then, we aggregate the estimated local watermark mask in the server to generate the global watermark mask with a weighted averaging model. Finally, we perform Morozov regularization-based local training by *actively* adjusting the regularization parameters with the estimated local and global mask. Intuitively, if the model training leads to worse overfitting to shortcut features, the regularization parameter will be increased, i.e., to aggressively mitigate overfitting introduced by the watermark; and vice versa. We evaluate federated Morozov regularization through experiments in real-world settings by deploying it on a test network of 40 Jetson devices, each with varying computational capabilities. We also evaluate our method on a real-world federated watermarked dataset, COVID-FL [54], where watermark heterogeneity is present. Evaluation results demonstrate the superior performance of our method compared to the baselines. federated Morozov regularization improves the accuracy of the learned model by up to 11.22%. We also conducted an ablation study of federated Morozov regularization to validate the contribution of each component to FL model performance in watermarked datasets.

The contributions of this paper can be summarized as:

- We are the first to formulate the shortcut learning problem arising from watermarked datasets in federated learning and find that watermark heterogeneity can further degrade the learning performance.
- We propose federated Morozov regularization, a new regularization method that can automatically adjust the regularization parameters based on the watermark knowledge of all clients in a privacy-preserving way.
- We evaluate federated Morozov regularization by deploying a real-world testbed of 40 Jetson devices with diverse computational capacities and comparing it to several baselines with real-world datasets. Our evaluations show that federated Morozov regularization outperforms existing baselines, achieving 11.22% higher accuracy.

## 2 Background & Related Work

### 2.1 FL for Multimedia Application

The integration of federated learning (FL) with multimedia applications is fundamentally motivated by the need to safeguard privacy [24, 30, 57]. This approach has facilitated the advancement of multimedia applications involving personal data, such as image classification [28], anomaly detection in public safety surveillance [61], and sentiment analysis in social media content [59]. The bulk of current research in this area has been concentrated on tackling data-centric challenges [33, 60], including non-iid data [27, 60], data imbalance [46], and the presence of noise [49].

However, a relatively unexplored issue in this domain is the influence of watermarked data in federated learning. Digital watermarking, a strategy widely adopted in multimedia applications for asserting data ownership [13, 53] and copyright protection [5, 20], has found extensive application in data involving privacy and copyright issues, such as medical images [43], surveillance videos [13], and social media [40]. Despite its primary intent, watermarking

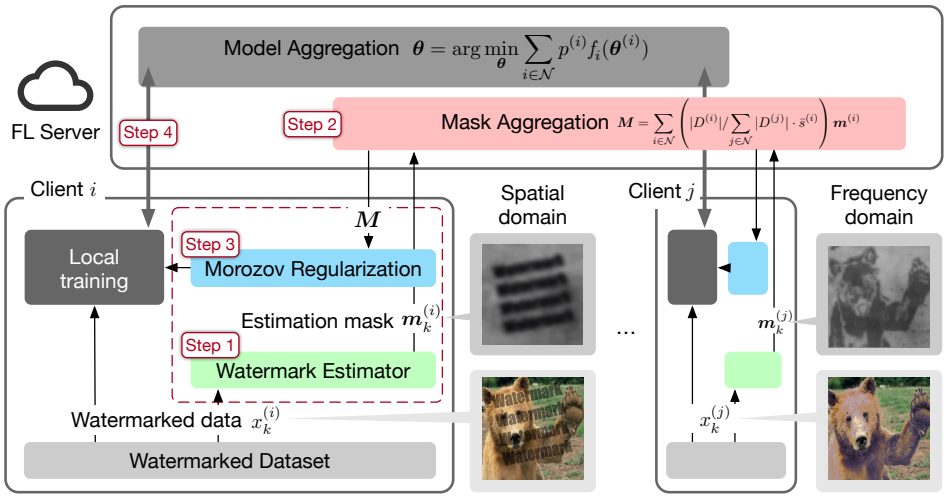



**Figure 2: Overview of the federated Morozov regularization in federated learning.**

unintentionally introduces detectable patterns into the data, precipitating a phenomenon known as shortcut learning.

## 2.2 Shortcut Learning

Shortcut learning refers to a phenomenon where deep learning models, during training, preferentially latch onto simple, detectable features—termed as shortcut features—instead of grappling with the more complex, core features of the data [10, 16]. This inclination can lead to models that perform well on training and in-distribution test data but falter significantly when faced with out-of-distribution inputs. Examples of shortcut learning include models relying on background elements [2] or specific textures for image classification [15], and even the presence of watermarks [4, 8].

Solutions to shortcut learning have primarily focused on data prepossessing [32, 36, 38, 44] and regularization techniques [18, 34]. Data prepossessing often involve the removal of shortcut features [36, 38] or data augmentation [32, 44]. However, in federated learning scenarios, watermarks are added due to a lack of trust in the federated learning applications or to embed ownership directly into the training model, making their removal impractical. The regularization method often views shortcuts as a consequence of model overparameterization [34]. Techniques like FD [18] emphasize high-frequency shortcut features, while methods like wMMD-T [34] focus on background elements as shortcut features. Yet, these shortcut features do not align with those introduced by watermarks.

Moreover, these approaches often require prior knowledge of the shortcut features from client data, such as labels or filter parameters, which contradicts the privacy-preserving nature of federated learning. Utilizing global information from the server side also fails to address the challenges brought by watermark heterogeneity.

## 2.3 Morozov Regularization

Morozov regularization [41] is one type of tool to adjust regularization parameters actively. One key principle of these methods is the discrepancy principle [45]. The rationale is that for a good regularized solution, the norm of the residual should match the noise level of the data.

Morozov regularization has been used in many applications in the past, e.g., to regularize noises from satellite sounder measurements for atmospheric profiling applications [25], to regularize sensor noises in digital images [6] and machine learning [19, 41].

The suitability of Morozov regularization for our problem lies in its precision in targeting specific distributions or explicitly formulated noise, offering localized regularization rather than a blanket, global approach. This characteristic is particularly aligned with the challenges posed by watermarks, which introduce shortcut features localized within parts of an image, rather than affecting it uniformly. Unlike other regularization methods that might operate under broad assumptions about noise or apply regularization uniformly across the entire data set, Morozov regularization provides an adaptive mechanism to fine-tune the regularization parameter, thereby mitigating the shortcut learning effect.

As compared to other regularization, Morozov regularization is simple and has less assumptions on noise approximation [1], practical a-posteriori rules [37], and/or convergence rate [39]. We choose Morozov regularization for its widely applicability and leave other types of regularization into future works.

# 3 Federated Morozov Regularization

## 3.1 Problem Definition

Federated Learning leverages a set of distributed clients $\mathcal{N} = \{1, \ldots, N\}$ to iteratively learn a global model $\theta$ without leaking any private local data to the central server [35]. In each client $i$, the local dataset is defined as $D^{(i)}$. For data ownership identification, each data sample $x_k^{(i)}$ is embedded with a digital watermark $n_k^{(i)}$. Let $\theta^{(i)}$ be the local model of client $i$, and the global model $\theta$ is learned by solving the following optimization problem:

$$F(\theta) := \underset{\theta}{\text{argmin}} \sum_{i=1}^{N} \sum_{k=1}^{D^{(i)}} f^{(i)}(\theta; WM(x_k^{(i)}, n_k^{(i)}), y_k^{(i)}), \quad (1)$$

where $f^{(i)}(\theta^{(i)}) = \frac{1}{|D^{(i)}|} \sum_{(x,y) \in D^{(i)}} \ell(x, y; \theta^{(i)})$, $|D^{(i)}|$ is the number of data sample in client $i$,$WM(\cdot)$ is the watermark embedding function, and $\ell(\cdot)$ is the loss function.

The integration of digital watermarks into image data for ownership identification introduces several challenges in the federated learning environment. Firstly, accurately modeling the embedded watermarks ($n_k^{(i)}$) within the data requires sophisticated techniques to distinguish and quantify their impact on the learning process. Secondly, the distributed nature of federated learning complicates the task of addressing the variability in watermark characteristics across different clients while maintaining global model performance. Thirdly, it is crucial to mitigate the performance degradation caused by watermarks without compromising the privacy of the data.

We propose a federated Morozov regularization method to address these challenges, and Fig. 2 shows the overview of federated Morozov regularization during an FL model training round. The complete training process include four steps. After obtaining user consent for training data, the client preprocesses the data using the *watermark estimator*. This step involves individually inputting watermarked data $x_k^{(i)}$ to obtain the watermark estimation mask $\boldsymbol{m}_k^{(i)}$ for the client's dataset (Step 1). *Mask aggregation* involves aggregating local watermark estimation masks $\boldsymbol{m}_k^{(i)}$ from each client into a global watermark estimation mask $\boldsymbol{M}$. This step synthesizes collective watermark characteristics from all participating clients (Step 2). Local training using watermarked data is conducted under the *Morozov regularization* module. This module automatically selects regularization parameters based on the watermark estimation mask corresponding to the training data, thereby adjusting the local model parameters to ignore watermarks in the data (Step 3). Finally, the server aggregates the local models from the selected clients to form a new global model for the next round of training (Step 4).

## 3.2 MAP-basd Watermark Mask Estimator

In our federated learning setting, the watermarking techniques applied by each client may be different and unknown to the server, making most existing watermark estimation techniques that are designed based on specific watermarking methods unsuitable, such as NN-based approaches for explicit watermark estimation or hash-based methods for LSB watermark estimation. However, watermarks generated by different techniques often present statistical commonalities, such as similar positions or patterns. To address this uncertainty while leveraging these commonalities, we adopt a stochastic approach based on maximum a posteriori (MAP) estimation, as described in [50]. This statistical technique incorporates prior distribution to estimate watermarks under uncertainty.

Consider the classical problem of watermark embedding, which involves embedding a watermark into an image without considering the image content. In the most general form in communication codec theory [51], the process of $WM(\cdot)$ can be modeled as $x' = x + n$, where $x'$ represents the watermarked data, $x$ is the original data, $x \in \mathcal{R}^N$ with $N = M \times M$, and $n$ denotes the watermark. Our goal is to estimate $\hat{n}$, which is an estimate of the watermark $n$.

Under the general assumption [51], we model the watermark as a Gaussian random variable. Let watermark sample $n_{u,v} (1 \le u, v \le M)$ and image sample $x_{u,v} (1 \le u, v \le M)$ be defined on the vertices of a grid $M \times M$. Let all samples be independent and identically distributed, we have conditional probability density of $n_{u,v}$:

$$p_{n_{u,v}}(x_{u,v} \mid n_{u,v}) = \frac{1}{\sqrt{\left(2\pi\sigma^2_{n_{u,v}}\right)^N}} \exp\left\{-\frac{1}{2\sigma^2_{n_{u,v}}}\Delta_n^T\Delta_n\right\}, \quad (2)$$

where the $\sigma_{n_{u,v}}$ of the watermark in the $(u, v)$ location signifies its *intensity*, $\Delta_n = x_{u,v} - n_{u,v}$. Higher variance indicates a more noticeable watermark (albeit with possible image distortion), whereas lower variance results in a subtler watermark.

To estimate the watermark throughout an image, we use a *local estimation mask* $\boldsymbol{m} = [\hat{n}_{u,v}]_{1 \le u,v \le M}$, which is a matrix represent each client's watermark information in local dataset. The index $(u, v)$ in this mask represents the watermark *location*. Each $\hat{n}_{u,v}$ is determined by the MAP criterion:

$$\hat{n}_{u,v} = \text{argmax}_{\tilde{n}_{u,v} \in \mathcal{R}^N} \left(\ln p_{x_{u,v}}(x' \mid \tilde{n}_{u,v}) + \ln p_{n_{u,v}}(\tilde{n}_{u,v})\right), \quad (3)$$

where $\tilde{n}$ represents a hypothetical watermark value being considered during the optimization process to maximize the posterior probability. The estimation accuracy enhancement is due to MAP estimation's statistical convergence towards the true watermark distribution as the dataset grows.

## 3.3 Global Watermark Mask Aggregation

In FL environments, the aggregation of local models is a crucial step for synthesizing a global model that benefits from the distributed learning process. Analogously, the aggregation of local watermark estimation masks is essential for constructing comprehensive global watermark knowledge.

The aggregation of the global watermark estimation mask, denoted by $\boldsymbol{M}$, incorporates contributions from local watermark estimation masks $\boldsymbol{m}^{(i)}$ from each client $i$ within the network $\mathcal{N}$. The aggregation process is governed by the equation:

$$\boldsymbol{M} = \sum_{i \in \mathcal{N}} \left( \frac{|D^{(i)}|}{\sum_{j \in \mathcal{N}} |D^{(j)}|} \cdot \bar{s}^{(i)} \right) \boldsymbol{m}^{(i)}, \quad (4)$$

where the weight for each client's local mask $\boldsymbol{m}^{(i)}$ is determined by the product of two key factors. The first factor, $\frac{|D^{(i)}|}{\sum_{j \in \mathcal{N}} |D^{(j)}|}$, consider the relative data sample size $|D^{(i)}|$ of the $i$-th client, indicating the proportion of data contributed by this client in comparison to the total data volume across all clients in $\mathcal{N}$. The second factor, $\bar{s}^{(i)}$, corresponds to the average size of the watermark estimation mask for the $i$-th client, which is computed as the mean of the dimensions of the mask $\boldsymbol{m}^{(i)}$. This measure reflects the spatial extent of the watermark information present within the client's data.

## 3.4 Morozov Regularization

After obtaining the global watermark estimation mask $\boldsymbol{M}$, it is a *mask integration* with the local masks to refine the watermark knowledge for each client. The refined local mask for client $i$, denoted by $\boldsymbol{m}^{*(i)}$, is achieved by blending $\boldsymbol{M}$ with $\boldsymbol{m}^{(i)}$:

$$\boldsymbol{m}^{*(i)} = \beta^{(i)}\boldsymbol{M} + (1 - \beta^{(i)})\boldsymbol{m}^{(i)}, \quad (5)$$

where $\beta^{(i)} \in [0, 1]$ is an adaptive hyperparameter that controls the degree to which the global mask influences the refined local

mask. The value of $\beta^{(i)}$ is dynamically adjusted based solely on the training performance difference, $\Delta\text{Acc}^{(i)}$, which is the difference between the highest validation accuracy among all clients and the validation accuracy of the current client $i$. To ensure $\beta^{(i)}$ scales appropriately between 0 and 1, it is calculated as follows:

$$\beta^{(i)} = \frac{\exp(-\Delta\text{Acc}^{(i)})}{\max_{j\in\mathcal{N}} \exp(-\Delta\text{Acc}^{(j)})}, \tag{6}$$

This formula uses an exponential function to decrease the influence of $\Delta\text{Acc}^{(i)}$ as it increases, ensuring that $\beta^{(i)}$ remains within the desired range and effectively balances the contribution of the global mask based on the relative performance of each client.

Regularization adds a term $reg(\cdot)$ to the loss function $f(\boldsymbol{\theta}; x, y)$, comprising a parameter matrix $\boldsymbol{\alpha}$ and norm $R(\boldsymbol{\theta})$, formulated as $reg(\boldsymbol{\theta}) = \boldsymbol{\alpha}R(\boldsymbol{\theta})$. The matrix $\boldsymbol{\alpha}$ balances regularization's importance, with higher values increasing bias and reducing overfitting, and lower values doing the opposite. This balance is captured by $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\delta)$, where $\delta$ measures deviation from real data.

Mathematically, refer to [26], the loss function with regularization is formulated as below,

$$F(\boldsymbol{\theta}^*) := \underset{\boldsymbol{\theta},\boldsymbol{\alpha}}{\textbf{argmin}} \sum_{i=1}^{N} \sum_{k=1}^{D^{(i)}} \left( f(\boldsymbol{\theta}; x_k + n_k, y_k) + \alpha ||\boldsymbol{\theta}||_2^2 \right). \tag{7}$$

Morozov regularization is a principle for choosing a regularization parameter, i.e., $\boldsymbol{\alpha}$, to stabilize the machine learning model to be trained. Specifically, let $x_\alpha$ be

$$x_\alpha = \arg\min_x \left\{ \frac{1}{2} \|\theta(x) - y\|^2 + \boldsymbol{\alpha} R(\boldsymbol{\theta}) \right\}. \tag{8}$$

$\boldsymbol{\alpha}$ can be considered as a control parameter. If $\boldsymbol{\alpha}$ is too small, the model overfits the watermarked in the data; and if $\boldsymbol{\alpha}$ is too big, the model loses the essential details. If $y^\delta$ is the watermarked data and assume that $\delta$ is the known noise level introduced by the watermarked data, then $\boldsymbol{\alpha}$ is chosen such that:

$$\|\theta(x_\alpha) - y^\delta\| = \delta = \boldsymbol{m}_k^*. \tag{9}$$

In other words, Morozov regularization chooses the value of $\boldsymbol{\alpha}$ that can make the norm $||\cdot||$ equal to the noise level (also called the Morozov's discrepancy principle [45]).

The *federated Morozov regularization* for FL in Alg. 1 operates in three main phases: watermark estimation, mask aggregation, and regularization parameter computation. Initially, each client's model parameters $\boldsymbol{\theta}^{(i)}$ are initialized. The watermark estimation phase involves using MAP-based method to estimate the $\hat{n}_k^{(i)}$ in each data point of client $i$'s dataset $d^{(i)}$ and get the $\boldsymbol{m}^{(i)}$ (Lines 4–7).

The watermark aggregation use the clients' watermark estimation mask to aggregate a global mask with Eq. (4) before federated learning training process (Lines 10–11).

During the federated learning process, each client refines the watermark estimation mask as outlined in Line 14. In this phase, the learning module of each client employs Morozov regularization to compute the regularization parameters. This involves setting an initial discrepancy tolerance and $\boldsymbol{\alpha}_k^{(i)}$, which are iteratively refined based on model predictions $\hat{y}_k^{(i)}$, residuals, and discrepancy measures until they converge within the set tolerance (Lines 15–21).

**Algorithm 1:** Federated Morozov Regularization

**Data:** Loss function $f(\cdot)$ in client $i$, data $x^i$ in client $i$.
**Result:** Regularized loss function.

1 **Initialization:**
2 **for** *client i in* $\mathcal{N}$ **do**
3  **for** *data* $x_k^{(i)}$ *in* $D^{(i)}$ **do**
4  Compute $\boldsymbol{m}_k^{(i)}$ using Eq.(3);
5  **end**
6  Combine dataset mask $\boldsymbol{m}^{(i)} = \frac{1}{|D^{(i)}|} \sum_{k=1}^{|D^{(i)}|} \boldsymbol{m}_k^{(i)}$;
7  Upload $\boldsymbol{m}^{(i)}$ to server;
8 **end**
9 **Server:**
10  Aggregate $\boldsymbol{M}$ using Eq. (4) and $\{\boldsymbol{m}^{(i)}\}_{i=1}^{\mathcal{N}}$ from clients;
11  Broadcast global watermark mask $\boldsymbol{M}$ to each client;
12 **Start FL training**:
13 **for** *each client i in* $\mathcal{N}$ **do**
14  Refine $\boldsymbol{m}^{*(i)}$ based on $\Delta\text{Acc}^{(i)}$ and Eq. (5);
15  Apply watermark estimation $\hat{n}_k^{(i)} \leftarrow \boldsymbol{m}_k^{*(i)}$;
16  Initialize $\boldsymbol{\alpha}_k^{(i)}$ and set tolerance $tol$;
17  **while** *discrepancy* $>$ *tol* **do**
18  Compute the model prediction $\hat{y}_k^{(i)} \leftarrow f(\boldsymbol{\theta}^{(i)}; x_k^{(i)})$;
19  Compute the residual: $residual_k^{(i)} = \left\| y_k^{(i)} - \hat{y}_k^{(i)} \right\|_2^2$;
20  Compute the discrepancy:
 $discrepancy = residual_k^{(i)} - \left\| \hat{n}_k^{(i)} \right\|_2^2$;
21  Update $\boldsymbol{\alpha}_k^{(i)}$;
22  **end**
23  **return** $\boldsymbol{\alpha}_k^{(i)}$;
24  **Update Regularized Loss:**
 $F_{reg}(\boldsymbol{\theta}^{(i)}) \leftarrow f(\boldsymbol{\theta}^{(i)}; x_k^{(i)} + \hat{n}_k^{(i)}, y_k^{(i)}) + \boldsymbol{\alpha}_k^{(i)} ||\boldsymbol{\theta}^{(i)}||_2^2$;
25 **end**

Finally, the algorithm utilizes the refined $\boldsymbol{\alpha}_k^{(i)}$ to adjust each client's model parameters $\boldsymbol{\theta}^{(i)}$. This adjustment considers the loss function $f(\cdot)$, regularization parameter $\boldsymbol{\alpha}_k^{(i)}$, and the estimated watermark $\hat{n}_k^{(i)}$. Consequently, the regularized loss $F_{reg}(\boldsymbol{\theta}^{(i)})$ is updated to reflect these changes, ensuring that the model parameters are optimized in alignment with the FL objectives and constraints.

Subsequently, each client performs local model training and adheres to the FL training protocol depicted in Step 4 of Fig. 2. Throughout the FL cycles, Alg. 1 systematically incorporates these updates into the overall FL training scheme.

# 4 Evaluation

## 4.1 Evaluation Settings

We assess the performance of our technique in a client-server testbed. The server is equipped with an Nvidia RTX 3090 GPU and an AMD Ryzen 9 5900X CPU, running on Ubuntu 20.04 LTS. For client devices, we employ 40 Nvidia Jetson. The performance

and quantity detail can be seen in Table. 1. Our testbed equation[1] have been shown in Fig. 3 to understand the efficacy of federated Morozov regularization in heterogeneity edge clients. We connected each client device to switches via an Ethernet cable. Data exchange in federated learning, including metadata and models, is facilitated by accessing the IP bound to each device. The communications protocol uses sockets. The underlying Jetson driver is supported by Jetpack 5.1.

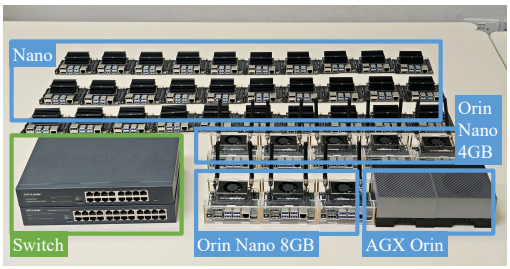



**Figure 3: Evaluation testbed in the lab.**

**Table 1: Performance Comparison of Edge Devices**

| Device | Quantity | GPU | CPU |
|---|---|---|---|
| AGX Orin | 2 | 248 TOPS | 8-core, 2.2 GHz |
| Orin Nano-8 | 3 | 40 TOPs | 6-core, 1.5 GHz |
| Orin Nano-4 | 5 | 20 TOPS | 6-core, 1.5 GHz |
| Nano | 30 | 472 GFLOPS | 4-core, 1.4 GHz |

**Models and datasets.** In the experimental phase, our investigation employed a range of neural network architectures to perform the image recognition task. The first is a Lite-CNN[2], characterized by its simplicity yet effectiveness. Alongside our custom CNN, we integrated two well-established models: VGG [47] and ResNet-18 [14]. The initial learning rate is set to 0.1, and the batch size is set to 64 by default. Using the SSP [17] synchronization strategy, the local epochs are set to 5 by default. In our experimental setup, we evaluated the performance of our proposed method using widely recognized image classification datasets, including MNIST [9], Cifar-10 [23], Tiny-ImageNet [7] and COVID-FL [54]. We chose these datasets to cover various image types and complexities, from simple handwritten digits to more complex real-world images that is originally collected with watermarks (i.e., COVID-FL). We used Lite-CNN for MNIST, VGG for Cifar-10, and ResNet for both Tiny-ImageNet and COVID-FL. These models were chosen to match the complexity and scale of each dataset, from simple tasks to more challenging multi-classification problems.

**Watermark setting.** In the creation of our watermarked dataset, various watermark embedding techniques, including frequency watermark: DWT, DCT, DFT, LSB, and spatial watermark: LSB, explicit watermarking [3, 43], are employed. Specifically, for each clean dataset (i.e., MNIST, CIFAR-10, and Tiny-ImageNet), we simulate the watermarked dataset with four watermarking techniques(i.e., DWT, DCT, DFT, and LSB), and the amount of data processed by each watermarking technique is determined by the watermark heterogeneity parameter $\beta$. The COVID-FL dataset was originally collected with different watermarks from the real world.

Although the adjustment parameters for watermark intensity vary across different methods (for instance, the intensity in explicit watermarks refers to transparency, while in some frequency domain watermarks, like DFT, it refers to modulation amplitude), we normalize the intensity of all watermarks to a 0-1 scale. The embedding location in spatial domain watermarks denotes the position of the watermark within the image (such as the center or edges), whereas in frequency domain watermarks, it refers to the frequency within the image spectrum (like high, mid, or low frequency; in DCT, this ranges from the LL to HH domain).

The watermark embedding intensity is adjusted between 0.01 and 1 for our experiments. Each client in the FL employs the same watermarking method and parameters, ensuring consistency across the dataset.

**Evaluation metrics.** Our performance evaluation focuses FL performance. Referring to the evaluation metric of the shortcut learning research [18, 34, 38], FL performance is assessed using task accuracy, which measures the percentage of correct predictions by the FL models on a distributed dataset, and loss, indicating the prediction error with lower values signifying better performance.

**Benchmark methods.** We compare the federated Morozov regularization with the following peer robust training methods in FL [42] [11], generalized regularization [56] and regularization for shortcut learning [18].

- FedAvg [35]: Used to establish a performance baseline in our experiments, serving as a foundation for comparison with other FL algorithms.
- GroupLasso [56]: a generalized regularization for machine learning by adding a penalty term. We modified GroupLasso to federated learning version based on the client-level profiling setting.
- AFL [11]: Using global model transmission, local gradient calculations, and averaging, with hyperparameters set to $\alpha_1 = 0.75, \alpha_2 = 0.01, \alpha_3 = 0.1$ in our experiments.
- RFA [42]: A Roubstness aggregation method for corrupted data. We applied with hyperparameters as per the original paper: R = 3 and $v = 10^{-6}$.
- FD [18]: A feature regularization with frequency filter tools. We modified FD to federated learning version (Fed-FD) based on the client-level profiling setting.

## 4.2 Evaluation Results & Analysis

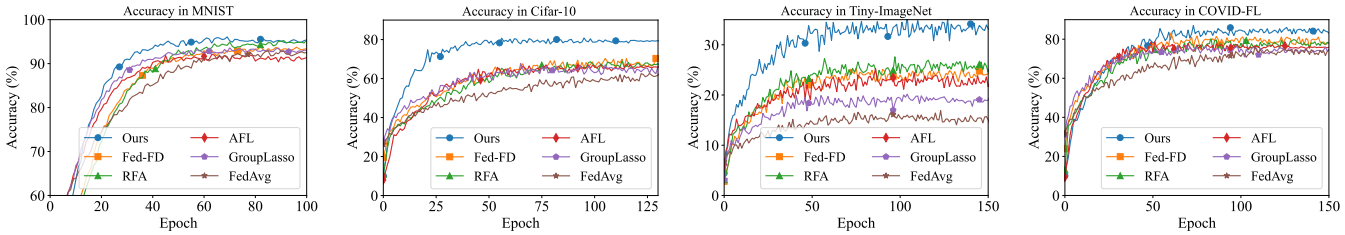
*4.2.1 Improvement with federated Morozov regularization.* The evaluation metric was task accuracy in FL, compared under two different training and inference conditions: with watermarked data but clean inference, and with both watermarked training and inference.

The experimental design in Table 2, bifurcates the analysis into two scenarios: inference on clean data and inference under watermarked conditions. This distinction aims to uncover the impact of shortcut learning induced by watermarks, which affects not only the inference with watermarked features but also the performance on clean data, highlighting the pervasive influence of watermarks on model behavior. The settings for data and watermark heterogeneity are set to $\alpha = 0.5$, $\beta = 0.5$, which be defined in Sec. 4.2.2.

[1]The 4GB and 8GB Jetson Orin Nano boards have the same appearance.
[2]Lite-CNN consists of two 5×5 convolutional layers (64 channels each) with 3×3 max pooling, followed by two dense layers (384 and 192 units) and a softmax output layer.

**Table 2: FL method benchmark accuracy(%) comparison under different settings.**

| Setting | Watermarked Dataset & Clean Inference | | | | Watermarked Dataset & Inference | | | |
|---|---|---|---|---|---|---|---|---|
| | MNIST | Cifar-10 | Tiny. | COVID-FL | MNIST | Cifar-10 | Tiny. | COVID-FL |
| FedAvg | 95.35±0.04 | 71.29±0.72 | 25.56±1.06 | 77.30±1.42 | 92.10±0.02 | 64.07±0.83 | 15.45±1.14 | 74.43±1.23 |
| GroupLasso | 96.44±0.02 | 71.30±0.71 | 28.94±1.00 | 81.43±1.53 | 92.59±0.01 | 64.14±0.51 | 19.22±0.83 | 74.29±1.43 |
| AFL | 96.05±0.02 | 72.52±0.35 | 30.42±0.53 | 83.41±2.56 | 91.53±1.03 | 65.10±0.07 | 22.52±0.51 | 75.41±1.51 |
| RFA | 96.73±0.03 | 72.54±0.87 | 30.62±1.03 | 84.52±1.98 | 94.14±0.01 | 67.89±1.12 | 25.70±0.97 | 77.62±1.83 |
| Fed-FD | 96.86±0.01 | 76.89±0.75 | 33.80±0.71 | 84.09±1.27 | 93.83±0.02 | 68.04±0.90 | 24.03±0.95 | 79.93±1.25 |
| **Ours** | **97.35±0.03** | **80.86±1.04** | **35.43±0.73** | **87.14±1.03** | **95.24±0.02** | **79.26±1.09** | **33.29±1.01** | **84.10±1.68** |




**(a) MNIST dataset** **(b) Cifar-10 dataset** **(c) Tiny-ImageNet dataset** **(d) COVID-FL dataset**

**Figure 4: Task accuracy and convergence with epoch growing of method benchmark in different watermarked datasets.**

Our method demonstrates superior accuracy across all datasets and settings, underscoring its effectiveness in mitigating the adverse effects of shortcut learning in FL. Specifically, in the clean inference setting, our approach achieves an accuracy of 97.35% on MNIST, 80.86% on Cifar-10, 35.43% on Tiny-ImageNet (denoted as Tiny.), and 87.14% on COVID-FL. These results are notably higher than those obtained with other methods, such as *FedAvg*, *GroupLasso*, *AFL*, *RFA*, and *Fed-FD*. The improvement is even more pronounced in the watermarked dataset & inference setting, with scores of 95.24% on MNIST, 79.26% on Cifar-10, 33.29% on Tiny-ImageNet, and 84.10% on COVID-FL. The detail learning performance with epoch growing can be seen in Fig. 4.

The underperformance of other methods can be attributed to their inability to effectively address the dual challenge posed by non-IID data and the presence of watermarks. Methods like *FedAvg* and *GroupLasso*, while foundational in FL, lack specific mechanisms to counteract the nuanced effects of watermarked data, leading to compromised accuracy. *AFL* and *RFA*, despite introducing robustness in aggregation, do not directly tackle the issue of shortcut learning induced by watermarks. *Fed-FD*, which applies feature regularization, shows promise but still falls short of fully mitigating the impact of watermarks on model learning.

*4.2.2 Results on Data and Watermark Heterogeneity.* In FL, data-level heterogeneity is primarily manifested through the presence of non-IID (independent and identically distributed) data challenges. For the non-IID problem in the FL experiment, we define the degree of non-IID data and non-IID watermark as follows:

In a multi-client training scenario, each client's data is independently sampled with class labels from $N$ classes, following a categorical distribution with vector $q$ ($q_i \geq 0$, $i \in [1, N]$, $\|q\|_1 = 1$). Non-IID client data is simulated by sampling $q$ from a *Dirichlet distribution*, $\mathrm{Dir}(\alpha\mathbf{p})$, where $\mathbf{p}$ is the prior class distribution, and $\alpha > 0$ determines client similarity. An infinite $\alpha$ implies uniform client distributions, while $\alpha$ near zero results in maximum divergence among clients.

In the context of non-IID watermark settings, we adopt a distribution similar to the *Dirichlet distribution* to manage the variability in watermark characteristics such as intensity ($I$) and location ($L$). Intensity ranges from 0 (no watermark) to 1 (maximum intensity), while location varies from low-frequency areas or image edges to high-frequency areas or central regions. We introduce a parameter $\beta$ in $\mathrm{Dir}(\beta\mathbf{p})$ to control the degree of non-IID in the watermark distribution. A higher $\beta$ indicates more uniformity in watermark characteristics across clients, leading to similar intensity and location settings. Conversely, a lower $\beta$ results in greater diversity, with each client having distinct watermark intensity and placement. This approach allows us to simulate a spectrum of watermark patterns across different clients, reflecting various degrees of intensity and placement. For the real-world dataset COVID-FL, the data is already divided among different clients by medical institutions, thus we utilize the official non-IID configuration distribution to proceed. The variation in equipment used by different medical institutions, along with their respective watermark design preferences, inherently introduces non-IID watermarks. Therefore, COVID-FL, as a more realistic watermarked dataset, can be considered a reference for real-world issues and does not require additional non-IID watermark design and settings.

In our experimental analysis, the combined impact of non-IID data and non-IID watermark on the federated Morozov regularization technique is depicted through heatmaps, revealing a compounded decrease in accuracy with the simultaneous presence of both non-IID conditions. We have selected *Fed-FD* as the benchmark for testing our method based on its superior performance as demonstrated in Sec. 4.2.1. When the non-IID degree for both data and watermark is at its highest, we observe a notable reduction in accuracy, illustrating the challenges posed by these conditions. For example, with $\alpha$ of 0.5 and $\beta$ of 0.5, the accuracy drops to around

68.04%. Modifications to the technique, as reflected in the second heatmap, show improvements in this challenging scenario with a notable increase in accuracy. Under the same high non-IID conditions, the accuracy improves to 79.26%. The third heatmap, which focuses on the percentage of improvement, highlights the effectiveness of our modifications. In scenarios with non-IID data and watermark, our method achieves a substantial improvement, with the most pronounced increase in accuracy reaching up to 11.22%.

**Table 3: Accuracy(%) comparison for different datasets on varying non-IID degrees.**

| Dataset | $\beta$ \ $\alpha$ | Fed-FD | | | **Ours** | | |
|---|---|---|---|---|---|---|---|
| | | 100.0 | 5.0 | 0.5 | 100.0 | 5.0 | 0.5 |
| Cifar10 | 0.5 | 72.01 | 71.16 | 68.04 | 81.29 | 80.71 | 79.26 |
| | 10.0 | 75.43 | 72.21 | 68.51 | 82.55 | 81.42 | 81.73 |
| | 100.0 | 83.86 | 74.99 | 69.07 | 83.92 | 81.85 | 81.14 |
| | w/o | 85.13 | 78.61 | 75.63 | 85.46 | 82.50 | 81.26 |
| MNIST | 0.5 | 96.54 | 95.69 | 93.83 | 96.75 | 96.43 | 95.24 |
| | 10.0 | 96.81 | 96.10 | 95.61 | 97.57 | 97.00 | 96.34 |
| | 100.0 | 97.85 | 96.72 | 96.24 | 98.21 | 97.41 | 96.83 |
| | w/o | 98.64 | 97.31 | 97.20 | 98.63 | 97.59 | 97.16 |
| Tiny. | 0.5 | 28.12 | 23.85 | 24.03 | 34.65 | 34.28 | 33.29 |
| | 10.0 | 33.43 | 29.19 | 26.92 | 35.93 | 34.47 | 33.82 |
| | 100.0 | 36.59 | 34.72 | 33.74 | 38.14 | 37.72 | 35.46 |
| | w/o | 39.73 | 38.98 | 36.80 | 39.51 | 39.27 | 36.75 |

**Table 4: Accuracy(%) comparison in ablation study. The bottom line is the component of our method.**

**(a) Study on watermark estimation.**

| Method | MNIST | Cifar-10 | Tiny. | COVID-FL |
|---|---|---|---|---|
| Blind | 93.86 | 64.89 | 26.80 | 79.00 |
| Stacking | 93.55 | 66.58 | 26.29 | 81.26 |
| **MAP** | **95.24** | **79.26** | **33.29** | **84.10** |

**(b) Study on estimation mask aggregation.**

| Method | MNIST | Cifar-10 | Tiny. | COVID-FL |
|---|---|---|---|---|
| w/o. | 94.45 | 72.52 | 30.50 | 83.23 |
| Avg. | 95.05 | 73.57 | **33.42** | 83.50 |
| **Aggr.** | **95.24** | **79.26** | 33.29 | **84.10** |

**(c) Study on feature extractor regularization.**

| Method | MNIST | Cifar-10 | Tiny. | COVID-FL |
|---|---|---|---|---|
| Tik. | 94.24 | 73.44 | 31.46 | 80.29 |
| L1 | 93.93 | 76.25 | 30.21 | 80.41 |
| **Moro.** | **95.24** | **79.26** | **33.29** | **84.10** |

*4.2.3 Ablation Study.* We analyze three components designed for such environments: MAP-based watermark estimation (MAP), watermark estimation aggregation (Aggr.) and Morozov regularization (Moro.) in Table. 4. The goal is to evaluate how effectively these components, can counteract the reduction in accuracy often caused by watermarking, compared to alternative methods or variations.

**Study on watermark estimation.** Our exploration delved into the efficacy of MAP-based watermark estimation by comparing it against both its variants and analogous statistical methodologies. One is Blind Image Quality Measurement (denoted as Blind) [48], a technique predicated on leveraging statistical attributes to gauge image quality. Another is the strategy of stacking all dataset images to generate a uniformly weighted mask, tantamount to an averaged weighted MAP-based estimation (denoted as Stacking). As shown in Table 4a, the MAP approach manifested a notably superior accuracy enhancement relative to its counterparts, with a 12.68% increment over Stacking within the Cifar-10 dataset. Such findings underscore that methodologies centered on image quality estimation (Blind) and indiscriminate estimation of images and watermarks (Stacking) are ineffectual in procuring a robust watermark estimation.

**Study on estimation mask aggregation.** We delve into the efficacy of watermark mask aggregation by both omitting this component (denoted as w/o.) and evaluating its variants, specifically average aggregation (denoted as Avg.), where the local masks from all clients undergo aggregation with equal weighting. As evidenced in the Table. 4b, aggregation demonstrates enhanced performance in the Cifar-10 dataset, characterized by strong heterogeneity and a smaller quantity of images. Conversely, for datasets with a larger volume and more uniform data, such as Tiny-ImageNet and COVID-FL, the performance difference compared to average aggregation is minimal. This phenomenon can be attributed to the intrinsic purpose of mask aggregation, which is to furnish a global mask that aids clients with less data in obtaining a more applicable mask. Therefore, if the local datasets of clients are sufficiently large, the improvement brought about by aggregation may be marginal.

**Study on Morozov regularization.** In our ablation study focusing on Morozov regularization, we maintained identical inputs for the estimation mask while employing a simplified form of regularization. Morozov regularization, conceptualized as a variant of Tikhonov regularization, introduces parameter adjustments that are more finely tuned to the noise levels encountered. Thus, Tikhonov regularization (denoted as Tik.) is utilized as a comparative measure to ascertain the significance of adjustments in regularization parameters. Furthermore, we investigate whether L1 regularization, a widely referenced regularization technique, also demonstrates improvements in the context of prior information on watermark estimation (denoted as L1). Insights from Table. 4c reveal that the enhancements attributed to Morozov Regularization are predominantly observed in datasets with smaller capacities, such as Cifar-10, and in datasets where the watermark patterns are relatively fixed, such as COVID-FL. It is also observed that, although other forms of regularization exhibit limited improvements over the baseline, their compatibility with watermark estimation is not as pronounced.

## 5 Conclusion

Our paper introduces federated Morozov regularization, a technique for federated learning on watermarked data. federated Morozov regularization addresses the challenges of diverse watermarking across FL participants without prior knowledge of watermark specifics. It probes watermark details and uses Morozov regularization to adapt local model training. Experiments on 40 Jetson edge devices show federated Morozov regularization improves accuracy by 11.22%. An ablation study validates each component's contribution to FL model performance on watermarked datasets.

## Acknowledgments

Dan Wang's work is supported by RGC GRF 15200321, 15201322, 15230624, RGC-CRF C5018-20G, ITC ITF-ITS/056/22MX, and PolyU 1-CDKK. Hao Wang's work is supported in part by the National Science Foundation (NSF) grants 2315612, 2327480, and 2332638. Dan Wang is the corresponding author.

## References

[1] Frank Bauer and Mark A. Lukas. 2011. Comparingparameter Choice Methods for Regularization of Ill-Posed Problems. *Mathematics and Computers in Simulation* 81, 9 (2011), 1795–1841.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in Terra Incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018). 456–473.

[3] Mahbuba Begum and Mohammad Shorif Uddin. 2022. Towards the Development of an Effective Image Watermarking System. *SECURITY AND PRIVACY* 5, 2 (2022), e196.

[4] Kirill Bykov, Klaus-Robert Müller, and Marina M-C Höhne. 2023. *Mark my words: Dangers of watermarked images in imagenet.* Springer.

[5] Qinwei Chang, Leichao Huang, Shaoteng Liu, Hualuo Liu, Tianshu Yang, and Yexin Wang. 2022. Blind Robust Video Watermarking Based on Adaptive Region Selection and Channel Reference. In *Proceedings of the 30th ACM International Conference on Multimedia* (New York, NY, USA, 2022-10-10) *(MM '22)*. 2344–2350.

[6] A. Zhibin Chen, B. Xiao-Mei Huo, and C. You-Wei Wen. 2013. Adaptive Regularization for Color Image Restoration Using Discrepancy Principle. In *2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2013)*. 1–6.

[7] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. 2017. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819* (2017).

[8] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. 2021. AI for Radiographic COVID-19 Detection Selects Shortcuts over Signal. *Nature Machine Intelligence* 3, 7 (2021), 610–619.

[9] Li Deng. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142.

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[11] Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu, and Anuj Kumar. 2019. Active Federated Learning. arXiv:1909.12641

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*.

[13] Kangshuai Guo, Zhijian Xu, Shichao Luo, Feigao Wei, Yan Wang, and Yanru Zhang. 2023. Invisible Video Watermark Method Based on Maximum Voting and Probabilistic Superposition. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 9446–9450.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Katherine Hermann, Ting Chen, and Simon Kornblith. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 19000–19015.

[16] Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael Curtis Mozer. 2023. On the Foundations of Shortcut Learning. In *The Twelfth International Conference on Learning Representations* (2023-10-13).

[17] Qirong Ho, James Cipar, Henggang Cui, Jin Kyu Kim, Seunghak Lee, Phillip B. Gibbons, Garth A. Gibson, Gregory R. Ganger, and Eric P. Xing. 2013. More Effective Distributed ML via a Stale Synchronous Parallel Parameter Server. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1* (Red Hook, NY, USA, 2013-12-05) *(NIPS'13)*. 1223–1231.

[18] Mobarakol Islam and Ben Glocker. 2023. Frequency Dropout: Feature-Level Regularization via Randomized Filtering. In *Computer Vision – ECCV 2022 Workshops* (Cham). 281–295.

[19] Tim Jahn and Bangti Jin. 2020. On the Discrepancy Principle for Stochastic Gradient Descent. *Inverse Problems* 36, 9 (2020), 095009.

[20] Zhaoyang Jia, Han Fang, and Weiming Zhang. 2021. MBRS: Enhancing Robustness of DNN-based Watermarking by Mini-Batch of Real and Simulated JPEG Compression. In *Proceedings of the 29th ACM International Conference on Multimedia* (New York, NY, USA, 2021-10-17) *(MM '21)*. 41–49.

[21] Changsong Jiang, Chunxiang Xu, and Yuan Zhang. 2021. PFLM: Privacy-preserving Federated Learning with Membership Proof. *Information Sciences* 576 (2021), 288–311.

[22] Wan Jiang, Yunfeng Diao, He Wang, Jianxin Sun, Meng Wang, and Richang Hong. 2023. Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 8910–8921.

[23] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[24] Mingrui Lao, Nan Pu, Zhun Zhong, Nicu Sebe, and Michael S. Lew. 2023. FedVQA: Personalized Federated Visual Question Answering over Heterogeneous Scenes. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 7796–7807.

[25] Jun Li and Hung-Lung Huang. 1999. Retrieval of Atmospheric Profiles from Satellite Sounder Measurements by Use of the Discrepancy Principle. *Applied Optics* 38, 6 (1999), 916–923.

[26] Lingfeng Li, Jiang Yang, et al. 2022. Generalization Error Analysis of Neural Networks with Gradient Based Regularization. *Communications in Computational Physics* 32, 4 (2022), 1007–1038.

[27] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. 2022. Federated Learning on Non-IID Data Silos: An Experimental Study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 965–978.

[28] Qiushi Li, Wenwu Zhu, Chao Wu, Xinglin Pan, Fan Yang, Yuezhi Zhou, and Yaoxue Zhang. 2020. InvisibleFL: Federated Learning over Non-Informative Intermediate Updates against Multimedia Privacy Leakages. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. New York, NY, USA, 753–762.

[29] Shiqin Liu, Shiyuan Feng, Jinxia Wu, Wei Ren, Weiqi Wang, and Wenwen Zheng. 2021. Exploration of the Influence on Training Deep Learning Models by Watermarked Image Dataset. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. 421–428.

[30] Yunfei Long, Zhe Xue, Lingyang Chu, Tianlong Zhang, Junjiang Wu, Yu Zang, and Junping Du. 2023. FedCD: A Classifier Debiased Federated Learning Framework for Non-IID Data. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 8994–9002.

[31] Jianghu Lu, Shikun Li, Kexin Bao, Pengju Wang, Zhenxing Qian, and Shiming Ge. 2023. Federated Learning with Label-Masking Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 222–232.

[32] Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. 2021. Rectifying the Shortcut Learning of Background for Few-Shot Learning. In *Advances in Neural Information Processing Systems* (2021), Vol. 34. 13073–13085.

[33] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. 2023. A State-of-the-Art Survey on Solving Non-IID Data in Federated Learning. *Future Generation Computer Systems* 135 (2023), 244–258.

[34] Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. 2022. Causally Motivated Shortcut Removal Using Auxiliary Labels. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. 739–766.

[35] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 1273–1282.

[36] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. 2020. Automatic Shortcut Removal for Self-Supervised Representation Learning. In *Proceedings of the 37th International Conference on Machine Learning* (2020-11-21). 6927–6937.

[37] Reza Moradi, Reza Berangi, and Behrouz Minaei. 2020. A Survey of Regularization Strategies for Deep Models. *Artificial Intelligence Review* 53, 6 (2020), 3947–3986.

[38] Nicolas M. Müller, Jochen Jacobs, Jennifer Williams, and Konstantin Böttinger. 2023. Localized Shortcut Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023). 3721–3725.

[39] Arnold Neumaier. 1998. Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization. *SIAM Rev.* 40, 3 (1998), 636–666.

[40] Jordi Nin and Sergio Ricciardi. 2013. Digital Watermarking Techniques and Security Issues in the Information and Communication Society. In *2013 27th International Conference on Advanced Information Networking and Applications Workshops*. 1553–1558.

[41] Luca Oneto, Sandro Ridella, and Davide Anguita. 2016. Tikhonov, Ivanov and Morozov Regularization for Support Vector Machine Learning. *Machine Learning* 103, 1 (2016), 103–136.

[42] Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. 2022. Robust Aggregation for Federated Learning. *IEEE Transactions on Signal Processing* 70 (2022), 1142–1154.

[43] Asaad F. Qasim, Farid Meziane, and Rob Aspin. 2018. Digital Watermarking: Applicability for Developing Trust in Medical Imaging Workflows State of the Art Review. *Computer Science Review* 27 (2018), 45–60.

[44] Piyapat Saranrittichai, Chaithanya Kumar Mummadi, Claudia Blaiotta, Mauricio Munoz, and Volker Fischer. 2022. Overcoming Shortcut Learning in a Target Domain by Generalizing Basic Visual Factors from a Source Domain. In *Computer*

*Vision – ECCV 2022* (Cham, 2022). 294–309.

[45] O. Scherzer. 1993. The Use of Morozov's Discrepancy Principle for Tikhonov Regularization for Solving Nonlinear Ill-Posed Problems. *Computing* 51, 1 (1993), 45–60.

[46] Jaemin Shin, Yuanchun Li, Yunxin Liu, and Sung-Ju Lee. 2022. FedBalancer: Data and Pace Control for Efficient Federated Learning on Heterogeneous Clients. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services* (New York, NY, USA, 2022-06-27) *(MobiSys '22)*. 436–449.

[47] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[48] Huixuan Tang, Neel Joshi, and Ashish Kapoor. 2011. Learning a Blind Measure of Perceptual Image Quality. In *CVPR 2011* (2011-06). 305–312.

[49] Tiffany Tuor, Shiqiang Wang, Bong Jun Ko, Changchang Liu, and Kin K. Leung. 2021. Overcoming Noisy and Irrelevant Data in Federated Learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 5020–5027.

[50] Sviatoslav Voloshynovskiy, Alexander Herrigel, Nazanin Baumgaertner, and Thierry Pun. 2000. A Stochastic Approach to Content Adaptive Digital Image Watermarking. In *Information Hiding*. Berlin, Heidelberg, 211–236.

[51] S. Voloshynovskiy, S. Pereira, T. Pun, J.J. Eggers, and J.K. Su. 2001. Attacks on Digital Watermarks: Classification, Estimation Based Attacks, and Benchmarks. *IEEE Communications Magazine* 39, 8 (2001), 118–126.

[52] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. 2019. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. *IEEE Journal on Selected Areas in Communications* 37, 6 (2019), 1205–1221.

[53] Xiaoshuai Wu, Xin Liao, and Bo Ou. 2023. SepMark: Deep Separable Watermarking for Unified Source Tracing and Deepfake Detection. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 1190–1201.

[54] Rui Yan, Liangqiong Qu, Qingyue Wei, Shih-Cheng Huang, Liyue Shen, Daniel L. Rubin, Lei Xing, and Yuyin Zhou. 2023. Label-Efficient Self-Supervised Federated Learning for Tackling Data Heterogeneity in Medical Imaging. *IEEE Transactions on Medical Imaging* 42, 7 (2023), 1932–1943.

[55] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLOS Medicine* 15, 11 (2018), e1002683.

[56] Huaqing Zhang, Jian Wang, Zhanquan Sun, Jacek M Zurada, and Nikhil R Pal. 2019. Feature selection for neural networks using group lasso regularization. *IEEE Transactions on Knowledge and Data Engineering* 32, 4 (2019), 659–673.

[57] Pengling Zhang, Huibin Yan, Wenhui Wu, and Shuoyao Wang. 2023. Improving Federated Person Re-Identification through Feature-Aware Proximity and Aggregation. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 2498–2506.

[58] Yi Zhang, Jitao Sang, Junyang Wang, Dongmei Jiang, and Yaowei Wang. 2023. Benign shortcut for debiasing: Fair visual recognition via intervention with shortcut features. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA) *(MM '23)*. 8860–8868.

[59] Zixin Zhang, Fan Qi, Shuai Li, and Changsheng Xu. 2023. AffectFAL: Federated Active Affective Computing with Non-IID Data. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. New York, NY, USA, 871–882.

[60] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. 2021. Federated Learning on Non-IID Data: A Survey. *Neurocomputing* 465 (2021), 371–390.

[61] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. 2021. Joint Optimization in Edge-Cloud Continuum for Federated Unsupervised Person Re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. New York, NY, USA, 433–441.