南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Federated Noisy Label Learning

——Noisy clients/samples selection-based method

Reporter: Tong Jin

Federated Learning with Noisy Labels (F-LNL)

FL has emerged as a powerful framework for decentralized ML enabling multiple clients to collaboratively train models without sharing raw data, thus preserving privacy.

However, some of the labels can be wrong due to carelessness or lack of expert knowledge.

Directly learning with such noisy labels

Wrong knowledge

Degrading the generalization performance

- **Setting:**

  (1) **Some clients are clean while others are not**

  (2) **Each client has partially noisy data**

Centralized Noisy Label Learning

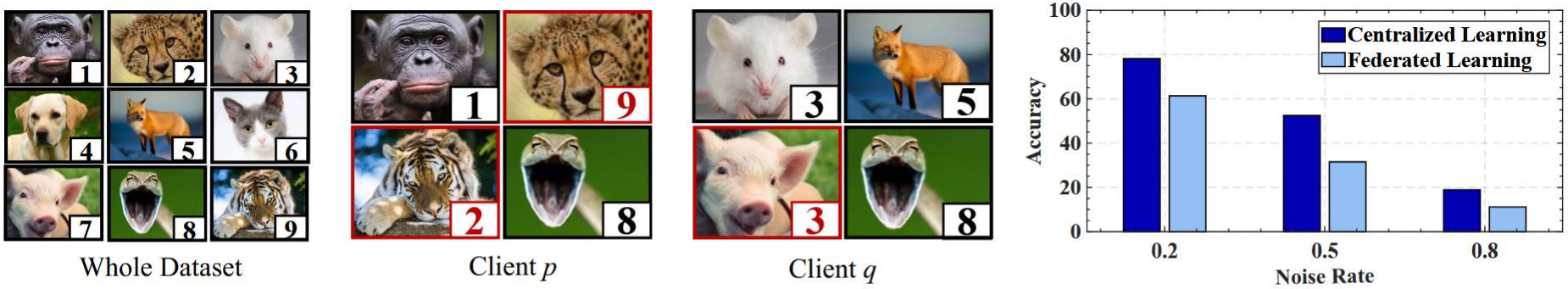
- **Method：**

(1) **Loss correction**: aims to correct the loss by estimating the noise transition matrix, adjusting the example labels or weights.

(2) **Example selection**: separate clean examples from noisy ones, and further treat the identified mislabeled examples as unlabeled data for semi-supervised learning or noisy-label correction.

Limitations

- **Data Privacy and Inaccessibility**: the restricted size and insufficient diversity of the local datasets in clients dramatically degrade the effectiveness of noise processing methods.
- **Data Heterogeneity:** There are large differences in the distribution of client data. Traditional methods assume that the data is IID, which leads to the failure of noise detection and correction.
- **Noise Heterogeneity:** The difference in the proportion or type of noise between different clients is significant. The global unified noise processing strategy cannot adapt to all clients.



(a) Label noise distribution on client p and q. Black are correct labels and Red are noisy labels.

(b) CL with noisy labels & FL with heterogeneous noisy labels.

Figure 1: (a): The heterogeneous label noise distributions encompass diverse true class samples or varying label noise transitions. (b): The performance comparison between CL and FL on CIFAR-10 with the rate of label noise 0.2, 0.5, and 0.8.

Federated Noisy Label Learning

- **Method：**

  (1) **Loss correction**

  (2) **Noisy clients/samples selection-based method**

  ① selection ② correction ③ reweight

FedDiv: Collaborative Noise Filtering for Federated Learning with Noisy Labels

Jichang Li[1,2], Guanbin Li[1,3*], Hui Cheng[4], Zicheng Liao[2,5], Yizhou Yu[2*]

[1]School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou, China
[2]Department of Computer Science, The University of Hong Kong, Hong Kong
[3]Guangdong Province Key Laboratory of Information Security Technology
[4]UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai, China
[5]Zhejiang University, Hangzhou, China
csjcli@connect.hku.hk, liguanbin@mail.sysu.edu.cn, chenghui_123@sjtu.edu.cn
zichengliao@gmail.com, yizhouy@acm.org

AAAI 2024

Motivation

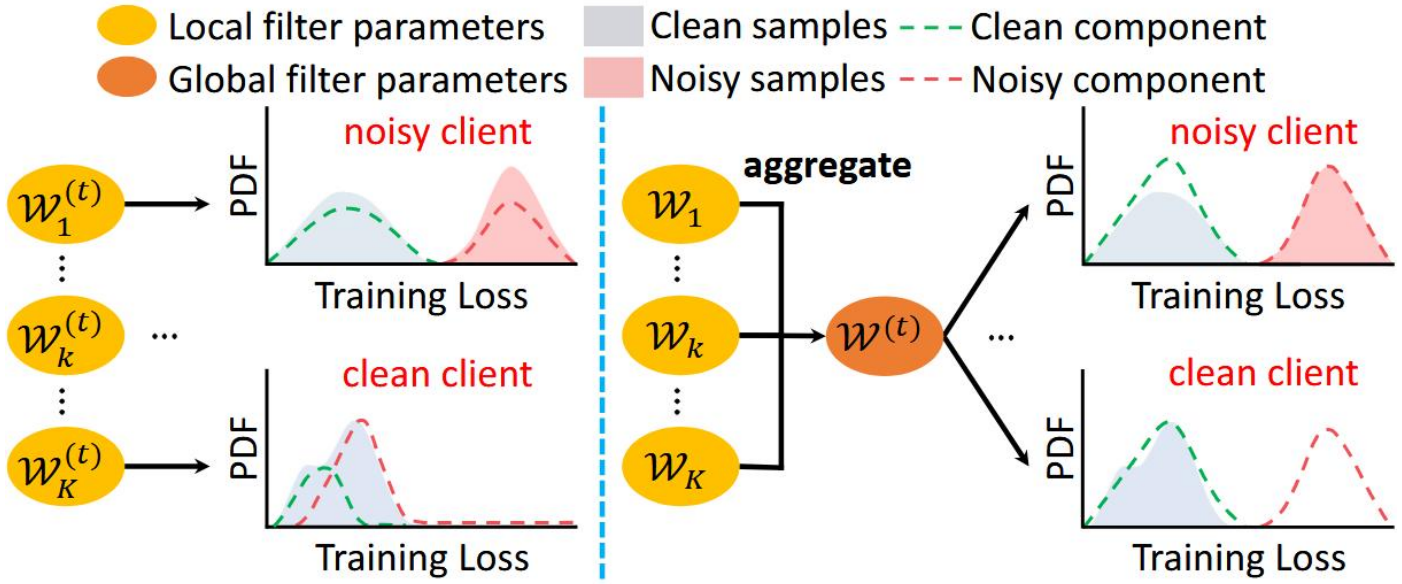


Figure 1: (a) Local noise filtering may have limited capabilities as each client develops its own **local noise filter using its own private data only**. Especially on clean clients, such filters would incorrectly identify a subset of clean samples to be noisy. (b) Collaborative noise filtering proposed by us significantly improves the performance of label noise filtering on each client as a **federated noise filter is learned by distilling knowledge from all clients.** PDF: Probability density function.

Overview of the Training Procedure

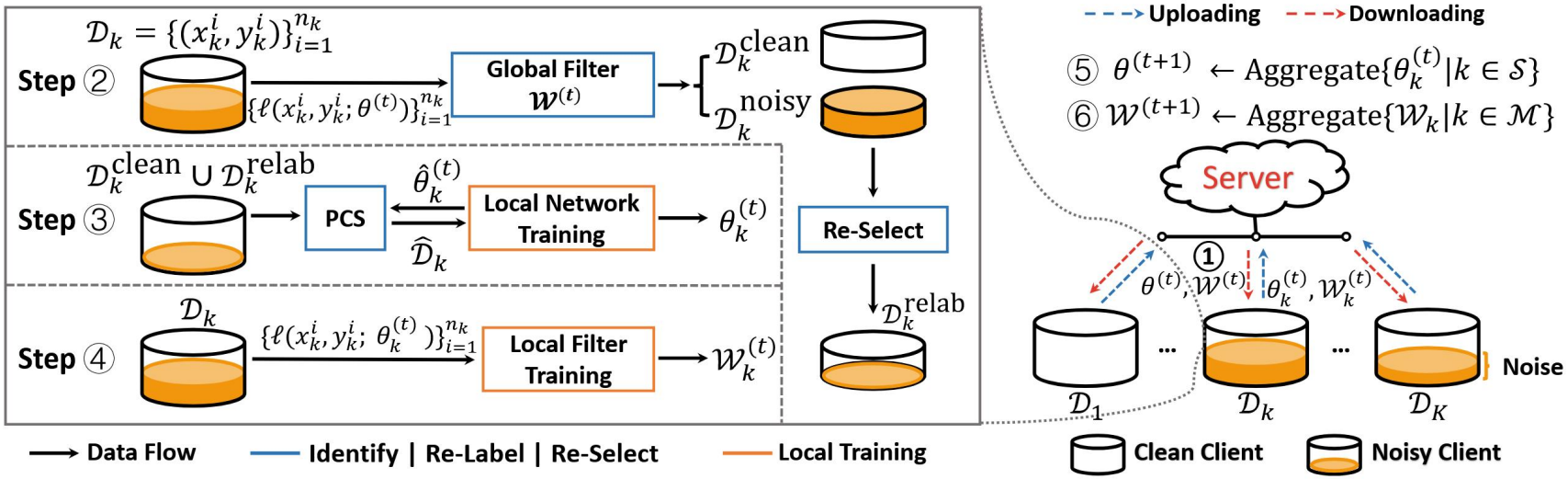


Figure 2: An overview of the training procedure proposed by FedDiv. In this work, the parameters of a local neural model and a local noise filter are simultaneously learned on each client during the local training sessions, while both types of parameters are aggregated on the server.

Federated Noise Filter

Local filter training. A local GMM to model the local distribution of clean and noisy samples by fitting the per-sample loss distribution.

Per-sample cross-entropy loss: $\{\ell(x, y; \theta_k^{(t)}) | (x, y) \in \mathcal{D}_k\}$,

Two-component GMM model: $\mathcal{W}_k^{(t)} = (\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\sigma}_k^{(t)}, \boldsymbol{\pi}_k^{(t)})$

$P(\ell(x, y; \theta_k^{(t)}) | z = g)$ is modeled as a Gaussian distribution $\mathcal{N}(\ell(x, y; \theta_k^{(t)}); \mu_{kg}^{(t)}, \sigma_{kg}^{(t)})$

g = 1 to represent the "clean" Gaussian component, i.e., the Gaussian component with a smaller loss, while g = 2 denotes the "noisy" one.

Federated filter aggregation. After parameter uploading, the federated filter model can be constructed by aggregating the local filter parameters corresponding to all the clients.

$$\mu_g^{(t+1)} = \sum_{k \in \mathcal{M}} \frac{n_k}{\sum_{k \in \mathcal{M}} n_k} \mu_{kg}, \quad \sigma_g^{(t+1)} = \sum_{k \in \mathcal{M}} \frac{n_k}{\sum_{k \in \mathcal{M}} n_k} \sigma_{kg}, \quad \pi_g^{(t+1)} = \sum_{k \in \mathcal{M}} \frac{n_k}{\sum_{k \in \mathcal{M}} n_k} \pi_{kg},$$

$$\mathcal{W}^{(t+1)} = (\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\sigma}^{(t+1)}, \boldsymbol{\pi}^{(t+1)})$$

Label Noise Filtering

The client receives the parameters of the global model and the federated filter model. The probability of being clean can be estimated through its posterior probability for the "clean" component

$$\mathbf{p}(\text{“clean”}|x, y; \theta^{(t)}) = P(z = 1|x, y; \theta^{(t)}).$$

We can divide the samples of $\mathcal{D}_k$ into a clean subset and a noisy subset by thresholding their probabilities of being clean with the threshold 0.5

$$\mathcal{D}_k^{\text{clean}} \leftarrow \{(x, y) \,|\mathbf{p}(\text{“clean”}|x, y; \theta_k^{(t)}) \geq 0.5, \forall\, (x, y) \in \mathcal{D}_k\},$$
$$\mathcal{D}_k^{\text{noisy}} \leftarrow \{(x,) \,|\mathbf{p}(\text{“clean”}|x, y; \theta_k^{(t)}) < 0.5, \forall\, (x, y) \in \mathcal{D}_k\}.$$

Noisy Sample Relabeling

Noise level of the k-th client: $\hat{\delta}_k = |\mathcal{D}_k^{\text{noisy}}|/|\mathcal{D}_k|$

$\hat{\delta}_k > 0.1$ as a noisy client.

We relabel those samples with high prediction confidence (by setting a confidence threshold ζ) by assigning predicted labels from the global model

$$\mathcal{D}_k^{\text{relab}} \leftarrow \{(x, \hat{y}) \,|\, \max(\mathbf{p}(x; \theta^{(t)})) \geq \zeta, \forall x \in \mathcal{D}_k^{\text{noisy}}\}, \qquad \hat{y} = \hat{y}(x) = \arg\max \mathbf{p}(x; \theta^{(t)})$$

Predictive Consistency Based Sampler

However, the complete elimination of label noise among clients during noise filtering and relabeling is unattainable; Relabeling inevitably introduces new label noise, causing instability in local model training.

Class-unbalanced local data would contribute to the cause of the local model bias towards the dominant classes.

$$\textbf{De-bias:}\quad F(x) \leftarrow f(x;\hat{\theta}_k^{(t)}) - \xi\log(\hat{p}_k),$$

$$\hat{p}_k^{(t)} \leftarrow m\hat{p}_k + (1-m)\frac{1}{n_k}\sum_{x\in\mathcal{D}_k}\mathbf{p}(x;\theta_k^{(t)}),$$

PCS is used to re-select labeled training samples to perform local training, enforcing the **consistency** of class labels respectively **predicted by global and local models**.

$$\mathcal{D}_k^{\text{resel}} \leftarrow \{(x,y)\,|\hat{y}(x) = \tilde{y}(x), \forall(x,y) \in \mathcal{D}_k^{\text{clean}} \cup \mathcal{D}_k^{\text{relab}}\}.$$

We update the training dataset for optimizing the local model

$$\hat{\mathcal{D}}_k = \begin{cases} \mathcal{D}_k^{\text{resel}}, & \text{if } \hat{\delta}_k \geq 0.1, \\ \mathcal{D}_k, & \text{if } \hat{\delta}_k < 0.1. \end{cases}$$

Objectives for Local Model Training

MixUp regularization: $\ddot{x} = \lambda x_i + (1-\lambda)x_j$ and $\ddot{y} = \lambda p_y(y_i) + (1-\lambda)p_y(y_j)$

Cross-entropy loss: $\mathcal{L}_{mix} = -\sum_{b=1}^{B} \ddot{y}_b \log \mathbf{p}(\ddot{x}_b; \hat{\theta}_k^{(t)})$

Regularizing the average prediction of a local model over every mini-batch using a uniform prior distribution is a viable solution to overcome the non-IID problem

$$\mathcal{L}_{reg} = \sum_{c=1}^{C} \hat{\mathbf{q}}^c \log\left(\frac{\hat{\mathbf{q}}^c}{\mathbf{q}^c}\right), \text{ where } \mathbf{q} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{p}(\ddot{x}_b; \hat{\theta}_k^{(t)}),$$

The overall loss function: $\mathcal{L}_{final} = \mathcal{L}_{mix} + \eta\mathcal{L}_{reg}$

Experimental Results

Performance Results

Method	Best Test Accuracy ± Standard Deviation					
	ρ=0.4		ρ=0.6		ρ=0.8	
	τ=0.0	τ=0.5	τ=0.0	τ=0.5	τ=0.0	τ=0.5
FedAvg	89.46±0.39	88.31±0.80	86.09±0.50	81.22±1.72	82.91±1.35	72.00±2.76
RoFL	88.25±0.33	87.20±0.26	87.77±0.83	83.40±1.20	87.08±0.65	74.13±3.90
ARFL	85.87±1.85	83.14±3.45	76.77±1.90	64.31±3.73	73.22±1.48	53.23±1.67
JointOpt	84.42±0.70	83.01±0.88	80.82±1.19	74.09±1.43	76.13±1.15	66.16±1.71
DivideMix	77.35±0.20	74.40±2.69	72.67±3.39	72.83±0.30	68.66±0.51	68.04±1.38
FedCorr	94.01±0.22	94.15±0.18	92.93±0.25	92.50±0.28	91.52±0.50	90.59±0.70
FedDiv (Ours)	**94.42±0.29**	**94.30±0.19**	**93.67±0.22**	**93.41±0.21**	**92.98±0.60**	**91.44±0.25**

Table 1: Best test accuracy (%) of FedDiv and existing SOTA methods on CIFAR-10 with IID setting at diverse noise levels.

Method	Best Test Accuracy ± Standard Deviation		
	p = 0.7 α_{Dir} = 10	p = 0.7 α_{Dir} = 1	p = 0.3 α_{Dir} = 10
FedAvg	78.88±2.34	75.98±2.92	67.75±4.38
RoFL	79.56±1.39	72.75±2.21	60.72±3.23
ARFL	60.19±3.33	55.86±3.30	45.78±2.84
JointOpt	72.19±1.59	66.92±1.89	58.08±2.18
DivideMix	65.70±0.35	61.68±0.56	56.67±1.73
FedCorr	90.52±0.89	88.03±1.08	81.57±3.68
FedDiv (Ours)	**93.18±0.42**	**91.95±0.26**	**85.31±2.28**

Table 2: Best test accuracy (%) of FedDiv and existing SOTA methods on CIFAR-10 with non-IID setting at the noise level $(\rho, \tau) = (6.0, 0.5)$.

Method	Best Test Accuracy ± Standard Deviation		
	ρ=0.4 τ=0.5	ρ=0.6 τ=0.5	ρ=0.8 τ=0.5
FedAvg	64.41±1.79	53.51±2.85	44.45±2.86
RoFL	59.42±2.69	46.24±3.59	36.65±3.36
ARFL	51.53±4.38	33.03±1.81	27.47±1.08
JointOpt	58.43±1.88	44.54±2.87	35.25±3.02
DivideMix	43.25±1.01	40.72±1.41	38.91±1.25
FedCorr	74.43±0.72	66.78±4.65	59.10±5.12
FedDiv (Ours)	**74.86±0.91**	**72.37±1.12**	**65.49±2.20**

Table 3: Best test accuracy (%) of FedDiv and existing SOTA methods on CIFAR-100 with IID setting at diverse noise levels.

Dataset	**CIFAR-100**	**Clothing1M**
Noise level (ρ, τ)	(0.4, 0.0)	-
Method\(p, α_{Dir})	(0.7, 10)	-
FedAvg	64.75±1.75	70.49
RoFL	59.31±4.14	70.39
ARFL	48.03±4.39	70.91
JointOpt	59.84±1.99	71.78
DivideMix	39.76±1.18	68.83
FedCorr	72.73±1.02	72.55
FedDiv (Ours)	**74.47±0.34**	**72.96±0.43**

Table 4: Best test accuracy (%) of FedDiv and existing SOTA methods on CIFAR-100 and Clothing1M under the non-IID data partitions.

FedFixer: Mitigating Heterogeneous Label Noise in Federated Learning

Xinyuan Ji[1, 2], Zhaowei Zhu[3], Wei Xi[1*], Olga Gadyatskaya[2], Zilong Song[1], Yong Cai[4], Yang Liu[5*]

[1]Xi'an Jiaotong University
[2]Leiden University
[3]Docta.ai
[4]IQVIA Inc & California State University, Monterey Bay
[5]University of California, Santa Cruz
{jixinyuan1996, weixi.cs}@gmail.com, yangliu@ucsc.edu

AAAI 2024

Methodology

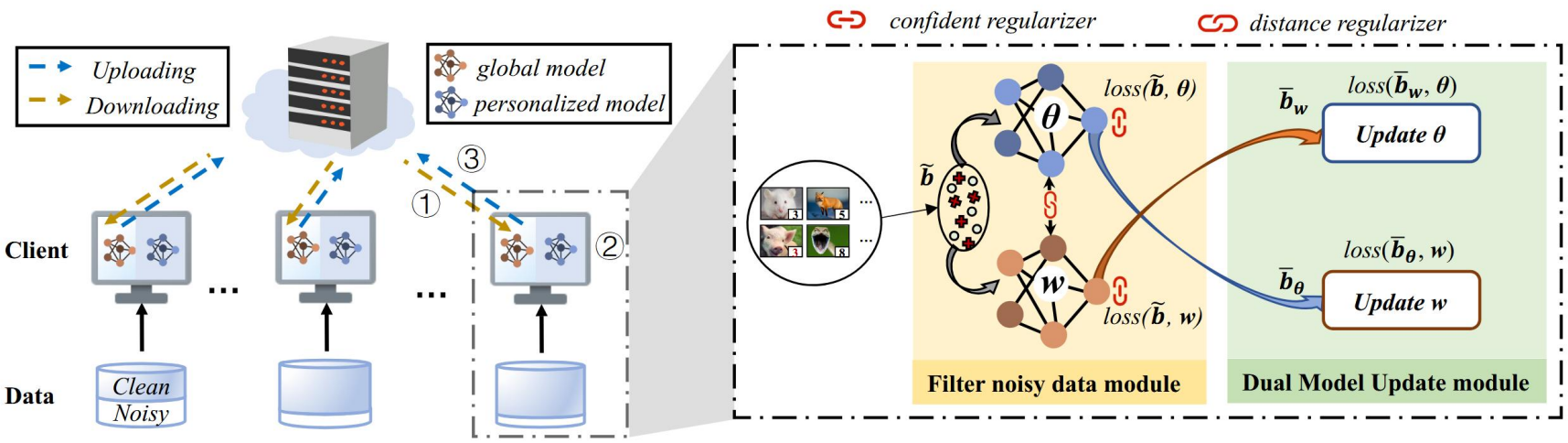


Figure 2: Overall architecture. The training process of the proposed FedFixer has three stages: ① deployment of the global model, ② the local model updates, and ③ the global model aggregation. In the second stage, dual models are alternately updated (in the Dual Model Update module) based on the selected samples (in the Filter Noisy Data module) by each other.

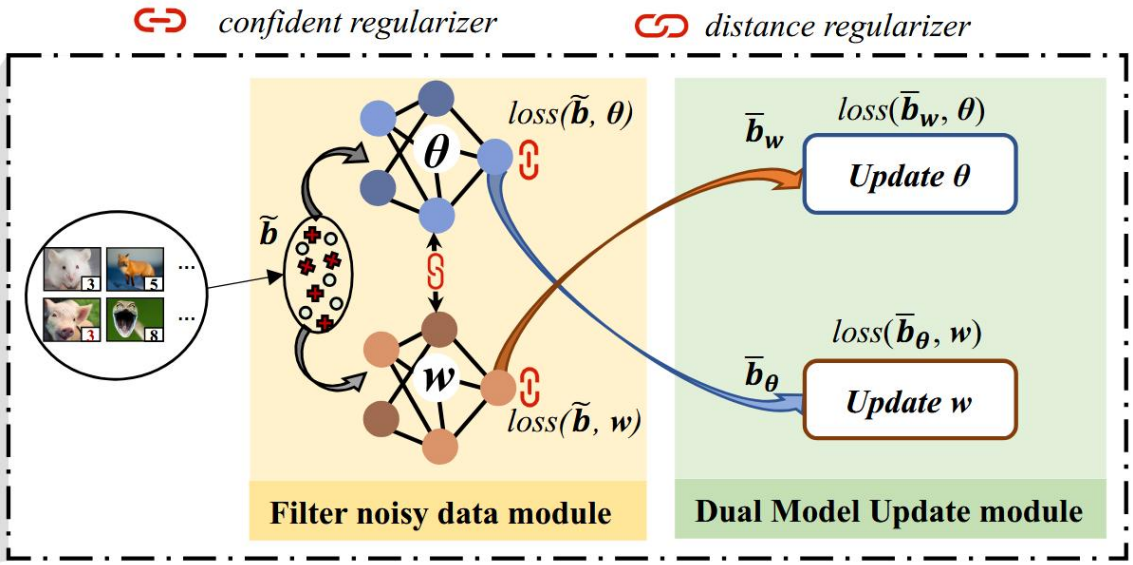
Filter Noisy Data

On client k, the dual models are denoted the global model w_k and personalized model θ_k.

When a batch $\mathcal{B}$ arrives, the personalized model and global model respectively select a small proportion of **low loss samples** out of the mini-batch $\mathcal{B}$ by sample selector

$$v_n = \mathbb{1}\left(\ell_{CE}\left(f\left(x_n\right), \tilde{y}_n\right) - \frac{1}{L}\sum_{\tilde{y}\in[L]} \ell_{CE}\left(f\left(x_n\right), \tilde{y}\right) < 0\right)$$

The selected instances are **fed into its peer model** to generate the loss in the "Dual Model Update" module for alternate parameter updates.



Regularizers

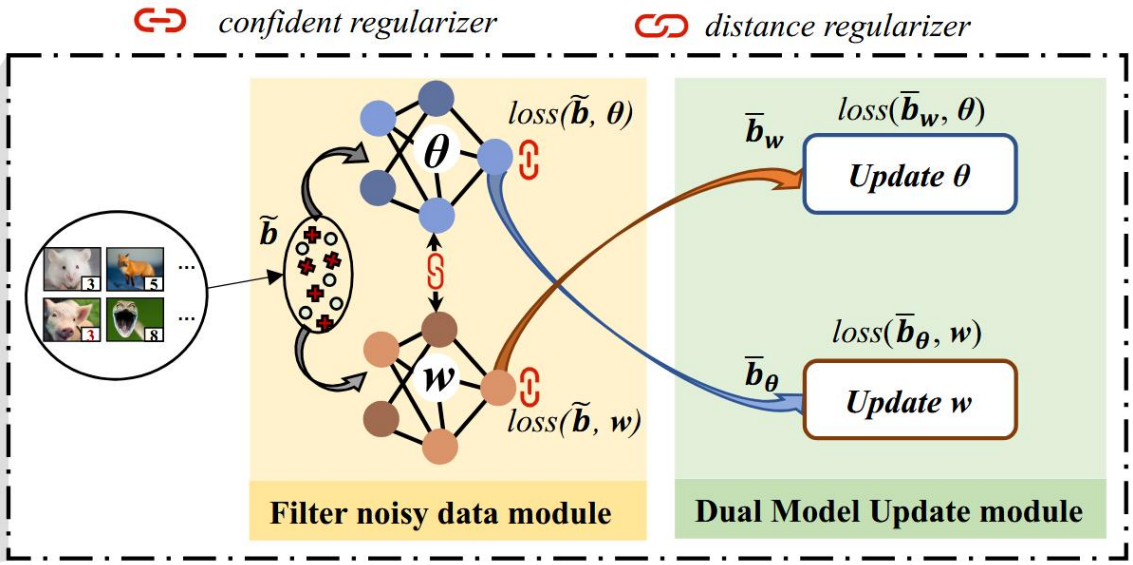
The personalized model θ is only updated solely based on the local client's data, it is **prone to overfitting due to the limited amount of data available**.

Confidence Regularizer: aims to guide the model towards better fitting clean datasets:

$$\ell_{\mathrm{CR}}\left(f\left(x_n\right)\right) := -\beta \cdot \mathbb{E}_{\mathcal{D}_{\widetilde{Y}|\widetilde{D}}}\left[\ell_{CE}\left(f\left(x_n\right), \widetilde{Y}\right)\right]$$

$$\mathbb{P}(\tilde{Y} = i) = \#\text{Label\_}i/N$$

Distance Regularizer: $\frac{\lambda}{2}\left\|\theta_k - w\right\|^2$



Experimental Results

Performance Results

Datasets	Methods	IID			non-IID		
		$\rho = 0.0$ $\tau = 0.0$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 1$ $\tau = 0.5$	$\rho = 0.0$ $\tau = 0.0$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 1$ $\tau = 0.5$
MNIST	Local + CORES^2	96.79 ± 0.05	62.38 ± 3.88	37.30 ± 1.62	97.52 ± 0.17	90.31 ± 3.29	55.80 ± 6.45
	Global + CORES^2	98.05 ± 0.05	97.39 ± 0.13	80.98 ± 4.78	97.46 ± 0.53	97.38 ± 0.12	87.45 ± 0.48
	FedAvg	98.39 ± 0.04	97.66 ± 0.09	93.90 ± 0.31	97.46 ± 0.57	97.73 ± 0.05	95.57 ± 0.28
	FedProx	98.22 ± 0.08	96.49 ± 0.08	93.90 ± 0.64	93.69 ± 4.98	97.25 ± 0.16	95.22 ± 0.37
	RFL*	90.70 ± 0.54	96.54 ± 0.12	96.64 ± 0.08	90.51 ± 0.36	95.61 ± 0.34	91.56 ± 8.37
	MR	97.41 ± 0.21	95.98 ± 0.36	88.47 ± 1.48	97.09 ± 0.54	95.35 ± 0.54	90.53 ± 2.26
	FedCorr*	$\mathbf{98.68 \pm 0.16}$	$\mathbf{98.09 \pm 0.22}$	95.67 ± 0.22	97.49 ± 0.82	97.75 ± 0.17	95.75 ± 0.46
	FedFixer	98.07 ± 0.02	97.80 ± 0.18	$\mathbf{96.79 \pm 0.92}$	$\mathbf{98.05 \pm 0.06}$	$\mathbf{98.01 \pm 0.18}$	$\mathbf{96.00 \pm 0.24}$
CIFAR-10	Local + CORES^2	84.23 ± 0.26	67.96 ± 0.60	22.51 ± 1.79	86.16 ± 0.71	68.46 ± 2.38	26.96 ± 1.21
	Global + CORES^2	91.31 ± 0.09	85.81 ± 0.26	54.22 ± 3.18	90.27 ± 0.17	86.23 ± 0.20	35.31 ± 3.51
	FedAvg	90.33 ± 0.11	77.93 ± 0.29	33.87 ± 0.19	89.82 ± 0.34	78.30 ± 0.24	28.77 ± 0.59
	FedProx	91.12 ± 0.20	79.33 ± 0.21	35.38 ± 0.31	$\mathbf{90.50 \pm 0.27}$	80.13 ± 0.40	29.63 ± 1.17
	RFL*	85.54 ± 0.26	84.06 ± 0.24	54.83 ± 1.06	83.83 ± 0.41	83.68 ± 0.32	49.49 ± 1.77
	MR	70.65 ± 1.61	50.02 ± 2.13	23.67 ± 1.37	68.80 ± 0.52	51.46 ± 1.29	22.23 ± 1.72
	FedCorr*	$\mathbf{91.82 \pm 0.22}$	$\mathbf{88.05 \pm 0.69}$	52.30 ± 0.91	81.07 ± 1.06	87.63 ± 0.64	45.43 ± 3.36
	FedFixer	90.72 ± 0.47	87.06 ± 0.30	$\mathbf{62.87 \pm 0.17}$	89.76 ± 0.32	$\mathbf{87.82 \pm 0.22}$	$\mathbf{59.01 \pm 0.55}$

Table 2: Average (5 trials) accuracies (%) of various methods on MNIST and CIFAR-10 datasets with IID and non-IID settings at different noise levels (ρ: ratio of noisy clients, τ: lower bound of client noise level). The best results are highlighted in bold. Methods trained with label correction are marked by $\star$.

Algorithm	Local + CORES^2	Global + CORES^2	FedAvg	FedProx	FedCorr*	RFL*	MR	FedFixer
Acc	53.32	65.18	66.75	66.76	67.50	65.41	50.01	**70.52**

Table 3: Accuracies (%) of various methods on Clothing1M with non-IID setting.

FedClean: A General Robust Label Noise Correction for Federated Learning

Xiaoqian Jiang [1] **Jing Zhang** [1] *

ICML 2025

Methodology

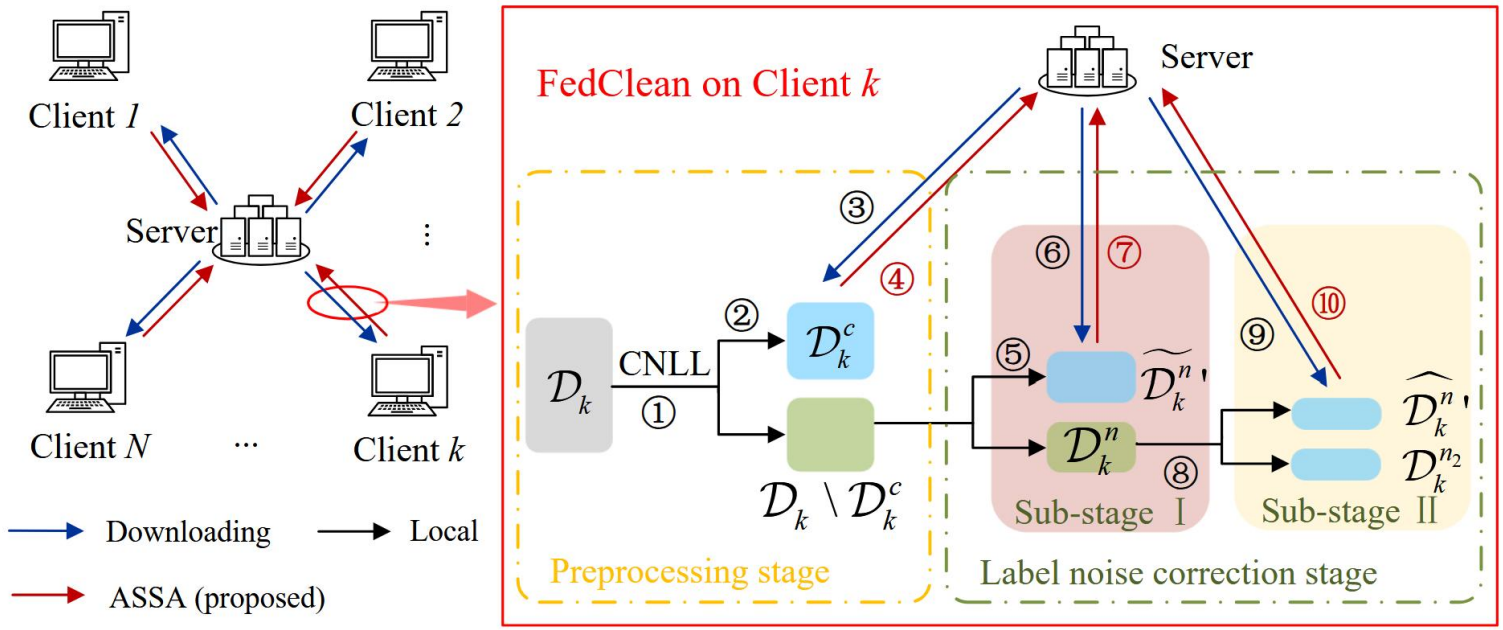
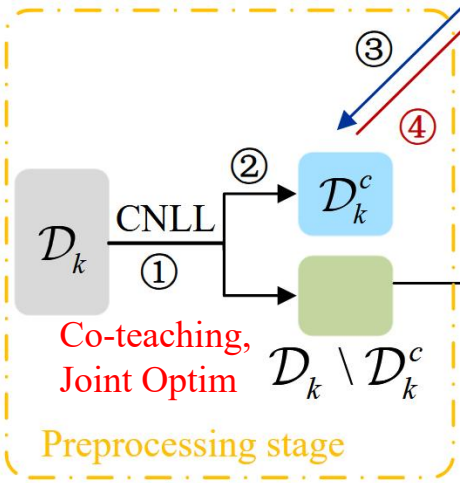


Figure 1. Framework of FedClean. Algorithm steps are numbered accordingly.

Preprocessing Stage



Add inferred labels. Let each client k execute CNLL locally to train local model θ_k. Each sample x_k^i is assigned an additional inferred label $\bar{y}_k^i$ based on the predictions of this local model.

Local dataset: $\mathcal{D}_k = \{(x_k^i, y_k^i, \bar{y}_k^i) | \bar{y}_k^i = \theta_k(x_k^i)\}$

Select clean samples. We select the samples whose annotation labels and inferred labels are identical.

Clean samples: $\mathcal{D}_k^c = \{(x_k^i, y_k^i, \bar{y}_k^i) \in \mathcal{D}_k | y_k^i = \bar{y}_k^i\}.$

Noisy samples: $\mathcal{D}_k \setminus \mathcal{D}_k^c$

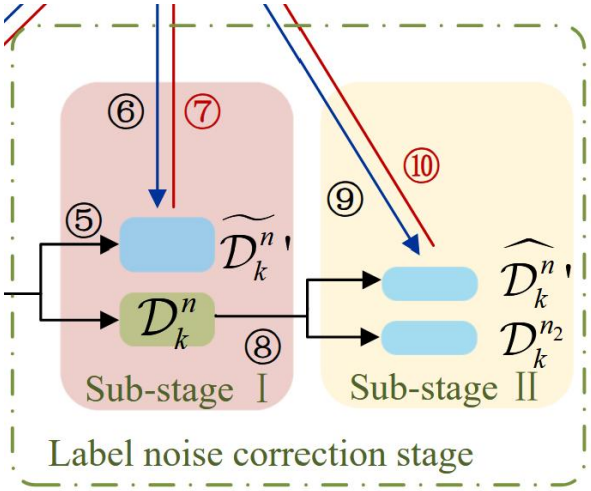
Adaptive Sample Size-weighted Aggregation.

We train the global model over T1 rounds using these **selected clean samples** on all clients.

The weighting coefficients are determined by **the size of the clean dataset** selected by each client.

$$w^t \leftarrow \sum_{k \in \mathcal{N}^t} \frac{|\mathcal{D}_k^c|}{\sum_{i \in \mathcal{N}^t} |\mathcal{D}_i^c|} \cdot w_k^t.$$

Label Noise Correction Stage



Sub-stage I: Correction dominated by inferred labels

For controversial samples (samples whose the annotation labels conflict with their inferred labels), we propose a **collaborative per-sample loss**:

$$\mathcal{L}_{\text{co}}(y_k^i, \bar{y}_k^i, \hat{y}_k^i; \theta_G^{T_1}) = \mathcal{L}_{\text{an}}(y_k^i, \hat{y}_k^i; \theta_G^{T_1}) - \mathcal{L}_{\text{in}}(\bar{y}_k^i, \hat{y}_k^i; \theta_G^{T_1}).$$

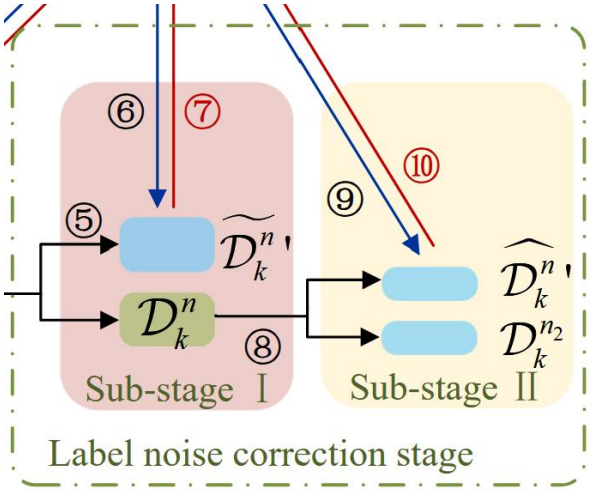
The annotation label loss $\mathcal{L}_{\text{an}}(y_k^i, \hat{y}_k^i; \theta_G^{T_1})$, measuring the discrepancy between the **annotation label** and **the global model's prediction**.

The inferred label loss $\mathcal{L}_{\text{in}}(\bar{y}_k^i, \hat{y}_k^i; \theta_G^{T_1})$ quantifying the misalignment between the sample's annotation and its assigned inferred label.

Only meaningful if the global model's prediction is consistent with the inferred label.

If the model's predicted label aligns with the inferred label, it indicates that the **global model has effectively learned the pattern of the inferred label**. The inferred label likely represents the true class of the sample or is closer to it.

Sub-stage I: Correction dominated by inferred labels



Calculate the collaborative per-sample loss.

$$\widetilde{\mathcal{D}}_k^n = \{(x_k^i, y_k^i, \bar{y}_k^i) \in \mathcal{D}_k \setminus \mathcal{D}_k^c | \bar{y}_k^i = \hat{y}_k^i\}.$$

GMM divide: $\widetilde{\mathcal{D}}_k^{n_1}$: corrected by inferred labels

$\widetilde{\mathcal{D}}_k^{n_2}$: disputed subset

Filter samples for correction.

We select the top σ1-percent of samples from $\widetilde{\mathcal{D}}_k^{n_1}$ that have the highest collaborative per-sample loss. These samples are then relabeled using the inferred label

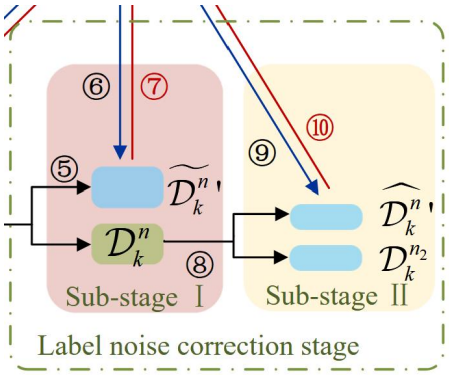
$$\widetilde{\mathcal{D}}_k^{n\prime} = \arg \max_{\substack{\tilde{\mathcal{D}} \subseteq \widetilde{\mathcal{D}}_k^{n_1} \\ |\tilde{\mathcal{D}}| = \sigma_1 \cdot |\widetilde{\mathcal{D}}_k^{n_1}|}} \mathcal{L}_{\text{co}}(\tilde{\mathcal{D}}; \theta_G^{T_1}).$$

Adaptive Sample Size-weighted Aggregation.

The global model is improved using the **corrected samples** by updating the parameter over T2 rounds.

$$w^t \leftarrow \sum_{k \in \mathcal{N}^t} \frac{|\mathcal{D}_k^c| + |\widetilde{\mathcal{D}}_k^{n\prime}|}{\sum_{i \in \mathcal{N}^t} (|\mathcal{D}_i^c| + |\widetilde{\mathcal{D}}_k^{n\prime}|)} \cdot w_k^t.$$

Sub-stage II: Correction dominated by global labels



The global model may not be fully optimized during the initial training phase, there still be a small number of samples that are not corrected.

Calculate the per-sample loss.

$$\mathcal{D}_k^n = \{(x_k^i, y_k^i) | (x_k^i, y_k^i, \bar{y}_k^i) \in \mathcal{D}_k \setminus (\mathcal{D}_k^c \cup \widetilde{\mathcal{D}_k^n}')\}$$

$$\mathcal{L}_{\text{an}}(\mathcal{D}_k^n; \theta_G^{T_1+T_2})$$

GMM divide：$\mathcal{D}_k^{n_1}$: noisy subset $\mathcal{D}_k^{n_2}$: clean subset

Select samples for correction.

We identify the top σ2-percent of samples from $\mathcal{D}_k^{n_1}$ that have the highest loss

$$\widehat{\mathcal{D}_k^n} = \arg \max_{\substack{\hat{\mathcal{D}} \subseteq \mathcal{D}_k^{n_1} \\ |\hat{\mathcal{D}}| = \sigma_2 \cdot |\mathcal{D}_k^{n_1}|}} \mathcal{L}_{\text{an}}(\hat{\mathcal{D}}; \theta_G^{T_1+T_2})$$

The subset of samples to be re-labeled: $\widehat{\mathcal{D}_k^n}' = \{(x_k^i, y_k^i) \in \widehat{\mathcal{D}_k^n} | \max(\theta_G^{T_1+T_2}(x_k^i)) \geq \varepsilon\}$

Adaptive Sample Size-weighted Aggregation.

We train the global model over T3 rounds using the **corrected labels and the clean samples**.

$$w^t \leftarrow \sum_{k \in \mathcal{N}^t} \frac{|\mathcal{D}_k^c| + |\widetilde{\mathcal{D}_k^n}'| + |\widehat{\mathcal{D}_k^n}'| + |\mathcal{D}_k^{n_2}|}{\sum_{i \in \mathcal{N}^t} (|\mathcal{D}_i^c| + |\widetilde{\mathcal{D}_k^n}'| + |\widehat{\mathcal{D}_k^n}'| + |\mathcal{D}_k^{n_2}|)} \cdot w_k^t$$

Performance Results

Table 2. Average (5 trials) accuracies (%) of various methods on CIFAR-10 dataset with IID and non-IID settings at different noise levels (ρ: ratio of noisy clients, τ: lower bound of client noise level). The best results are highlighted in bold.

Methods	IID			non-IID		
	$\rho = 0$ $\tau = 0$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 1$ $\tau = 0.5$	$\rho = 0$ $\tau = 0$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 1$ $\tau = 0.5$
FedAvg	91.74 ± 0.19	83.16 ± 0.31	38.36 ± 2.21	90.04 ± 0.17	82.61 ± 0.26	34.65 ± 1.53
FedProx	91.52 ± 0.22	82.45 ± 0.27	35.21 ± 1.75	**90.82 ± 0.18**	81.76 ± 0.22	32.84 ± 1.65
FedCorr	**91.83 ± 0.21**	**91.12 ± 0.30**	47.49 ± 1.98	90.21 ± 0.16	**89.11 ± 0.25**	39.40 ± 1.51
FedNoRo	90.05 ± 0.19	88.48 ± 0.24	32.18 ± 1.89	88.91 ± 0.20	86.99 ± 0.21	30.21 ± 1.72
FedBeat	89.28 ± 0.23	85.92 ± 0.28	36.13 ± 2.03	89.55 ± 0.19	83.92 ± 0.24	33.20 ± 1.62
FedELC	85.62 ± 0.20	87.60 ± 0.29	35.72 ± 2.10	89.90 ± 0.17	83.75 ± 0.23	31.95 ± 1.80
FedFixer	90.72 ± 0.47	87.06 ± 0.30	62.87 ± 0.17	89.76 ± 0.32	87.82 ± 0.22	59.01 ± 0.55
FedClean[1]	88.77 ± 0.17	85.25 ± 0.29	81.68 ± 2.10	87.79 ± 0.20	86.53 ± 0.25	77.12 ± 1.95
FedClean[2]	91.14 ± 0.19	88.41 ± 0.27	**83.75 ± 2.03**	89.34 ± 0.21	86.79 ± 0.23	**80.55 ± 1.82**

Table 3. Average (5 trials) accuracies (%) of various methods on CIFAR-100 dataset with IID setting at different noise levels (ρ: ratio of noisy clients, τ: lower bound of client noise level). The best results are highlighted in bold.

Methods	$\rho = 0$ $\tau = 0$	$\rho = 0.5$ $\tau = 0.3$	$\rho = 1$ $\tau = 0.5$
FedAvg	**72.36 ± 0.19**	62.12 ± 0.25	31.34 ± 0.91
FedProx	72.04 ± 0.12	63.53 ± 0.20	32.51 ± 0.88
FedCorr	72.33 ± 0.16	**72.40 ± 0.19**	40.92 ± 0.79
FedNoRo	71.78 ± 0.22	67.02 ± 0.24	38.12 ± 0.85
FedBeat	70.82 ± 0.28	68.01 ± 0.26	30.74 ± 0.92
FedELC	71.82 ± 0.18	70.16 ± 0.22	31.45 ± 0.93
FedClean[1]	69.86 ± 0.33	68.75 ± 0.21	63.11 ± 0.92
FedClean[2]	70.94 ± 0.25	71.20 ± 0.18	**66.54 ± 0.84**

Table 4. Accuracies (%) of various methods on Clothing1M with non-IID setting.

Methods	FedAvg	FedProx	FedCorr	FedNoRo	FedBeat	FedELC	FedFixer	FedClean[1]	FedClean[2]
Acc	68.63	69.15	69.02	69.21	67.05	69.24	70.52	70.17	**72.39**

THANKS