Federated Semi-Supervised Learning (FSSL)

Introduction

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

• Objective of FSSL

The objective of FSSL is to optimize a model under the non-independent and identically distributed (Non-IID) data sets.

• Challenges of FSSL

distributed data and privacy protection

• Broadly three lines of FSSL methods

1) there are only limited labeled data in the central server
2) each client has partially labeled data
3) few clients have fully labeled data and the training datasets in other clients are fully unlabeled

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Class Balanced Adaptive Pseudo Labeling for Federated Semi-Supervised Learning

Ming Li Qingli Li Yan Wang*
Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University
lm1640362161@gmail.com, qlli@cs.ecnu.edu.cn, ywang@cee.ecnu.edu.cn

CVPR 2023

Method——CBAFed

Step 1) Warm up stage: train fully supervised models on only labeled clients using residual weight connection in a normal federated learning manner.

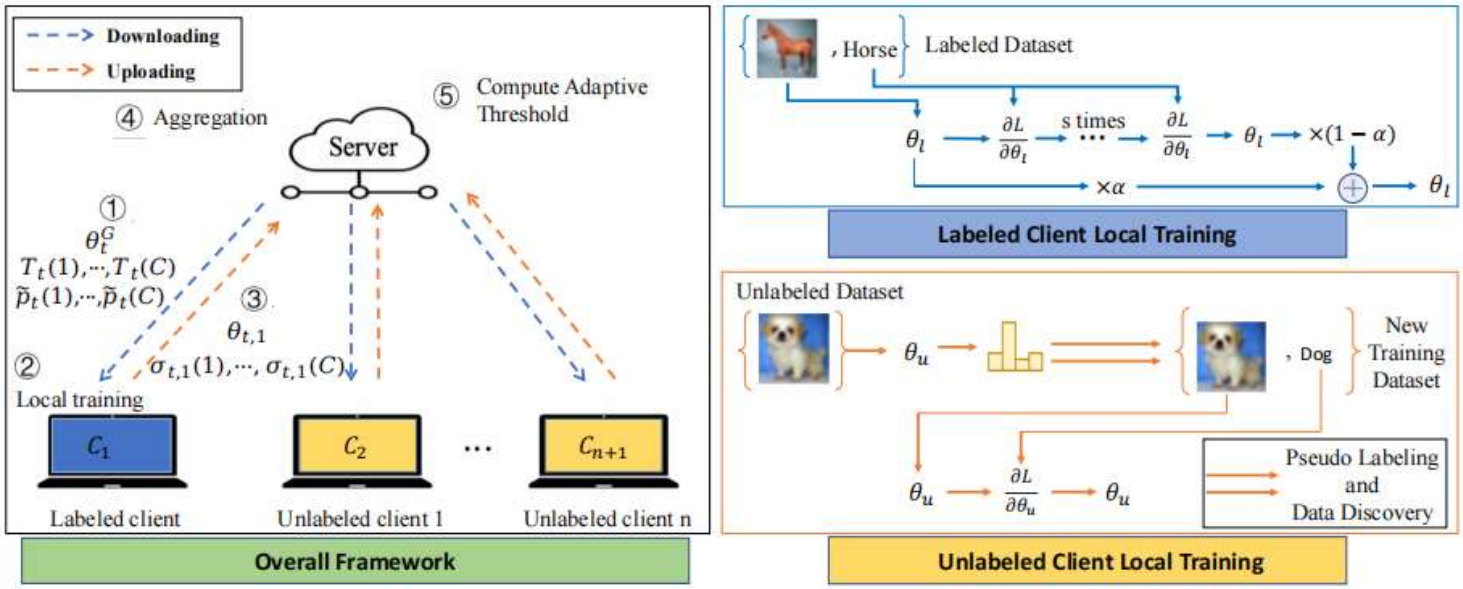
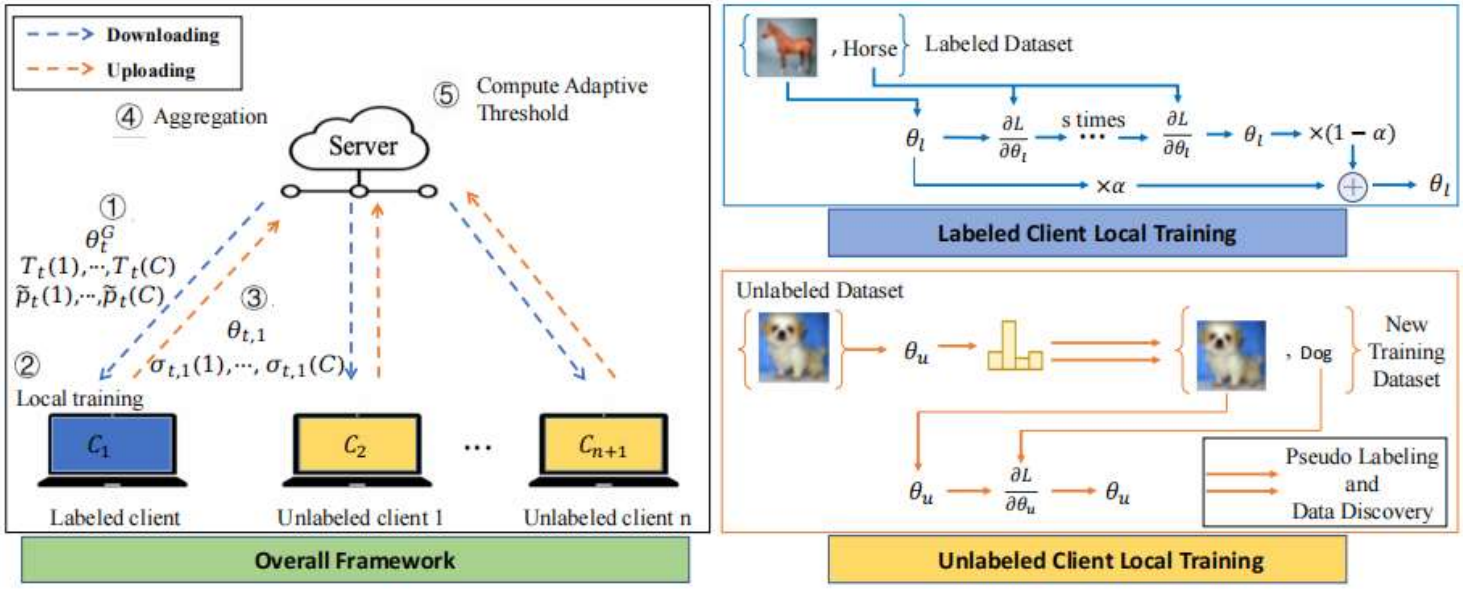


Figure 1. An overview of our CBAFed. In the central server (left side), the global model is aggregated with the returned local models (step ④) and the adaptive thresholds are calculated by the returned training data statistics (step ⑤). Then central server passes the global model, adaptive thresholds and class distribution to all local clients (step ①). After downloading these data, local clients perform local training on the right side (step ②). Labeled clients use labeled data to train the model with residual weight connection. Unlabeled clients obtain the new training dataset by adaptive pseudo labeling and tail class data discovery and use it to train the model. After local training, local clients return trained models and number of data in each class back to central server (step ③).

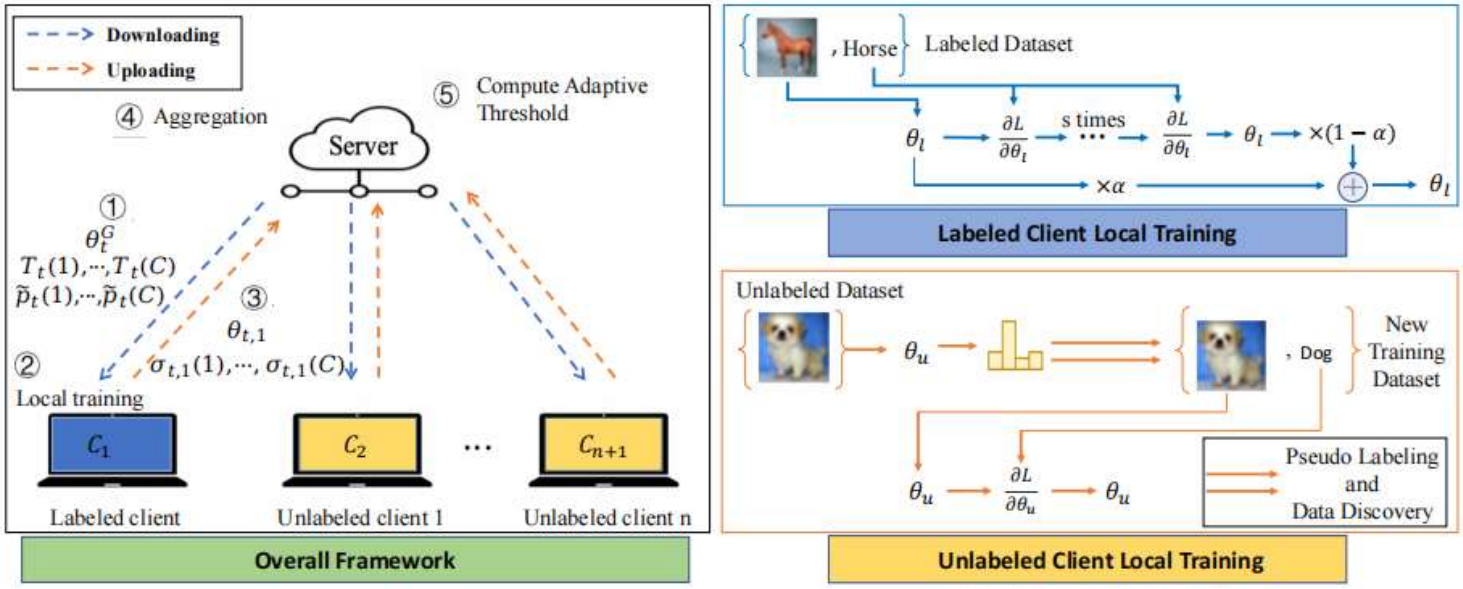
Method——CBAFed

Step 2) The central server computes the empirical class distribution and obtains the class balanced adaptive thresholds, then passes them to local clients.
Step 3) All local clients update local models, adaptive threshold and class distribution.
Labeled clients: train local models on all the data using proposed residual weight connection.
Unlabeled clients: acquire the fixed training set by the threshold and the tail class datasets, and train local models on the newly obtained training dataset.



Method——CBAFed

Step 4) The central server aggregates a new model with residual weight connection, computes the class distribution, and obtains the class balanced adaptive threshold. Then, the central server passes them to local clients.
Step 5) Repeat step (3)-(4) until the specified number of communication round is reached.



Method——CBAFed

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

• Residual Weight Connection

In ResNet, there is a skip connection between every layer.
There is a skip connection of model's parameters between training epochs (or communication rounds).

$$\theta^E = \begin{cases} \theta^E & E\%s \neq 0 \\ \alpha_1 \theta^{E-s} + (1-\alpha_1)\theta^E & E\%s = 0 \end{cases}$$

→ **Averaging model weights over training steps tends to produce a more accurate model than using the final weights directly.**

• Pseudo Labeling Methods

warm up:

$$\mathcal{L}_\ell = \frac{1}{|D_\ell|} \sum_{(X^\ell, y^\ell) \in D_\ell} H(y^\ell, p_m(y|\theta_t^\ell(X^\ell))),$$

$$\theta_{t+1}^G = \begin{cases} \sum_{\ell=1}^m \frac{|D_\ell|}{\sum_{i=1}^m |D_i|} \theta_t^\ell & t\%s \neq 0 \\ \alpha_1 \theta_{t+1-s}^G + (1-\alpha_1) \sum_{\ell=1}^m \frac{|D_\ell|}{\sum_{i=1}^m |D_i|} \theta_t^\ell & t\%s = 0 \end{cases} \quad (3)$$

fixed pseudo labeling:

$$\hat{y}_i^\mu = \arg\max p_m(y|\theta_t^\mu(X_i^\mu), i = 1, 2, \cdots, N_\mu.$$

$$\widetilde{\mathcal{D}}_\mu = \{(X_i^\mu, \hat{y}_i^\mu) \mid X_i^\mu \in D_\mu \wedge \max(p_m(y|\theta_t^\mu(X_i^\mu))) > \tau\}_{i=1}^{N_\mu}. \quad (5)$$

$$\mathcal{L}_\mu = \frac{1}{|\widetilde{\mathcal{D}}_\mu|} \sum_{(X^\mu, \hat{y}^\mu) \in \widetilde{\mathcal{D}}_\mu} H(\hat{y}^\mu, p_m(y|\theta_t^\mu(X^\mu))).$$

Method——CBAFed

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

• Class Balanced Adaptive Threshold for Pseudo Labeling（CBAPL）

Setting a fixed threshold usually makes the model fail to consider different learning status and learning difficulties of different classes.

• Curriculum Pseudo Labeling

$$\mathcal{T}_t(c) = \beta_t(c) \cdot \tau$$

Due to the Non-IID partition, the labeled data are not balanced, so purely using the number of selected unlabeled data to design threshold is improper.

⟶ Introduce many noisy labels into training

• CBAPL

$$\sigma_t^{\mu}(c) = \sum_{i=1}^{N_\mu} \mathbf{1}(\max(p_m(y|\theta_t^{\mu}(X_i^{\mu}))) > \mathcal{T}_t(c)) \mathbf{1}(\hat{y}_i^{\mu} = c)$$

$$\sigma_t^{\ell}(c) = \sum_{i=1}^{N_\ell} \mathbf{1}(y_i^{\ell} = c).$$

$$\sigma_t(c) = \sum_{\ell=1}^{m} \sigma_t^{\ell}(c) + \sum_{\mu=m+1}^{n+m} \sigma_t^{\mu}(c).$$

Method——CBAFed

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

• Class Balanced Adaptive Threshold for Pseudo Labeling（CBAPL）

empirical distribution

$$\widetilde{p}_t(c) = \frac{\sigma_t(c)}{\sum_{i=1}^{C} \sigma_t(i)}.$$

standard deviation

$$std(\widetilde{p}_t) = \sqrt{\frac{1}{C-1}\sum_{c=1}^{C}(\widetilde{p}_t(c) - \overline{p}_t)^2},$$

$$\overline{p}_t = \frac{1}{C}\sum_{c=1}^{C}\widetilde{p}_t(c).$$

threshold of class c

$$\tau_{t,c} = \widetilde{p}_t(c) + \tau - std(\widetilde{p}_t)$$

• upper bound of threshold

$$\mathcal{T}_{t+1}(c) = \begin{cases} \tau_{t,c}, & \tau_{t,c} < \tau_h \\ \tau_h, & \tau_{t,c} \geq \tau_h \end{cases}$$

fixed pseudo label training dataset

$$\widetilde{\mathcal{D}}_{t+1,\mu} = \{(X_i^\mu, \hat{y}_i^\mu) | X_i^\mu \in D_\mu \wedge \max(p_m(y|\theta_t^\mu(X_i^\mu))) > \mathcal{T}_{t+1}(\hat{y}_i^\mu)\}_{i=1}^{N_\mu}$$

• lower bound of threshold

Theorem 3.1.

$$\tau + \widetilde{p}_t(c) - \sqrt{\frac{1}{C}} \leq \mathcal{T}_t(c) \leq \tau + \widetilde{p}_t(c).$$

Since $\tau >> \sqrt{\frac{1}{C}}$, $\mathcal{T}_t(c)$ will have a high lower bound

modified $\widetilde{p}_t(c)$

$$\widetilde{p}_t(c) = \frac{\sigma_t(c)}{\sum_{i=1}^{C}\sigma_t(i)} \times \frac{C}{10}.$$

Method——CBAFed

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

• Discovery of Unlabeled Data from Tail Classes

For warm up stage in labeled clients, it is similar to long-tailed classification, so the problems in long-tailed classification will also exist in our pseudo labeling process : models tend to classify tail (rare) classes as head (common) classes

mask function

$$\mathcal{M}_i(p) = \begin{cases} p_i & i \neq \arg\max p \\ 0 & i = \arg\max p \end{cases}$$

analyze the second largest confidence score

$$\hat{y}_i^{u\prime} = \arg\max \mathcal{M}(p_m(y|\theta_t^\mu(X_i^\mu)))$$

misclassfied data

$$D_\mu^{tail} = \{(X_i^\mu, \hat{y}_i^{\mu\prime}) | X_i^\mu \in D_\mu$$
$$\wedge \max(p_m(y|\theta_t^\mu(X_i^\mu))) \leq \mathcal{T}_t(\hat{y}_i^\mu) \wedge \widetilde{p}_t(\hat{y}_i^{\mu\prime}) < \frac{\beta}{C}\}$$

$$D_\mu^{train} = D_\mu^{tail} \cup \widetilde{D}_\mu$$

$$\mathcal{L}_\mu = \frac{1}{|D_\mu^{train}|} \sum_{(X^\mu, \hat{y}^\mu) \in D_\mu^{train}} H(\hat{y}^\mu, p_m(y|\theta_t^\mu(X^\mu))).$$

$$\sigma_t^\mu(c) = \sum_{(X^\mu, \hat{y}^\mu) \in D_\mu^{train}} \mathbf{1}(\hat{y}^\mu = c).$$

Method——CBAFed

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- **Aggregation of local models**

$$w_t^i = \begin{cases} \frac{|D_i|}{|D_t^{train}|} & \text{if } i \in \{1, \cdots, m\} \\ \frac{|D_{t,i}^{train}|}{|D_t^{train}|} & \text{if } i \in \{m+1, \cdots, m+n\} \end{cases}$$

$$|D_t^{train}| = \sum_{\ell=1}^{m} |D_\ell| + \sum_{\mu=m+1}^{m+n} |D_{t,\mu}^{train}|$$

$$\theta_{t+1}^G = \begin{cases} \sum_{i=1}^{m+n} w_t^i \theta_t^i & t\%s \neq 0 \\ \alpha_2 \theta_{t+1-s}^G + (1-\alpha_2) \sum_{i=1}^{m+n} w_t^i \theta_t^i & t\%s = 0, \end{cases}$$

Experiments

Table 1. Results on SVHN, CIFAR-10/100, Fashion MNIST and ISIC 2018 datasets under heterogeneous data partition with ResNet18. FedAVG$^+$ means FedAvg [19] trained with all one labeled clients using our residual weight connection. Fed-consist$^+$ means Fed-Consist [31] using our proposed fixed pseudo labeling without enlarging the weight of labeled client.

Labeling Strategy	Method	Client Num.		Dataset				
		labeled	unlabeled	SVHN	CIFAR10	CIFAR100	Fashion-MNIST	ISIC 2018
Fully supervised	FedAvg [19](upper-bound)	10	0	91.83	80.89	51.38	90.14	81.32
	FedAvg [19](lower-bound)	1	0	67.71	54.66	20.49	74.87	65.13
	FedAvg$^+$ [19]	1	0	76.98	58.21	24.84	78.26	66.69
Semi supervised	FedIRM [18]	1	9	69.22	52.84	20.20	76.83	64.85
	Fed-Consist [31]	1	9	70.56	54.23	21.81	76.57	65.20
	Fed-Consist$^+$ [31]	1	9	86.57	56.35	23.25	78.35	65.50
	RSCFed [14]	1	9	76.74	57.07	28.46	78.40	67.21
	CBAFed(ours)	1	9	**88.07**	**67.08**	**30.18**	**85.49**	**68.29**

local training epoch： 11 (labeled client) / 1 (unlabeled client)

Experiments

Table 2. Comparison of our method against RSCFed [14], Fed-Consist [31] and FedAVG [19] in SVHN dataset on ViT [5] as the backbone, with one labeled and nine unlabeled clients.

Method	Client Num.		Accuracy
	labeled	unlabeled	
FedAVG [19](upper bound)	10	0	96.81
FedAVG [19](lower bound)	1	0	81.68
FedAVG$^+$ [19]	1	0	88.93
FedIRM [18]	1	9	79.44
Fed-Consist [31]	1	9	85.91
Fed-Consist$^+$ [31]	1	9	93.21
RSCFed [14]	1	9	89.43
CBAFed(ours)	1	9	**95.09**

Table 3. Comparison of our method against RSCFed [14], Fed-Consist [31], FedIRM [18] and FedAVG [19] with the number of labeled and unlabeled client set to 2 and 8.

Method	Client Num.		Accuracy
	labeled	unlabeled	
FedAVG [19](upper bound)	10	0	80.89
FedAVG [19](lower bound)	2	0	61.85
FedAVG$^+$ [19]	2	0	66.55
FedIRM [18]	2	8	62.62
Fed-Consist [31]	2	8	61.67
Fed-Consist$^+$ [31]	2	8	68.04
RSCFed [14]	2	8	64.25
CBAFed(ours)	2	8	**72.01**

Experiments

Table 4. Ablation Study of CBAFed in CIFAR-10/100 and Fashion MNIST Datasets. Fixed PL: fixed pseudo labeling, CBA: class balanced adaptive pseudo labeling, DD: tail class data discovery.

Dataset	Fixed PL	CBA	DD	Res-Weight	Accuracy
CIFAR-10	✓				59.16
	✓	✓			64.29
	✓	✓	✓		65.15
	✓	✓	✓	✓	67.08
CIFAR-100	✓				27.64
	✓	✓			29.41
	✓	✓	✓		29.86
	✓	✓	✓	✓	30.18
Fashion-MNIST	✓				79.99
	✓	✓			80.87
	✓	✓	✓		84.37
	✓	✓	✓	✓	85.49

FedCD: Federated Semi-Supervised Learning with Class Awareness Balance via Dual Teachers

Yuzhi Liu[1], **Huisi Wu**[1*], **Jing Qin**[2]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

AAAI 2024

Method——FedCD

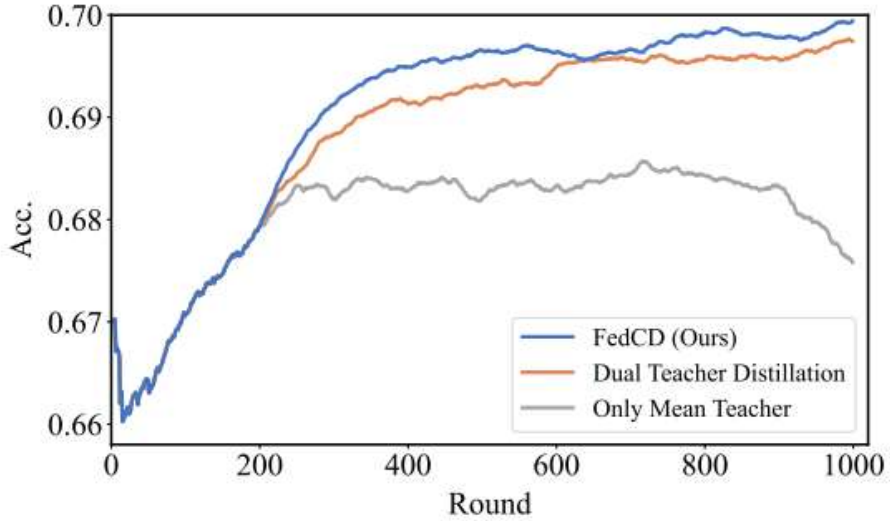


Figure 1: Comparisons of the test accuracy curves showed that our proposed FedCD method with dual teacher distillation outperformed the variant with only mean teacher distillation. The performance of the model relying solely on the mean teacher declined due to the inherent limitations of local knowledge. However, after incorporating the dual teacher distillation and class awareness balance modules, the issue of localized knowledge limitation was substantially mitigated, resulting in remarkable performance improvements.

Method——FedCD

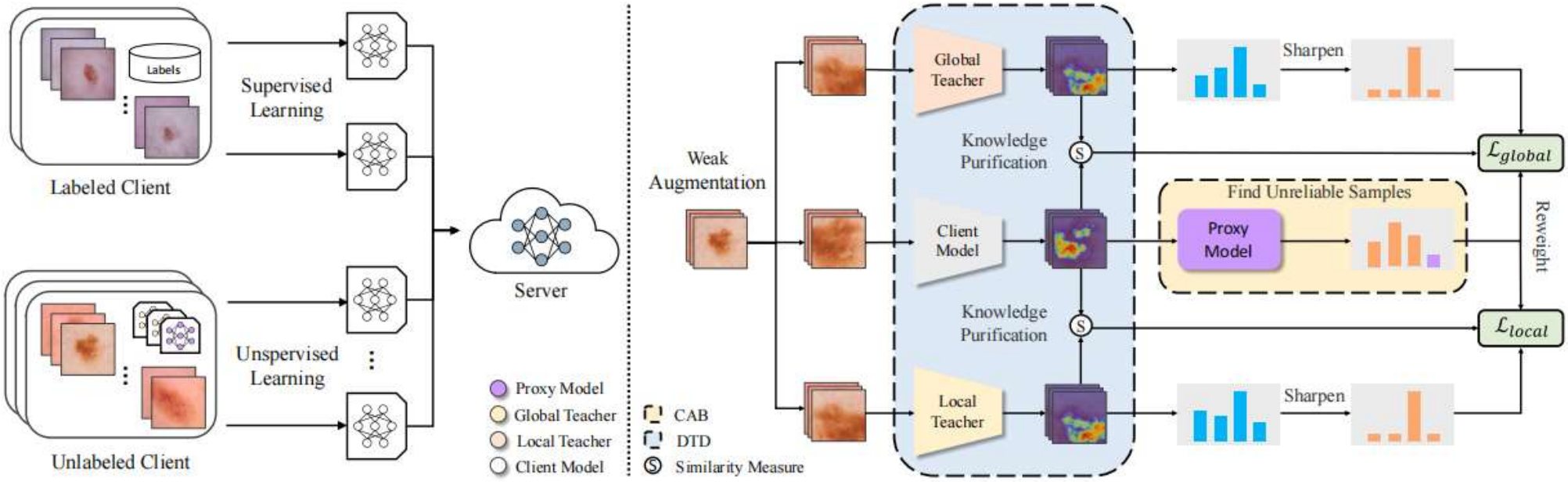


Figure 3: An overview of our proposed FedCD framework. The left side shows our overall architecture, where we introduce a proxy model and global-local teacher framework to assist unlabeled clients. The right side depicts the training process within unlabeled clients. We propose the dual teacher distillation and class awareness balance module for effective balanced learning.

Method——FedCD

- **Problem Setting**

$$\arg\min_{\theta_s} \mathcal{L}(\theta_s) = \sum_{i=1}^{m} \frac{|\mathcal{S}_i^l|}{|\mathcal{S}|} \mathcal{L}_{ce}(\theta_c) + \sum_{i=1}^{n} \frac{|\mathcal{S}_i^u|}{|\mathcal{S}|} \mathcal{L}_u(\theta_c)$$

- **Dual Teacher Distillation**

$$\mathcal{L}_{mse-local} = \left\| \hat{P}_l - P_c \right\|$$

$$\mathcal{L}_{mse-global} = \left\| \hat{P}_g - P_c \right\|$$

$$\hat{P} = P_i^{\frac{1}{\tau}} / \sum_j P_j^{\frac{1}{\tau}}$$

$$V_{local} = \mathrm{KL}(F_c \,||\, F_l)$$

$$V_{global} = \mathrm{KL}(F_c \,||\, F_g)$$

$\longrightarrow$

$$\mathcal{L}_{local} = \mathrm{e}^{-V_{local}} * \mathcal{L}_{mse-local}$$

$$\mathcal{L}_{global} = \mathrm{e}^{-V_{global}} * \mathcal{L}_{mse-global}$$

$$\mathcal{L}_u = \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{local}$$

Method——FedCD

- **Class Awareness Balance**

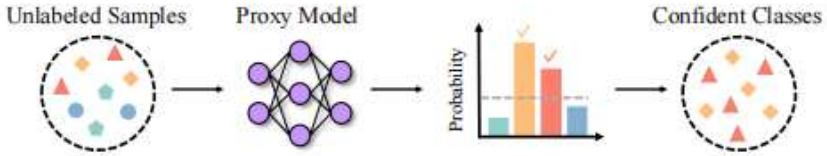


Figure 4: The pipeline of exploiting confident classes. We leverage global information to mine the confidence classes of each unlabeled client.

Exploit Confdent Classes

$$s_k = \sum_{i}^{n_u} \mathrm{G}_k\left(\theta_p\left(x_i\right)\right)$$

$$s_k = \frac{s_k - \min(s)}{\max(s) - \min(s)} \qquad s = [s_0, s_1, \ldots, s_{k-1}] \qquad s_k > \beta$$

Identify Unreliable Samples

$$D_u = \{x \mid t_l < T(k) < t_h\}$$

$T(\cdot)$ is the order operation argsort (P_l)

Recalibrate Loss Function

$$\mathcal{L}_u = \alpha w * (\lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{local})$$

where

$$w = \begin{cases} \frac{1}{N_w}, & \text{Unreliable} \\ \frac{1}{N_u - N_w}, & \text{others} \end{cases}$$

$$\alpha_t = \alpha_0 + (\alpha_n - \alpha_0)\frac{t}{\text{Rounds}}$$

Experiments

Labeling Strategy	Method	Client Num.		Metrics			
		labeled	unlabeled	Acc. (%)	AUC (%)	Precision (%)	Recall (%)
		Task 1: Skin Lesion Diagnosis					
Fully supervised	FedAvg (upper-bound)	10	0	80.42	93.47	71.57	54.39
	FedAvg(lower-bound)	1	0	68.07	79.02	34.86	31.37
Semi supervised	Fed-Consist	1	9	67.84	81.25	37.49	29.08
	FedIRM	1	9	68.39	81.6	37.49	31.81
	RSCFed	1	9	69.09	82.59	37.94	32.59
	CBAFed	1	9	69.79	83.06	37.99	32.75
	Ours	1	9	**70.99**	**83.64**	**42.22**	**35.63**
		Task 2: Intracranial Hemorrhage Diagnosis					
Fully supervised	FedAvg(upper-bound)	10	0	72.03	88.19	62.85	59.86
	FedAvg(lower-bound)	1	0	59.27	77.45	46.49	42.27
Semi supervised	Fed-Consist	1	9	58.96	75.86	46.07	42.04
	FedIRM	1	9	58.98	74.79	45.37	42.88
	RSCFed	1	9	59.32	77.51	47.53	43.04
	CBAFed	1	9	59.34	78.17	47.56	43.01
	Ours	1	9	**63.10**	**79.55**	**47.77**	**46.93**

Table 1: Resluts on the HAM10000 and RSNA ICH datasets under heterogeneous data partition. We employ four commonly used metrics for method comparison, including Accuracy(Acc.), Area under the ROC Curve (AUC), Precision, and Recall. The best results are in bold. It reports that our method achieves the best performance among all methods.

Experiments

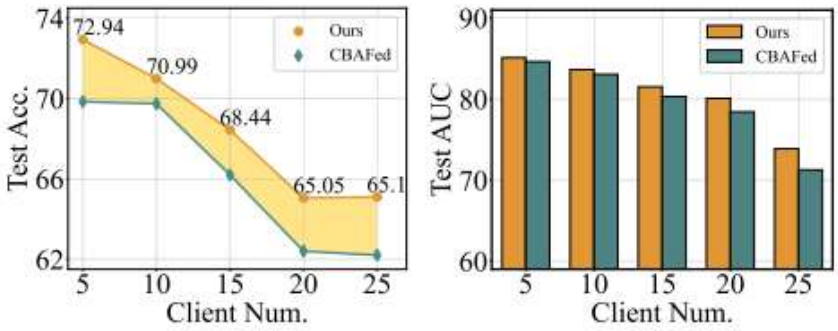


Figure 5: The accuracy score and AUC vary with the change in the number of unlabeled clients. Noted that the number of labeled clients remains at 1.

	CAB	DTD	Acc.(%)	AUC(%)
Basic	×	×	68.07	79.02
Basic+CAB	✓	×	70.29	82.41
Basic+DTD	×	✓	70.44	83.53
Ours	✓	✓	**70.99**	**83.64**

Table 4: Ablation studies on the effectiveness of dual teacher distillation and class awareness balance.

Local Teacher	Global Teacher	Knowledge Purification	Metrics	
			Acc.(%)	AUC(%)
✓	×	×	68.07	79.02
✓	×	✓	70.24	82.63
✓	✓	×	69.94	82.79
✓	✓	✓	**70.44**	**83.53**

Table 5: Ablation studies on the effectiveness of dual teachers and knowledge purification.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Exploring One-Shot Semi-supervised Federated Learning with Pre-trained Diffusion Models

Mingzhao Yang *, **Shangchao Su** *, **Bin Li** †, **Xiangyang Xue** †

Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University
{mzyang20,scsu20,libin,xyxue}@fudan.edu.cn

AAAI 2024

Method——FedDISC

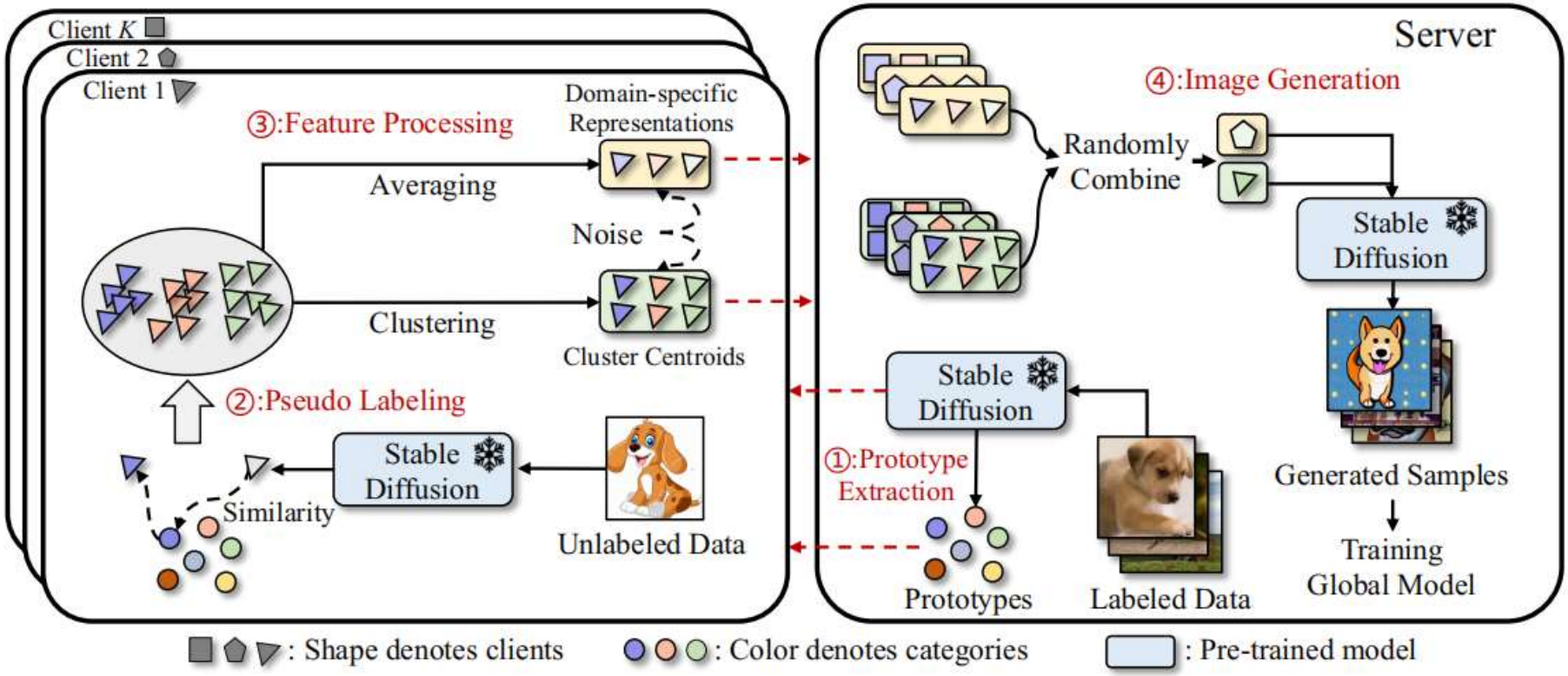


Figure 1: The framework of FedDISC. The overall method consists of four steps: Prototype Extraction, Pseudo Labeling, Feature Processing, and Image Generation.

Method——FedDISC

- **Prototype Extraction**

$$\mathbf{p}_j = \frac{\sum_{(\mathbf{x}_i^s, y_i) \in \mathcal{D}_s} E_\theta(\mathbf{x}_i^s) * \mathbb{I}(y_i = j)}{\sum_{(\mathbf{x}_i^s, y_i) \in \mathcal{D}_s} \mathbb{I}(y_i = j)}$$

noise-adding process

$$\bar{\mathbf{z}}_{j,l}^k = \sqrt{\alpha_n}\mathbf{z}_{j,l}^k + \sqrt{1-\alpha_n}\varepsilon_1, \bar{\mathbf{g}}_j^k = \sqrt{\alpha_n}\mathbf{g}_j^k + \sqrt{1-\alpha_n}\varepsilon_2$$

$$\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, \mathcal{I})$$

- **Pseudo Labeling**

$$sim(E_\theta(\mathbf{x}_i^k), \mathbf{p}_j) = \frac{E_\theta(\mathbf{x}_i^k)^\top \mathbf{p}_j}{\|E_\theta(\mathbf{x}_i^k)\| \, \|\mathbf{p}_j\|}, \mathbf{x}_i^k \in \mathcal{D}_k$$

$$\hat{y}_i^k = \arg\max_j sim(E_\theta(\mathbf{x}_i^k), \mathbf{p}_j)$$

- **Feature Processing** select L cluster centroids $\{\mathbf{z}_{j,l}^k\}_{l=1}^L$ obtain the domain-specific representations $\{\mathbf{g}_j^k\}_{j=1}^M$

$$\arg\min_{\mathbf{z}_{j,l}^k} \sum_{l=1}^{L} \sum_{\mathbf{x}_i^k \in \mathcal{D}_k} \left\| E_\theta(\mathbf{x}_i^k) - \mathbf{z}_{j,l}^k \right\|^2 * \mathbb{I}(\hat{y}_i^k = j)$$

Method——FedDISC

- **Image Generation**

$$\mathbf{s}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{s}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{s}_t, t|p, q)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(\mathbf{s}_t, t|p, q) + \sigma_t\varepsilon_t$$

conditional probability distribution of the sample **s**

$$p(\mathbf{s}|\bar{\mathbf{z}}_{j,l}^k, \bar{\mathbf{g}}_j^{k_i}, \mathcal{C}_j) \propto p(\mathbf{s}|\mathcal{C}_j)\frac{p(\mathbf{s}|\bar{\mathbf{z}}_{j,l}^k, \mathcal{C}_j)}{p(\mathbf{s}|\mathcal{C}_j)}\frac{p(\mathbf{s}|\bar{\mathbf{g}}_j^{k_i}, \mathcal{C}_j)}{p(\mathbf{s}|\mathcal{C}_j)}$$

final predicted noise

$$\hat{\epsilon}_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}_{j,l}^k, \bar{\mathbf{g}}_c^{k_i}, \mathcal{C}_j) = \epsilon_\theta(\mathbf{s}_t, t|\mathcal{C}_j) + w_f(\epsilon_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}_{j,l}^k, \mathcal{C}_j) - \epsilon_\theta(\mathbf{s}_t, t|\mathcal{C}_j)) + w_g(\epsilon_\theta(\mathbf{s}_t, t|\bar{\mathbf{g}}_j^{k_i}, \mathcal{C}_j) - \epsilon_\theta(\mathbf{s}_t, t|\mathcal{C}_j))$$

fine-tune a classifcation model $h = F_\theta \circ E_\theta$

$$\mathbf{s}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{s}_t - \sqrt{1-\alpha_t}\hat{\epsilon}_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}_{j,l}^k, \bar{\mathbf{g}}_j^{k_i}, \mathcal{C}_j)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\hat{\epsilon}_\theta(\mathbf{s}_t, t|\bar{\mathbf{z}}_{j,l}^k, \bar{\mathbf{g}}_j^{k_i}, \mathcal{C}_j) + \sigma_t\varepsilon_t$$

Experiments

	OpenImage						DomainNet					
	client0	client1	client2	client3	client4	average	clipart	infograph	painting	quickdraw	sketch	average
Ceiling	*54.05*	*58.42*	*62.59*	*63.21*	*64.79*	*60.61*	*81.54*	*52.49*	*73.54*	*30.11*	*72.34*	*62.01*
Fine-tune	36.67	46.81	45.43	47.17	42.1	43.64	67.57	45.47	65.28	10.42	62.14	50.17
Zero-shot	56.03	40.61	40.28	44.06	**61.45**	48.47	65.86	40.5	62.25	13.36	57.92	47.98
Prompt	48.61	54.03	59.07	58.42	53.49	54.72	66.42	37.45	59.62	10.73	63.92	47.63
FedAvg	41.11	44.06	46.57	47.45	37.63	43.36	49.95	30.67	51.07	1.74	38.46	34.38
SemiFL	48.15	52.78	61.05	55.23	46.16	52.67	69.55	**47.16**	64.54	7.02	63.32	50.32
RSCFed	28.97	38.04	40.82	33.98	36.35	35.63	71.5	45.73	61.96	11.53	65.03	51.15
FedDISC	**56.11**	**62.49**	**62.53**	**59.16**	56.77	**59.42**	**72.54**	43.47	**67.42**	**17.71**	**67.25**	**53.68**
	NICO++_C						NICO++_U					
	client0	client1	client2	client3	client4	average	client0	client1	client2	client3	client4	average
Ceiling	*89.19*	*91.9*	*89.51*	*90.47*	*85.1*	*89.23*	*96.35*	*96.42*	*96.88*	*97.01*	97.26	*96.78*
Fine-tune	86.5	89.39	83.61	87.21	76.95	84.73	84.75	79.08	81.48	86.58	83.52	83.08
Zero-shot	78.66	85.26	80.01	80.7	72.14	79.35	89.2	89.24	87.19	85.5	88.6	87.94
Prompt	86.94	87.41	**89.73**	82.69	73.51	84.05	90.61	87.14	**89.96**	87.48	88.16	88.67
FedAvg	86.98	90.82	82.68	87.57	74.48	84.51	83.26	73.3	77.93	80.8	79.28	78.91
SemiFL	87.55	89.27	81.93	87.16	77.01	84.58	78.21	74.7	79.87	80.69	77.02	78.09
RSCFed	52.08	60.15	52.6	55.35	43.89	52.81	71.88	64.14	70.82	69.71	69.67	69.24
FedDISC	**87.97**	**92.09**	86.44	**90.52**	**84.17**	**88.24**	**91.73**	**90.82**	89.63	**92.83**	**90.15**	**91.03**

Table 1: The performances of different methods on OpenImage, DomainNet, and NICO++, where the italicized texts represent the inaccessible supervised ceiling performance used solely as a reference, and bold texts represent the best performance excluding the supervised ceiling performance.

Experiments

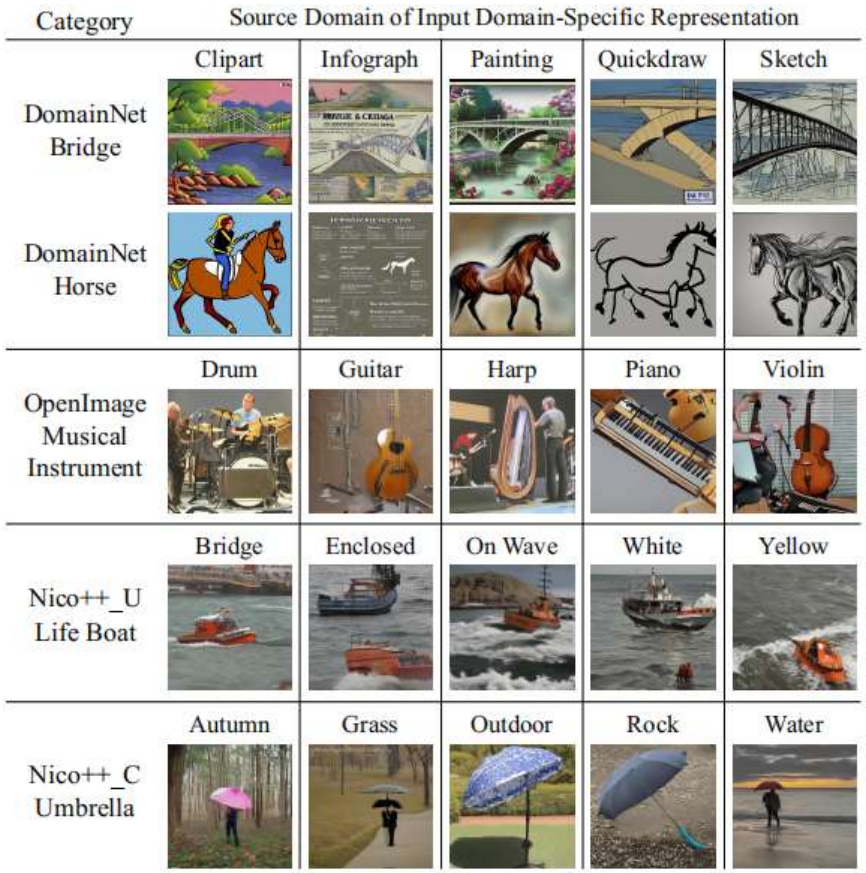


Figure 2: Generated images comply with different distributions on different datasets.

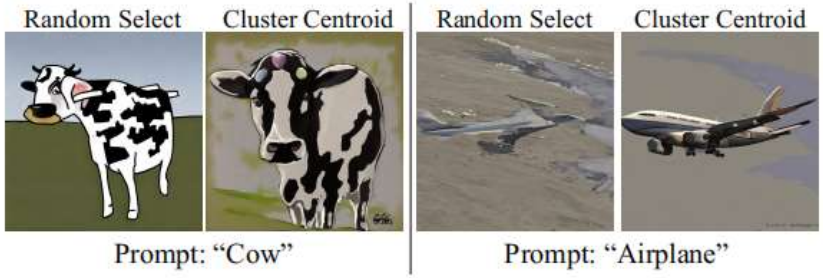


Figure 3: Comparison between generating using clustering centroids and the randomly selected client representations. With the provision of clustering centroids, the introduction of more representative semantic information leads to a significant improvement in the stability of the generated outputs.

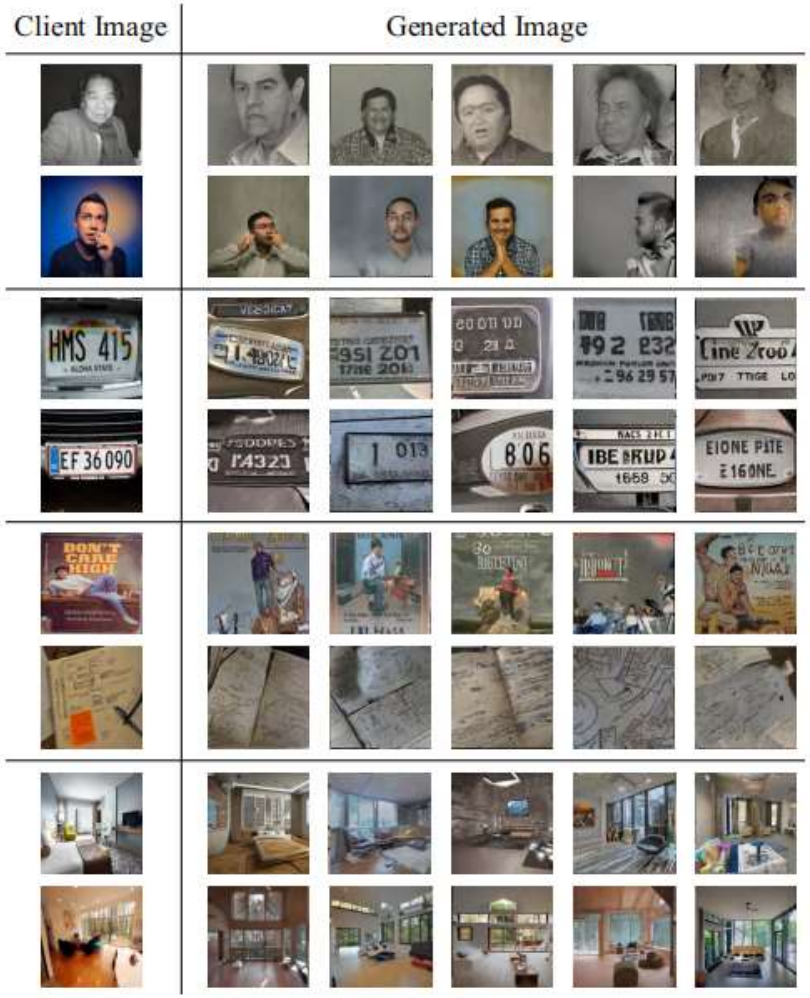
Experiments

	client0	client1	client2	client3	client4
$L = 3$ Fine-tune	36.01	46.01	44.59	45.55	41.87
$L = 3$ FedDISC	**56.33**	**61.93**	**58.62**	**56.71**	**58.74**
$L = 5$ Fine-tune	36.67	46.81	45.43	47.17	42.10
$L = 5$ FedDISC	**56.11**	**62.49**	**62.53**	**59.16**	**56.77**
$L = 10$ Fine-tune	37.55	45.86	44.85	46.01	42.15
$L = 10$ FedDISC	**57.16**	**63.84**	**61.12**	**57.91**	**59.13**

Table 2: The influence of the number of cluster centroids.

DR	CC	client0	client1	client2	client3	client4
		66.42	37.45	59.62	10.73	63.92
✓		67.79	40.02	63.59	13.77	60.57
	✓	65.83	38.27	64.56	14.30	60.37
✓	✓	**72.54**	**43.47**	**67.42**	**17.71**	**67.25**

Table 4: The influence of different conditions.



南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Towards Unbiased Training in Federated Open-world Semi-supervised Learning

Jie Zhang [1] **Xiaosong Ma** [1] **Song Guo** [1] **Wenchao Xu** [1]

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

ICML 2023

Method——FedoSSL

- **Problem Definition**

FedoSSL: $\mathcal{C}^l \neq \mathcal{C}^u$

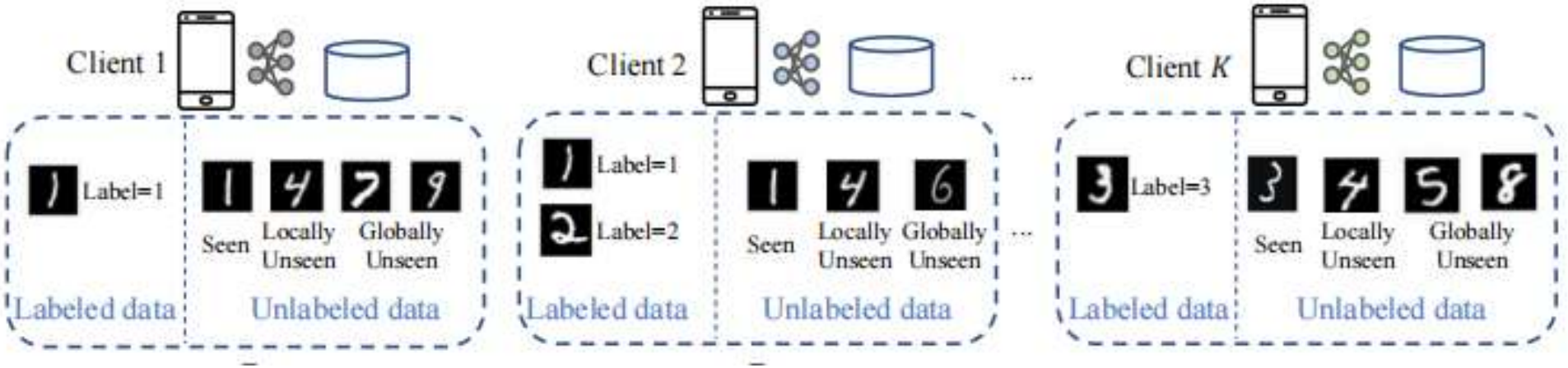
$$\mathcal{C}_{seen} = \mathcal{C}^l \cap \mathcal{C}^u$$

$$\mathcal{C}_{unseen} = \mathcal{C}^u \setminus \mathcal{C}_{seen}$$

- **Inconsistent data distribution on different clients raises another new problem: some unseen classes may exist in more than one client, resulting in biased training among different unseen classes**

- **More fine-grained definition on unseen classes:**

Definition 1 (locally unseen & globally unseen class). In FedoSSL, the unseen classes $\mathcal{C}_{i,unseen}$ on client i can be divided into two types: locally unseen classes $\mathcal{C}_{i,lu}$, in which $\mathcal{C}_{i,lu} = \mathcal{C}_{1,unseen} \cap \cdots \cap \mathcal{C}_{K,unseen}$; and globally unseen classes $\mathcal{C}_{i,gu}$, in which $\mathcal{C}_{i,gu} = \mathcal{C}_{i,unseen} \setminus \mathcal{C}_{i,lu}$.



Method——FedoSSL

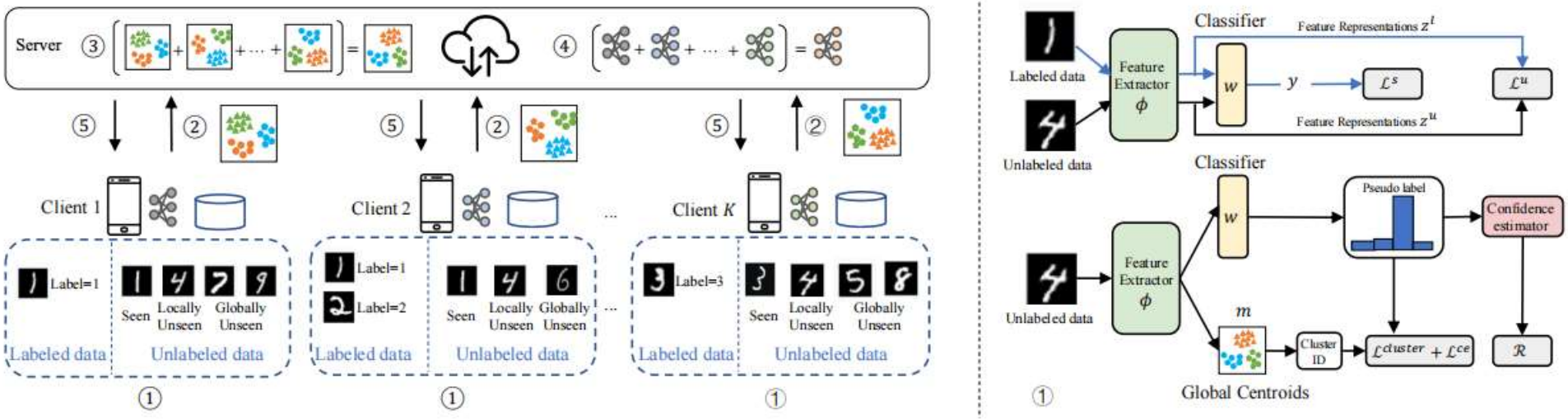
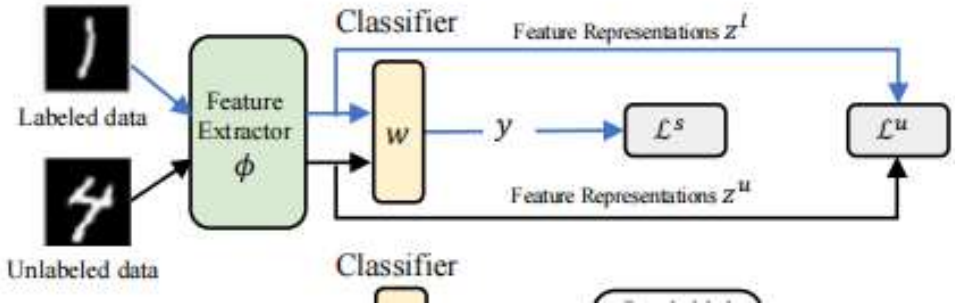


Figure 1: Framework of the proposed FedoSSL algorithm. **Pipeline**: ① **Local Training:** Each client first performs local training on its private dataset for several epochs (i.e., via optimizing loss function in Eq. (1)), and then computes local centroids via a Sinkhorn-Knopp based clustering algorithm (Genevay et al., 2019). ② Upload model parameters and local centroids to the server. ③ The server performs standard model aggregation. ④ The server performs centroids aggregation by again using Sinkhorn-Knopp clustering to obtain global centroids. ⑤ The global model and global centroids are returned to the clients, who use them for local training.

Method——FedoSSL



Two typical forms of unsupervisied loss $\mathcal{L}_i^u$:
1) pseudo-labels
2) consistency regularization
fail to classify seen classes and unseen classes

$$\min_\theta \mathcal{L}(\theta) := \sum_{i=1}^{K} \frac{n_i}{n} \mathcal{L}_i(\theta)$$

$$n_i = n_i^l + n_i^u$$

$$\mathcal{L}_i = \mathcal{L}_i^s + \alpha \mathcal{L}_i^u$$

$$\mathcal{L}_i^s = \frac{1}{n_i^l} \sum_{(x_j, y_j) \in \mathcal{D}_i^l} \mathcal{H}(y_j, p(x_j; \theta))$$

Based on **ORCA and NACH**, use **pairwise objective** as unsupervised loss on unlabeled data:

$$\mathcal{L}_i^u = -\frac{1}{n_i^l + n_i^u} \sum_{\substack{z_j, \bar{z}_j \in \\ (Z_i^l \cup Z_i^u, \bar{Z}_i^l \cup \bar{Z}_i^u)}} \mathcal{H}(p(w^\top \cdot z_j), p(w^\top \cdot \bar{z}_j))$$

Method——FedoSSL

two main challenges:

a) locally unseen classes may be learned faster than globally unseen classes
existing unsupervised pairwise loss treats each class equally → a big bias on pseudo-label generation

b) both labeled data and unlabeled data are required to feed into the same model classifier → generated cluster/class id heterogeneous among different clients

$$\mathcal{L}_i^* = \mathcal{L}_i + \beta \mathcal{R}_i + \gamma \mathcal{L}_i^{cal}$$

The overall objective consists of three parts:
1) fundamental semi-supervised loss for all data;
2) an uncertainty-aware regularization loss to reduce the training gap among locally unseen and globally unseen classes;
3) a calibration loss to achieve efficient model aggregation

Method——FedoSSL

UNCERTAINTY-AWARE LOSS

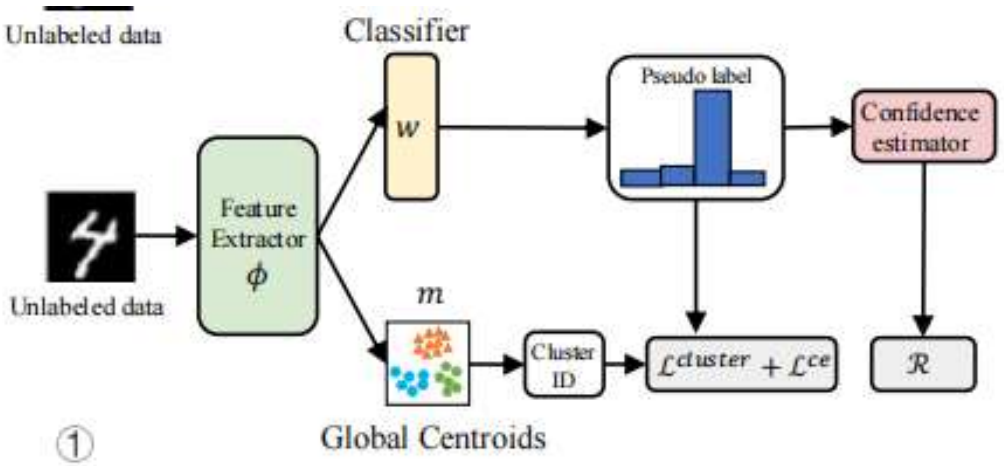
$$\mathcal{R}_i = \frac{1}{n_i^u} \sum_{x_j^u \in \mathcal{D}_i^u} |\pi(x_j^u)|$$

$$\pi(x_j^u) = \rho(n^c | \arg\max_c p(x_j^u; \theta))[1 - \max_c p(x_j^u; \theta)]$$

$$\rho(n^c) = -\tau^{1-\frac{n^c}{n_{\max}}}$$

CALIBRATION MODULE

$$\mathcal{L}_i^{cal} = \mathcal{L}_i^{ce} + \mathcal{L}_i^{cluster}$$



$$\mathcal{L}_i^{ce} = \frac{1}{n_i^u} \sum_{z_j \in Z_i^u} \mathcal{H}(q(z_j; m), p(w^\top \cdot z_j))$$

corresponding cluster assignments computed by matching representations with global centroids

$$\mathcal{L}_i^{cluster} = \frac{1}{n_i^u} \sum_{z_j, \bar{z}_j \in Z_i^u, \bar{Z}_i^u} \mathcal{H}(q(z_j; m), q(\bar{z}_j; m))$$

Experiments

Table 2: Classification accuracy of compared methods on seen, unseen and all classes with 10 clients over three benchmark datasets. Asterisk (∗) in *SemiFL denotes that the original methods cannot classify unseen classes (and we had to extend it). On unseen classes, *LU.* denotes locally unseen classes, while *GU.* denotes globally unseen classes. *AU.* represents the overall accuracy of all unseen classes. Gray rows indicate the upper bound of the model performance of FedoSSL.

#Method	CIFAR-10 (%)					CIFAR-100 (%)					CINIC-10 (%)				
	All	Seen	Unseen			All	Seen	Unseen			All	Seen	Unseen		
			LU.	GU.	AU.			LU.	GU.	AU.			LU.	GU.	AU.
Cen-O	78.26	86.63	-	-	71.95	56.92	73.68	-	-	44.28	69.32	83.18	-	-	58.86
Cen-N	81.02	89.47	-	-	74.64	58.98	75.10	-	-	46.82	71.89	83.82	-	-	62.89
Local-O	65.98	79.57	-	-	45.60	43.10	54.33	-	-	26.25	55.33	65.23	-	-	40.48
Local-N	67.67	83.95	-	-	43.26	45.28	57.24	-	-	27.34	57.31	65.70	-	-	44.73
Fed-AO	69.46	81.01	89.38	42.03	52.15	47.91	59.67	38.07	29.12	30.26	54.85	63.22	71.31	37.88	42.29
Fed-RO	71.72	82.22	89.84	53.43	55.96	47.72	59.79	44.13	28.86	29.62	57.16	62.26	72.24	42.09	49.50
Fed-AN	66.58	84.18	78.76	37.58	40.15	47.25	58.24	42.11	30.44	30.77	53.49	63.61	66.78	36.06	38.32
Fed-RN	68.83	**85.52**	79.84	41.79	43.81	48.02	59.4	**48.77**	30.36	30.96	58.11	65.97	68.81	39.01	46.33
*SemiFL	64.91	81.57	86.33	31.16	39.92	42.28	54.94	31.68	21.46	23.29	52.27	62.72	64.53	37.21	37.34
FedoSSL	**76.26**	84.29	**90.68**	**59.69**	**64.22**	**51.58**	**61.12**	45.76	**33.82**	**31.13**	**63.82**	**68.40**	**79.79**	**47.78**	**56.96**

O:ORCA
N:NACH
Fed-A:FedAvg
Fed-R:FedRep

Experiments

Table 3: Analysis of Loss function: classification accuracy on CIFAR-10 (the number of clients: 10).

METHOD	SEEN	UNSEEN	ALL
FED-AO	81.01	52.15	69.46
FEDOSSL-$\mathcal{R}_i$-$\mathcal{L}_i^{ce}$	83.53	52.24	71.01
FEDOSSL-$\mathcal{R}_i$	83.13	62.98	75.07
FEDOSSL	84.29	64.22	76.26

Table 4: Analysis of Loss function: classification accuracy on CINIC-10 (the number of clients: 10).

METHOD	SEEN	UNSEEN	ALL
FED-AO	63.22	42.29	54.85
FEDOSSL-$\mathcal{R}_i$-$\mathcal{L}_i^{ce}$	69.10	40.31	57.58
FEDOSSL-$\mathcal{R}_i$	67.59	47.73	59.65
FEDOSSL	68.40	56.69	63.82

Table 5: Classification accuracy of compared methods on seen, unseen and all classes with 50 clients over three benchmark datasets.

	CIFAR-10 (%)			CIFAR-100 (%)			CINIC-10 (%)		
#Method	All	Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen
Fed-AO	70.22	83.34	50.54	45.63	56.25	29.69	53.81	60.49	43.80
Fed-RO	71.36	84.31	51.93	45.18	56.78	27.79	57.26	61.70	50.61
Fed-AN	69.89	85.36	46.68	45.22	56.30	28.59	53.42	63.62	38.13
Fed-RN	71.49	**86.28**	49.30	45.57	56.79	28.73	57.81	65.29	46.60
FedoSSL	**76.41**	85.71	**62.46**	**47.01**	**58.34**	**30.17**	**64.02**	**69.56**	**55.71**

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Thanks