International Journal of Computer Vision (2025) 133:3312–3326
https://doi.org/10.1007/s11263-024-02330-1

Relation-Guided Versatile Regularization for Federated Semi-Supervised Learning

Qiushi Yang[1] · Zhen Chen[2] · Zhe Peng[3] · Yixuan Yuan[4]

Received: 30 May 2024 / Accepted: 10 December 2024 / Published online: 5 January 2025

Abstract

Federated semi-supervised learning (FSSL) target to address the increasing privacy concerns for the practical scenarios, where data holders are limited in labeling capability. Latest FSSL approaches leverage the prediction consistency between the local model and global model to exploit knowledge from partially labeled or completely unlabeled clients. However, they merely utilize data-level augmentation for prediction consistency and simply aggregate model parameters through the weighted average at the server, which leads to biased classifiers and suffers from skewed unlabeled clients. To remedy these issues, we present a novel FSSL framework, Relation-guided Versatile Regularization (FedRVR), consisting of versatile regularization at clients and relation-guided directional aggregation strategy at the server. In versatile regularization, we propose the model-guided regularization together with the data-guided one, and encourage the prediction of the local model invariant to two extreme global models with different abilities, which provides richer consistency supervision for local training. Moreover, we devise a relation-guided directional aggregation at the server, in which a parametric relation predictor is introduced to yield pairwise model relation and obtain a model ranking. In this manner, the server can provide a superior global model by aggregating relative dependable client models, and further produce an inferior global model via reverse aggregation to promote the versatile regularization at clients. Extensive experiments on three FSSL benchmarks verify the superiority of FedRVR over state-of-the-art counterparts across various federated learning settings.

Keywords Federated semi-supervised learning · Versatile regularization · Relation-guided aggregation.

1 Introduction

Federated learning (FL) (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017; Mendieta et al., 2022; X. Li, Jiang, Zhang, Kamp, and Dou, 2021; L. Zhang, Luo, Bai, Du, and Duan, 2021; T. Li et al., 2020; X- C. Li et al., 2022; Y. Huang et al., 2021) is a decentralized machine learning paradigm, where multiple clients collaboratively train a global model without data sharing. FL leverages the models trained at various clients and yields an aggregated model at the dependable server. Due to the data privacy-preserving issue, FL has demonstrated incredible success in a wide range of scenarios, including clinical diagnosis, public security and digital finance (Z. Chen, Yang, Zhu, Peng, and Yuan, 2022; Dou et al., 2021; McMahan, Moore, Ramage, Hampson, and y Arcas, 2017; L. Zhang, Shen, Ding, Tao, and Duan, 2022).

Most current FL methods (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017; Mendieta et al., 2022; X. Li, Jiang, Zhang, Kamp, and Dou, 2021; L. Zhang, Luo, Bai, Du, and Duan, 2021; T. Li et al., 2020; X- C. Li et al., 2022;

Communicated by Gunhee Kim.

✉ Zhe Peng
zhepeng@polyu.edu.hk

✉ Yixuan Yuan
yxyuan@ee.cuhk.edu.hk

Qiushi Yang
qsyang2-c@my.cityu.edu.hk

Zhen Chen
zhen.chen@cair-cas.org.hk

[1] Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China

[2] Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong SAR, China

[3] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[4] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

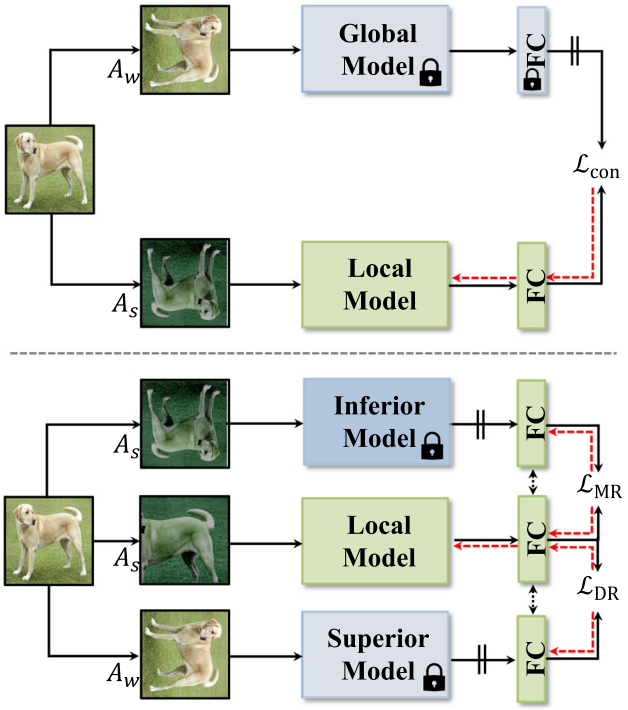

Fig. 1 Different training frameworks on unlabeled data. **Top**: Existing FSSL works enforce the predictions to be similar across input variants under data augmentation via consistency regularization $\mathcal{L}_{con}$. **Bottom**: Our versatile regularization aligns predictions of the local model and global models via both data-guided and model-guided constraints $\mathcal{L}_{DR}$ and $\mathcal{L}_{MR}$

Y. Huang et al., 2021) assume that the local data at each client are fully annotated, which is laborious or even unrealistic in many FL applications with limited labeling capability. To ameliorate these issues, federated semi-supervised learning (FSSL) (Jeong, Yoon, Yang, and Hwang, 2020; Liang, Lin, Fu, Zhu, and Li, 2022; Fan, Hu, and Huang, 2022; Q. Liu, Yang, Dou, and Heng, 2021; Jiang et al., 2022) is formulated to exploit knowledge from a large amount of unlabeled data across clients. As illustrated in Fig. 1 (left), these FSSL methods (Kuo, Ma, Huang, and Kira, 2020; Fan, Hu, and Huang, 2022; Liang, Lin, Fu, Zhu, and Li, 2022; Q. Liu, Yang, Dou, and Heng, 2021) usually adopt data-guided consistency to train the local model. They leverage a global model received from the server and encourage prediction alignment between the global and local models under data augmentation.

Although previous works have achieved decent performance, they merely constrain the consistency between the local model and global model through the *data-level* augmentation. Different from the centralized semi-supervised learning, the clients in FSSL scenarios can leverage multiple global models to perform diverse *feature-level* augmentations and introduce more effective regularization for training local classifier, thereby fully exploiting knowledge from unlabeled data. In this regard, we adopt two global models with extremely different abilities to generate perturbated features, and perform the consistency among these perturbed features as the model-guided regularization to complement the constraints for local training. As illustrated in Fig. 1 (right), the *superior* global model as the feature extractor outputs the features, which can yield credible pseudo labels to guide the local model training. Different from merely using strong data augmentation for regularization (Sohn et al., 2020; J. Li, Xiong, and Hoi, 2021; Hu, Yang, Hu, and Nevatia, 2021; B. Zhang et al., 2021), this work represents the first effort to introduce an *inferior* global model as an *aggressive feature augmentor* to generate highly perturbed features for the classifier, and achieves aggressive model-guided regularization. With both data-guided and model-guided versatile regularizations, the client can efficiently train local model and improve FSSL performance.

Moreover, existing studies treat all unlabeled clients equally, whilst different unlabeled clients perform different untrustworthy local training, which impede model aggregation and further mislead decentralized training. Latest studies in network performance estimation (Wen et al., 2020; Xu, Wang, et al., 2021; Y. Chen et al., 2021) indicate that the potential pairwise model relations can be estimated via a parametric predictor. To this end, we introduce a *relation predictor* to capture pairwise model relations at the server, which can yield a directional path for model aggregation via a relation-aware model ranking. In this way, we can obtain a superior global model to promote decentralized training. Towards stronger regularization for local training, we further produce an inferior global model via reverse aggregation. This relation-guided directional aggregation not only produces robust global models, but also facilitates local training.

In this work, we propose a novel FSSL framework, Relation-guided Versatile Regularization (FedRVR), comprising versatile regularization and relation-guided directional aggregation, to efficiently exploit knowledge from unlabeled data across clients. At each client, we present versatile regularization, which introduces two global models including one with superior ability and another with inferior ability to provide more efficient and aggressive constraints for local model training. Specifically, the superior global model renders data-guided regularization, while the inferior global model yields diverse features towards the model-guided regularization. The predictions of the local model are encouraged to be invariant with two global models simultaneously. These two global models offer compatible regularizations and enable the local model to achieve more compelling FSSL performance. Moreover, at the server, we design a relation-guided directional aggregation to acquire the robust global model against skewed unlabeled clients. It introduces a parametric relation predictor to obtain pairwise model relation and gathers client models according to a model ranking yielded by pairwise relations among local

models. This relation-guided directional aggregation can produce more robust global models and further promote local training at decentralized clients.

To summarize, our contributions are fourfold:

- In this work, we present FedRVR to exploit knowledge from unlabeled data, which significantly improves local training efficiency and the robustness of server model aggregation.
- Different from prior works merely relying on data-level augmentation consistency, to render versatile regularization towards better local training at each client, we introduce both data-guided and model-guided constraints via two global models.
- At the server, we design a relation-guided aggregation strategy via model relation learning, which endows the server to produce the superior global model for robust decentralized training and the inferior global model for efficient local training.
- Extensive experiments on three FSSL benchmarks demonstrate that our framework FedRVR outperforms state-of-the-art methods across various FL settings.

2 Related Work

2.1 Semi-Supervised Learning

Semi-supervised learning (SSL) is dominated by two popular directions: (1) Pseudo labeling-based approaches (Xie, Dai, Hovy, Luong, and Le, 2020; Cascante-Bonilla, Tan, Qi, and Ordonez, 2020; Xu, Shang, et al., 2021; Sohn et al., 2020; Xie, Luong, Hovy, and Le, 2020; Berthelot, Carlini, Goodfellow, et al., 2019; Berthelot, Carlini, Cubuk, et al., 2019) produce artificial labels as supervision for unlabeled samples, and then train models with these labels in an ad-hoc fashion. In recent years, many works (Xie, Dai, Hovy, Luong, and Le, 2020; Cascante-Bonilla, Tan, Qi, and Ordonez, 2020; Xu, Shang, et al., 2021; Sohn et al., 2020; Xie, Luong, Hovy, and Le, 2020; Berthelot, Carlini, Goodfellow, et al., 2019; Berthelot, Carlini, Cubuk, et al., 2019) adopt pseudo labeling process in SSL frameworks. (2) Consistency-based methods (Laine and Aila, 2016; J. Li, Xiong, and Hoi, 2021; Gong, Wang, and Liu, 2021; Yang, Chen, and Yuan, 2023; Hu, Yang, Hu, and Nevatia, 2021; Sohn et al., 2020; B. Zhang et al., 2021; Berthelot, Carlini, Goodfellow, et al., 2019; Yang, Liu, Chen, Ibragimov, and Yuan, 2022) depend on the low-density assumption that the decision boundary usually lies on the low data density regions. Under this assumption, these works encourage the consistency between different views of samples, and enable the trained models to capture information related to the decision boundary. Most consistency-based SSL methods (J. Li, Xiong, and Hoi, 2021; Gong, Wang, and Liu, 2021; Berthelot, Carlini, Goodfellow, et al., 2019; Sohn et al., 2020; B. Zhang et al., 2021; Miyato, Maeda, Koyama, and Ishii, 2018; Yang, Liu, Chen, Ibragimov, and Yuan, 2022) perform the consistency constraints of sample predictions using different data augmentations. Others (Laine and Aila, 2016; Tarvainen and Valpola, 2017; Zhou, Wang, and Bilmes, 2020; T. Huang, Sun, Wang, Yao, and Zhang, 2021) enforce the sample predictions to be consistent across different training generations.

2.2 Federated Learning

Federated learning (FL) (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017; Mendieta et al., 2022; T. Li et al., 2020; Y. Huang et al., 2021; X- C. Li et al., 2022; L. Zhang, Luo, Bai, Du, and Duan, 2021; L. Zhang, Shen, Ding, Tao, and Duan, 2022; Z. Chen, Yang, Zhu, Peng, and Yuan, 2022; Zhu, Liao, Liu, and Yuan, 2023) emerges to collaboratively optimize models across multiple clients with labeled private data. The recent works mainly focus on promoting local training at clients and model aggregation at the server. (1) The local training methods (Z. Chen, Yang, Zhu, Peng, and Yuan, 2022; L. Zhang, Shen, Ding, Tao, and Duan, 2022; L. Zhang, Luo, Bai, Du, and Duan, 2021) aim to facilitate the local model of each client by leveraging the global model and information from other clients. For example, FedFTG (L. Zhang, Shen, Ding, Tao, and Duan, 2022) utilizes a dynamic global model for each client to boost local training via a data-free knowledge distillation strategy at the server. FedBN (X. Li, Jiang, Zhang, Kamp, and Dou, 2021) maintains the batch normalization layers updated locally and upload other layers to the server towards the issue of personalized information within each client. (2) For the model aggregation, FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017) gathers the knowledge of model parameters with a simple average. Many works (Mendieta et al., 2022; T. Li et al., 2020; Y. Huang et al., 2021; X- C. Li et al., 2022) improve the sub-optimal FedAvg and design re-weighting strategies for each client model. FedAMP (Y. Huang et al., 2021) performs federated attentive message passing to improve similar clients to be more collaborative. FedProx (T. Li et al., 2020) proposes a tolerating partial strategy to pick up a set of reliable models to achieve model aggregation by minimizing the knowledge divergence across all clients. Compared with these methods developed for fully-supervised FL, we focus on more practical FSSL scenarios and propose the FedRVR framework to improve both local training and model aggregation.

2.3 Federated Semi-Supervised Learning

Federated Semi-Supervised Learning (FSSL) (Jeong, Yoon, Yang, and Hwang, 2020; Fan, Hu, and Huang, 2022; M. Li,

Li, and Wang, 2023; Liang, Lin, Fu, Zhu, and Li, 2022; Q. Liu, Yang, Dou, and Heng, 2021; Long et al., 2020; Jiang et al., 2022; Z. Zhang et al., 2020; Cho, Joshi, and Dimitriadis, 2023; Y. Liu, Wu, and Qin, 2024) considers more practical scenarios where data holders are limited in labeling capability. The FSSL settings can be divided into three lines: (1) Labeled-Unlabeled FSSL assumes that the majority of clients contain fully unlabeled data while few clients have fully labeled data. FedSSL (Fan, Hu, and Huang, 2022) designs a unified mixup strategy to generate mixed samples for better training in a shared global data space. RSCFed (Liang, Lin, Fu, Zhu, and Li, 2022) performs random sub-sampling across clients to exploit model aggregation consensus with a distance-based re-weighting term. FedCD (Y. Liu, Wu, and Qin, 2024) targets at the class-imbalanced issue and proposes a global-local distillation framework with a class-aware balancing strategy for better rare categories recognition. (2) Partial-Labeled FSSL considers each client has mostly unlabeled data and a small number of labeled data. FedMatch (Jeong, Yoon, Yang, and Hwang, 2020) exploits knowledge of unlabeled data through inter-client prediction consistency. (3) Labels-at-Server FSSL assumes that the client only contains unlabeled data and the server has a small amount of labeled data. SemiFL (Diao, Ding, and Tarokh, 2021) fine-tunes the global model using the labeled samples stored at the server, and performs self-training at clients in an alternate manner. In this work, we follow the mainstream FSSL works (Jeong, Yoon, Yang, and Hwang, 2020; Liang, Lin, Fu, Zhu, and Li, 2022; Fan, Hu, and Huang, 2022; Q. Liu, Yang, Dou, and Heng, 2021; Cho, Joshi, and Dimitriadis, 2023; Y. Liu, Wu, and Qin, 2024) and focus on first two settings, since the server in FL usually coordinates the client collaboration rather than storing data.

3 Motivation and Preliminaries

3.1 Motivation

Existing FSSL approaches typically employ a *single* global model with *data-level* augmentation to maintain prediction consistency, overlooking the potential of server-side resources to provide multiple global models and enhance regularization efficiency. To address this limitation, we propose to leverage *diverse* global models based on *different* aggregation strategies at the server. These global models serve as feature augmentors, producing highly perturbed features to facilitate model-guided regularization for local classifier training. Our approach requires the server to evaluate and rank model performance based on their inter-relationships. This process aims to identify an *inferior* global model with relative worse ability than local models to produce highly perturbed features. By integrating the strengths of FL with SSL techniques, we introduce the FedRVR framework, which consists of two key components, including a model relation-guided aggregation to provide superior and inferior global models, and a versatile regularization that exploits more efficient regularization.

3.2 Preliminaries

Federated learning. In standard FL, given K clients, the local model $f_k \in \{f_1, ..., f_K\}$ at each client is trained by local labeled dataset. At each communication round, client models $\{f_k\}_{k=1}^K$ are uploaded to the server and yield a global model g by parameter aggregation. The decentralized training optimizes the global objective: $\min_f \sum_{k=1}^K \frac{N_k}{N} \mathcal{L}_k(f_k)$, where N_k refers to the number of data at k-th client, and N is the number of all data. $\mathcal{L}_k$ denotes the loss function at k-th client. This objective can be solved by FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017), where the server averages client models to produce a global model $g = \frac{1}{K} \sum_{k=1}^K f_k$. The global model g is then broadcast to clients and updates local models for further local training. The framework repeats these procedures until training convergence.

Federated semi-supervised learning. We consider the fully-labeled and fully-unlabeled clients in FSSL. We assume all clients $\{C_1, ..., C_K\}$ contain K_L labeled clients $\{C_1, ..., C_{K_L}\}$ and K_U unlabeled clients $\{C_{K_L+1}, ..., C_{K_L+K_U}\}$, $K = K_L + K_U$, in general, $K_L \ll K_U$. Each labeled client has a dataset $D^k = \{x_i^l, y_i^l\}_{i=1}^{N_K}$ including N_k labeled samples, and each unlabeled client has a dataset $D^k = \{x_i^u\}_{i=1}^{N_K}$ containing N_k unlabeled examples. The objective of FSSL is minimizing the global loss function $\mathcal{L}_{\text{global}} = \mathcal{L}_l + \mathcal{L}_u$, where $\mathcal{L}_l$ and $\mathcal{L}_u$ are supervised and unsupervised losses yielded by labeled and unlabeled client models respectively.

4 Methodology

In this section, we first describe the overview of our FedRVR framework, and then introduce its components including the versatile regularization at clients and the relation-guided directional aggregation at the server.

4.1 Overview of FedRVR

As illustrated in Fig. 2, FedRVR is composed of the versatile regularization to promote local training, and the relation-guided directional aggregation to facilitate decentralized training at the server. At each unlabeled client, the local model f is jointly trained by two proposed regularizations via two global models g_I and g_S. At the server, the relations

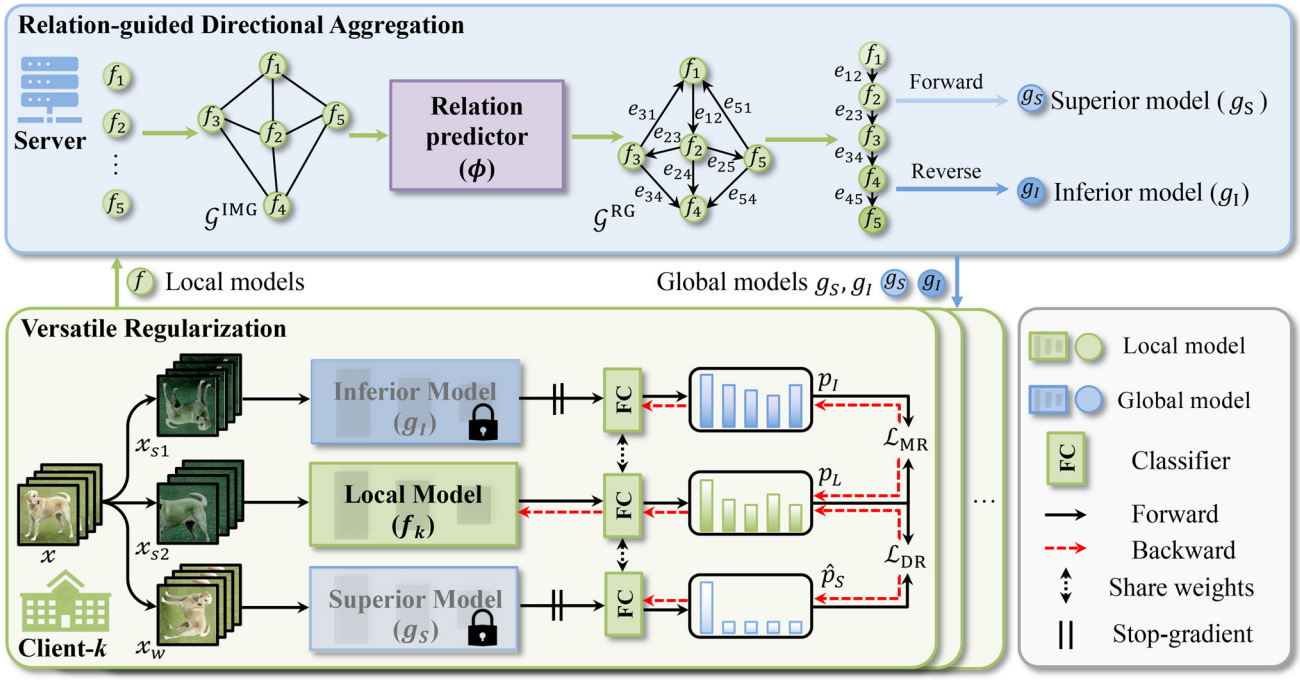

Fig. 2 The proposed FedRVR framework. It consists of versatile regularization for efficient consistency regularizations towards local self-training at the client, and a relation-guided directional model aggregation strategy to produce global models at the server. In each communication round, all clients send the local models and the labeled clients upload the relation predictor to the server. Then, the server aggregates local models and produces a superior global model and an inferior global model for each client. After each client receives two global models, the versatile regularization guarantees efficient self-training of the local model.

among client models are captured by a relation predictor ϕ, and two global models g_I and g_S are produced by relation-guided directional aggregation via predictor ϕ and a relation graph $\mathcal{G}^{\text{RG}}$. The relation predictor ϕ is trained at labeled clients and average aggregated at the server.

4.2 Versatile Regularization for Unlabeled Data

Given k-th unlabeled clients with dataset $D^k = \{x_i^u\}_{i=1}^{N_k}$ containing N_k unlabeled samples, we aim to leverage it to train a local model f with the support of two global models g_S and g_I. To provide efficient regularization for local training, versatile regularization involves data-guided and model-guided regularizations to boost the local model training in different aspects, respectively.

Data-guided regularization. To update the local model f in a self-training fashion, we introduce the data-guided regularization, a data augmentation-based consistency to regularize the local model with the superior aggregated global model g_S. Specifically, with the weak augmented data $\mathcal{A}_w(x_i^u)$, where $\mathcal{A}_w()$ represents the weak data augmentation, we can obtain the prediction $p_S^i = g_S(\mathcal{A}_w(x_i^u))$ using the global model g_S. Then, the relative reliable prediction p_S^i with high prediction confidence (i.e., $\max p_S^i \geq \tau$, where τ is the threshold) is selected to generate pseudo label $\hat{p}_S^i = \arg\max p_S^i$. Meanwhile, the local model f utilizes the strong augmented data $\mathcal{A}_s(x_i^u)$, where $\mathcal{A}_s()$ denotes the strong data augmentation, to infer prediction $p_L^i = f(\mathcal{A}_s(x_i^u))$, which is encouraged to be consistent with the pseudo label $\hat{p}_S^i$. The data-guided regularization can be formulated as:

$$\mathcal{L}_{\text{DR}} = \frac{1}{N_k^B} \sum_{i=1}^{N_k^B} H(\hat{p}_S^i, p_L^i), \tag{1}$$

where $H(\cdot, \cdot)$ refers to cross-entropy loss and N_k^B is the number of unlabeled samples within a mini-batch at the k-th client. The data-guided regularization enforces the predictions to be invariant towards different data augmentations.

Model-guided regularization. In addition to data-guided regularization, the global models received from the server can generate diverse features, and provide model-guided regularization as the compatible constraint. To this end, we further propose the model-guided regularization via another global model g_I with inferior ability as an aggressive feature augmentor to generate highly perturbed features. Particularly, to yield aggressive feature augmentations, we download the inferior global model g_I and leverage its feature encoder g_I^{enc} to extract diverse features $h_i = g_I^{\text{enc}}(\mathcal{A}_s(x_i^u))$ of unlabeled data, and we will explain the aggregation strategy towards the

inferior global model in the next Sect. 4.3. With the generated aggressive features h_i, we feed them into the local classifier f^{cls} to produce prediction $p_I^i = f^{\text{cls}}(h_i)$, which is then enforced to be invariant to the prediction $p_L^i = f(\mathcal{A}_s(x_i^u))$ from the local model f via the model-guided regularization:

$$\mathcal{L}_{\text{MR}} = \frac{1}{N_k^B} \sum_{i=1}^{N_k^B} H(p_I^i, p_L^i). \tag{2}$$

In this way, the data-guided and model-guided regularizations offer compatible constraints via aggressive augmentations on both data and features to promote local training.

Local training. The local model f at each unlabeled client is jointly trained by the superior global model g_S via the proposed data-guided and model-guided regularizations , respectively. Note that the local classifier f^{cls} of the client model f is shared across both local and global models. To avoid incurring degenerate solution, two global models are frozen except the classifiers. The overall loss function of local training for unlabeled clients can be expressed as:

$$\mathcal{L}_u = \mathcal{L}_{\text{DR}} + \lambda \mathcal{L}_{\text{MR}}, \tag{3}$$

where λ is a trade-off factor to balance the contributions of two regularizations. With respect to the labeled clients, the local training is supervised by cross-entropy loss $\mathcal{L}_l = \frac{1}{N_k^B} \sum_{i=1}^{N_k^B} H(y_i, \hat{y}_i)$, where y_i and $\hat{y}_i$ denote the label and prediction, respectively.

4.3 Relation-Guided Directional Aggregation

Most existing aggregation strategies (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017; Jeong, Yoon, Yang, and Hwang, 2020; Q. Liu, Yang, Dou, and Heng, 2021) treat each unlabeled client model equally at the server, while they ignore the potential relation among models, which is beneficial to robust model aggregation. To this end, we propose the relation-guided directional aggregation by building a model relation graph via pairwise model relation learning.

Intrinsic model graph. In the server with K client models $\{f_1, ..., f_K\}$, we construct an undirected intrinsic model graph $\mathcal{G}^{\text{IMG}} = \{\mathcal{V}^{\text{IMG}}, \mathcal{E}^{\text{IMG}}\}$, where the vertices $\mathcal{V}^{\text{IMG}}$ represent client models and the edges $\mathcal{E}^{\text{IMG}}$ are defined as the consine similarity between the parameters of two models $e_{i,j}^{\text{IMG}} = \frac{f_i \cdot f_j}{\|f_i\|_2 \|f_j\|_2}$. This graph represents the intrinsic relation among all client models, which can be regarded as an undirected weighted graph.

Relation predictor. To capture the pairwise relation towards potential discrimination ability among client models, as illustrated in Fig. 3, we introduce a parametric relation predictor ϕ to make comparisons between two models, which is trained at each labeled client after each communication round and average aggregated at the server. We formulate an inter-model relation learning task at each labeled client. In particular, we construct the model pairs $\{f_i, f_j\}$, consisting of two local models, as the training data, and set the true classification accuracy comparison as the label $y^{\text{RP}} \in \{0, 1\}$, where $y^{\text{RP}} = 1$ if f_i is superior to f_j, and vice versa. Since the model parameter space is continuous and the number of parameters is extremely large, it is inefficient to directly encode all parameters and train the relation predictor. To alleviate this issue, we first sample part of layers (the classifier and penultimate layer) within the model. Then, we prune the selected layers by setting the unimportant parameters smaller than a threshold θ_{pru} as zero. Afterward, we randomly pick two pruned models f_i and f_j as the inputs to feed the relation predictor ϕ. The predictor produces the binary comparison result $p^{\text{RP}} = \phi(f_i, f_j)$, $p^{\text{RP}} \in [0, 1]$. With the true comparison result y^{RP} as the supervision, the relation predictor is trained as a binary classification task via binary cross-entropy loss $\mathcal{L}_{\text{RP}} = -[y^{\text{RP}} \log(p^{\text{RP}}) + (1 - y^{\text{RP}}) \log(1 - p^{\text{RP}})]$. Note that all historical local and global models from the past training iterations of the labeled client are utilized to train the relation predictor. The well-trained relation predictors ϕ from labeled clients are sent to the server to average and yield a global relation predictor $\phi = \frac{1}{K_L} \sum_{k=1}^{K_L} \phi_k$, where ϕ_k denotes the relation predictor of k-th labeled client. Obtained the global relation predictor, the server can evaluate the relation among all client models and construct a directed relation graph $\mathcal{G}^{\text{RG}}$, which inherits the weights of edges $\mathcal{E}^{\text{IMG}}$ within the undirected intrinsic model graph $\mathcal{G}^{\text{IMG}}$. Then, the relation graph $\mathcal{G}^{\text{RG}}$ can be utilized to achieve superior model aggregation illustrated in next directional model aggregation.

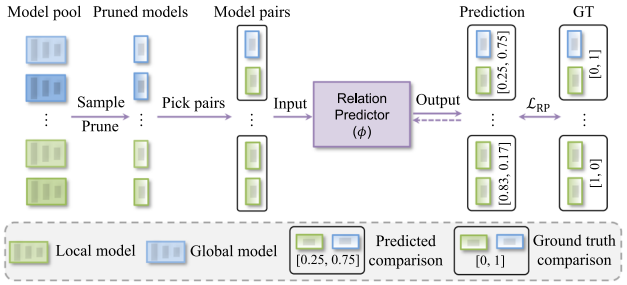

Fig. 3 Training of model relation predictor. With the pairwise models as inputs from the labeled client, the predictor is trained by the true performance comparison of the pair models. All historical local and global models from the past training iterations of the labeled client are utilized to train the relation predictor. Note that the relation predictor is trained at labeled clients and deployed at the server.

Directional model aggregation. Given the estimated relations with the directional graph $\mathcal{G}^{\text{RG}}$ from the relation predictor, we first leverage the approximation algorithm proposed in Arch-Graph (M. Huang et al., 2022) to find the maximal acyclic subgraph $\mathcal{G}^{\text{MAS}} = (\mathcal{V}, \mathcal{E}^{\text{MAS}})$ weighted by

Algorithm 1 Pipeline of our FedRVR framework.

Input: Unlabeled data x_u at unlabeled clients $C_1, \ldots, C_{K_U}$, client models $f_1, \ldots, f_K$, learning rate η; **Output:** Global model g_S;

procedure SERVERUPDATE
  for each round in MaxRound **do**
    for each client in K_U **do**
      $\{f_1, \ldots, f_K\}, \{\phi_1, \ldots, \phi_{K_L}\} \leftarrow$ **LocalUpdate**(g_S, g_I)
    end for
    Build intrinsic model graph $\mathcal{G}^{\text{IMG}}$, and obtain similarity score as the edges $\mathcal{E}^{\text{IMG}}$
    Feed the pairwise client models to relation predictor, and yield relation graph $\mathcal{G}^{\text{RG}} = \phi(\mathcal{G}^{\text{IMG}})$
    Search maximal acyclic subgraph $\mathcal{G}^{\text{MAS}}$
    Forward aggregation to produce superior global models g_S
    Reverse aggregation to produce inferior global models g_I
    Return two global models g_S and g_I
  end for
end procedure
procedure LOCALUPDATE(g_S, g_I)
  for each epoch in MaxLocalEpoch **do**
    Data-guided regularization $\mathcal{L}_{\text{DR}}$
    Model-guided regularization $\mathcal{L}_{\text{MR}}$
    Update the local model: $f \leftarrow f - \eta\nabla(\mathcal{L}_{\text{DR}} + \lambda\mathcal{L}_{\text{MR}})$
  end for
end procedure
return f and ϕ

the similarity scores in the directional relation graph $\mathcal{G}^{\text{RG}}$, i.e., the edges $\mathcal{E}^{\text{IMG}}$ of intrinsic model graph $\mathcal{G}^{\text{IMG}}$. Then, we rank these models from the best ability one to the worst ability one Rank $= \{f_{r_1}, ..., f_{r_M}\}$, where M represents the node number of the $\mathcal{G}^{\text{MAS}}$ and r_i means the index of the i-th best client model. This model ranking can yield a directional path towards model aggregation, which contains the information of model relation, and produce the superior global model and inferior global model via the intrinsic similarity scores within $\mathcal{G}^{\text{IMG}}$. Afterward, considering that the maximal acyclic subgraph $\mathcal{G}^{\text{MAS}}$ represents the directional relation among client models, we calculate the intrinsic similarity score $e^{\text{IMG}}_{r_i,r_1}$ between the client model f_{r_i} and the *best* model f_{r_1}. Then, we normalize it across the subgraph $\mathcal{G}^{\text{MAS}}$ as the weight $\hat{e}^{\text{IMG}}_{r_i,r_1}$ of the model f_{r_i} to achieve sequential model aggregation. Then, the superior global model g_S is obtained through the aggregation via the forward path of the $\mathcal{G}^{\text{MAS}}$ and $\hat{\mathcal{E}}^{\text{IMG}}$, as follows:

$$g_S = \frac{1}{M}\sum_{i=1}^{M} \hat{e}^{\text{IMG}}_{r_i,r_1} f_{r_i}. \tag{4}$$

In this manner, with the relation-guided model ranking, we can determine the best client model and assign it the largest weight. The weights of other models are obtained via the similarity between the best client model. In addition, to produce highly perturbed features towards strong model-guided regularization at the client, we compute the intrinsic similarity score $e^{\text{IMG}}_{r_i,r_M}$ between the client model f_{r_i} and the *worst* model f_{r_M}. The normalized score $\hat{e}^{\text{IMG}}_{r_i,r_M}$ is used as the weight of the model f_{r_i} to yield an inferior global model g_I, as follows:

$$g_I = \frac{1}{M}\sum_{i=1}^{M} \hat{e}^{\text{IMG}}_{r_i,r_M} f_{r_i}, \tag{5}$$

where the worst client model f_{r_M} is assigned the largest weight towards reverse aggregation. Hence, the server provides two extreme global models g_S and g_I to perform the data-guided and model-guided regularizations in Sect. 4.2, and further render better consistency constraints for local training in FSSL.

4.4 Optimization Pipeline

In each communication round, each client sends the local model f and the labeled clients upload the relation predictor ϕ to the server. The server aggregates local models via relation-guided directional aggregation, and produces a superior global model g_S and an inferior global model g_I. After each client receives g_S and g_I, the versatile regularization guarantees efficient self-training of the local model with data-guided and model-guided constraints. Finally, the local models are optimized by decentralized training, and the superior global model g_S serves as the output model of FedRVR for inference. We summarize the overall pipeline of the proposed FedRVR framework in Algorithm 1.

5 Experiments

In this section, we evaluate the FedRVR on two FSSL settings (i.e., labeled-unlabeled and partial-labeled clients) with different data heterogeneity types of IID and Non-IID.

5.1 Experimental Setup

Datasets. To evaluate the effectiveness of our model, we conduct experiments on three image classification benchmarks, including CIFAR-10 (Krizhevsky, Hinton, et al., 2009), CIFAR-100 (Krizhevsky, Hinton, et al., 2009) and ISIC-2018 (Codella et al., 2019). Following common practice (Liang, Lin, Fu, Zhu, and Li, 2022), we use official training and test sets of CIFAR-10 and CIFAR-100, and randomly split ISIC dataset into 80% as training set and 20% as test set. The images in ISIC dataset are resized as 240×240 in the pre-processing for efficient training.

FSSL settings. In the labeled-unlabeled FSSL, we set 10 clients, where 1 client is fully labeled and the other 9 clients only contain unlabeled data. This task is challenging since the majority of unlabeled clients may mislead the training, especially in the early phase of FSSL. Moreover, we further

evaluate FedRVR in the partial-labeled setting, where each client holds 10% labeled and 90% unlabeled data.

Data heterogeneity. We consider three cases in terms of data heterogeneity across different clients, including one IID and two Non-IID settings. **(1) IID:** It endows an identical ratio of class distribution over all clients. We set the uniform class number for each client. **(2) Non-IID:** In this setting, the class distributions are distinct across clients as the heterogeneous FL scenarios. We randomly perform the number of data per category for each client via Dirichlet distribution Dir(α), where each client contains a subset of categories. The coefficient α set as 0.5 and 0.8 constitutes two Non-IID settings in our experiments, respectively.

Implementation details. All models are optimized by stochastic gradient descent (SGD) (Sutskever, Martens, Dahl, and Hinton, 2013) with a momentum of 0.9 and a weight decay (Loshchilov & Hutter, 2016) of 5×10^{-4}. We use PyTorch (Paszke et al., 2019) to implement our models, and train all models for 1000 rounds with 1 local training epoch per round on each dataset. The batch size for CIFAR-10 and CIFAR-100 is set as 64, and 12 for ISIC. We use the crop-and-flip as the weak data augmentation $\mathcal{A}_w()$ and the standard RandAugmentation (Cubuk, Zoph, Shlens, and Le, 2020) as the strong data augmentation $\mathcal{A}_s()$. The trade-off factor λ in Eq. (3) is set as 1.0, and the threshold θ_{pru} towards model pruning is equal to the median of the selected layer parameters. To facilitate fair evaluation, we adopt the same backbones as the previous work (Liang, Lin, Fu, Zhu, and Li, 2022), where a simple CNN consisting of two convolution layers and two fully-connected layers is used for CIFAR-10 and CIFAR-100, and ResNet-18 (He, Zhang, Ren, and Sun, 2016) is utilized for ISIC. We will release the source code for reproduction.

5.2 Comparison with State-of-the-Arts

Labeled-unlabeled FSSL. To verify the effectiveness of the FedRVR framework, we conduct comparison experiments for three runs with state-of-the-art FSSL methods in both IID and Non-IID cases. As shown in Table 1, our FedRVR outperforms other FSSL works with a clear improvement in IID, Non-IID with Dir(0.5) and Non-IID with Dir(0.8) settings across three datasets. Specifically, compared with the second-best FedLabel (Cho, Joshi, and Dimitriadis, 2023) in three settings, FedRVR outperforms with 1.21%, 2.02%, 1.33% average accuracy on CIFAR-10, 1.67%, 1.58%, 1.36% accuracy on CIFAR-100, and 1.62%, 1.36%, 1.64% accuracy on ISIC. These advantages demonstrate the effectiveness of FedRVR. Particularly, in Non-IID setting with Dir(0.5), FedRVR achieves 58.24%, 15.80%, 69.77% accuracy, and 86.81%, 80.76%, 87.56% AUC score on CIFAR-10, CIFAR-100 and ISIC, respectively, indicating that FedRVR is still powerful even the data heterogeneity is largely skewed across clients.

Partial-labeled FSSL. We further consider another FSSL scenario, i.e., partial-labeled FSSL, where each client contains 10% of labeled data and 90% of unlabeled data. In Table 2, FedRVR on CIFAR-10 delivers 68.43% accuracy and 93.40% AUC score in IID setting, outperforming all other competitors. In Non-IID with Dir(0.5) setting, FedRVR exhibits 65.46% accuracy and 88.93% AUC score, with a significant advantage of 1.56% in accuracy and 1.55% AUC score over the second-best FedLabel (Cho, Joshi, and Dimitriadis, 2023). These results confirm that the FedRVR framework consistently performs better in different FSSL scenarios.

5.3 Ablation Study

We conduct extensive ablation studies on CIFAR-10 to investigate each module in the FedRVR framework.

Effectiveness of relation-guided directional aggregation. We study the effectiveness of the proposed relation-guided directional aggregation (RDA) at the server. Firstly, we only use the intrinsic model graph (IMG) to obtain the similarity scores $\mathcal{E}^{\text{IMG}}$ with all other clients for each client to calculate the re-weighting term for model aggregation. From Table 3, the accuracy increases from 54.31% and 51.66% to 55.67% and 53.27% in IID and Non-IID with Dir(0.5), respectively. Then, we adopt directional aggregation (DA), i.e., only using $\mathcal{G}^{\text{RG}}$ to perform client ranking and average aggregating models to produce a single global model. As shown in 5^{th} row of Table 3, the accuracy suggests 1.04% and 1.23% improve in two FL settings respectively. We finally simultaneously perform the IMG and DA via relation predictor to gather the global model. The performance constantly increases by 1.26% and 0.88% of accuracy. These advantages indicate that the server model aggregation strategy with intrinsic model graph and relation learning can yield a robust global model, boosting decentralized training and local training.

Impact of versatile regularization. To observe the impact of the versatile regularization (VR), we employ the data-guided and model-guided regularizations on the framework with the relation-guided directional aggregation. As seen in Table 3, the data-guided regularization ($\mathcal{L}_{\text{DR}}$) brings 1.55% and 2.78% accuracy improvements in IID and Non-IID cases, respectively. The model-guided regularization ($\mathcal{L}_{\text{MR}}$) can further lead to a 2.06% and 1.31% rise in two cases. These results imply that both two regularizations facilitate the FSSL performance, and they can offer complementary constraints for local training. Particularly, the versatile regularization is robust towards data heterogeneity with consistent performance gain.

Table 1 Comparison with state-of-the-art FSSL algorithms on CIFAR-10, CIFAR-100 and ISIC-2018 datasets across IID and Non-IID settings over three runs. “ub" and “lb" denote the upper and lower bound respectively. “L" and “UL" mean the number of labeled and unlabeled clients.

Dataset	Method	Num.		IID		Non-IID w/ Dir(0.5)		Non-IID w/ Dir(0.8)	
		L	UL	Acc.	AUC	Acc.	AUC	Acc.	AUC
CIFAR-10	FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017) (ub)	10	0	72.37±1.15	92.67±0.52	68.15±2.46	88.43±0.82	69.59±2.17	90.33±0.68
	FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017) (lb)	1	0	55.42±1.29	85.23±0.62	52.13±2.06	81.55±0.94	53.74±1.74	83.87±0.82
	FedMatch (Jeong, Yoon, Yang, and Hwang, 2020)	1	9	55.94±1.37	86.10±0.82	52.90±1.83	82.10±0.84	54.46±1.51	84.56±0.80
	FedIRM (Q. Liu, Yang, Dou, and Heng, 2021)	1	9	56.57±1.02	86.85±0.77	53.37±1.58	82.85±0.77	54.91±1.20	85.06±0.73
	FedSSL (Fan, Hu, and Huang, 2022)	1	9	56.97±1.18	87.58±0.92	54.52±1.64	83.52±1.15	55.58±1.65	85.77±0.98
	RSCFed (Liang, Lin, Fu, Zhu, and Li, 2022)	1	9	58.63±0.73	88.75±0.80	55.68±1.32	84.37±0.69	56.43±0.90	86.91±0.67
	FedLabel (Cho, Joshi, and Dimitriadis, 2023)	1	9	58.92±0.58	88.95±0.71	55.87±1.25	84.70±0.58	56.72±0.88	87.31±0.59
	FedCD (Y. Liu, Wu, and Qin, 2024)	1	9	59.33±0.64	89.27±0.74	56.22±1.17	85.41±0.55	57.17±0.74	87.66±0.55
	FedRVR	1	9	**60.54±0.60**	**90.46±0.72**	**58.24±1.10**	**86.81±0.65**	**58.50±0.83**	**88.77±0.59**
CIFAR-100	FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017) (ub)	10	0	28.23±0.49	92.75±0.52	21.69±1.14	87.15±0.86	25.48±0.85	90.15±0.66
	FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017) (lb)	1	0	14.52±0.67	79.33±0.92	10.08±0.30	72.17±1.27	12.31±0.45	76.39±1.13
	FedMatch (Jeong, Yoon, Yang, and Hwang, 2020)	1	9	17.58±0.66	82.58±0.93	11.39±0.55	75.83±1.59	13.67±0.85	78.42±1.26
	FedIRM (Q. Liu, Yang, Dou, and Heng, 2021)	1	9	18.39±0.51	84.31±0.50	12.35±0.37	76.48±1.27	14.09±0.77	79.08±0.73
	FedSSL (Fan, Hu, and Huang, 2022)	1	9	19.26±0.46	84.69±0.52	13.06±0.75	77.21±0.91	14.88±0.65	79.76±0.79
	RSCFed (Liang, Lin, Fu, Zhu, and Li, 2022)	1	9	20.68±0.57	86.12±0.61	13.85±0.64	78.60±0.76	16.17±0.53	81.81±0.95
	FedLabel (Cho, Joshi, and Dimitriadis, 2023)	1	9	20.89±0.51	86.34±0.57	13.96±0.60	78.90±0.68	16.51±0.50	82.28±0.84
	FedCD (Y. Liu, Wu, and Qin, 2024)	1	9	21.27±0.49	86.60±0.50	14.22±0.59	79.30±0.63	16.88±0.47	82.54±0.81
	FedRVR	1	9	**22.94±0.53**	**89.37±0.74**	**15.80±0.70**	**80.76±0.63**	**18.24±0.59**	**84.15±1.07**
ISIC	FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017) (ub)	10	0	87.18±0.53	97.66±0.45	83.06±1.62	93.48±0.95	84.32±1.30	95.82±0.74
	FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017) (lb)	1	0	69.77±1.27	85.72±1.01	63.06±0.94	79.48±1.41	68.56±1.03	84.85±1.20
	FedMatch (Jeong, Yoon, Yang, and Hwang, 2020)	1	9	70.26±0.68	85.22±0.70	64.25±1.03	80.60±0.94	67.36±0.80	83.21±0.72
	FedIRM (Q. Liu, Yang, Dou, and Heng, 2021)	1	9	70.87±0.83	84.78±0.86	64.77±1.21	81.25±1.48	67.67±0.92	83.70±0.64
	FedSSL (Fan, Hu, and Huang, 2022)	1	9	71.36±0.75	86.65±0.82	65.37±1.46	81.94±1.30	68.34±1.24	85.63±0.66
	RSCFed (Liang, Lin, Fu, Zhu, and Li, 2022)	1	9	72.37±0.49	88.67±0.63	67.58±0.72	84.06±0.58	70.63±0.88	86.35±0.50
	FedLabel (Cho, Joshi, and Dimitriadis, 2023)	1	9	72.56±0.45	88.95±0.69	67.91±0.70	84.37±0.64	70.97±1.14	86.73±0.63
	FedCD (Y. Liu, Wu, and Qin, 2024)	1	9	72.95±0.50	89.48±0.60	68.41±0.81	84.78±0.59	71.31±0.97	87.40±0.58
	FedRVR	1	9	**74.57±0.60**	**91.48±0.58**	**69.77±0.43**	**87.56±0.82**	**72.95±1.22**	**89.74±0.70**

The result with best performance are highlighted with bold

Table 2 Comparison with partial-labeled FSSL on CIFAR-10

Method	IID		Non-IID w/ Dir(0.5)	
	Acc.	AUC	Acc.	AUC
FedMatch	63.60	89.13	59.21	84.73
FedIRM	64.43	89.95	60.72	85.15
FedSSL	65.14	90.39	61.47	85.77
RSCFed	65.97	91.03	62.84	86.44
FedLabel	66.46	91.50	63.40	86.93
FedCD	66.88	91.95	63.90	87.38
FedRVR	**68.43**	**93.40**	**65.46**	**88.93**

The result with best performance are highlighted with bold

Table 3 Ablation study of the proposed FedRVR on CIFAR-10

RDA (Server)		VR (Client)		IID		Non-IID w/ Dir(0.5)	
IMG	DA	$\mathcal{L}_{DR}$	$\mathcal{L}_{MR}$	Acc.	AUC	Acc.	AUC
				54.31	84.59	51.66	81.03
✓				55.67	85.48	53.27	82.66
	✓			55.35	85.33	52.89	82.38
✓	✓			56.93	87.21	54.15	83.75
✓	✓	✓		58.48	88.87	56.93	85.16
✓	✓		✓	58.10	88.45	56.09	84.52
✓	✓	✓	✓	60.54	90.46	58.24	86.81

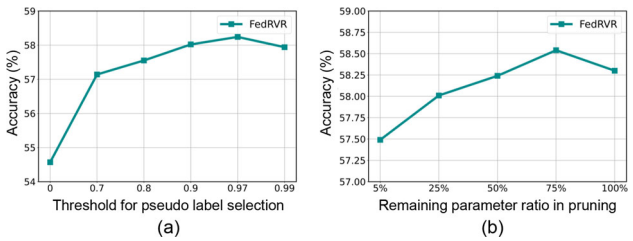

Fig. 4 Threshold of pseudo label selection and ratio of parameter pruning. **a** Threshold for pseudo label selection. **b** Remaining parameter ratio in pruning

5.4 Sensitivity Analysis

To investigate the robustness of each proposed component, including the thresholds for pseudo label selection and parameter pruning, we further conduct detailed sensitivity analysis in this section.

Threshold τ_s for Pseudo Label Selection. To indicate the effect of the threshold τ_s used to pick up reliable predictions to generate pseudo labels, we train FedRVR with different thresholds τ_s from the set $\{0, 0.7, 0.8, 0.9, 0.97, 0.99\}$ on CIFAR-10, Non-IID with Dir(0.5). From Fig. 4a, we observe that the accuracy increases with larger threshold τ_s, while it shows a drop when τ_s is up to 0.99. The accuracy shows the best with the threshold as 0.97, and we therefore use $\tau_s = 0.97$ as the choice in FedRVR. Moreover, the accuracy performs worst when the threshold $\tau_s = 0$, i.e., nominating all predictions as pseudo labels. The reason is that it is hard to train models with many noisy pseudo labels without the filtering strategy.

Threshold τ_p for Parameter Pruning. To study the impact of the threshold τ_p towards parameter pruning when training the relation predictor ϕ, we conduct an ablation study on different thresholds τ_p, which can hold part of parameters via a remaining parameter ratio from the set $\{5\%, 25\%, 50\%, 75\%, 100\%\}$ on CIFAR-10, Non-IID with Dir(0.5). From Fig. 4b, the accuracy achieves the best when remaining 75% of parameters, and the comparable result with the remaining parameter ratio as 50%. For efficient training with a smaller amount of parameters, we adopt the threshold with the remaining parameter ratio as 50% in FedRVR. Moreover, we notice that either too few or too many parameters lead to performance slides, which suggests that choosing a suitable scale of model parameters is important to train the relation predictor ϕ.

5.5 Further Analysis

To understand the robustness of FedRVR, we analyze the extention on label-at-server scenario, the accuracy of relation predictor, the ratio of labeled and unlabeled clients, the trade-off of two regularizations, the communication cost, and the application on standard FL.

Evaluation on Label-at-Server FSSL. To evaluate the proposed FedRVR framework, we follow previous works(Diao,

Table 4 Comparison with label-at-server FSSL on CIFAR-10. "FS" and "PS" refer to fully- and partially-supervised federated learning

Method	IID		Non-IID w/ Dir(0.5)	
	Acc.	AUC	Acc.	AUC
FedAvg (FS)	72.37	92.67	68.15	88.43
FedAvg (PS)	55.42	85.23	52.13	81.55
FedMatch	54.37	86.96	52.63	81.59
+FedRVR	**56.72**	**88.60**	**53.49**	**82.77**
SemiFL	57.83	88.12	55.48	83.44
+FedRVR	**58.34**	**88.95**	**56.18**	**84.30**

The result with best performance are highlighted with bold

Table 5 Comparison with hybrid-labeled FSSL on CIFAR-10

Method	IID		Non-IID w/ Dir(0.5)	
	Acc.	AUC	Acc.	AUC
FedMatch	56.47	86.63	53.35	82.80
FedIRM	57.12	87.20	53.82	83.17
FedSSL	57.29	87.91	54.86	83.67
RSCFed	59.04	88.97	55.90	84.83
FedCD	59.60	89.55	56.46	85.81
FedRVR	**60.86**	**90.91**	**58.75**	**87.22**

The result with best performance are highlighted with bold

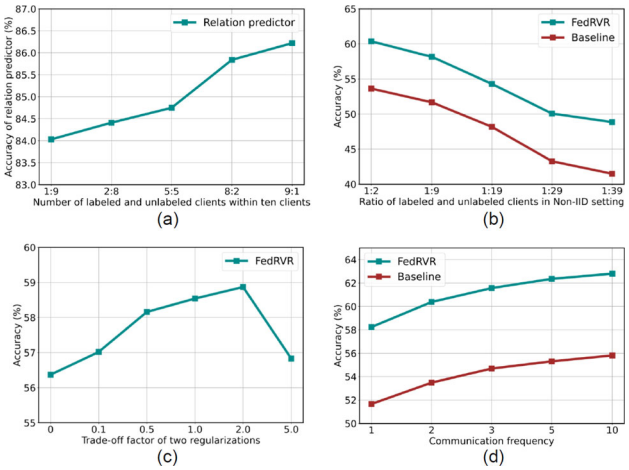

Fig. 5 Detailed analysis. **a** Accuracy of the relation predictor under different labeled clients numbers. **b** Ratio of labeled and unlabeled clients. **c** Trade-off factor λ of two regularizations. **d** Communication frequency

Ding, and Tarokh, 2021; Jeong, Yoon, Yang, and Hwang, 2020) and consider the label-at-server FSSL scenario, where the 10% labeled data are located at the server and the clients only contain remaining 90% unlabeled data. In this scenario, we employ the proposed FedRVR framework on the basis of FedMatch(Gao et al., 2022) and SemiFL(Diao, Ding, and Tarokh, 2021) respectively. From Table 4, with the FedRVR, the accuracy of FedMatch(Jeong, Yoon, Yang, and Hwang, 2020) can increase from 54.37% and 52.63% to 56.72% and 53.49% in IID and Non-IID with Dir(0.5) settings, respectively. The FedRVR can boost the SemiFL(Diao, Ding, and Tarokh, 2021) with 0.51% and 0.70% accuracy in IID and Non-IID with Dir(0.5) settings, respectively. These results verify the benefits of the proposed FedRVR framework on label-at-server FSSL, indicating the effectiveness and robustness of FedRVR across multiple FSSL scenarios.

Evaluation on Hybrid-Labeled FSSL. We further consider another setting termed hybrid-Labeled FSSL, in which there is 1 labeled client, 1 semi-labeled clients with 10% labeled samples and 90% unlabeled samples, and 8 unlabeled clients. We conduct the experiments on CIFAR-10 in both IID and No-IID with Dir(0.5) FL settings to observe the ability the different models and study the impact of semi-labeled clients. From the Table 5, the proposed FedRVR exhibits 60.86%, 58.75% accuracy scores in IID and No-IID settings, respectively, superior over other competitors, which demonstrate that our FedRVR can be generalized on hybrid-labeled FSSL by efficiently utilizing both labeled and unlabeled samples.

Analysis on relation predictor. To investigate the robustness and generalization of the proposed relation predictor, we calculate the accuracy of the relation predictor at the server on multiple FL settings with different ratios and numbers of labeled clients respectively. In Fig. 5a, on test set of ISIC, the accuracy is 84.03% in 1:9 labeled and unlabeled client FL setting under Non-IID with Dir(0.5), and the accuracy shows a slight increase when the predictor is trained by more labeled clients. Additionally, the relation predictor consisting of 3 FC layers brings 3.7% extra training time at FL setup with 1 labeled client and 9 unlabeled clients, which can be negligible towards extra computation cost. These results verify the effectiveness and robustness of the relation predictor and suggest that the predictor can be well applied on unlabeled clients even though it is trained by the labeled client.

Analysis on the ratio of labeled and unlabeled clients. We compare multiple ratios of labeled and unlabeled clients on CIFAR-10. From Fig. 5b, we have consistent observations with the performance decrease when diminishing the ratio of labeled clients. As the ratio reduces from 1:2 to 1:39, the accuracy of FedRVR drops significantly by 11.52% in Non-IID case. Although it brings a performance deterioration, FedRVR still surpasses the baseline with a clear margin. This observation confirms that the proposed method is leading to relatively optimal training and is less vulnerable to scenarios with dominant unlabeled clients.

Analysis on trade-off factor λ. We study the sensitivity of the FedRVR towards the trade-off factor λ of two regularizations $\mathcal{L}_{\text{DR}}$ and $\mathcal{L}_{\text{MR}}$ on CIFAR-10 under Non-IID with Dir(0.5). From Fig. 5c, the performance grows with the trade-off factor λ increasing from 0.0 to 2.0, while it plunges when λ is larger than 2.0. These suggest that both data-guided and model-guided regularizations lead to beneficial constraints

Table 6 Comparison with state-of-the-art FSSL algorithms on CIFAR-100 with WRN-28-2 as the backbone. "FS" and "PS" refer to fully- and partially-supervised federated learning

Method	IID		Non-IID w/ Dir(0.5)	
	Acc.	AUC	Acc.	AUC
FedAvg (FS)	65.48	95.88	63.50	92.16
FedAvg (PS)	33.53	85.28	30.07	83.67
FedMatch	35.26	87.52	31.73	85.66
FedIRM	36.61	88.24	32.57	86.45
FedSSL	37.95	88.94	33.46	87.72
RSCFed	38.44	89.62	34.29	88.51
FedLabel	38.80	89.96	34.51	88.67
FedCD	39.11	90.22	34.83	88.75
FedRVR	**40.17**	**90.67**	**36.71**	**89.15**

The result with best performance are highlighted with bold

Table 7 Study on scalability of different client numbers

Method	Num.		IID		Non-IID w/ Dir(0.5)	
	L	UL	Acc.	AUC	Acc.	AUC
FedMatch	1	9	54.37	85.11	49.67	80.14
FedMatch	10	90	47.64 (↓ 6.73)	78.70 (↓ 6.41)	40.09 (↓ 9.58)	70.50 (↓ 9.64)
RSCFed	1	9	57.69	87.52	52.18	82.95
RSCFed	10	90	51.44 (↓ 6.25)	80.90 (↓ 6.62)	44.50 (↓ 7.68)	73.63 (↓ 9.32)
FedRVR	1	9	59.36	89.40	55.66	84.59
FedRVR	10	90	**53.85** (↓ 5.51)	**84.28** (↓ 5.12)	**49.27** (↓ 6.34)	**76.95** (↓ 7.64)

The result with best performance are highlighted with bold

with the appropriate range of λ, while too strong or too weak of the model-guided regularization may degrade the FSSL performance.

Privacy and communication cost. Following standard privacy-preserving protocols in FL (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017; Y. Huang et al., 2021; L. Zhang, Luo, Bai, Du, and Duan, 2021), FedRVR maintains private data at clients and avoids privacy leakage. Moreover, two global models contain global information towards the clients and the relation predictor involves the knowledge of model relation, excluding identifiable private information. Therefore, FedRVR can ensure privacy-preserving issue during the communication process. In each communication round of our framework, each client sends one local model to the server and the server sends two global models to each client (3 models totally), additionally, each labeled client uploads the lightweight relation predictor to the server (can be ignored). In total, compared with previous FL methods (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017; Wang, Yurochkin, Sun, Papailiopoulos, and Khazaeni, 2020) that send a single model between the server and each client (2 models totally), our FedRVR suggests nearly 1.5 times communication cost than them. To maintain a fair comparison with the comparable communication cost for different methods, we reduce the communication frequency with a $\frac{1}{2}$ time ratio. Specifically, our FedRVR performs each communication every two local training epochs and other methods performs communication every one training epoch. In this manner, our framework yields less total communication cost than other competitors. Note that all models are trained by the same epochs under the fair protocol. From Fig. 5d, with the less communication cost, FedRVR still exhibits noticeably superior performance over the baseline, which verify the efficiency of the FedRVR.

Analysis on Backbone Scalability. To investigate the influence of the backbone scalability for our FedRVR, we replace the small network Simple-CNN with a large one, i.e., Wide-ResNet-28-2 (WRN-28-2) (Zagoruyko, 2016), as the backbone to train FedRVR and other competitors on CIFAR-100 in IID and Non-IID with Dir(0.5) settings. From Table 6, FedRVR delivers 40.17% and 36.71% accuracy in IID and Non-IID, respectively, outperforming other methods. These results prove that the proposed FedRVR framework is robust and superior towards the scalable backbone.

Analysis on Client Number Scalability. To study the scalability of the proposed FedRVR framework, we enlarge the client numbers with 10 labeled clients and 90 unlabeled clients to train the model on CIFAR-10 dataset with IID and Non-IID, Dir(0.5) FL settings respectively. Considering the training efficiency issue, we employ a smaller CNN network containing one convolution layer and one fully-connected layer for each client model. From the following Table 7, as

Table 8 Results comparison of the relation-guided directional aggregation (RDA) on standard FL on CIFAR-10

Method	IID		Non-IID w/ Dir(0.5)	
	Acc.	AUC	Acc.	AUC
FedAvg	71.80	94.62	68.15	92.33
FedProx	72.66	95.11	69.40	92.94
FedAMP	73.24	95.73	70.55	93.60
RDA	**73.91**	**96.26**	**71.53**	**94.57**

The result with best performance are highlighted with bold

the client number of labeled and unlabeled clients increasing from 1:9 to 10:90, our FedRVR suggests 5.51%, 6.34% accuracy drops and 5.12%, 7.64% AUC drops in IID and Non-IID settings, respectively, which are smaller than other methods. In the scalable scenario of 10:90 labeled and unlabeled clients, FedRVR achieves 53.85%, 49.27% accuracy in two FL settings, outperforming other previous approaches, which verifies that our framework can be well scalable to a high number of clients.

Extension to standard FL. The proposed relation-guided directional aggregation can yield the superior global model, which can also improve the standard FL. We further apply it on standard FL with ten labeled clients in CIFAR-10. From Table 8, with the relation-guided directional aggregation, the model achieves 73.91% and 71.53% accuracy in IID and Non-IID with Dir(0.5), surpassing other state-of-the-art comparisons. This indicates that the relation-guided directional aggregation can also be applied in standard FL to achieve competitive results.

6 Qualitative Results

6.1 Visualization of the Feature Distributions

The t-SNE(Van der Maaten and Hinton, 2008) visualizations of the feature distributions are presented in Fig. 6. The features are extracted from a well-trained framework on CIFAR-10, Non-IID with Dir(0.5), and we reduce the feature dimension to 2 for clearer illustration. From Fig. 6a–d, FedRVR shows the best ability to distinguish different categories, and retains the compactness of intra-class distribution. These results verify that the proposed FedRVR framework can encourage the features belonging to the same class to be more similar, and the features from different categories to be more discriminative.

6.2 Visualization of the Confusion Matrix

The confusion matrix of model prediction can reflect the classification performance, and represent the confirmation bias (Sohn et al., 2020) via inter-class correlation of predictions. We therefore analyze the confusion matrices across CIFAR-10, Non-IID with Dir(0.5) to perform the classification performance and the mitigation effect of confirmation bias. From Fig. 7a–d, compared with FedAvg (McMahan, Moore, Ramage, Hampson, and y Arcas, 2017), FedIRM (Q. Liu, Yang, Dou, and Heng, 2021) and RSCFed (Liang, Lin, Fu, Zhu, and Li, 2022), the confusion matrix of our FedRVR exhibits more active on diagonal entries (marked in red), implying more accurate for classification prediction. Moreover, the confusion matrices obtained by FedAvg, FedIRM and RSCFed models show that some data tend to be mis-classified into other categories and generate noisy pseudo labels. The proposed FedRVR framework can handle this problem, as shown in Fig. 7d, which delivers less inter-class confusion and can perform more reliable pseudo labeling.

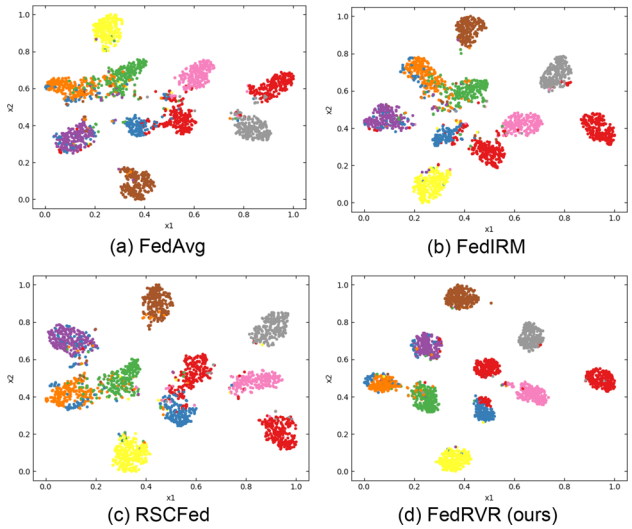

Fig. 6 Qualitative comparison of t-SNE visualization among FedAvg, RSCFed and FedRVR. Compared with other methods, the feature distribution of the FedRVR is more compact within each category, and more discriminative across classes

7 Conclusion

In this work, we target at the federated semi-supervised learning task, and present the FedRVR framework, consisting of versatile regularization and relation-guided directional aggregation. In versatile regularization, we propose both the data-guided and model-guided regularizations, and enforce the predictions of the local model to be aligned with two extreme global models, which can offer richer consistency supervision for local training. Moreover, we devise a relation-guided directional aggregation at the server to produce robust global models with a directional model relation graph. As

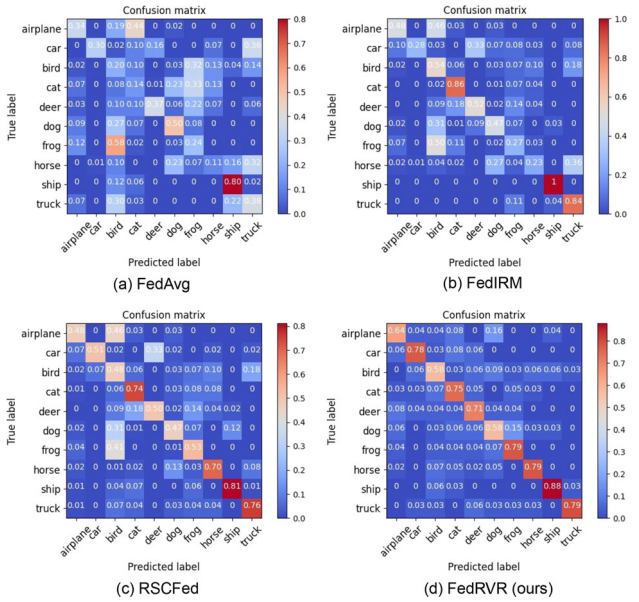

Fig. 7 The confusion matrices among FedAvg, RSCFed and FedRVR on CIFAR-10, Non-IID with Dir(0.5). FedRVR reduces the inter-class confusion and boosts the classification confidence

such, our aggregation relies more on dependable client models to provide a superior global model for FSSL, and further yields an inferior global model to promote the versatile regularization at clients. Extensive experiments on three FSSL benchmarks verify the superiority of the proposed FedRVR over state-of-the-art approaches across various FL settings, and detailed analysis on the proposed component, hyper-parameters and data privacy reflects the robustness of FedRVR.

Acknowledgements This work was supported by the Hong Kong Research Grants Council (RGC) General Research Fund under Grant 14220622 and Innovation and Technology Commission Innovation and Technology Fund ITS/229/22.

Data Availability All datasets used in this work are publicly available. CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton, et al., 2009) are available at https://www.cs.toronto.edu/ kriz/cifar.html. ISIC (Codella et al., 2019) is available at https://challenge.isic-archive.com/data.

References

Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C. (2019). Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *Proc. ICLR* Vol. abs/1911.09785.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Proceeding NeurIPS, 32*, 5050–5060.

Cascante-Bonilla, P., Tan, F., Qi, Y., & Ordonez, V. (2020). Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *Proceeding AAAI, 35*, 6912–6920.

Chen, Y., Guo, Y., Chen, Q., Li, M., Zeng, W., Wang, Y., & Tan, M. (2021). Contrastive neural architecture search with neural architecture comparators. - *Proc. CVPR* pp. 9502–9511.

Chen, Z., Li, W., Xing, X., & Yuan, Y. (2023). Medical federated learning with joint graph purification for noisy label learning. *Med image anal* Vol. 90, pp. 102976–102988. Elsevier.

Chen, Z., Yang, C., Zhu, M., Peng, Z., & Yuan, Y. (2022). Personalized retrogress-resilient federated learning towards imbalanced medical data. *IEEE Transactions on Medical Imaging, 41*, 3663–3674.

Chen, Z., Zhu, M., Yang, C., & Yuan, Y. (2021). Personalized retrogress-resilient framework for real-world medical federated learning. *Proceeding MICCAI, 24*, 347–356.

Cho, Y.J., Joshi, G., &Dimitriadis, D. (2023). Local or global: Selective knowledge assimilation for federated learning with limited labels. *Proc. ICCV* pp. 17087–17096.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., & Gutman, D. others (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368.

Cubuk, E.D., Zoph, B., Shlens, J., & Le, Q.V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. *CVPR Workshops* pp. 702–703.

Diao, E., Ding, J., & Tarokh, V. (2021). - Semifl: Communication efficient semi-supervised federated learning with unlabeled clients. *Proc. NeurIPS* pp. 17871–17884.

Dou, Q., So, T. Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., et al. (2021). Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *NPJ Digital Medicine, 4*, 1–11.

Fan, C., Hu, J., & Huang, J. (2022). - Private semi-supervised federated learning. *Proc. IJCAI* pp. 2009–2015.

Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., & Xu, C- Z. (2022). Feddc: Federated learning with non-iid data via local drift decoupling and correction. *Proc. CVPR* pp. 10112–10121.

Gong, C., Wang, D.,& Liu, Q. (2021). - Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. *Proc. CVPR* pp. 13683–13692.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. - *Proc. CVPR* pp. 770–778.

Hu, Z., Yang, Z., Hu, X., & Nevatia, R. (2021). Simple: Similar pseudo label exploitation for semi-supervised classification. *Proc. CVPR* pp. 15099–15108.

Huang, M., Huang, Z., Li, C., Chen, X., Xu, H., Li, Z., & Liang, X. (2022). Arch-graph: Acyclic architecture relation predictor for task-transferable neural architecture search. *Proc. CVPR* pp. 11881–11891.

Huang, T., Sun, Y., Wang, X., Yao, H., & Zhang, C. (2021). Spatial ensemble: a novel model smoothing mechanism for student-teacher framework. *Proceeding NeurIPS* pp. 15957–15968.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., & Zhang, Y. (2021). Personalized cross-silo federated learning on non-iid data. *Proceeding AAAI* pp. 7865–7873.

Jeong, W., Yoon, J., Yang, E., Hwang, S.J. (2020). Federated semi-supervised learning with inter-client consistency & disjoint learning. arxiv preprint arxiv:2006.12097.

Jiang, M., Yang, H., Li, X., Liu, Q., Heng, P- A., & Dou, Q. (2022). Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. - *Proceeding MICCAI* Vol. 13433, pp. 196–206.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. - *Technical Report.*

Kuo, C- W., Ma, C- Y., Huang, J- B., & Kira, Z. (2020). Featmatch: Feature-based augmentation for semi-supervised learning. *ECCV, 12363*, 479–495.

Laine, S., & Aila, T. (2016). Temporal ensembling for semi-supervised learning. - arxiv preprint arxiv:1610.02242.

Li, J., Xiong, C., & Hoi, S.C. (2021). - Comatch: Semi-supervised learning with contrastive graph regularization. *ICCV* pp. 9475–9484.

Li, M., Li, Q., & Wang, Y. (2023). Class balanced adaptive pseudo labeling for federated semi-supervised learning. *Proceeding CVPR* pp. 16292–16301.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proc. MLSys, 2*, 429–450.

Li, X., Jiang, M., Zhang, X., Kamp, M.,& Dou, Q. (2021). Fedbn: Federated learning on non-iid features via local batch normalization. arxiv preprint arxiv:2102.07623.

Li, X- C., Xu, Y- C., Song, S., Li, B., Li, Y., Shao, Y., & Zhan, D- C. (2022). Federated learning with position-aware neurons. *Proceeding CVPR* pp. 10082–10091.

Liang, X., Lin, Y., Fu, H., Zhu, L., Li, X. (2022). Rscfed: Random sampling consensus federated semi-supervised learning. *Proceeding CVPR* pp. 10154–10163.

Liu, Q., Yang, H., Dou, Q., & Heng, P.-A. (2021). Federated semi-supervised medical image classification via inter-client relation matching. *Proceeding MICCAI, 64*, 325–335.

Liu, Y., Wu, H., & Qin, J. (2024). Fedcd: Federated semi-supervised learning with class awareness balance via dual teachers. *Proceedings AAAI, 38*, 3837–3845.

Long, Z., Che, L., Wang, Y., Ye, M., Luo, W.J.X.H., & Junyu, Ma, F. (2020). Fedsemi: An adaptive federated semi-supervised learning framework. arXiv preprint arXiv:2012.03292.

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. - arxiv preprint arxiv:1608.03983.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B.A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings AISTATS* pp. 1273–1282.

Mendieta, M., Yang, T., Wang, P., Lee, M., Ding, Z., & Chen, C. (2022). Local learning matters: Rethinking data heterogeneity in federated learning. *Proc. CVPR* pp. 8397–8406.

Miyato, T., Maeda, S- I., Koyama, M., Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE TPAMI,41*, 1979–1993.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Proceedings NeurIPS, 32*, 8024–8035.

Sohn, K., Berthelot, D., Li, C-L., Zhang, Z., Carlini, N., Cubuk, E.D., Raffel, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Proc. NeurIPS,33*, 596–608.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *Proceedings ICML, 28*, 1139-1147.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. - Proc. *NeurIPS, 30*, 1195–1204.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *JMLR, 9*, 2579–2605.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D.,& Khazaeni, Y. (2020). Federated learning with matched averaging. arxiv preprint arxiv:2002.06440.

Wen, W., Liu, H., Chen, Y., Li, H., Bender, G., & Kindermans, P-J. (2020). Neural predictor for neural architecture search. *Proc. ECCV,12374*, 660–676.

Xie, Q., Dai, Z., Hovy, E., Luong, M- T., & Le, Q.V. (2020). Unsupervised data augmentation for consistency training. *Proc. NeurIPS,33*, 6256–6268.

Xie, Q., Luong, M- T., Hovy, E., & Le, Q.V. (2020). Self-training with noisy student improves imagenet classification. *Proc. CVPR* pp. 10687–10698.

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y- F., Sun, B., & Jin, R. (2021). Dash: Semi-supervised learning with dynamic thresholding. *Proceedings ICML,139*, 11525–11536.

Xu, Y., Wang, Y., Han, K., Tang, Y., Jui, S., Xu, C., & Xu, C. (2021). Renas: Relativistic evaluation of neural architecture search. *Proc. CVPR* pp. 4411–4420.

Yang, Q., Chen, Z., & Yuan, Y. (2023). Hierarchical bias mitigation for semi-supervised medical image classification. *IEEE Trans Med Imag, 42*, 2200–2210.

Yang, Q., Liu, X., Chen, Z., Ibragimov, B., & Yuan, Y. (2022). Semi-supervised medical image classification with temporal knowledge-aware regularization. *Proc. MICCAI, 13438*, 119–129.

Zagoruyko, S. (2016). Wide residual networks. arxiv preprint arxiv:1605.07146.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., & Shinozaki, T. (2021). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. - Proc. *NeurIPS, 34*, 18408–18419.

Zhang, L., Luo, Y., Bai, Y., Du, B.,& Duan, L- Y. (2021). Federated learning for non-iid data via unified feature learning and optimization objective alignment. - *Proc. ICCV* pp. 4420–4428.

Zhang, L., Shen, L., Ding, L., Tao, D.,& Duan, L-Y. (2022). Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. *Proc. CVPR* pp. 10174–10183.

Zhang, Z., Yao, Z., Yang, Y., Yan, Y., Gonzalez, J.E., & Mahoney, M.W. (2020). Benchmarking semi-supervised federated learning. arXiv preprint arXiv:2008.11364.

Zhou, T., Wang, S., & Bilmes, J. (2020). Time-consistent self-supervision for semi-supervised learning. *Proc. ICML, 119*, 11523–11533.

Zhu, M., Liao, J., Liu, J., & Yuan, Y. (2023). Fedoss: Federated open set recognition via inter-client discrepancy and collaboration. *IEEE Trans Med Imag, 43*, 190–202.