FedLTF: Linear Probing Teaches Fine-tuning to Mitigate Noisy Labels in Federated Learning

Shaojie Zhan[*] ZHANSHAOJIE@MAIL.YNU.EDU.CN
Lixing Yu[*] YULIXING@YNU.EDU.CN
Hanqi Chen CHENHANQI@MAIL.YNU.EDU.CN
School of Information Science and Engineering, Yunnan University, Yunnan, China
Tianxi Ji[†] TIJI@TTU.EDU
Department of Computer Science, Texas Tech University, Lubbock, USA

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

The presence of noisy labels has always been a primary factor affecting the effectiveness of federated learning (FL). Conventional FL approaches relying on Supervised Learning (SL) tend to overfit the noise labels, resulting in suboptimal Feature Extractor (FE). In this paper, we exploit models obtained in Self-Supervised Learning (SSL) to mitigate the impact of noisy labels in FL. In addition, we explore two popular methods to transfer to downstream tasks: linear probing, which updates only the last classification layers, and fine-tuning, which updates all model parameters. We empirically observe that, although fine-tuning typically yields higher accuracy than linear probing, in the presence of noise, it is very sensitive to noisy labels and will cause performance degradation. To achieve the best of both worlds (i.e., high accuracy and robustness against noisy labels), we "teach" fine-tuning to control overfitting. In particular, we leverage SSL to obtain a robust FE that is unaffected by noisy labels, and employ linear probing to train the classifiers. The FE and classifiers are integrated to construct a teacher model, which undergoes knowledge distillation to instruct the fine-tuning process of the student model. Extensive experimental evaluations conducted on multiple datasets demonstrate the effectiveness and robustness of our proposed framework against noisy labels in FL, outperforming state-of-the-art methods. The code is available at https://github.com/ss3b3/FedLTF.

Keywords: Federated Learning, Noisy Label, Robustness, Self-Supervised Learning

1. Introduction

Federated Learning (FL), a distributed machine learning paradigm, allows clients to collaboratively learn a global model while preserving data privacy. Existing FL approaches tackling heterogeneity, communication efficiency, and privacy issues (Li et al., 2020) heavily rely on the assumption of high-quality annotations of client data. However, this noise-free assumption poses practical challenges due to the difficulty and cost of manual annotation.

In deep neural network (DNN) architectures, the impact of noisy labels is primarily observed in the representation learning phase rather than affecting the classification process (Zhang and Yao, 2020). We have observed similar problem in the context of FL. In particular, we find out that the noise labels in FL can distort the local models' representation

[*] Equal contribution.
[†] Corresponding author.

learning, hence undermine the feature extractor (FE). To corroborate this, we conducted experiments on the CIFAR-10 dataset using the FedAvg settings. We introduced symmetric noise, which involves randomly mislabeled samples with equal probability across classes, into the datasets of local agents at different rates, i.e., the percentage of mislabeled samples ranges from 0.3 to 0.7. We show the results in Figure 1 by visualizing output of the FE. Clearly, Figure 1 shows that, as the noise rate increases, the attainment of a well-clustered FE poses a formidable challenge for the model.

To alleviate the adverse effects of noisy labels in FL, several studies (Yang et al., 2022a; Chen et al., 2020) have employed auxiliary datasets with perfectly accurate labels to identify noisy samples prior to training. However, acquiring such auxiliary datasets is challenging in practical scenarios. Some other researches (Lu et al., 2023; Fang and Ye, 2022; Fang et al., 2023; Yang et al., 2022a) assume that noisy labels are confined to a small subset of clients and aims to identify and handle those clients before incorporating them into the FL process. However, all these methods are based on supervised learning (SL) and it is impossible to avoid the continuous influence of noisy labels during network training.

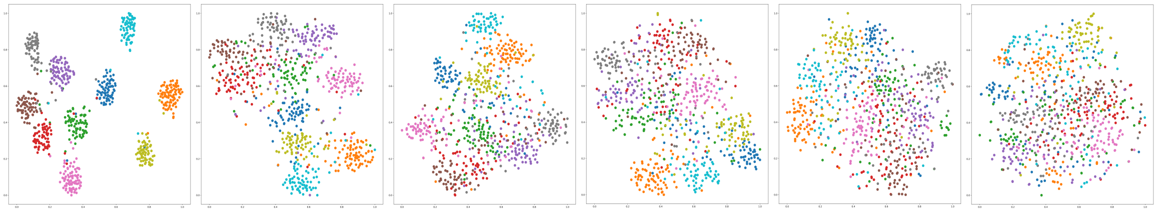

Figure 1: The t-SNE visualization results of the FedAvg under varying noise rates. Left to right: no noise, noise rates of 0.3 to 0.7, respectively. Clearly, it is challenging for a model to acquire a well-clustered FE as the noise rate increases.

A few studies (Yao et al., 2021; Zheltonozhskii et al., 2022) have attempted to address the noisy labels within centralized learning scenarios by applying Self-Supervised Learning (SSL), especially Contrastive Learning (CL), to initialize representations. Similarly, some prior studies (Zhuang et al., 2021, 2022) have explored SSL in FL, and these studies verified the feasibility of using SSL in FL. Additionally, Self-Supervised representations exhibit improved robustness to class imbalance compared to supervised representations, as they capture both label-relevant features and intrinsic properties of the input distribution (Liu et al., 2022). This robustness is particularly suitable for the common non-i.i.d. scenarios encountered in FL.

Besides obtaining well-performing FE through Self-Supervised Contrastive Learning (SSCL), the process of transferring FE to downstream tasks is inevitably affected by the presence of noisy labels. Typically, two widely employed methods are used for further training the classifier: 1) linear probing, which freezes the FE and only trains the classifier, and 2) fine-tuning, which further trains the entire model, including both the FE and classifier. We have also observed a significant performance disparity between these two methods in FL as shown in Figure 2. The former demonstrates higher resilience to noise labeling and achieves superior performance, while the latter is susceptible to overfitting noise labels, resulting in degraded performance. We thoroughly analyze these approaches and provide the experimental results in Section 3, demonstrating the aforementioned disparities.

Targeting the disruptive impact of noise labels on FE performance in SL methods and considering the varying training outcomes of linear probing and fine-tuning under noise labels, we propose FedLTF, a simple yet effective multi-Stage FL framework designed to address the challenges of training on noisy labeled data in FL. Specifically, our framework divides the training process into three stages. In Stage 1, we employ SSCL to train the FE, avoiding the performance degradation caused by the traditional SL-based FL framework where the local model's FE overfits to noise labels. In Stage 2, we utilize linear probing to train the classifier, aiming to minimize the influence of noise labels to the greatest extent possible compared to fine-tuning. Lastly, in Stage 3, we combine the obtained FE and classifier to create a teacher model. This teacher model is then used to conduct knowledge distillation on the local student model, controlling the fine-tuning process of it, thereby ensuring that the student model learns accurate data feature representations and is resilient to the impact of noise labels. Our approach aims to achieve well-performing local models and ultimately enhance the overall performance of the global model in FL when dealing with noisy labeled data. Our contributions can be summarized as follows:

- We conducted preliminary experiments and found empirical evidence that, similar to centralized learning scenarios, the presence of noisy labels in FL hinders the model's ability to perform effective representational learning, resulting in the failure to obtain a well-performing FE.
- We propose a multi-stage FL framework that ensures privacy preservation without auxiliary datasets. Our framework leverages SSCL and individually trained classifiers through linear probing to obtain a teacher model. Subsequently, the local models are fine-tuned using the teacher model, integrating knowledge distillation and variance regularization.
- We validate the efficacy of our approach through comprehensive experiments conducted on four benchmark datasets, considering a range of noise levels, and compare the results with state-of-the-art techniques. Furthermore, we conduct thorough ablation experiments to ascertain the indispensability of the different components in our proposed method.

2. Related Work

2.1. Noisy Label Learning

Research on noisy label learning has focused on centralized model scenarios. Prior studies have presented various frameworks and algorithms for the detection of noisy labels. Notably, the studies (Hu et al., 2023; Li et al., 2023) introduce a Weibull mixture model-based approach and apply the Stochastic Featured Averaging (SFA) method to identify noisy samples. In parallel, Gui et al. (2021); Xiao et al. (2023) propose a Small-Loss Criterion-based mechanism and a matched high confidence selection technique, respectively, to identify clean data samples. In addition, various studies have addressed the challenge of noisy labels by proposing robust loss functions. For instance, Wang et al. (2019); Zhou et al. (2021) propose modifications or enhancements to the loss function to augment model performance. These approaches involve introducing regularization techniques or additional terms to reformulate the loss function, thereby promoting more accurate and diverse predictions. By doing so, effectively alleviating the impact of overfitting and insufficient learning induced by noisy labels. There are also studies that propose methods to correct noisy labels, such as Lu and He (2022), which proposes updating the original noise labels using the overall prediction formed by the exponential moving average of the network output. Additionally, Li et al. (2019) designs a method for generating pseudo-labels for potentially noisy samples. However, in FL, privacy concerns and limited data per client may hinder the effectiveness of applying these methods designed for centralized models.

2.2. Federated Noisy Label Learning

Federated Noisy Label Learning (FNLL) is an emerging research topic with pioneering works that focus on designing training methods for mitigating the impact of noisy labels in FL. FedCorr, FedNoiL and FedNoRo (Xu et al., 2022; Wang et al., 2022; Wu et al., 2023), propose to identify the clean clients for training, and relabel the noisy clients with the obtained model or uses knowledge distillation and a distance-aware aggregation function together to perform FL model updating. However, such methods depending on the stringent assumption that the clients can be classified as clean and noisy. In studies as Fang and Ye (2022); Fang et al. (2023); Yang et al. (2022a), the individual clients of FL are heterogeneous, and there exists clean label data in the server of FL to cope with the noisy labels present in the individual clients. However, this scenario is not practical in real-world FL tasks. FedLSR (Jiang et al., 2022) prevents local models from overfitting noisy labels by minimising the difference in model outputs on the original and augmented data through self-distillation. RoFL (Yang et al., 2022b) mitigates noise by exchanging centroids of local class features and forming clean global class features. Overall, previous studies on noisy labels in FL have predominantly relied on conventional SL approaches, making it challenging to entirely mitigate the impact of noisy labels on the model across all methods employed.

3. Proposed Method

In this section, we delineate the problem by providing a formal definition. Subsequently, we elucidate our underlying motivation by performing a preliminary experiment. Finally, we expound upon our FedLTF framework in a comprehensive manner, providing detailed explanations and insights. The FedLTF framework is illustrated in Figure 4.

3.1. Preliminaries

Settings and Notations. We consider a FL system with K clients, equipped with a neural network (NN) ϕ, having their own datasets denote as $\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^K$. We divide ϕ into two parts: the FE f_{e} and the classifier f_{c}. f_{e} extracts the features of each input sample x and output a d-dimensional vector while the f_{c} employ such features vectors obtained from f_{e} to compute logits that represents the class confidence scores. Consequently, $\phi = \{f_{\mathrm{e}}, f_{\mathrm{c}}\}$. The model parameters of an arbitrary local client k at round r is represented as W_r^k, while the global model is denoted as W_r.

Basic FedAvg Algorithm. In conventional FedAvg (McMahan et al., 2017),during round r, the central server disseminates the global model W_r to all participating clients. Each client subsequently updates their respective local model by applying their local dataset $\mathcal{D}^k, k = 1, ..., K$

$$W_{r+1}^k \leftarrow W_r^k - \eta \nabla_W \ell(W_r; \mathcal{D}^k). \tag{1}$$

Following the local model updates, the clients transmit their respective local models back to the central server, which conducts model aggregation to obtain the global model for the subsequent round, i.e., round $r+1$

$$W_{r+1} = \sum_k \frac{|\mathcal{D}^k|}{\sum_k |\mathcal{D}^k|} W_{r+1}^k. \tag{2}$$

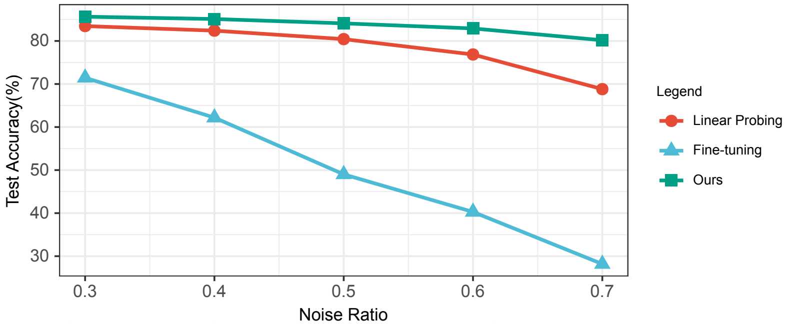

Figure 2: The performance of linear probing and fine-tuning is evaluated under varying noise rates in the CIFAR-10 dataset. It becomes evident that as the noise level increases, linear probing exhibits greater resilience compared to fine-tuning.

3.2. Empirical Finding and Motivation

SSL has garnered significant attention in the domain of addressing noisy labels (Yao et al., 2023; Ghosh and Lan, 2021). Additionally, SSL-based CL, namely SSCL, has proven to be equally applicable in the context of FL (Zhuang et al., 2021, 2022). Empirical evidence substantiates that the SSCL approach can effectively tackle the combined challenge of FL under noisy labels.

In our preliminary FL setup, we utilize ResNet18 as the FE of the clients' models f_e and the global model f_e^g, and we train them using the BYOL (Grill et al., 2020). The CIFAR-10 dataset is evenly partitioned among 20 clients for training. We apply SSCL to locally train f_e on the 20 clients and aggregate them for f_e^g at the central server. This f_e^g is then broadcasted back to update f_e as the initial FEs. Each local model is formed by adding a classifier, i.e., f_c to f_e. Afterward, we train $\phi = \{f_e, f_c\}$ using both linear probing (which trains only the classifier f_c) and fine-tuning (which trains the entire model ϕ). Symmetric noise is introduced to all clients' data, and we increase the noise level from 0.3 to 0.7.

The experimental findings in Figure 2 demonstrate that when noisy labels are present in the data, the aggregated global model derived from clients trained with linear probing shows lower susceptibility to the influence of noise labels compared to the model aggregated from locally fine-tuned models.

On the contrary, fine-tuning yields better accuracy than linear probing under noise-free datasets since the FE can learn better feature representation (Kumar et al., 2021). Considering both the existed conclusion and our experimental results, we argue that the fine-tuning

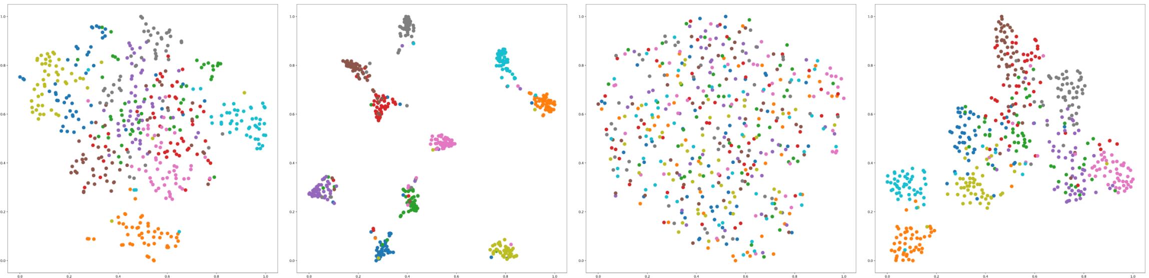

Figure 3: The t-SNE visualization results are compared among different approaches under a noise rate of 0.7 in the CIFAR-10 dataset. The visualizations, from left to right, depict SSL, fine-tuning with clean data, fine-tuning with noisy data and the proposed method.

process corrupts the FE acquired by SSL due to noisy labels. Figure 3 further substantiates our argument and demonstrates the superiority of our method's capability to mitigate the impact of noisy labels during fine-tuning, resulting in superior feature representations compared to SSL and direct fine-tuning.

3.3. Stage 1: Federated Contrastive Representation Learning

At this stage, we use SSL for FL training so that no labels are required. In this study, the BYOL method (Grill et al., 2020) was employed. Following the BYOL settings, the model consists of two structurally identical branches: the online network and the target network, whose model parameters are denoted as W_o^k and W_t^k, respectively. Both the online network and the target network consist of an FE and a projection head.

Local Update. In every communication round, each client, such as client k, receives the identical global model W^g, which consists of global encoder W_o^g and global predictor W_p^g, respectively, from the server. Each clients initial its W_o^k and W_p^k to the received W_o^g and W_p^g, correspondingly. Each round, the client updates its local W_o^k and W_p^k. An image $x \sim \mathcal{D}^k$ sampled uniformly from private dataset $\mathcal{D}^k$. After image augmentation two images x_1 and x_2 are obtained. x_1 is input to the online network to get output z_o and x_2 is input to the target network to get output z_t. Then we output a prediction $W_p^k(z_o)$ of z_t and ℓ_2-normalize both $W_p^k(z_o)$ and z_t to $\overline{W_p^k}(z_o) \triangleq W_p^k(z_o)/\|W_p^k(z_o)\|_2$ and $\overline{z_t} \triangleq z_t/\|z_t\|_2$. Finally, the BYOL loss is defined as,

$$L_{BYOL} \triangleq \|\overline{W_p^k}(z_o) - \overline{z_t}\|_2^2 = 2 - 2 \cdot \frac{\langle W_p^k(z_o), z_t\rangle}{\|W_p^k(z_o)\|_2 \cdot \|z_t\|_2}, \tag{3}$$

After the W_o^k updating, BYOL employs Exponential Moving Average (EMA) on it, updating the W_t^k in each small batch as: $W_\mathrm{t}^k = mW_\mathrm{o}^k + (1-m)W_\mathrm{o}^k$, where m represents the momentum value conventionally set to 0.99.

Model Communication. When clients k finish the local update, the newly obtained W_o^k is uploaded to the central server, alone with the updated W_p^k for the global model W^g updating, together with other clients' models. As W_t^k is updated through EMA in BYOL, there is no necessity to transmit W_t^k to the server. Instead, at the commencement of each round, the parameters of W_t^k can be initialized to the new round's initial model W_o^k. Upon receiving the online networks and predictors from clients, the server conducts aggregation using Equation (2) to acquire an updated global model W_{o}^{g} and W_{p}^{g}. Subsequently, the server broadcasts the new W_{o}^{g} (the new W_{e}^{g} and W_{h}^{g}) and W_{p}^{g} obtained through aggregation to the clients for next round of updating.

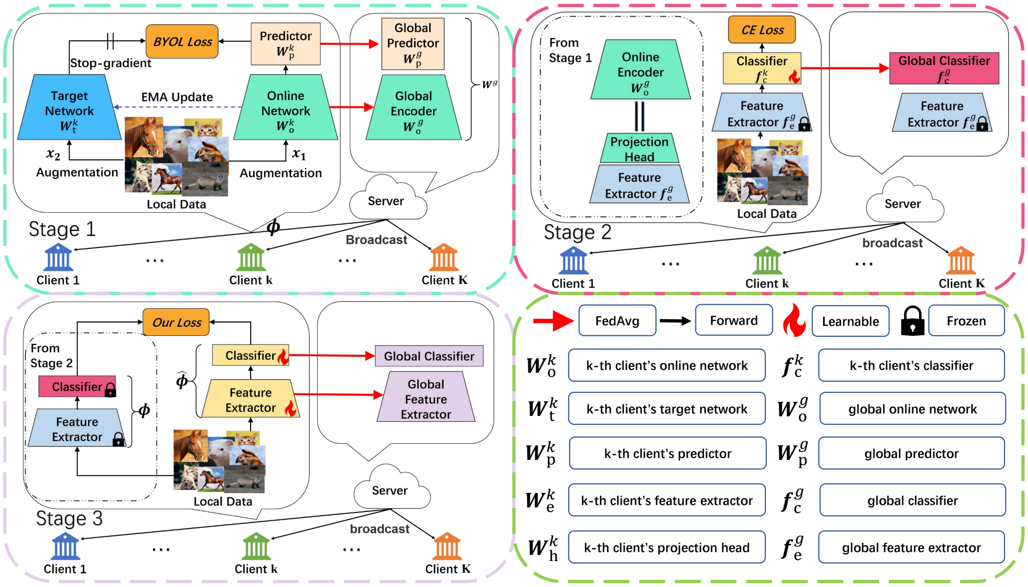
Stage 1
BYOL Loss
Stop-gradient
Predictor W_{p}^{k}
Target Network W_{t}^{k}
EMA Update
Online Network W_{o}^{k}
x_2
Augmentation
x_1
Augmentation
Local Data
Global Predictor W_{p}^{g}
Global Encoder W_{o}^{g}
W^{g}
Server
ϕ
Broadcast
Client 1
Client k
Client K
Stage 2
From Stage 1
Online Encoder W_{o}^{g}
Projection Head
Feature Extractor f_{e}^{g}
CE Loss
Classifier f_{c}^{k}
Feature Extractor f_{e}^{g}
Local Data
Global Classifier f_{c}^{g}
Feature Extractor f_{e}^{g}
Server
broadcast
Client 1
Client k
Client K
Stage 3
Our Loss
From Stage 2
Classifier
Feature Extractor
ϕ
$\hat{\phi}$
Classifier
Feature Extractor
Local Data
Global Classifier
Global Feature Extractor
Server
broadcast
Client 1
Client k
Client K
FedAvg
Forward
Learnable
Frozen
W_{o}^{k} k-th client's online network
f_{c}^{k} k-th client's classifier
W_{t}^{k} k-th client's target network
W_{o}^{g} global online network
W_{p}^{k} k-th client's predictor
W_{p}^{g} global predictor
W_{e}^{k} k-th client's feature extractor
f_{c}^{g} global classifier
W_{h}^{k} k-th client's projection head
f_{e}^{g} global feature extractor

Figure 4: The framework of the proposed method. Our method proceeds in the order of Stage 1, Stage 2, and Stage 3. Irrespective of the Stage, in each round clients send updated local models to the server, and the server sends the aggregated global model back to clients.

3.4. Stage 2: Federated Linear Probing

In the second stage, each local client trains a classifier f_{c} using linear probing. Specifically, after obtaining and broadcasting W_{o}^{g} to each client, the parameters of W_{e}^{g} are employed to initialize the local FE parameters W_{e}^{k}. Different from fine-tuning, each client freeze the parameters of the W_{e}^{k} and randomly initialize a classifier f_{c}, following the linear evaluation protocol (Kolesnikov et al., 2019; Grill et al., 2020). All local classifiers are consequently trained using their own datasets separately. And the trained local f_{c}s are aggregated following Equation (2) for obtaining a global f_{c}^{g}. Note that at this stage, the cross-entropy loss function is exclusively utilized for the purpose of classifier training. At this time we obtained a complete network $\phi = \{f_{\mathrm{e}}^{g}, f_{\mathrm{c}}^{g}\}$.

3.5. Stage 3: Fine-tuning Under Noisy Label

Currently, we have a usable network $\phi = \{f_{\mathrm{e}}^{g}, f_{\mathrm{c}}^{g}\}$. However, the self-supervised learning of FE f_{e}^{g} in Stage 1 may be sub-optimal in representation learning for clean data, despite its robustness to noisy labels. Fine-tuning the network is necessary (Mahajan et al., 2021; Raffel et al., 2019; Xie et al., 2020). Meanwhile, our previous findings (e.g., Figure 2) show that directly fine-tuning on dataset with noisy labels yields worse results than linear probing. However, fine-tuning excels in clean data. Such discrepancy arises from overfitting to noisy labels, distorting the feature representation, especially in the FE, leading to inferior performance.

To address this question, we utilize ϕ as the teacher model to facilitate knowledge distillation on the fine-tuning model $\hat{\phi}$, with the aim of mitigating the issue of overfitting to noisy labels during direct fine-tuning. In detail, the knowledge distillation serves as a regularization term, preventing the feature representation of $\hat{\phi}$ from drifting too far away.

More specifically, we integrate the hard label supervision, knowledge distillation and variance regularization to form our teacher-student fine-tuning method. It enables the client to leverage both correctly and noisily labeled data for optimizing the feature representation of $\hat{\phi}$, thereby achieving enhanced performance surpassing that of its teacher model ϕ.

Hard Label Supervision. We apply the standard cross-entropy loss L_{ce} for executing the hard-label supervision.

$$L_{ce}(y, \hat{\phi}(x)) = -\sum_i y_i \log(\hat{\phi}(x_i)), \tag{4}$$

where y is the label (possibly noisy), x is the sample of the image, and $\hat{\phi}$ is the student model being fine-tuned.

Knowledge Distillation. Knowledge distillation (Hinton et al., 2015) is conducted to transfer knowledge acquired in teacher model ϕ to the student model $\hat{\phi}$, giving the right direction to the fine-tuning, ensuring $\hat{\phi}$ is as effective as ϕ.

Soft labels $y_i^{(t)}$ are derived from the teacher model ϕ by utilizing the output logits,

$$y_i^{(t)} = \frac{\exp(\phi(x_i)/T)}{\sum_j \exp(\phi(x_j)/T)}. \tag{5}$$

Similarly, the logits output of the student model $\hat{\phi}$ undergoes the same process,

$$y_i^{(s)} = \frac{\hat{\phi}(x_i)/T)}{\sum_j \exp(\hat{\phi}(x_j)/T)}. \tag{6}$$

The probabilistic prediction vectors $y_i^{(s)}$ and $y_i^{(t)}$ are obtained, where T represents the temperature controlling the softness of the logits, empirically set to 2. The knowledge distillation loss L_{kd} can be expressed as:

$$L_{kd}(y_i^{(t)}, y_i^{(s)}) = T^2 \sum y_i^{(t)} \log\left(y_i^{(t)}/y_i^{(s)}\right). \tag{7}$$

Variance Regularization. We employ the variance regularization (Zheng and Yang, 2021) to determine whether a given sample is noisy or clean by measuring the variance through estimating the uncertainty in the predictions of the student model $\hat{\phi}$ and the teacher model ϕ. The variance can be approximated as:

$$Var(\hat{\phi}(x_i)) \approx \mathbb{E}[(\hat{\phi}(x_i) - \phi(x_i))^2], \tag{8}$$

where the output of ϕ is applied to replace the true label y_i since it is difficult to acquire in practice FL.

Furthermore, the KL-divergence between the predictions of the two models is utilized as the measure of variance,

$$D_{kl} = \mathbb{E}[\hat{\phi}(x_i)\log(\frac{\hat{\phi}(x_i)}{\phi(x_i)})]. \tag{9}$$

The value obtained from Equation (9) will be large, if a substantial disparity exists between the predictions of ϕ and $\hat{\phi}$, indicating a high likelihood that the sample is noisy.

Next, we apply regularization to the standard cross-entropy loss L_{ce} by incorporating the approximate variance as:

$$L_{rect} = \mathbb{E}[\exp(-D_{kl})L_{ce} + D_{kl}]. \tag{10}$$

When D_{kl} assumes a larger value, we refrain from imposing penalties. To avoid the model consistently favoring larger D_{kl} values, we introduce a regularization term to mitigate D_{kl}.

Ultimately, we construct the overall training loss for the fine-tuning in Stage 3 expressed as Equation (11):

$$L = \alpha L_{rect} + \beta L_{kd}, \tag{11}$$

where L_{rect} is the hard label supervision after Variance Regularization, L_{kd} is vanilla Knowledge Distillation loss, while α and β are hyperparameters used to adjust the weights between L_{rect} and L_{kd} obtained through training.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on four artificially corrupted datasets MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), CIFAR-10/100 (Krizhevsky et al., 2009), and one real-world dataset Clothing1M (Xiao et al., 2015).

Noise Setting. For artificially corrupted datasets, we incorporate two commonly employed types of synthetic noise, i.e., symmetric noise and asymmetric noise.

- **Symmetric Noise**: Each class's label has an equal probability of flipping to another class's label.
- **Asymmetric Noise**: Labels in some classes are flipped to labels in a set of similar classes. For MNIST, flipping 2 → 7, 3 → 8, 5 ↔ 6. For Fashion-MNIST, flipping T-SHIRT → SHIRT, PULLOVER → COAT, SANDALS → SNEAKER. For CIFAR-10, TRUCK → AUTOMOBILE, BIRD → AIRPLANE, DEER → HORSE, CAT ↔ DOG. For CIFAR-100, each category flips to the next category in the same super-category.

Baselines. We conduct a comparative analysis of our method against the following baseline approaches: FedAvg (McMahan et al., 2017), Co-teaching (Han et al., 2018), Symmetric CE (Wang et al., 2019), FedLSR (Jiang et al., 2022), FedCorr (Xu et al., 2022) and FedNoRo (Wu et al., 2023).

Implementation Details. All experiments are implemented, using PyTorch-2.0 , on Nvidia GeForce RTX4090 GPUs. We use ResNet18 as the base model f_{e} to train and test on artificially corrupted dataset. To ensure a fair comparison, we also utilize the same model for the other baseline methods. We set up a total of 20 clients, all of which participate in training during each communication round. Each client performs local training for 5 epochs per communication round.

Table 1: Top-1 test accuracy on MNIST dataset and FMNIST dataset with different noise levels.

Dataset	**Method**	Test Accuracy(%)								
	Noise Type	Symmetric					Asymmetric			Avg.
	Noise Ratio	0.30	0.40	0.50	0.60	0.70	0.20	0.30	0.40	
MNIST	FedAvg (McMahan et al., 2017)	90.83	84.14	78.70	69.55	58.64	96.46	92.47	87.19	82.25
	Symmetric CE (Wang et al., 2019)	92.32	86.24	73.20	60.48	47.43	95.89	92.94	86.98	79.44
	Co-teaching (Han et al., 2018)	98.97	98.87	98.64	98.10	97.58	98.70	98.42	85.22	96.81
	FedLSR (Jiang et al., 2022)	99.24	99.02	98.77	98.35	98.15	99.39	79.94	79.32	94.02
	FedCorr (Xu et al., 2022)	98.57	97.70	97.42	95.96	95.26	99.01	98.46	95.47	97.23
	FedNoRo (Wu et al., 2023)	98.84	97.50	96.59	95.45	94.31	99.17	98.68	96.05	97.07
	Ours(Stage 2)	98.14	97.13	96.89	96.08	95.54	96.98	95.18	90.37	95.79
	Ours(Stage 3)	**99.34**	**99.12**	**98.81**	**98.63**	**98.21**	**99.42**	**98.71**	**96.27**	**98.56**
FMNIST	FedAvg (McMahan et al., 2017)	79.66	75.56	69.58	63.47	53.81	85.54	81.48	74.93	73.00
	Symmetric CE (Wang et al., 2019)	81.90	79.03	71.73	62.20	47.95	87.46	83.65	76.94	73.86
	Co-teaching (Han et al., 2018)	91.18	90.86	90.05	88.90	87.18	91.98	89.28	86.24	89.46
	FedLSR (Jiang et al., 2022)	91.39	91.18	90.90	89.99	88.22	83.14	82.45	77.63	86.86
	FedCorr (Xu et al., 2022)	91.60	90.34	88.85	86.22	83.61	91.03	89.30	87.77	88.59
	FedNoRo (Wu et al., 2023)	91.04	90.24	90.17	86.53	84.73	99.17	98.68	96.05	89.10
	Ours(Stage 2)	90.04	89.63	88.46	86.86	82.29	90.44	88.51	82.67	87.36
	Ours(Stage 3)	**91.91**	**91.42**	**91.00**	**90.24**	**89.41**	**92.70**	**91.28**	**90.33**	**91.04**

In Stage 1, we set the batch size to 128, utilize an SGD optimizer with an initial learning rate of 0.03, and employ cosine annealing (Loshchilov and Hutter, 2016) to reduce the learning rate. For Stage 2 and Stage 3, we set the batch size to 512 and utilize a simple single linear layer as the classifier f_{c}. Specifically, in Stage 2, we freeze f_{e} and train the classifier f_{c} in the FL setting, following the Linear evaluation protocol. To prevent overfitting, we employ a relatively large learning rate of 0.1. In Stage 3, we set the learning rate to 0.03 to train the entire model $\hat{\phi}(f_{\mathrm{e}}, f_{\mathrm{c}})$.

4.2. IID Results

MNIST & Fashion-MNIST Results. The experimental outcomes are presented in Table 1, encompassing the MNIST and FMNIST datasets. The findings demonstrate the effective and robust performance of most methods on these datasets (excluding FedAvg and Symmetric CE). Notably, our proposed method attains the highest test accuracy. Furthermore, a comparative analysis of Stage 2 results against alternative methods reveals that solely employing linear probing falls short of surpassing the performance of conventional SL approaches in the context of simple datasets.

CIFAR Results. Our method was further evaluated on CIFAR-10 and CIFAR-100 datasets, as shown in Table 2. Notably, we did not conduct experiments with FedLSR on CIFAR-100 due to its potential limitations with datasets containing numerous categories. When the noise scenario becomes harder (i.e., Symmetric 50%, and Asymmetric 40%), model performance inevitably starts to drop, especially for Symmetric CE. However, our method is still effective and outperforms other methods. And it can be seen that, as the noise level increases, our method always outperforms other methods by a large margin, which demonstrated the superiority in the robustness of our method.

4.3. Non-IID Results

We assessed the efficacy of our method on non-IID data setting(CIFAR-10, Clothing), ensuring comprehensive evaluation. We follow the widely adopted approach (Yurochkin et al.,

Table 2: Top-1 test accuracy on CIFAR-10/100 with different noise levels.

Dataset	Method	Test Accuracy(%)								
	Noise Type	Symmetric					Asymmetric			Avg.
	Noise Ratio	0.30	0.40	0.50	0.60	0.70	0.20	0.30	0.40	
CIFAR-10	FedAvg (McMahan et al., 2017)	52.34	41.32	32.90	25.01	17.68	50.20	47.81	46.95	39.28
	Symmetric CE (Wang et al., 2019)	53.75	44.35	36.84	26.26	18.65	73.24	69.41	64.92	48.43
	Co-teaching (Han et al., 2018)	62.39	60.77	57.70	36.91	25.20	73.07	70.45	54.04	55.07
	FedLSR (Jiang et al., 2022)	73.61	70.14	67.02	62.17	54.95	76.50	70.99	61.20	63.95
	FedCorr (Xu et al., 2022)	73.26	70.66	66.36	60.44	52.62	78.55	72.24	67.01	67.64
	FedNoRo (Wu et al., 2023)	78.12	75.38	72.77	68.05	61.08	81.66	80.29	78.92	74.53
	Ours(Stage 2)	83.46	82.40	80.43	76.86	69.81	83.40	80.44	74.92	78.96
	Ours(Stage 3)	**85.64**	**85.10**	**84.09**	**82.91**	**80.17**	**86.15**	**84.27**	**79.28**	**83.45**
CIFAR-100	FedAvg (McMahan et al., 2017)	16.75	14.12	12.78	10.47	8.13	18.85	16.33	13.02	13.81
	Symmetric CE (Wang et al., 2019)	16.99	13.97	12.63	10.06	8.53	26.14	21.51	16.64	15.81
	Co-teaching (Han et al., 2018)	34.21	31.24	21.87	16.75	11.29	34.19	27.32	22.83	24.96
	FedCorr (Xu et al., 2022)	32.15	27.81	23.60	18.76	12.01	41.12	35.58	28.45	27.43
	FedNoRo (Wu et al., 2023)	38.58	34.93	31.03	24.93	21.85	45.42	40.96	33.17	33.86
	Ours(Stage 2)	55.23	53.17	50.41	48.50	43.50	52.63	46.87	38.68	48.62
	Ours(Stage 3)	**58.43**	**56.27**	**53.03**	**49.87**	**46.21**	**57.78**	**52.75**	**43.45**	**52.22**

2019) to generate non-IID clients. To maintain consistency with other FL methods (Jiang et al., 2022; Xu et al., 2022), we set the parameter α_{Dir} to 0.5 and use a pre-trained ResNet50 for Clothing1M. For Clothing1M, we randomly sample a subset of 32K images from the noisy train dataset and test on 10K images.

CIFAR-10 Results. We employed the linear evaluation protocol (Kolesnikov et al., 2019; Grill et al., 2020) and conducted a semi-supervised learning evaluation to assess the performance of the model obtained in Stage 1. Specifically, for linear evaluation, we froze the model obtained in Stage 1 and trained a new linear classifier using the full dataset (data held by all clients). For the semi-supervised learning evaluation, we fine-tuned the entire model using only 10% of the full dataset. Additionally, we reported the results of fine-tuning using the full dataset. For Stage 2 and Stage 3, we assessed the effectiveness of our method in handling extreme noise environments, including symmetric noise with a noise rate of 0.7 and asymmetric noise with a noise rate of 0.4. The results presented in Table 4 demonstrate that our method achieves superior performance compared to other baselines when applied to non-IID data with highly noisy labels.

Clothing1M Results. Table 5 shows the results on real-world datasets Clothing1M and our method outperforms the other baseline.

Table 3: Top-1 accuracy on CIFAR-10 datasets in different data partitioning

Data Partitioning	Linear Probing(%)	Semi-supervised(%)	Fine-tuning(%)
IID	87.17	76.71	88.24
Non-IID	85.85	74.03	86.35

4.4. Ablation Study

Our ablation study, conducted on the CIFAR-10 dataset with i.i.d partitioning, emphasizes the assessment of our method's three core components: 1) Linear probing before fine-tuning, 2) Knowledge distillation, and 3) Variance regularization. By systematically removing each component individually, we evaluate the consequent performance degradation. Table 6 presents a comprehensive overview of the effects of the components employed in our method. Subsequently, we consolidate key insights into the factors contributing to the effectiveness of FedLTF:

- All components help to improve accuracy.

Table 4: Test accuracy (%) on non-IID CIFAR-10 dataset with extreme noisy labels (Symmetric noise at 0.7, asymmetric noise at 0.4).

Method	Symmetric 0.7	Asymmetric 0.4
FedAvg	22.16	53.17
Symmetric CE	25.41	61.05
Co-teaching	34.86	67.42
FedLSR	45.63	64.53
FedCorr	47.54	59.11
Ours(Stage2)	68.06	64.97
Ours(Stage3)	**77.45**	**77.90**

Table 5: Test Acc on Clothing 1M

Methods	FedAvg	Symmetric CE	Co-teaching	FedLSR	FedCorr	FedNoRo	Ours
Test ACC	70.70	71.60	71.52	71.23	72.45	73.21	73.81

• Linear probing preceding fine-tuning is irreplaceable, as there is an extremely large performance gap between linear probing and fine-tuning in the presence of noisy labels.

• The superior performance of our method stems primarily from knowledge distillation, which is increasingly critical for performance improvement as noise levels increase.

• The reason why the role of Variance Regularization seems to be insignificant is that Variance Regularization serves as an additional refinement when used in conjunction with knowledge distillation. However, in the absence of knowledge distillation(w/o KD), performance improvements can be observed by solely employing Variance Regularization over Fine-tuning with Cross Entropy Loss.

Table 6: Ablation study results on CIFAR-10

Noise Type	Noise Ratio	Linear Probing	Fine-tuning	Ours w/o KD	Ours w/o VR	Ours
Symmetric	0.3	83.46(-2.18)	71.45(-14.19)	83.52(-2.12)	85.19(-0.45)	85.64
	0.4	82.40(-2.7)	62.19(-22.91)	80.10(-5.54)	84.23(-0.87)	85.10
	0.5	80.43(-3.66)	49.01(-34.08)	75.96(-8.13)	83.21(-0.88)	84.09
	0.6	76.86(-6.05)	40.28(-42.63)	67.37(-15.54)	82.10(-0.81)	82.91
	0.7	69.81(-10.36)	28.18(-51.99)	53.21(-26.96)	79.00(-1.17)	80.17
Asymmetric	0.2	83.40(-2.75)	82.14(-4.01)	85.72(-0.43)	85.68(-0.47)	86.15
	0.3	80.44(-3.83)	77.93(-6.34)	83.37(-0.9)	83.71(-0.56)	84.27
	0.4	74.92(-4.36)	72.52(-6.76)	77.07(-2.21)	78.56(-0.72)	79.28

4.5. Other contrastive learning framworks

Table 7 indicates that our approach is not limited to BYOL methods, but can also use other contrastive learning frameworks

Table 7: Comparison of different SSL methods on CIFAR-10 with 50% Symmetric Noise

Methods	SimCLR	SimSiam	MoCoV1	MoCoV2	BYOL
Ours(Stage 2)	77.94	76.13	78.42	79.81	80.43
Ours(Stage 3)	83.71	80.03	83.21	83.12	84.09

4.6. Combination with other methods

Furthermore, our method can be combined with label correction techniques used in other FL methods (Xu et al., 2022; Wang et al., 2022). For example, we use the same basic label correction strategy as in (Xu et al., 2022), i.e., replacing labels with high confidence model predictions, before Our Stage 3. In Tabel 8, we show the improved performance of enabling simple label correction technique.

Table 8: ✓/ ✗indicates the label correction technique is enable/disable

Label Correction	CIFAR-10				
Technique	30%	40%	50%	60%	70%
✗	85.65	85.10	84.09	82.91	80.17
✓	86.01	85.54	84.88	83.79	81.43

5. Conclusion

In this study, we initially conducted a preliminary experiment to demonstrate how noise labels can impede representation learning and corrupt the FE of local models, consequently undermining the performance of the aggregated global model in FL. To address this issue, we proposed a simple yet effective three-stage training framework for FL. Experimental validation on multiple datasets substantiated the effectiveness of our approach in mitigating the impact of noisy labels and improving the overall FL performance.

Acknowledgments

The work is partially supported by the National Natural Science Foundation of China(No. 62106217), the Yunnan Provincial Department of Science and Technology Basic Research Program (Grant No.202201AU070112) and Program of Yunnan Key Laboratory of Intelligent Systems and Computing (Grant No.202405AV340009).

References

Yiqiang Chen, Xiaodong Yang, Xin Qin, Han Yu, Biao Chen, and Zhiqi Shen. Focus: Dealing with label quality disparity in federated learning. *arXiv preprint arXiv:2001.11359*, 2020.

Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10072–10081, June 2022.

Xiuwen Fang, Mang Ye, and Xiyuan Yang. Robust heterogeneous federated learning under data corruption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5020–5030, October 2023.

Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2703–2708, June 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Xian-Jin Gui, Wei Wang, and Zhang-Hao Tian. Towards understanding deep learning from noisy labels with small-loss criterion. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2469–2475, 8 2021.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Chuanyang Hu, Shipeng Yan, Zhitong Gao, and Xuming He. Mild: Modeling the instance learning dynamics for learning with noisy labels. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 828–836, 8 2023.

Xuefeng Jiang, Sheng Sun, Yuwei Wang, and Min Liu. Towards federated learning against noisy labels via local self-regularization. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 862–873, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Hao-Tian Li, Tong Wei, Hao Yang, Kun Hu, Chong Peng, Li-Bo Sun, Xun-Liang Cai, and Min-Ling Zhang. Stochastic feature averaging for learning with long-tailed noisy labels. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3902–3910, 2023.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2019.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60, 2020. doi: 10.1109/MSP.2020.2975749.

Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*, 2022.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016.

Yang Lu, Lin Chen, Yonggang Zhang, Yiliang Zhang, Bo Han, Yiu-ming Cheung, and Hanzi Wang. Federated learning with extremely noisy clients via negative distillation. *arXiv preprint arXiv:2312.12703*, 2023.

Yangdi Lu and Wenbo He. Selc: Self-ensemble label correction improves learning with noisy labels. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3278–3284, 7 2022.

Dhruv Mahajan, Kaiming He, Ross Girshick, and Abhinav Gupta. Rethinking pre-training and fine-tuning in self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Mateusz Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zhuowei Wang, Tianyi Zhou, Guodong Long, Bo Han, and Jing Jiang. Fednoil: A simple two-level sampling method for federated learning with noisy labels. *arXiv preprint arXiv:2205.10110*, 2022.

Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4424–4432, 8 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Ruixuan Xiao, Yiwen Dong, Haobo Wang, Lei Feng, Runze Wu, Gang Chen, and Junbo Zhao. Promix: Combating label noise via maximizing clean sample utility. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4442–4450, 8 2023.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jingyi Xu, Zihan Chen, Tony Q.S. Quek, and Kai Fong Ernest Chong. Fedcorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10184–10193, June 2022.

Miao Yang, Hua Qian, Ximin Wang, Yong Zhou, and Hongbin Zhu. Client selection for federated learning with label noise. *IEEE Transactions on Vehicular Technology*, 71(2): 2193–2197, 2022a.

Seunghan Yang, Hyoungseob Park, Junyoung Byun, and Changick Kim. Robust federated learning with noisy labels. *IEEE Intelligent Systems*, 37(2):35–43, 2022b. doi: 10.1109/MIS.2022.3151466.

Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5192–5201, June 2021.

Yu Yao, Mingming Gong, Yuxuan Du, Jun Yu, Bo Han, Kun Zhang, and Tongliang Liu. Which is better for learning with noisy labels: the semi-supervised method or modeling label noise? In *International Conference on Machine Learning*, pages 39660–39673. PMLR, 2023.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019.

Hui Zhang and Quanming Yao. Decoupling representation and classifier for noisy label learning. *arXiv preprint arXiv:2011.08145*, 2020.

Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M. Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1657–1667, January 2022.

Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 2021.

Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 72–81, October 2021.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4912–4921, 2021.

Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. In *International Conference on Learning Representations*, 2022.