Learning Locally, Revising Globally: Global Reviser for Federated Learning with Noisy Labels

Anonymous Authors[1]

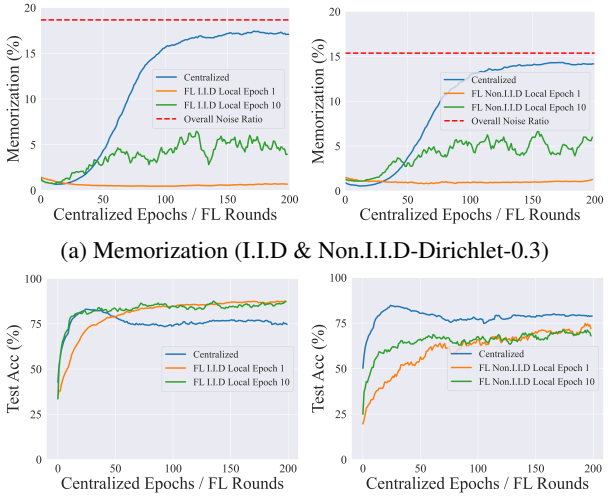
Abstract

Conventioanl federated learning (FL) heavily depends on high-quality labels, which are often impractical in the real world, leading to the federated label-noise (F-LN) problem. Worsely, the F-LN problem is exacerbated by the heterogeneity of FL, whereas clients experience different label-noise types, ratios, and data distribution. In this study, we first observe an intriguing phenomenon that the global model of FL exhibits a slow memorization of noisy labels, suggesting its ability to maintain reliable predictions and robust representations in FL. Motivated on this, we propose a novel method termed Federated Global Reviser (FedGR), a straightforward yet effective method comprising three modules that collaboratively rectify noisy labels and regularize local training. By exploiting above inherent property, FedGR improve the label-noise robustness of FL in a self-contained manner. Extensive experiments on three widely used F-LN benchmarks demonstrate the superior performance of FedGR, consistently outperforming seven state-of-the-art baselines even in severe label-noise and data heterogeneity. Code will be released upon acceptance.

1. Introduction

Federated learning (FL) facilitates privacy-preserving collaborative training across clients for applications like healthcare (Kaissis et al., 2020) and recommendation systems (Sun et al., 2024). Despite promising performance (McMahan et al., 2017; Li et al., 2020b; Meng et al., 2024), FL heavily relies on high-quality annotated data. However, precisely annotating decentralized datasets is impractical (Irvin et al., 2019), inevitably leading to the federated label noise (F-LN) problem (Yang et al., 2022; Xu et al., 2022). Unlike centralized label-noise (C-LN) problem (Han et al., 2018; Li et al., 2020a), F-LN is more challenging due to the label-noise and data heterogeneity, encompassing diverse noise patterns (*e.g.*, varying ratios/types) and data heterogeneity causing class imbalance (Li et al., 2022; Qi et al., 2023; Wu et al., 2023; Qi et al., 2025). This heterogeneity significantly hinders the direct application of centralized learning with noisy labels (C-LNL) methods (Li et al., 2022; Wei et al., 2021). Thus, it is highly expected to customize a federated learning with noisy labels (F-LNL) approach to tackle the F-LN problem.

Existing F-LNL approaches typically treat the F-LN problem as a distributed extension of learning with noisy labels, focusing on refining client-side training algorithms (Jiang et al., 2022; Wang et al., 2022; Xu et al., 2022; Ji et al., 2024; Kim et al.) or detecting and isolating noisy clients (Xu et al., 2022; Lu et al., 2024) to mitigate negative impacts. How-



(a) Memorization (I.I.D & Non.I.I.D-Dirichlet-0.3)

(b) Test Acc (I.I.D & Non.I.I.D-Dirichlet-0.3)

Figure 1. **(a) Slower Memorization Effect**: On CIFAR-10, the global FL model memorizes $\leq$30% of noisy labels, while significantly lower than that of centralized training. **(b) Preservation of Test Performance**: The global model in FL avoids the test performance degradation typically observed in centralized training under noisy labels. Please see `appendix` for more results and discussions, which indicates that such a phenomenon is non-trivial.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ever, the dual heterogeneity inherent in F-LNL often renders these methods ineffective. While recent studies (Yang et al., 2022; Kim et al., 2022; Wu et al., 2023; Tam et al., 2023; Li et al., 2024) have attempted to address this challenge by constructing consensus among clients, such mechanisms frequently involve local data statistics, thereby posing significant risks of privacy leakage. In contrast to these methods, this work tackles the F-LN problem from a novel perspective. Specifically, as illustrated in Figure 1, we anatomize the memorization phenomenon of FL and observe that the global model exhibits significantly slower overfitting to label noise compared to centralized training—a phenomenon we term the intrinsic label-noise robustness of FL. Harnessing this previously not well-explored characteristic enables us to enhance the label-noise robustness of FL in a self-contained and privacy-preserving manner.

In othe words, we propose a novel method termed Federated Global Reviser (FedGR) to mitigate the adverse effect of the F-LN problem. To be specific, FedGR comprises three modules and takes advantage of the robust global model in two aspects: noisy label correction and local model regularization. First, FedGR introduces a sieving-and-refining module to partition clean and noisy samples for each client and subsequently rectify noisy labels. To address the quality and quantity issues of refined labels under the dual-heterogeneity of the F-LN problem, FedGR introduces a globally revised exponential moving average (EMA) distillation module and a global representation regularization module to further regularize the local training. To sum up, the contributions of this study are outlined as follows:

- This study provides an insightful observation that the global model of FL has a slower tendency to overfit noisy labels, which we refer to as the intrinsic label-noise robustness of FL. To the best of our knowledge, this phenomenon has not been well-explored in previous works, and motivates us to enhance the label-noise robustness of FL in a self-contained and privacy-preserving manner.
- Motivated by this insight, we introduce FedGR, a novel method designed to enhance robustness in a self-contained and privacy-preserving manner. FedGRemploys three modules: sieving-and-refining for sample selection, globally revised EMA distillation for robust pseudo labeling and regularization, and global representation regularization to further prevent label-noise overfitting. Additionally, we provide a theoretical analysis to guarantee the convergence.
- Comprehensive experiments on three public F-LN benchmarks, under diverse noise levels and distribution settings, show that FedGR consistently surpasses seven state-of-the-art baselines, delivering substantial gains in both accuracy and robustness.

2. Related Work

Centralized Learning with Noisy Labels. To address the C-LN problem, most existing C-LNL studies leverage the *memorization effect* (Arpit et al., 2017) to design robust training strategies for sample selection (Han et al., 2018; Yu et al., 2019) and noisy label correction (Berthelot et al., 2019; Li et al., 2020a; Xiao et al., 2023; Zhang et al., 2024). However, due to the following two reasons, it is undesirable for FL to directly adopt these C-LNL methods to tackle the F-LN problem. On the one hand, the data heterogeneity (Li et al., 2022) of FL makes the class-balanced assumption used by almost all the C-LNL approaches unattainable (Li et al., 2020a; Wei et al., 2021). On the other hand, these methods induce sophisticated learning algorithms, such as two peer networks (Han et al., 2018; Yu et al., 2019; Li et al., 2020a), which involve high computation and communication overhead for FL. In contrast, FedGR performs sample sieving on the server instead of each client, which mitigates the adverse impacts of the dual heterogeneity and introduces moderate computation and communication overhead.[2]

Federated Learning with Noisy Labels. By treating the F-LN problem as a distributed extension of learning with noisy labels, existing F-LNL methods often adopt client-side independent sample selection (Wang et al., 2022; Xu et al., 2022; Ji et al., 2024; Jiang et al., 2024; Jiang & Zhang, 2025) or noisy client detection (Xu et al., 2022; Lu et al., 2024; Jiang et al., 2024; Morafah et al.) to alleviate the negative effects of noisy labels. Additionally, some works even straightforwardly employ regularization techniques to regularize the local training of client (Jiang et al., 2022; Kim et al.; Zhou & Wang, 2024). However, these methods often struggle to handle the dual heterogeneity of the F-LN problem effectively. Thus, most recent studies (Yang et al., 2022; Kim et al., 2022; Wu et al., 2023; Tam et al., 2023; Li et al., 2024) attempt to construct consensus among clients to address the dual heterogeneity of the F-LN problem. Even so, these F-LNL approaches exhibit several limitations. To be specific, both client-side independent sample selection and noise client detection, which relies on the memorization effect (Arpit et al., 2017), proves unreliable for clients presenting dual heterogeneity. Moreover, techniques that construct consensus based on local data statistics pose potential privacy risks, as they necessitate the transmission of sensitive information (Yang et al., 2022; Kim et al., 2022; Tam et al., 2023). In contrast, we exploit the intrinsic label-noise robustness of the global model to promote the performance of FL faced with noisy labels, which is irrelevant to data distribution and thus privacy-preserving. And the effectiveness of the proposed FedGR is substantiated through comprehensive experimental evaluations.

[2]Please see `appendix` for analysis.

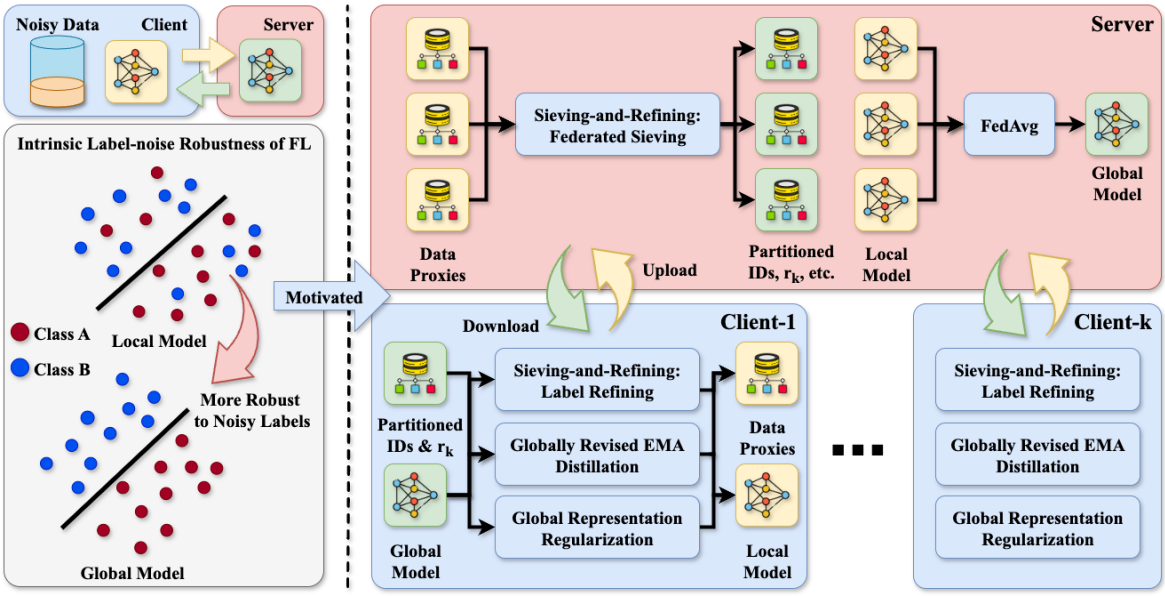


Figure 2. FedGR operates in three modules.[1] Initially, the Sieving-and-Refining module employs Federated Sieving (FS) to model global noise patterns via client-side data proxies, followed by Label Refining (LR) to correct noisy labels using the robust global model. Second, to tackle the quality and quantity issues of refined labels under the dual-heterogeneity, Globally Revised EMA Distillation is introduced to distills knowledge from a local EMA model revised by the global parameters. Lastly, Global Representation Regularization is proposed to further prevent local overfitting to cumulative EMA errors by enforcing global-to-local representation consistency.

3. Method

3.1. Problem Definition

A typical FL system (McMahan et al., 2017; Xu et al., 2022) maintains a server for model parameters aggregation and K clients which train their local models on its local low-quality dataset $\hat{\mathcal{D}}_k = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^{n_k}$, where $\hat{y}_i$ and n_k represents the one-hot vector of the label and the size of the local dataset on client k, respectively. The local objective of client k with a loss function $\ell(\cdot,\cdot)$ on $\hat{\mathcal{D}}_k$ at t-th round could be:

$$\mathcal{L}_k = \mathbb{E}_{\hat{\mathcal{D}}_k}\left[\ell(\mathbf{p}_i^l, \hat{y}_i)\right] \text{ s.t. } \mathbf{p}_i^l = (h \circ f)\left(\mathbf{x}_i; \mathbf{w}_k^t\right), k \in \mathcal{S}\left(t\right), \tag{1}$$

where $\mathbf{p}_i^l$ is the logits output by the head $h(\cdot; \mathbf{w}_{k,h}^t)$ and backbone $f(\cdot; \mathbf{w}_{k,f}^t)$ with the local model parameters $\mathbf{w}_k^t = \{\mathbf{w}_{k,h}^t, \mathbf{w}_{k,f}^t\}$. The global model parameters $\mathbf{w}_g^t$ at communication round[3] t are computed as the importance-weighted average of the aggregated local model parameters:

$$\mathbf{w}_g^t = \sum_{k\in\mathcal{S}(t)} a_k \mathbf{w}_k^t \quad \text{s.t.} \quad \sum_{k\in\mathcal{S}(t)} a_k = 1, \tag{2}$$

where $\mathcal{S}(t)$ is the set of selected clients at round t and $a_k = n_k / \sum_{i\in\mathcal{S}(t)} n_i$ is the corresponding importance weight. Finally, the global objective $\mathcal{L}$ of F-LNL can be formulated as:

$$\min_{\mathbf{w}_g} \mathcal{L}(\mathbf{w}_g) = \sum_{k\in\mathcal{S}} a_k \mathcal{L}_k(\mathbf{w}_k) \quad \text{s.t.} \quad \sum_{k\in\mathcal{S}} a_k = 1, \tag{3}$$

where $\mathcal{S}$ is the set of all clients and $\|\mathcal{S}\| = K$.

3.2. Overview of FedGR

As shown in Figure 1, the global model of FL exhibits a slower propensity to overfit noisy labels, indicating its ability to maintain reliable predictions and robust representations during training. Building upon this observation, we propose FedGR, which incorporates three specialized modules for local training in FL to enhance the label-noise robustness, as shown in Figure 2. In breief, FedGR first leverages the label-noise robust characteristic of global model to sieve and refine the noisy labels of each client with the sieving-and-refining module. It then regularizes local model training through globally revised EMA distillation module and global representation regularization module, with the help of the global model. By combining the objectives of these three modules, the local learning objective of the FedGR can be:

$$\mathcal{L}_k = \mathcal{L}_k^{SR} + \lambda_{\mathcal{B}}\mathcal{B}_k + \lambda_{\mathcal{R}}\mathcal{R}_k, \tag{4}$$

where $\mathcal{L}_k^{SR}$, $\mathcal{B}_k$, and $\mathcal{R}_k$ correspond to the sieving-and-refining objective, globally revised EMA distillation, and global representation regularization, respectively. The hyperparameters $\lambda_{\mathcal{B}}$ and $\lambda_{\mathcal{R}}$ control the relative importance of each term. Thoretical convergence analysis is presented in `appendix`. The following will elaborate on each module.

[3]Communication round and round are used interchangeably.

3.3. Sieving-and-Refining

Briefly, the sieving-and-refining module consists of two components: Federated Sieving (FS) and Label Refining (LR). In FS, the server employs a Gaussian Mixture Model (GMM) to model the label-noise patterns using aggregated instance-level data proxies (*e.g.,* loss). Based on this modeling, FS partitions each client's data proxies into clean and noisy subsets and estimates the client's label-noise ratio r_k. Then, these partitioning results are transmitted back to the clients, where LR is adopted to refine the noisy samples identified by FS, guided by the estimated r_k. In the following, we will introduce the objective of the sieving-and-refining module and then elaborate on the FS and LR.

According to the memorization effect (Arpit et al., 2017), the global model will undergo a α rounds warm-up phase before the LR is activated. Thus, the local learning objective of sieving-and-refining could be divided into two phases. For the first α rounds, the local objective of client k is to perform vanilla supervised learning on its local dataset $\hat{\mathcal{D}}_k$, *i.e.,*

$$\mathcal{L}_k^{SR} = \mathbb{E}_{\hat{\mathcal{D}}_k}\left[\mathcal{H}\left(\mathbf{p}_i^l, \hat{y}_i\right)\right], \text{ if } t < \alpha\,, \tag{5}$$

where $\mathbf{p}_i^l$ and $\mathcal{H}(\cdot)$ are the output logits and cross entropy loss. To resist the overfitting to noisy labels (Nishi et al., 2021), we adopt strong data augmentation on the input to get the logits, *i.e.,*

$$\mathbf{p}_i^l \to \mathbf{p}_i^{l,s} = (h \circ f)\left(\mathbf{x}_i^s; \mathbf{w}_k^t\right)\,. \tag{6}$$

Next, after α rounds, client k would adopt LR to refine the noisy labels detected by FS and the refined dataset $\tilde{\mathcal{D}}_k$ will be subsequently used for the local training, *i.e.,*

$$\mathcal{L}_k^{SR} = \mathbb{E}_{\tilde{\mathcal{D}}_k}\left[\mathcal{H}\left(\mathbf{p}_i^{l,s}, \tilde{y}_i\right)\right], \text{ if } t \geq \alpha\,. \tag{7}$$

To sum up, the objective of sieving-and-refining module is

$$\mathcal{L}_k^{SR} = \begin{cases} \mathbb{E}_{\hat{\mathcal{D}}_k}\left[\mathcal{H}\left(\mathbf{p}_i^{l,s}, \hat{y}_i\right)\right], & t < \alpha \\ \mathbb{E}_{\tilde{\mathcal{D}}_k}\left[\mathcal{H}\left(\mathbf{p}_i^{l,s}, \tilde{y}_i\right)\right], & t \geq \alpha \end{cases}. \tag{8}$$

Federated Sieving. The FS comprises two steps: client-side instance-level data proxy computation and server-side noisy sample partitioning. To be specific, for the first step, client $k \in \mathcal{S}(t)$ would adopt the label-noise robust global model $\mathbf{w}_g^{t-1}$ to compute a noise-distinguishable data proxy for each sample $\mathbf{x}_i \in \hat{\mathcal{D}}_k$ at the beginning of local training in each round. In order to preserve privacy, we define the data proxy of sample $\mathbf{x}_i \in \hat{\mathcal{D}}_k$ as its mean inference loss in the previous t rounds. To compute it, the client k will first maintain a loss observation set for each sample and the loss observation set $L_i^t = \{\ell_{i,p}\}_{p=1}^{T_k}$ of sample $\mathbf{x}_i$ at round t is updated by following:

$$\ell_{i,T_k} = \mathcal{H}\left(\mathbf{p}_i^g, \hat{y}_i\right) \text{ s.t. } \mathbf{p}_i^g = (h \circ f)\left(\mathbf{x}_i; \mathbf{w}_g^{t-1}\right)\,, \tag{9}$$

where T_k represents the number of selected times for client k in previous t rounds. Subsequently, the mean inference loss of sample $\mathbf{x}_i$ in the previous t rounds can be obtained as follows:

$$\bar{\ell}_i^t = \frac{1}{T_k}\sum_{p=1}^{T_k} \ell_{i,p}\,. \tag{10}$$

Then, the client k would upload its local parameters $\mathbf{w}_k^t$ and instance-level data proxies $\{(d_{i,k}, \bar{\ell}_i^t)\}_{i=1}^{n_k}$ to the server at the end of local training, where $d_{i,k}$ denotes a global unique identifier for sample $\mathbf{x}_i$ on client k. In the second step of FS, the server would aggregate these data proxies from all selected clients $\mathcal{S}(t)$ and model their distribution using a two-component GMM. By setting partitioning threshold for the the posterior probability $q_{i,k}$ (Li et al., 2020a) of sample $\mathbf{x}_i \in \hat{\mathcal{D}}_k$ belonging to the "clean" GMM component, the server can partition the identifiers into clean/noisy subsets. Subsequently, the client's noise ratio r_k can also be derived from the above partition. Finally, the partitioning results of client k and the aggregated global model parameters will be returned to it when it is selected for collaborative training in subsequent rounds. To obtain the partitioning results of all clients, FS follows (Xu et al., 2022) to adopt random sampling without replacement in the first α rounds (Xu et al., 2022), deviating from the standard FL setup. After that, the standard FL client sampling (McMahan et al., 2017) is adopted.

The benefits of FS can be two-fold. On the one hand, the proposed FS is a more reliable sample sieving under the dual heterogeneity of the F-LN problem, as it leverages larger-scale training dynamics (*e.g.,* loss) than any single client can provide for label-noise modeling. On the other hand, FS is privacy-preserving, as it only transmits the information irrelevant to the data distribution.[4]

Label Refining. At the beginning of local training on client $k \in \mathcal{S}(t)$ during the *t-th* round, the client can divide its local dataset $\hat{\mathcal{D}}_k$ into a clean subset $\hat{\mathcal{D}}_k^c$ and a noisy subset $\hat{\mathcal{D}}_k^n$, based on the returned partition results. Then, client k can employ LR to generate pseudo labels that correct the labels of the noisy subset $\hat{\mathcal{D}}_k^n$. Notably, due to the dual heterogeneity of the F-LN problem, we propose a label refinement strategy that is conditioned on the estimated r_k to obtain reliable refined labels. Specifically, the refined label $\tilde{y}_i$ of sample $\mathbf{x}_i \in \hat{\mathcal{D}}_k$ is defined as follows:

$$\tilde{y}_i = \begin{cases} \hat{y}_i, & r_k < \beta \text{ and } \mathbf{x}_i \in \hat{\mathcal{D}}_k^c \\ q_{i,k}\hat{y}_i + (1 - q_{i,k})y_i^{pse}, & r_k < \beta \text{ and } \mathbf{x}_i \in \hat{\mathcal{D}}_k^n \\ y_i^{pse}, & r_k \geq \beta \end{cases}, \tag{11}$$

where y_i^{pse} is the pseudo label and β is the label-noise ratio

[4] Loss is irrelevant to the distribution of $(\mathbf{x}, \mathbf{y})$, thus it is privacy preserving. See `appendix` for discussion.

threshold. According to the observation on the global model, we adopt FixMatch (Sohn et al., 2020) on the global model to generate the reliable pseudo label y_i^{pse} for sample $\mathbf{x}_i$.

The above strategy is predicated on the following two considerations. Firstly, for clients exhibiting simple label-noise patterns, the partitioning results are deemed relatively reliable. Therefore, we refine the noisy label set by leveraging the clean probability q_i derived from the GMM, following the methodology outlined in (Li et al., 2020a). Secondly, for the clients suffering from high label-noise ratios (*e.g.,* $r_k \geq \beta$) or complex label-noise types (*e.g.,* asymmetric), we consider all their provided labels to be untrustworthy and only use the pseudo labels as the refined labels.

3.4. Globally Revised EMA Distillation

Although the global model is relatively robust to noisy labels, it struggles to fit the local data distribution of each client due to dual heterogeneity. Thus, it often fails to produce a sufficient number of reliable pseudo labels for each client. Whilst the local model can fit the clients' distribution, but is easily corrupted by noisy labels, especially on high-noise clients, resulting in unreliable predictions. To address such a conflict between the global and local models under dual heterogeneity, we propose a globally revised EMA distillation module. Such a module resorts to two types of models, *i.e.,* the global model and the local EMA model, to regularize the local learning via knowledge distillation.

To be specific, as EMA inherently has stability and early-training robustness in learning with noisy labels (Zhou et al., 2021; Morales-Brotons et al., 2024), each client will maintain a local EMA model $\mathbf{w}_{k,ema}^t$ during the local training. To mitigate the cumulative effect of noisy labels on the local EMA model, we propose revising the local EMA model with the global model before distillation, as the global model is more robust to noisy labels. Formally, for the t-th round, the revising and usual updating step for the local EMA model $\mathbf{w}_{k,ema}^t$ on client k at m_k-th local training step could be:

$$\mathbf{w}_{k,ema}^{t,m_k} = \begin{cases} \gamma_g \mathbf{w}_{k,ema}^{t-1,m_k} + (1-\gamma_g)\mathbf{w}_g^{t-1}, & m_k = 0 \\ \gamma_l \mathbf{w}_{k,ema}^{t,m_k-1} + (1-\gamma_l)\mathbf{w}_k^{t,m_k}, & m_k \geq 1 \end{cases} \tag{12}$$

where m_k and $\gamma_{g/l}$ denote the local training step and the momentum decay for the global revised EMA step/local EMA step, respectively. Then, the proposed globally revised EMA distillation module would adopt the revised local EMA model $(h \circ f)\left(\cdot; \mathbf{w}_{k,ema}^{t,0}\right)$ as teacher to distill the knowledge to the local model $(h \circ f)\left(\cdot; \mathbf{w}_k^t\right)$ and the objective of it on client k can be formulated as:

$$\mathcal{B}_k = \mathbb{E}_{\hat{\mathcal{D}}_k}\left[KL\left(\frac{\mathbf{p}_i^{le,w}}{\tau}, \frac{\mathbf{p}_i^{l,s}}{\tau}\right)\right], \tag{13}$$

where $\mathbf{p}_i^{le/l,w}$, KL, and τ denote the output logits of revised local EMA model/local model on weakly augmented data, the Kullback-Leibler divergence loss, and the temperature, respectively. Notably, the logits predicted by the revised local EMA model is

$$\mathbf{p}_i^{le,w} = (h \circ f)\left(\mathbf{x}_i^w; \mathbf{w}_{k,ema}^{t,0}\right), \tag{14}$$

where $\mathbf{x}_i^w$ refers to the weakly augmented $\mathbf{x}_i \in \hat{\mathcal{D}}_k$. The benefit of distilling logits from the revised EMA model instead of the online EMA model $(h \circ f)\left(\cdot; \mathbf{w}_{k,ema}^{t,m_k}\right)$ could be two-fold. On the one hand, it lowers the forward computation cost, as the teacher's logits are computed only once at the beginning of local training. On the other hand, it improves the resilience of logits to the accumulated adverse effect of noisy labels and incorrect refined labels. The ablation in Table 4 also demonstrates that such a mechanism is more effective.

Similar to the label refinement strategy, γ_g of each client should be conditioned on the r_k and the results of sieving-and-refining module due to the dual heterogeneity of the F-LN problem. For instance, after warm up phase, if the client $k \in \mathcal{S}(t)$ suffers from high label-noise ratio ($r_k \geq \beta$) or fails to obtain a sufficient number of samples, the local EMA model is deemed unreliable and should be entirely replaced by the global model, *i.e.,* $\gamma_g = 0$. Otherwise, the local EMA model will be revised by the global model with momentum decay γ_g. Formally, γ_g could be adjusted as follows:

$$\gamma_g = \begin{cases} \gamma_g, & t \geq \alpha \\ 0, & \left(r_k \geq \beta \text{ and } \dfrac{\left\|\tilde{\mathcal{D}}_k^r\right\|}{\left\|\hat{\mathcal{D}}_k\right\|} < \mu\right) \text{ or } t < \alpha \end{cases}, \tag{15}$$

where $\tilde{\mathcal{D}}_k^r$ denotes the data subset with relatively reliable hard labels after LR which is initialized to $\emptyset$, *i.e.,*

$$\tilde{\mathcal{D}}_k^r = \tilde{\mathcal{D}}_k^r \cup \hat{\mathcal{D}}_k^c \cup \{y_i^{pse} | y_i^{pse} \neq \mathbf{0} \quad \forall i = 1, \cdots, n_k\}, \tag{16}$$

$\left\|\tilde{\mathcal{D}}_k^r\right\| / \left\|\hat{\mathcal{D}}_k\right\|$ refers to the proportion of the samples with reliable labels and μ is the corresponding threshold. Additionally, in order to not affect quality of the FS via regularization, the globally revised EMA distillation will be activated after α rounds *i.e.,* $\lambda_{\mathcal{B}} = 0$ if $t < \alpha$.

3.5. Global Representation Regularization

Though the globally revised EMA distillation module can effectively regularize the local training, the local model on the high label-noise ratio client is still inevitably overfitting the noisy labels. In light of the representation learning (Chen & He, 2021; Lubana et al., 2022) and our observation, we

Table 1. Results on CIFAR-10. The 1st/2nd-best results are in a gray box w/. and w/o. boldface.

Data Partition	I.I.D								Non.I.I.D-Dirichlet (0.3)							
Noise Type	Clean	Sym		Asym		Mixed		Avg	Clean	Sym		Asym		Mixed		
ϕ	0.0	0.6	1.0	0.6	1.0	0.6	1.0		0.0	0.6	1.0	0.6	1.0	0.6	1.0	Avg
$\mathcal{U}(\rho_{min}, \rho_{max})$	0.0-0.0	0.5-1.0	0.5-1.0	0.2-0.4	0.2-0.4	0.2-0.4	0.2-0.4		0.0-0.0	0.5-1.0	0.5-1.0	0.2-0.4	0.2-0.4	0.2-0.4	0.2-0.4	
FedAvg	92.55 ±0.14	61.60 ±3.73	23.89 ±3.62	83.46 ±0.89	70.44 ±1.80	81.90 ±0.73	70.66 ±0.69	65.33 ±1.91	**87.05 ±1.45**	39.46 ±3.02	17.32 ±1.07	69.98 ±1.58	53.74 ±1.61	66.10 ±2.09	51.92 ±1.70	49.75 ±1.84
FedProx	90.75 ±0.16	56.57 ±4.27	23.02 ±2.90	79.78 ±1.00	64.44 ±1.64	77.94 ±0.61	65.23 ±0.83	61.16 ±1.88	86.86 ±0.58	36.94 ±2.42	16.69 ±1.32	69.19 ±0.98	52.68 ±0.60	64.08 ±1.50	49.77 ±1.00	48.22 ±1.30
FL-Coteaching	-	66.43 ±2.65	47.28 ±5.03	87.40 ±0.58	82.61 ±0.75	86.51 ±0.43	83.99 ±0.54	75.70 ±1.66	-	44.42 ±2.11	33.49 ±1.00	76.20 ±1.99	72.93 ±1.93	74.40 ±1.67	72.42 ±1.56	62.31 ±1.71
FL-DivideMix	-	76.54 ±0.42	68.47 ±3.00	85.46 ±0.21	86.23 ±0.36	84.76 ±0.31	85.19 ±0.48	81.11 ±0.80	-	58.94 ±1.19	38.35 ±4.45	73.13 ±1.71	71.45 ±1.47	70.18 ±1.37	68.86 ±1.29	63.49 ±1.91
FedCorr [CVPR22]	92.55 ±0.71	92.04 ±0.17	55.12 ±1.55	83.76 ±0.74	83.06 ±1.36	84.15 ±0.52	84.00 ±0.17	80.36 ±0.75	77.14 ±8.44	78.85 ±4.74	29.42 ±2.43	57.91 ±8.15	55.67 ±6.81	83.33 ±1.43	67.85 ±3.75	62.17 ±4.55
FedNoRo [IJCAI23]	-	63.30 ±0.93	33.98 ±4.71	71.83 ±0.33	63.29 ±1.12	71.07 ±0.32	63.24 ±0.30	61.12 ±1.29	-	51.64 ±2.07	18.60 ±1.12	58.99 ±2.07	41.18 ±2.92	57.09 ±1.57	43.99 ±1.30	45.25 ±1.84
FedDiv [AAAI24]	-	90.36 ±0.57	33.14 ±21.83	**93.67 ±0.04**	**92.86 ±0.19**	**93.43 ±0.07**	**92.36 ±0.09**	82.64 ±3.80	-	20.76 ±7.76	14.22 ±3.65	26.85 ±2.49	26.28 ±11.31	35.71 ±6.23	23.20 ±6.28	24.50 ±6.29
FedFixer [AAAI24]	-	66.49 ±1.03	30.18 ±2.92	83.29 ±0.28	70.65 ±1.11	85.52 ±0.95	77.50 ±1.15	68.94 ±1.24	-	52.37 ±3.43	22.53 ±2.54	72.24 ±2.18	57.30 ±1.20	74.83 ±3.02	63.72 ±2.30	57.17 ±2.45
FedGR [Ours]	**93.95 ±0.10**	**92.09 ±0.25**	**83.91 ±1.32**	93.38 ±0.30	91.64 ±0.38	93.13 ±0.32	92.27 ±0.18	**91.07 ±0.46**	86.22 ±1.97	**82.04 ±2.23**	**63.64 ±5.39**	**86.79 ±2.68**	**83.67 ±5.02**	**86.50 ±2.36**	**84.65 ±2.38**	**81.22 ±3.43**

Table 2. Results on CIFAR-100. The 1st/2nd-best results are in a gray box w/. and w/o. boldface.

Data Partition	I.I.D								Non.I.I.D-Dirichlet (0.3)							
Noise Type	Clean	Sym		Asym		Mixed		Avg	Clean	Sym		Asym		Mixed		
ϕ	0.0	0.6	1.0	0.6	1.0	0.6	1.0		0.0	0.6	1.0	0.6	1.0	0.6	1.0	Avg
$\mathcal{U}(\rho_{min}, \rho_{max})$	0.0-0.0	0.5-1.0	0.5-1.0	0.2-0.4	0.2-0.4	0.2-0.4	0.2-0.4		0.0-0.0	0.5-1.0	0.5-1.0	0.2-0.4	0.2-0.4	0.2-0.4	0.2-0.4	
FedAvg	64.85 ±0.33	33.97 ±1.82	13.00 ±1.85	54.51 ±0.83	44.18 ±1.72	52.37 ±0.40	43.02 ±0.44	40.18 ±1.18	63.86 ±0.61	29.04 ±1.64	10.23 ±1.27	51.74 ±0.60	41.36 ±1.16	49.29 ±0.68	39.16 ±1.60	36.80 ±1.16
FedProx	56.85 ±0.55	30.22 ±1.68	11.94 ±1.91	45.97 ±0.92	37.83 ±0.76	45.08 ±0.50	36.82 ±0.47	34.64 ±1.04	58.83 ±0.63	26.19 ±1.52	9.31 ±1.27	46.12 ±0.87	36.86 ±1.16	44.26 ±0.80	35.40 ±1.23	37.77 ±0.98
FL-Coteaching	-	41.19 ±1.04	25.98 ±1.87	58.84 ±0.63	50.56 ±1.26	58.13 ±0.36	53.42 ±0.68	48.02 ±1.02	-	36.64 ±1.60	24.81 ±1.03	57.70 ±0.54	50.71 ±0.95	56.80 ±0.40	52.35 ±1.21	46.50 ±0.96
FL-DivideMix	-	49.48 ±0.52	**35.35 ±1.15**	59.26 ±0.25	55.12 ±0.31	59.40 ±0.25	57.53 ±0.20	52.69 ±0.45	-	44.25 ±0.81	27.29 ±2.50	57.93 ±0.35	52.53 ±1.33	57.70 ±0.23	55.47 ±0.95	49.20 ±1.03
FedCorr [CVPR22]	70.77 ±2.11	58.29 ±1.30	27.54 ±2.04	67.04 ±0.49	59.58 ±0.70	66.56 ±0.67	61.41 ±1.03	56.74 ±1.04	64.46 ±3.54	55.83 ±0.40	19.02 ±1.52	61.29 ±0.90	52.18 ±1.61	61.45 ±1.88	53.94 ±0.69	50.62 ±1.17
FedNoRo [IJCAI23]	-	29.61 ±0.40	16.00 ±0.39	35.48 ±0.27	30.57 ±0.52	34.56 ±0.36	30.81 ±0.98	29.50 ±0.49	-	26.66 ±0.57	12.43 ±0.61	34.00 ±0.37	27.00 ±0.61	32.64 ±0.22	27.02 ±0.33	26.63 ±0.45
FedDiv [AAAI24]	-	37.14 ±4.20	4.13 ±3.25	**69.37 ±0.42**	60.26 ±1.56	66.65 ±0.66	62.64 ±0.42	59.21 ±1.25	-	10.18 ±1.05	1.11 ±0.18	39.49 ±7.41	39.59 ±5.80	32.72 ±9.67	33.20 ±5.33	27.19 ±0.42
FedFixer [AAAI24]	-	34.46 ±1.03	13.93 ±0.60	53.17 ±0.56	44.11 ±0.33	53.86 ±0.98	46.08 ±0.07	40.94 ±0.60	-	29.26 ±0.22	11.61 ±0.34	51.43 ±0.71	43.01 ±0.58	51.67 ±1.30	43.63 ±0.38	38.44 ±0.59
FedGR [Ours]	**71.64 ±0.22**	**63.19 ±0.91**	35.28 ±1.47	69.10 ±0.20	**62.97 ±0.44**	**68.73 ±0.46**	**64.56 ±0.32**	**60.64 ±0.63**	**69.38 ±0.52**	**57.76 ±0.54**	**30.30 ±0.96**	**65.57 ±0.65**	**56.49 ±0.49**	**65.57 ±0.64**	**59.68 ±0.40**	**55.90 ±0.61**

further introduce a global representation regularization module to regularize the local learning of the local model. To be specific, we adopt an instance discriminative-like task, *i.e.,* the global representation of a weak augmented image should be consistent with the local representation of the strong augmented image, as the goal of regularization. Formally, the objective of regularization on k-th client at t-th round could be

$$\mathcal{R}_k = \mathbb{E}_{\hat{\mathcal{D}}_k}\left[KL\left(\frac{f\left(\mathbf{x}_i^w; \mathbf{w}_{g,f}^{t-1}\right)}{\tau}, \frac{f\left(\mathbf{x}_i^s; \mathbf{w}_{k,f}^{t}\right)}{\tau}\right)\right]. \tag{17}$$

4. Experiments

We conduct experiments against eight baselines under I.I.D. and Non-I.I.D. FL settings with various label-noise levels and types to evaluate the effectiveness of the proposed FedGR. Then we perform ablations on the three main modules to investigate their effects and further analysis to show the superiority of the proposed FS. Please refer to the `appendix` for the more experimental details, results, analysis, discussion, and data visualization.

Datasets & F-LNL setups. The experiments are conducted on CIFAR-10/100 and the real-world label-noise benchmark Clothing1M (Xiao et al., 2015) under various F-LNL settings. To simulate the label-noise heterogeneity, we first partition the CIFAR10, CIFAR100, and Clothing1M into 100, 100, and 500 FL clients under I.I.D and extreme Dirichlet-Non.I.I.D (Li et al., 2022)settings, respectively. Subsequently, we perform a noise label synthetic process. Specifically, ϕ is introduced to control the proportion of clients affected by noise label. Next, the parameter ρ_{min} and ρ_{max} are used to bound the a uniform distribution $\mathcal{U}(\rho_{min}, \rho_{max})$, where the client-$k$'s label-noise ratio is sampled. In addition to the label-noise ratio, the label-noise type (Song et al., 2023), *i.e.,* symmetric, asymmetric, or a mixture of thereof, is also controlled. For instance, the *Mixed* in Table 1, 2 and 4 denotes that the noise type of each client is randomly assigned as either *Sym* or *Asym*. Consequently, the I.I.D and Non.I.I.D. in all Tables in this study also denote the label-noise heterogeneity and dual heterogeneity involving

Table 3. Results on Clothing1M. The 1st/2nd-best results are in a gray box w/. and w/o. boldface.

Methods	FedAvg	FedProx	FL-Coteaching	FL-DivideMix	FedCorr [CVPR22]	FedDiv [AAAI24]	FedFixer [AAAI24]	FedGR [Ours]
I.I.D	69.60±0.27	69.69±0.25	69.48±0.20	69.13±0.22	69.84±0.20	67.71±0.30	70.61±0.28	**71.19±0.42**
Non.I.I.D-Dirichlet (0.3)	67.90±1.31	68.18±1.13	68.16±0.82	63.88±1.21	60.42±7.23	65.79±2.77	65.16±0.95	**68.52±1.11**

Table 4. Ablation studies.

Data Partition	I.I.D			Non.I.I.D-Dirichlet (0.3)		
Noise Type	Sym	Asym	Mixed	Sym	Asym	Mixed
ϕ	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{U}(\rho_{min}, \rho_{max})$	0.5-1.0	0.2-0.4	0.2-0.4	0.5-1.0	0.2-0.4	0.2-0.4
FedGR	83.91 ±1.32	91.64 ±0.38	92.27 ±0.18	63.64 ±5.39	83.67 ±5.02	84.65 ±2.38
w/o. FS	54.59 ±3.04	87.49 ±0.12	91.71 ±0.13	45.48 ±7.81	83.45 ±4.49	84.01 ±0.31
w/o. LR	75.23 ±2.86	90.42 ±0.13	90.46 ±0.10	59.48 ±5.85	82.92 ±3.43	83.21 ±1.89
w/o. $\mathcal{R}_k$	81.49 ±1.07	91.22 ±0.34	91.84 ±0.30	58.23 ±4.08	81.80 ±2.32	82.70 ±0.63
w/o. $\mathcal{B}_k$	78.14 ±1.48	91.54 ±0.24	91.24 ±0.13	51.07 ±5.16	80.07 ±3.36	79.44 ±2.79

Table 5. Further analysis.

Data Partition	I.I.D			Non.I.I.D-Dirichlet (0.3)		
Noise Type	Sym	Asym	Mixed	Sym	Asym	Mixed
ϕ	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{U}(\rho_{min}, \rho_{max})$	0.5-1.0	0.2-0.4	0.2-0.4	0.5-1.0	0.2-0.4	0.2-0.4
FedGR	83.91 ±1.32	91.64 ±0.38	92.27 ±0.18	63.64 ±5.39	83.67 ±5.02	84.65 ±2.38
w/o. weak aug	55.73 ±12.97	89.93 ±0.29	90.37 ±0.20	25.32 ±4.36	72.48 ±6.03	78.54 ±4.22
w/o. strong aug	34.51 ±3.26	80.53 ±0.41	78.72 ±0.68	18.43 ±1.47	59.74 ±2.01	58.60 ±1.09
Online EMA distill	82.80 ±1.44	91.34 ±0.27	92.04 ±0.29	62.18 ±4.13	82.50 ±5.79	83.56 ±3.63
Randomly sample clients w/. replacement in warmup	82.56 ±1.15	91.21 ±0.46	92.08 ±0.22	62.03 ±5.85	82.91 ±2.23	83.19 ±2.52

Table 6. Hyperparameter analysis for $\lambda_{\mathcal{B}}$ & $\lambda_{\mathcal{R}}$.

Data Partition	I.I.D			Non.I.I.D-Dirichlet (0.3)		
Noise Type	Sym	Asym	Mixed	Sym	Asym	Mixed
ϕ	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{U}(\rho_{min}, \rho_{max})$	0.5-1.0	0.2-0.4	0.2-0.4	0.5-1.0	0.2-0.4	0.2-0.4
$\lambda_{\mathcal{B}} = 1.0$ $\lambda_{\mathcal{R}} = 0.1$	83.91 ±1.32	91.64 ±0.38	92.27 ±0.18	63.64 ±5.39	83.67 ±5.02	84.65 ±2.38
$\lambda_{\mathcal{B}} = 1.0$ $\lambda_{\mathcal{R}} = 0.2$	73.16 ±11.26	91.32 ±0.40	92.26 ±0.17	24.49 ±4.04	82.44 ±4.33	83.84 ±3.39
$\lambda_{\mathcal{B}} = 1.0$ $\lambda_{\mathcal{R}} = 0.5$	52.74 ±9.33	91.87 ±0.20	92.53 ±0.17	32.94 ±5.57	82.10 ±4.83	84.47 ±2.93
$\lambda_{\mathcal{B}} = 0.5$ $\lambda_{\mathcal{R}} = 0.1$	83.18 ±0.80	90.93 ±2.40	91.73 ±0.20	62.22 ±4.32	79.60 ±4.95	79.59 ±3.41
$\lambda_{\mathcal{B}} = 1.0$ $\lambda_{\mathcal{R}} = 0.1$	83.91 ±1.32	91.64 ±0.38	91.73 ±0.20	63.64 ±5.39	83.67 ±5.02	84.65 ±2.38
$\lambda_{\mathcal{B}} = 2.0$ $\lambda_{\mathcal{R}} = 0.1$	83.68 ±0.45	91.08 ±0.30	92.13 ±0.14	49.62 ±6.36	83.66 ±5.87	85.23 ±3.13

both label-noise and data distribution, respectively.

Baselines. The experimental comparison employs eight baselines, which are divided into three groups: 1) the classic FL methods, *i.e.,* FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020b); 2) the typical C-LNL methods implemented in FL, *i.e.,* FL-Coteaching (Han et al., 2018) and FL-DivideMix (Li et al., 2020a); 3) the most recent F-LNL approaches, *i.e.,* FedCorr (Xu et al., 2022), FedNoRo (Wu et al., 2023), FedFixer (Ji et al., 2024), and FedDiv (Li et al., 2024).

Implementation Details. We report the mean test accuracy over the last 10 rounds instead of the best test accuracy, demonstrating the capability for preventing the overfitting of noisy labels and mitigating the substantial fluctuations. All the experiments are conducted three times with different random seeds and the mean and standard deviation are reported. For Non.I.I.D. data partition, we follow (Li et al., 2022) to use the Dirichlet distribution to partition the data where $\alpha_{dirichlet} = 0.3$. As for the data augmentation, we adopt RandAugmentation (Cubuk et al., 2020) and the augmentation in (Xu et al., 2022) as the strong and weak data augmentation, respectively. Following (Xu et al., 2022), we adopt SGD as optimizer with a constant learning rate and use ResNet-18, ResNet-34, and pre-trained ResNet-50 as the backbone for CIFAR-10, CIFAR-100, and Clothing1M, respectively. The local epochs are set to 10 and 2 for CIFAR-10/100 and Clothing1M, respectively. $\lambda_{\mathcal{B}}$ and $\lambda_{\mathcal{R}}$ are set to 1.0 and 0.2 as default, respectively. For CIFAR-10, we decrease $\lambda_{\mathcal{R}}$ to 0.1, as the dataset is relatively simple.

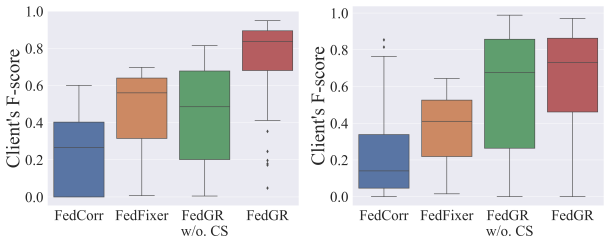
Comparison with State-Of-The-Arts. We compared FedGR against eight baselines under diverse label-noise and data heterogeneity setups (Kim et al., 2022; Li et al., 2022) (Table 1– 3). In brief, the proposed FedGR achieves eye-catching performance and outperforms all the baselines by a considerable margin. Specifically, on CIFAR-10/-100, whenever the data is I.I.D. or Non.I.I.D., the proposed FedGR could achieve the smallest performance gap to clean-data-trained models, if there are clean clients in a FL system ($\phi = 0.6$). Crucially, it remains robust in high-noise settings (*i.e.,* Sym, $\phi = 1.0$, and $\mathcal{U}(0.5, 1.0)$) where baselines like FedNoRo, FedDiv, and FedFixer fail. Notably, under the most challenging dual heterogeneity setups, the proposed FedGR outperforms baselines a considerable margin. Surprisingly, the proposed FedGR even outperforms FedAvg trained with clean data in some setups (*e.g.,* Mixed, $\phi = 0.6$, and $\mathcal{U}(0.2, 0.4)$) due to the regularization. Neverthless, the magnitude of this effect is relatively small in the clean setting, whereas on noisy labels, the gains of FedGR over FedAvg are much more pronounced. This indicates that the primary benefit of FedGR indeed comes from its ability to handle label noise rather than regularization. Going beyond, the results in Table 3 verify the effectiveness of the FedGR on a large-scale real-world label-noise dataset

Table 7. Hyperparameter analysis for γ_g & γ_l.

Data Partition	I.I.D			Non.I.I.D-Dirichlet (0.3)		
Noise Type	Sym	Asym	Mixed	Sym	Asym	Mixed
ϕ	1.0	1.0	1.0	1.0	1.0	1.0
$\mathcal{U}(\rho_{min}, \rho_{max})$	0.5-1.0	0.2-0.4	0.2-0.4	0.5-1.0	0.2-0.4	0.2-0.4
$\gamma_g = 0.9$ $\gamma_l = 0.99$	83.91 ±1.32	91.64 ±0.38	92.27 ±0.18	63.64 ±5.39	83.67 ±5.02	84.65 ±2.38
$\gamma_g = 0.95$ $\gamma_l = 0.99$	80.97 ±0.49	90.82 ±0.57	91.79 ±0.22	62.18 ±4.51	82.07 ±4.90	81.82 ±3.92
$\gamma_g = 0.99$ $\gamma_l = 0.99$	79.87 ±0.56	90.61 ±0.38	91.60 ±0.17	61.71 ±4.93	81.65 ±4.79	80.57 ±4.03
$\gamma_g = 0.9$ $\gamma_l = 0.9$	80.13 ±0.38	90.94 ±0.38	91.75 ±0.25	61.83 ±4.72	81.19 ±4.57	80.09 ±3.85
$\gamma_g = 0.9$ $\gamma_l = 0.99$	83.91 ±1.32	91.64 ±0.38	91.73 ±0.20	63.64 ±5.39	83.67 ±5.02	84.65 ±2.38
$\gamma_g = 0.9$ $\gamma_l = 0.999$	84.85 ±0.57	92.39 ±0.23	92.14 ±0.14	50.86 ±4.13	79.00 ±3.87	81.28 ±3.34

Cloting1M [5]. The F-LNL-oriented baselines (*e.g.,* FedCorr, FedDiv, and FedFixer) do not achieve superior results than the proposed FedGR , especially under the dual heterogeneity setups. In conclusion, the proposed FedGR achieves the a new state-of-the-art label-noise robustness.

(a) Pearson Coefficient Analysis
(I.I.D & Non.I.I.D-Dirichlet-0.3)



(b) Clients' F-score Distributions
(I.I.D & Non.I.I.D-Dirichlet-0.3)

Figure 3. The (a) is the pearson coefficient analysis of the $\{r_k | k \in \mathcal{S}\}$. The (b) is the clients' F-score distributions of different methods. The F-LNL setup: CIFAR-10, Sym, $\phi = 1.0$, and $\mathcal{U}(0.5, 1.0)$.

We conduct ablation studies and hyperparameter analyses on CIFAR-10 to assess module contributions and sensitivity.

Ablations. We evaluate the efficacy of FS by comparing it against a baseline that uses local model $\mathbf{w}_k^t$ for loss observation inference and perform GMM independently on each client. Then we compare LR against a vanilla strategy that trains solely on estimated local clean sets. We further set $\lambda_k = 0$ and $\lambda_b = 0$ to examine the contributions of $\mathcal{R}_k$ and $\mathcal{B}_k$, respectively. As reported in Table 4, FedGR achieves the best overall performance. The quality of sample sieving, *i.e.*, the use of FS, is particularly critical in challenging F-LNL settings (*e.g.,* Sym with $\phi = 1.0$ and $\mathcal{U}(0.5, 1.0)$). The LR component further improves performance by rectifying noisy labels. Additionally, both $\mathcal{R}_k$ and $\mathcal{B}_k$ yield further performance gains.

Further Analysis. We also conduct further analysis to study the effects of data augmentation, the online EMA distillation for the globally revised EMA distillation module, and the client sampling strategy during warm-up (*i.e.,* whether all clients are guaranteed to be sampled at least once). As shown in Table 5, weak-strong augmentation substantially mitigates overfitting to noisy labels in FL. Additionally, replacing our globally revised EMA with online EMA distillation, or removing the warm-up sampling guarantee, results in slight performance degradation.

Hyperparameters Analysis. From Table 6, one can find $\lambda_{\mathcal{R}}$ should not be too large, as excessively strong regularization leads to performance degradation, while $\lambda_{\mathcal{B}}$ must balance regularization strength against overfitting to label noise. As for γ_g and γ_l, it should satisfy $\gamma_g < \gamma_l$ as it is designed to resolve the conflict between the global and local models under dual heterogeneity. The analysis about them are shown in Table 7, and FedGR is quite robust to the choice of γ_l when $\gamma_l > 0.9$, while using a larger γ_g tends to degrade performance.

FS Quality. Figure 3 shows that FedGR can more accurately capture the relative noise ratios across clients (Pearson correlation > 0.9) and perform effective sample selection compared with FedCorr and FedFixer.

5. Conclusion

This study introduces FedGR, a novel approach for addressing the F-LN problem, inspired by that the global model in FL exhibits a reduced propensity for label-noise overfitting, which has not been explored to our best knowledge. By strategically leveraging the global model through our proposed sieving-and-refining, globally revised EMA distillation, and global representation regularization modules, FedGR effectively enhances label-noise robustness while respecting the privacy constraints of FL. The comprehensive experiments across diverse and realistic F-LNL scenarios underscore the significant effectiveness of FedGR compared to existing state-of-the-art methods. We leave the theoretical analysis of this phenomenon for future investigation and intend to transfer such a phenomenon to address C-LN problem.

[5]Here, the Clothing1M is less faithful to the realistic F-LN problem than CIFAR setups, which is discussed in `appendix`.

References

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. A Closer Look at Memorization in Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017.

Berthelot, D., Carlini, N., Goodfellow, I. J., Papernot, N., Oliver, A., and Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5050–5060, 2019.

Chen, X. and He, K. Exploring Simple Siamese Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021*, pp. 15750–15758. Computer Vision Foundation / IEEE, 2021.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, 2020.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8536–8546, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 590–597. AAAI Press, 2019.

Ji, X., Zhu, Z., Xi, W., Gadyatskaya, O., Song, Z., Cai, Y., and Liu, Y. FedFixer: Mitigating Heterogeneous Label Noise in Federated Learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 12830–12838. AAAI Press, 2024.

Jiang, X. and Zhang, J. FedClean: A General Robust Label Noise Correction for Federated Learning. In *Forty-Second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4kF2ZZcePc.

Jiang, X., Sun, S., Wang, Y., and Liu, M. Towards Federated Learning against Noisy Labels via Local Self-Regularization. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pp. 862–873. ACM, 2022.

Jiang, X., Sun, S., Li, J., Xue, J., Li, R., Wu, Z., Xu, G., Wang, Y., and Liu, M. Tackling Noisy Clients in Federated Learning with End-to-end Label Correction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1015–1026, Boise ID USA, October 2024. ACM. ISBN 979-8-4007-0436-9.

Kaissis, G., Makowski, M. R., Rueckert, D., and Braren, R. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.*, 2(6): 305–311, 2020.

Kim, S., Shin, W., Jang, S., Song, H., and Yun, S.-Y. FedRN: Exploiting k-Reliable Neighbors Towards Robust Federated Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pp. 972–981. ACM, 2022.

Kim, T., Kim, D., and Yun, S.-Y. Flr: Label-mixture regularization for federated learning with noisy labels. URL https://openreview.net/forum?id=Z8A3HDgS0E.

Li, J., Socher, R., and Hoi, S. C. H. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a.

Li, J., Li, G., Cheng, H., Liao, Z., and Yu, Y. FedDiv: Collaborative Noise Filtering for Federated Learning with Noisy Labels. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 3118–3126. AAAI Press, 2024.

Li, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 965–978. IEEE, 2022.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated Optimization in Heterogeneous Networks. In *Proceedings of the Third Conference on Machine Learning and Systems, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org, 2020b.

Lu, Y., Chen, L., Zhang, Y., Zhang, Y., Han, B., Cheung, Y.-m., and Wang, H. Federated Learning with Extremely Noisy Clients via Negative Distillation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 14184–14192. AAAI Press, 2024.

Lubana, E. S., Tang, C. I., Kawsar, F., Dick, R. P., and Mathur, A. Orchestra: Unsupervised Federated Learning via Globally Consistent Clustering. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14461–14484. PMLR, 2022.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Meng, L., Qi, Z., Wu, L., Du, X., Li, Z., Cui, L., and Meng, X. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Morafah, M., Chang, H., Chen, C., and Lin, B. Federated Learning Client Pruning for Noisy Labels. 10(2):1–25. ISSN 2376-3639, 2376-3647. URL https://dl.acm.org/doi/10.1145/3706058.

Morales-Brotons, D., Vogels, T., and Hendrikx, H. Exponential Moving Average of Weights in Deep Learning: Dynamics and Benefits. *Trans. Mach. Learn. Res.*, 2024, 2024.

Nishi, K., Ding, Y., Rich, A., and Höllerer, T. Augmentation Strategies for Learning With Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021*, pp. 8022–8031. Computer Vision Foundation / IEEE, 2021.

Qi, Z., Meng, L., Chen, Z., Hu, H., Lin, H., and Meng, X. Cross-Silo Prototypical Calibration for Federated Learning with Non-IID Data. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 3099–3107, New York, NY, USA, October 2023. Association for Computing Machinery. ISBN 979-8-4007-0108-5.

Qi, Z., Meng, L., Li, Z., Hu, H., and Meng, X. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19986–19994, 2025.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E. D., Kurakin, A., and Li, C.-L. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, 2020.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Trans. Neural Networks Learn. Syst.*, 34 (11):8135–8153, 2023.

Sun, Z., Xu, Y., Liu, Y., He, W., Kong, L., Wu, F., Jiang, Y., and Cui, L. A Survey on Federated Recommendation Systems. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024. ISSN 2162-2388.

Tam, K., Li, L., Zhao, Y., and Xu, C. FedCoop: Cooperative Federated Learning for Noisy Labels. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 2298–2306. IOS Press, 2023.

Wang, Z., Zhou, T., Long, G., Han, B., and Jiang, J. FedNoiL: A Simple Two-Level Sampling Method for Federated Learning with Noisy Labels. *CoRR*, abs/2205.10110, 2022.

Wei, T., Shi, J.-X., Tu, W.-W., and Li, Y. Robust Long-Tailed Learning under Label Noise. *CoRR*, abs/2108.11569, 2021.

Wu, N., Yu, L., Jiang, X., Cheng, K.-T., and Yan, Z. 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 2023, 19th-25th August 2023, Macao, SAR, China*, pp. 4424–4432. ijcai.org, 2023.

Xiao, R., Dong, Y., Wang, H., Feng, L., Wu, R., Chen, G., and Zhao, J. ProMix: Combating Label Noise via Maximizing Clean Sample Utility. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pp. 4442–4450. ijcai.org, 2023.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2691–2699, 2015.

Xu, J., Chen, Z., Quek, T. Q. S., and Chong, K. F. E. FedCorr: Multi-Stage Federated Learning for Label Noise Correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10174–10183. IEEE, 2022.

Yang, S., Park, H., Byun, J., and Kim, C. Robust Federated Learning With Noisy Labels. *IEEE Intell. Syst.*, 37(2): 35–43, 2022.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., and Sugiyama, M. How does Disagreement Help Generalization against Label Corruption? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7164–7173. PMLR, 2019.

Zhang, J., Song, B., Wang, H., Han, B., Liu, T., Liu, L., and Sugiyama, M. BadLabel: A Robust Perspective on Evaluating and Enhancing Label-Noise Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4398–4409, 2024.

Zhou, T., Wang, S., and Bilmes, J. A. Robust Curriculum Learning: From clean label detection to noisy label self-correction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Zhou, X. and Wang, X. Federated Label-Noise Learning with Local Diversity Product Regularization. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 17141–17149. AAAI Press, 2024.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. 32. URL https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html.