FedRG: Unleashing the Representation Geometry for Federated Learning with Noisy Clients

Tian Wen[1*] Zhiqin Yang[2*] Yonggang Zhang[4] Xuefeng Jiang[1] Hao Peng[3]
Yuwei Wang[1†] Bo Han[2†]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]TMLR Group, Hong Kong Baptist University [3]Beihang University
[4]The Hong Kong University of Science and Technology

Abstract

*Federated learning (FL) suffers from performance degradation due to the inevitable presence of noisy annotations in distributed scenarios. Existing approaches have advanced in distinguishing noisy samples from the dataset for label correction by leveraging loss values. However, noisy samples recognition relying on scalar loss lacks reliability for FL under heterogeneous scenarios. In this paper, we rethink this paradigm from a representation perspective and propose FedRG (**Fed**erated under **R**epresentation **G**emometry), which follows **the principle of "representation geometry priority"** to recognize noisy labels. Firstly, FedRG creates label-agnostic spherical representations by using self-supervision. It then iteratively fits a spherical von Mises-Fisher (vMF) mixture model to this geometry using previously identified clean samples to capture semantic clusters. This geometric evidence is integrated with a semantic-label soft mapping mechanism to derive a distribution divergence between the label-free and annotated label-conditioned feature space, which robustly identifies noisy samples and updates the vMF mixture model with the newly separated clean dataset. Lastly, we employ an additional personalized noise absorption matrix on noisy labels to achieve robust optimization. Extensive experimental results demonstrate that FedRG significantly outperforms state-of-the-art methods for FL with data heterogeneity under diverse noisy clients scenarios. Our code is publicly available at* https://github.com/Tianjoker/IFNLL

1. Introduction

Federated learning (FL) [1, 2] enables collaborative model training across multiple data sources without centralizing raw data. In FL, local models are trained on client data and subsequently aggregated on the server to obtain a robust global model [3, 4]. Nevertheless, the distributed architecture introduces critical challenges. Specifically, the data distribution across clients is inherently heterogeneous, also known as the Non-IID problem [5–7]. Furthermore, the variability or scarcity of labeling experts often results in the presence of noisy labels within these heterogeneous datasets [8, 9]. These combined factors critically hinder the performance and practical deployment of FL sys-

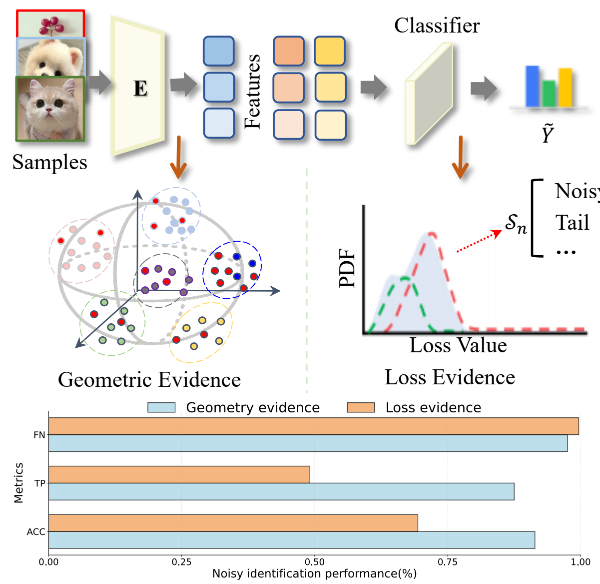


Figure 1. Comparison of noise identification performance between the existing method and FedRG under severe data heterogeneity. The existing methods mainly identify based on the distribution of scalar loss values. In heterogeneous scenarios, the loss value alone becomes unreliable, while FedRG provides more robust spatial geometry to filter out noise samples. Higher FN (False Negatives) but lower TP (True Positives) indicates that noise identification based on loss value wrongly classifies many correct labels as noise labels. The red circle denotes that the noisy samples we want to identify based on the geometric evidence.

*Equal contribution.
†Corresponding author: ywwang@ict.ac.cn,bhanml@comp.hkbu.edu.hk

tems [10, 11].

Existing methods have made notable progress in enhancing the robustness of FL against noisy labels [12–14]. Some approaches adopt robust aggregation or training strategies to alleviate the adverse impact of noisy samples [15–18]. In traditional machine learning, loss values serve as the reliable indicator for identifying noisy samples [19–21]. Drawing from this intuition, several FL studies distinguish between noisy and clean samples by leveraging insights from conventional label noise research [13, 15]. For example, some methods improve FL robustness in noisy environments by exploiting the distribution of losses [15, 22]. There are also other works utilize loss values in a more granular manner. For instance, FedCorr [13] employs a Gaussian mixture model (GMM) to separate noisy and clean data based on loss values, while FedClean [8] assesses the consistency between annotated and inferred labels using the same metric.

Built upon existing methods, it is natural that we raise a key ***Question 1:*** *Can the loss value be reliably used to identify noisy labels in heterogeneous FL scenarios?* Recall that label skew represents a primary form of heterogeneity in FL. This local label imbalance often results in a long-tailed distribution within client datasets [23, 24]. Training on such imbalanced data introduces a strong bias in the local model toward dominant classes, impairing the classifier's ability to accurately discern samples from rare classes [25]. As a result, tail-class samples inherently exhibit higher loss values, irrespective of label accuracy [26, 27]. Apart from the loss value, we turn our attention to the representation space. **To answer *Question 1*,** we shift our focus from mere loss values to the representation space itself. Furthermore, we propose the representation geometry principle to identify noisy labels: the intrinsic representation geometry, which is independent of labels, should align consistently with the geometry shaped by the constraints of annotated labels. We illustrate the differences between these two strategies in Fig. 1.

As a consequence, emerging another ***Question 2****: How to measure the geometric representation divergence with and without the annotated labels?* Revisit the self-supervised methods that have attracted substantial attention, owing to their capacity to extract data representations without dependence on labels [28–30]. **To address *Question 2*,** we leverage instance discrimination to ensure that the features of local data are distributed across a hypersphere [31]. To model the intrinsic semantic structure on this hypersphere, we adopt the von Mises-Fisher (vMF) distribution [32, 33]. Then the measurement of geometric divergence is then formulated through a cross-validation between two independently derived cluster distributions within vMF: label-free geometry representation Γ and annotated label-dependent geometry representation B. Drawing the intuition from above analysis, we propose **FedRG** (**Fed**erated under **R**epresentation **G**emometry), to tackle noisy clients under data heterogeneity in FL. Based on the recognized noisy labels, we train the whole model coupled with noise absorption matrix which personalized estimates the noisy transition probability for each client. Overall, our contributions are summarized as follows:

- We rethink the paradigm of existing federated learning with noisy clients beyond the loss value. From the representation perspective, we propose the principle of representation geometry priority to guide the noise identification procedure.
- We propose FedRG framework to distinguish noisy samples based on the geometric evidence under the vMF distribution. We draw the inspiration that a clean sample's intrinsic feature clustering should align with the clustering pattern implied by its annotated label, while mislabeled samples exhibit geometric inconsistency.
- We conduct comprehensive experiments across multiple datasets with 4 different noisy scenarios where 2 specific scenarios occur due to the nature of the distributed. The experimental results consistently demonstrate the efficacy of FedRG. Further ablation studies also shows the effectiveness and indispensability of each component.

2. Related Work

2.1. Federated Learning

In recent years, federated learning (FL) [34, 35] has been widely studied and applied in privacy-sensitive domains such as medical analysis [15, 36] and remote sensing [37]. FedAvg [3] aggregates local models through weighted averaging and converges efficiently under the IID assumption. To cope with the Non-IID data (i.e. statistical heterogeneity), FedProx [38] introduces a proximal term while SCAFFOLD [39] uses control variates for variance reduction. MOON [40] conducts representation alignment via contrastive learning [28]. While these methods demonstrate effectiveness on clean datasets, their robustness diminishes under noisy labels, which is a common concern for real-world scenarios.

2.2. Noisy Label Learning

Noisy label learning (NLL) [41] aims to mitigate the negative impact of noisy labels and train models in centralized scenarios. Regularization methods are commonly utilized by designing robust loss functions [42–44]. Symmetric cross-entropy [44] augments CE with reverse cross-entropy (RCE). Mean Absolute Error (MAE) [42] provides formal robustness properties relative to CE. Since these approaches operate entirely at the client side, they can be incorporated into FL by substituting the local training objective, without altering the FL protocol.

2.3. Noisy Labels in Federated Learning

Centralized noise-robust training is difficult to deploy in FL due to privacy and communication constraints [45]. Prior federated methods facing label noise fall into two streams [10]: *Client-level assessment* scores the reliability of each client and adapts training accordingly. FedNoRo [15] separates clean and noisy clients using a GMM-based criterion [46] and then aligns unreliable clients via knowledge distillation [47]. FedClean [8] alternates updates while explicitly differentiating client types; *Sample-level assessment* instead targets example credibility on-device. Despite their efficacy, these approaches share notable weaknesses. Most client or data detectors hinge on the *small-loss* heuristic[8, 13], which is unreliable. Deep models eventually memorize mislabeled data, driving their losses down and causing noisy samples to be kept, while truly hard but correctly labeled examples often retain large losses and are discarded [48]. More detailed related works are listed in Appendix Sec.5.

Motivated by these observations, we propose a federated noise-robust framework that, from a geometric representation perspective, identifies clean and noisy samples to guide robust optimization. Unlike previous approaches that rely solely on small-loss heuristics [8, 13], our method exploits geometric relationships within the feature space to capture structural consistency, enabling effective learning under heterogeneous federated conditions.

3. Methodology

3.1. Preliminary

Federated Learning. We investigate the classic federated learning framework in which K distributed devices (clients) collaboratively train a C-class classification model coordinated by the central server. Each client $k \in [K]$ holds a private dataset $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}, x \in \mathcal{X}, y \in \mathcal{Y} = \{1, \dots, C\}$. Without loss of generality, we explicitly allow statistical heterogeneity across clients, where local samples are drawn i.i.d. from client-specific distribution P_k on $\mathcal{X} \times \mathcal{Y}$ and these distributions differ across clients (i.e., Non-IID setting). The server aims to learn a global model parameterized as $\theta \in \Theta$ with a predictor $f_\theta : \mathcal{X} \to \Delta^{C-1}$, by minimizing the following objective:

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) = \sum_{k=1}^{K} \omega_k^{(t)} \mathbb{E}_{(x,y)\sim P_k}\Big[\ell\big(f_\theta(x), y\big)\Big], \quad (1)$$

where ℓ is the loss function and $\omega_k^{(t)}$ determines the aggregated weight of the k-th client. During each communication round $t \in \{1, \dots, T\}$, the server broadcasts the latest global parameters θ_t to a subset of clients $\mathcal{S}_t \subseteq [K]$. Each participating client performs E local epochs of stochastic optimization on private data to obtain the updated model $\theta_{t+1}^{(k)}$, and then the server aggregates $\theta_{t+1} = \sum_{k\in\mathcal{S}_t} \omega_k^{(t)} \theta_{t+1}^{(k)}$ for the next training round.

Federated Label Noise. In practical FL deployments, local annotations are often imperfect due to weak supervision or crowd sourcing. Following most previous works [8, 15], we adopt the instance-independent noise model and describe label corruption by a transition kernel $\mathcal{C}(\tilde{y} \mid y)$, where $\sum_{\tilde{y}=1}^{C} \mathcal{C}(\tilde{y} \mid y) = 1$ and $\mathcal{C}(\tilde{y} \mid y) \in [0, 1]$. For client k, the observed label is drawn as $\tilde{y} \sim \mathcal{C}(\cdot \mid y)$, and the local training set is $\tilde{\mathcal{D}}_k = \{(x_i^{(k)}, \tilde{y}_i^{(k)})\}_{i=1}^{n_k}$.

We study two kinds of corruption settings that labels may be corrupted in federated systems. (i) Globalized corruption: a *single* kernel $\mathcal{C}(\tilde{y} \mid y)$ governs flips for all clients; equivalently, $p(\tilde{y} = j \mid y = i) = \mathcal{C}(j \mid i)$ is identical throughout the federation, where $p(\tilde{y} = j \mid y = i)$ is the probability of flipping from the true label y of class i to the wrong label $\tilde{y}$ of class j. (ii) Localized corruption: each client k has specialized kernel $\mathcal{C}_k(\tilde{y} \mid y)$ that operates only within its local label support $\mathcal{Y}_k \subseteq \mathcal{Y}$. In particular, $\mathcal{C}_k(\tilde{y} \mid y) = 0$ whenever $\tilde{y} \notin \mathcal{Y}_k$. This captures realistic client-specific biases and distinguishes our setting from ordinary label-noise assumptions that allow flips to any global class regardless of local availability.

For both corruption settings, we respectively consider the symmetric flipping and pairwise flipping (i.e., pairflip) methods, aligning with prior work [10]. In the symmetric flipping, each true label can be randomly flipped to any other admissible alternative class with equal probability. By contrast, pairflip indicates that each class is primarily confused with one specific class. More details are provided in Sec. 4.2.

3.2. Motivation

Federated learning in the wild faces severe non-IID data and heterogeneous label noise, under which the prevalent small-loss heuristic often collapses because losses conflate mislabels, rare yet clean examples, and domain shift. Warm-up with noisy supervision further pulls adsorbs features toward erroneous prototypes, contaminates the manifold geometry, and degrades separability. These issues make prediction-space proxies unreliable from the outset. We therefore advocate the representation geometry priority principle, where robust assessment of sample cleanliness arises from evidence in the intrinsic representation manifold rather than from the easily contaminated prediction space, and we replace loss-based heuristics with a geometry-driven probabilistic model.

3.3. Label Decoupled Spherical Representation

To conduct the representation geometry priority principle, we require an initialization that sculpts a reliable manifold

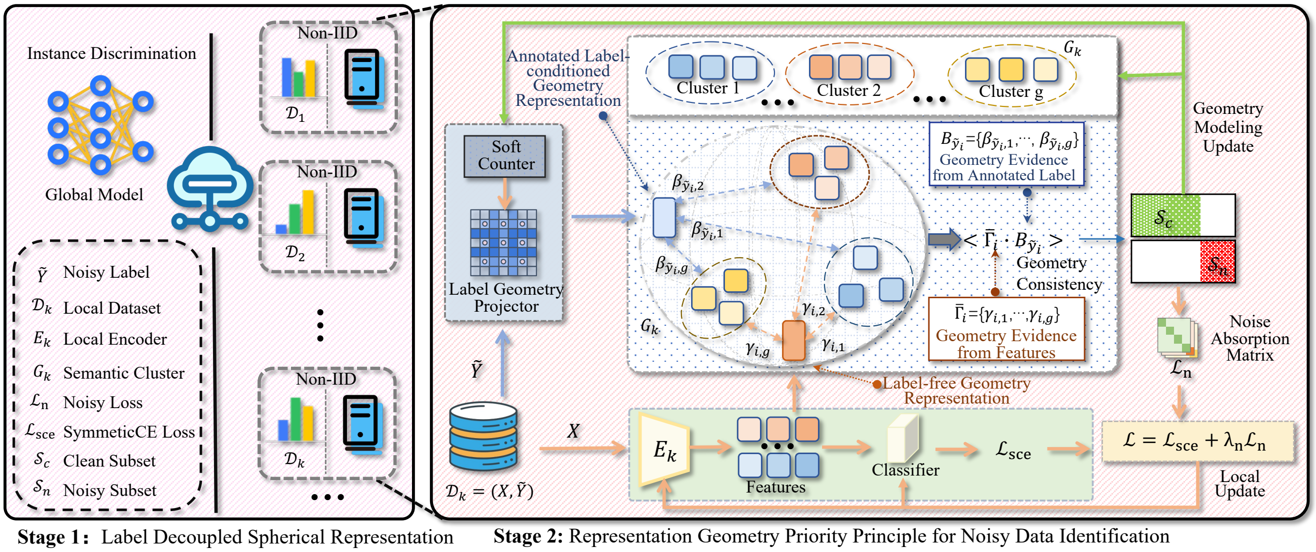


Figure 2. The overview of FedRG which is consisted by two stage. The label decoupled spherical representation stage gets the hypersphere of samples. Then, we follow the representation geometry priority principle to split the clean and noisy samples based on the distribution divergence in the vMF distribution. The pseudocode overview of FedRGis provided in Appendix Sec.2

without invoking any annotated supervision and whose inductive biases are commensurate with the spherical statistics used later. We therefore resort to instance discrimination (i.e., SimCLR) for label-free pretraining [31]. Each sample x is stochastically augmented twice and encoded to a unit-norm representation on the hypersphere as $z = \frac{f_\theta(x)}{\|f_\theta(x)\|} \in \mathbb{S}^{d-1}$, which is optimized with the symmetric NT-Xent (InfoNCE) objective over $2b$ views, where b is the batch size. For a positive pair (i, i^+), the training loss is shown as:

$$\mathcal{L}_{\text{SimCLR}} = \frac{1}{2b} \sum_{i=1}^{2b} -\log \frac{\exp\left(z_i^\top z_{i^+}/\tau\right)}{\sum_{\substack{a=1 \\ a \neq i}}^{2b} \exp\left(z_i^\top z_a/\tau\right)}, \quad (2)$$

where the temperature τ regulates the attraction–repulsion field on $\mathbb{S}^{d-1}$. This objective enforces *alignment* of positive views with similar semantics while promoting *uniformity* of features on the sphere, yielding a clean geometric substrate with dense, fine-grained semantic neighborhoods that emerge as directional concentrations, whereas anomaly or heavily corrupted instances gravitate toward the uniform background.

After T_1 rounds of label-free spherical representation learning, we obtain a well-conditioned spherical representation manifold on $\mathbb{S}^{d-1}$, which we subsequently exploit for robust training under label noise.

3.4. Representation Geometry Priority Principle

Given the label-decoupled spherical representations produced in Sec. 3.3, we now instantiate the representation geometry priority principle by placing an explicit probabilistic model on the hypersphere $\mathbb{S}^{d-1}$. The key idea is that assessment of sample cleanliness should arise from the intrinsic geometry of these normalized representations rather than from the easily contaminated prediction space.

3.4.1. Geometric Representation via vMF Distribution

On the SimCLR induced hyperspherical manifold, we expect that semantic regularities to manifest as directional concentrations: samples sharing stable, view-invariant factors (e.g., textures or poses) should aggregate around a few preferred directions, while out-of-distribution or heavily corrupted samples are approximately uniform. To turn this intuition into a probabilistic model that respects the representation geometry, we therefore seek a distribution on $\mathbb{S}^{d-1}$ that can represent such directional modes together with a uniform background in a geometrically consistent way. The von Mises–Fisher (vMF) distribution is a natural choice for the directional component. It is defined on the unit hypersphere and smoothly interpolates between a sharply peaked mode around a mean direction and the uniform distribution when no directional preference exists [32, 49]. We therefore model the distribution of normalized representations as a mixture of vMF components and a uniform background. Formally, letting $z \in \mathbb{S}^{d-1}$ denotes the normalized representation, we assume:

$$p(z) \;=\; \pi_0\, U(z) \;+\; (1-\pi_0)\, \text{vMF}(z \,|\, \mu, \kappa)\,, \quad (3)$$

where $U(z)$ is the uniform density on $\mathbb{S}^{d-1}$ and $\{\pi_g\}_{g=0}^{G}$ are mixture weights satisfying $\sum_{g=0}^{G} \pi_g = 1$. Each vMF component corresponds to a semantic cluster (indexed by g) on the sphere, capturing a directionally concentrated mode induced by stable, view-invariant factors, while the uniform component absorbs geometry-level outliers and severely noisy instances that do not align with any semantic mode. Importantly, clusters are *label-agnostic*, as they capture geometry-driven regularities that may refine a class into sub-modes or, conversely, span cross-class motifs. This decoupling plays a key role under noisy supervision, and in the next step, we turn this mixture into sample-wise geometric evidence.

3.4.2. Label-free Geometry Representation.

Given the vMF–uniform mixture in Eq. (3), we first extract label-free geometric evidence in the form of posterior responsibilities over semantic clusters. Let $p_0(z){=}U(z)$ and $p_g(z){=}\mathrm{vMF}(z \mid \mu_g, \kappa_g)$ for $g{\geq}1$, with mixture weights $\{\pi_g\}_{g=0}^{G}$. For a representation z_i, the posterior responsibility of belonging to semantic cluster $g \in \{0, 1, \ldots, G\}$ is:

$$\gamma_{i,g} \;=\; \frac{\pi_g \, p_g(z_i)}{\sum_{h=0}^{G} \pi_h \, p_h(z_i)}, \tag{4}$$

which constitute our initial label-free geometric evidence.

To further exploit semantic stability, we refine these responsibilities using multi-view consistency. For each training sample x_i, we draw two augmented views using standard SimCLR-style stochastic augmentations and obtain unit-norm features $z_{i,1}, z_{i,2} \in \mathbb{S}^{d-1}$. Their agreement yields a scalar consistency score, which is mapped to a precision-like tempering factor r_i; this factor is then used to flatten responsibilities for low-consistency samples and to preserve sharp assignments for well-aligned ones. The full augmentation pipeline and the resulting tempered likelihoods are detailed in Appendix Sec.3.

Using these tempered likelihoods, the responsibilities become a refined vector $\tilde{\gamma}_i = (\tilde{\gamma}_{i,0}, \ldots, \tilde{\gamma}_{i,G})$ as:

$$\tilde{\gamma}_{i,g} = \begin{cases} \dfrac{\pi_0 \, \tilde{p}_0(z_i; r_i)}{\pi_0 \, \tilde{p}_0(z_i; r_i) \;+\; \sum_{h=1}^{G} \pi_h \, \tilde{p}_h(z_i; r_i)}, & g = 0, \\ \dfrac{\pi_g \, \tilde{p}_g(z_i; r_i)}{\pi_0 \, \tilde{p}_0(z_i; r_i) \;+\; \sum_{h=1}^{G} \pi_h \, \tilde{p}_h(z_i; r_i)}, & g \neq 0. \end{cases} \tag{5}$$

which we refer to as our *label-free geometric representation*. This representation will serve as the primary evidence for noisy-label identification and robust training in subsequent sections. Crucially, to prevent noisy examples from biasing the established geometric model, we update the vMF parameters only with samples certified as clean by geometry, which follows standard practice and is referred to the Appendix Sec.3.4.

3.4.3. Annotated Label-conditioned Geometry Representation.

The preceding stage maps samples onto a spherical representation manifold, where the vMF–uniform mixture, together with the refined responsibilities $\tilde{\gamma}_{i,g}$, provides a geometry-driven notion of semantic evidence over components $g \in \{0, \ldots, G\}$. To incorporate the observed (potentially noisy) labels in a compatible way, we construct a complementary label-based evidence representation in the *same* geometric component space. Concretely, we treat the label-free semantic components as a stable reference and, for each class c, summarize how its labeled samples are distributed across the semantic components.

Formally, we estimate a class-to-geometry distribution via Dirichlet smoothing ($\alpha > 0$):

$$\beta_{c,g} \;=\; \frac{\sum_{i:\,\tilde{y}_i = c} \tilde{\gamma}_{i,g} + \alpha}{\sum_{g'=1}^{G} \Big(\sum_{i:\,\tilde{y}_i = c} \tilde{\gamma}_{i,g'} + \alpha \Big)}, \tag{6}$$

where $\tilde{y}_i$ denotes the observed (noisy) label of sample i, and we explicitly exclude the uniform background component $g = 0$ from the sums to avoid contamination by geometry-level outliers. Given the class c, the vector $\mathrm{B}_c = \{\beta_{c,1}, \ldots, \beta_{c|G}\}$ describes how that class is expressed across the semantic geometry.

In this way, $\beta_{c,g}$ embeds label evidence into the same semantic component space as the geometric responsibilities $\tilde{\gamma}_{i,g}$, aligning class space with the representation geometry and enabling a direct comparison between geometry-based and label-based evidence in subsequent stages.

3.4.4. Geometric Consistency for Noisy Detection

We have obtained two sources of evidence: the geometry-driven evidence from the feature representations, and the label-driven evidence from the observed (potentially noisy) labels. Here, we assess the consistency between these two types of evidence within the same geometric space. Specifically, we examine how well the label information aligns with the geometric representation of each sample.

We first re-normalize augmentation-aware responsibilities over *non-background* components to place every sample on the cluster simplex $\bar{\gamma}_{i,g} = \frac{\tilde{\gamma}_{i,g}}{\sum_{g'=1}^{G} \tilde{\gamma}_{i,g'}}$ where $g \in \{1, \ldots, G\}$. Then, given the class-to-geometry distribution, the geometry–label consistency of sample i with its observed label $\tilde{y}_i$ is defined as an inner product on the simplex:

$$\tilde{q}_i(\tilde{y}_i) = < \bar{\Gamma}_i \cdot \mathrm{B}_{\,\tilde{y}_i} > \in (0, 1) \tag{7}$$

where $\bar{\Gamma}_i = \{\tilde{\gamma}_{i,1}, \ldots \tilde{\gamma}_{i,g}\}$ is the distribution of responsibilities for feature z_i. A larger $\tilde{q}_i$ value indicates that the clusters which *geometrically* explain z_i are also *semantically* aligned with its observed class $\tilde{y}_i$, whereas smaller values flag potential mislabels, out-of-domain drift, or severe client shift.

Treat each sample as competing among (i) correctly labeled within the semantic geometry, (ii) mislabeled but still on-geometry, and (iii) background (uniform component).

Table 1. Experimental results on CIFAR10 with two label noise patterns. The **bold** denotes the best result while the underlined denotes the second-best result.

Noise Type	Symmetric Label Noise						Pairflip Label Noise					
Noise Pattern	Localized			Globalized			Localized			Globalized		
Metric	**Accuracy**	Precision	F-score	**Accuracy**	Precision	F-score	**Accuracy**	Precision	F-score	**Accuracy**	Precision	F-score
FedAvg [3]	34.20	46.17	30.96	42.58	42.64	41.57	44.44	50.46	41.43	43.89	45.50	42.39
FedProx [38]	38.88	62.70	34.34	46.18	47.16	44.48	49.65	57.35	46.42	46.62	49.22	44.83
MOON [40]	44.88	45.73	43.20	44.88	45.73	43.20	51.68	58.97	48.80	47.54	48.76	46.22
CCVR [50]	42.63	42.44	41.53	42.63	42.44	41.53	45.16	53.53	42.30	43.69	45.20	41.95
FedDisco [51]	43.07	43.19	41.85	43.07	43.19	41.85	44.71	49.47	41.58	43.59	45.30	42.22
FedSAM [52]	41.05	60.13	37.35	46.63	46.75	45.53	50.69	58.41	43.45	59.03	64.00	54.80
FedExp [53]	42.06	42.04	41.05	42.06	42.04	41.05	45.22	52.48	42.31	42.33	44.15	40.54
FedInit [54]	36.22	48.86	33.54	42.42	42.78	41.44	46.33	53.86	43.51	42.00	43.80	40.33
SymmetricCE [16]	55.01	63.90	49.78	59.23	64.55	54.38	45.50	51.95	41.04	48.25	48.31	44.74
FedLSR [55]	55.32	57.45	48.96	55.32	57.45	48.96	35.12	25.86	25.94	35.21	24.93	25.66
FedNoRo [15]	43.44	43.46	42.42	40.13	40.41	39.33	46.67	48.41	46.34	43.73	43.67	42.90
FedELC [56]	40.83	42.33	39.66	47.64	56.57	45.76	45.94	53.19	44.18	54.31	60.11	52.61
FedCorr [13]	39.18	44.51	42.57	35.83	48.79	41.05	40.39	39.01	37.68	36.82	34.24	41.34
FedClean [8]	45.82	42.33	43.55	39.87	38.66	41.25	47.82	46.53	43.05	41.66	42.61	42.88
FedRG	**59.99**	**70.32**	**55.31**	**63.29**	**72.94**	**59.14**	**61.52**	**73.01**	**57.09**	**64.88**	**64.88**	**60.98**

Normalizing these three hypotheses yields a *clean posterior*

$$P_i^{\text{clean}} = \frac{\tilde{q}_i(\tilde{y}_i)}{\tilde{q}_i(\tilde{y}_i) + \sum_{c\neq\tilde{y}_i}\sum_{g=1}^{G}\beta_{c,g}\,\bar{\gamma}_{i,g} + \tilde{\gamma}_{i,0}}. \tag{8}$$

As $\sum_c \beta_{c|g} = 1$ and $\sum_g \bar{\gamma}_{i,g} = 1$, the on-geometry terms sum to 1, giving the compact form $P_i^{\text{clean}} = \frac{\tilde{q}_i(y_i^{\text{obs}})}{1+\tilde{\gamma}_{i,0}}$,which naturally discounts samples with high background mass $\tilde{\gamma}_{i,0}$. Following the prior works [8, 13], we firstly employ a two-component Gaussian mixture model (GMM) on the per-sample P_i^{clean} to separate all samples into two distinct clusters: $\mathcal{S}_{clean}$ set denotes the clean samples while $\mathcal{S}_{noisy}$ indicates the noisy sample set.

3.5. Overview of Robust optimization

Geometry-model optimization. We initialize the vMF model and the class-to-geometry matrix at random. To project features and observed labels more faithfully onto the spherical representation space, we only update the vMF model and the class-to-geometry matrix with the clean dataset during local client training. Implementation details for the full pipelines are described in the Appendix Sec.3.5.

Global model optimization. The total training objective combines the batch-level robust supervised loss for supervised learning with the forward corrected loss for noise absorption:

$$\mathcal{L} = \lambda_s \mathcal{L}_{\text{SCE}} + \lambda_{\text{n}} \mathcal{L}_{\text{n}} \ . \tag{9}$$

Here, λ_s and λ_{n} are the trade-off hyperparameters. $\mathcal{L}_{\text{SCE}}$ is the noise-resilient loss symmetric cross-entropy (SCE) [16]. To explicitly model noisy samples, we introduce a noise absorption matrix $T \in \mathbb{R}^{C\times C}$, where $T_{c,c'} \approx P(\tilde{y} = c' \mid y^{\star} = c)$. Following prior work [57], T is implemented as an additional linear layer appended after the classifier head. It does not require the model to predict the unknown true label directly. Instead, the classifier outputs the label distribution p_i, and the noise absorption layer maps it to the observed noisy-label space via the forward correction $p_i T$. Concretely, for samples in the noisy set $\mathcal{S}_{\text{noisy}}$, we maximize the likelihood of the observed labels under the forward-corrected distribution:

$$\mathcal{L}_{\text{n}} = \frac{\sum_{i=1}^{B} m_i \cdot p_i T}{\sum_{i=1}^{B} m_i + \varepsilon},$$

where $m_i \in [0,1]$ is a per-sample weight (e.g., an indicator or a calibrated soft weight) and $\varepsilon > 0$ is a small constant for numerical stability. Ablation experiments in Sec. 4.4 demonstrate the advantage of the noise absorption matrix.

4. Experimental Results

Our evaluation reports the main results and targeted ablations; additional experiments and a detailed cost analysis are provided in the Supplementary Sec.4 and Sec.5.

4.1. Experimental Details

Datasets, Models, and Metrics: We evaluate our method on three representative datasets: CIFAR-10 [58], SVHN [59], and CIFAR-100 [58]. Among them, CIFAR-100 [58] contains richer category diversity with 100 fine-grained classes. For model architectures, we adopt the ResNet family [60] according to prior representative studies [8, 13]. We report multiple evaluation metrics, including accuracy (ACC), precision, and F1-score, following prior works [19, 55, 56]. Additionally, we include the Clean–Noise

Table 2. Experimental results on SVHN and CIFAR100 with globalized noise setting. The **bold** denotes the best result while the <u>underlined</u> denotes the second-best result.

Dataset	**SVHN**						**CIFAR100**					
Noise Type	Symmetric Label Noise			Pairflip Label Noise			Symmetric Label Noise			Pairflip Label Noise		
Metric	Accuracy	Precision	F-score	Accuracy	Precision	F-score	Accuracy	Precision	F-score	Accuracy	Precision	F-score
FedAvg [3]	43.52	42.79	42.29	45.73	41.04	38.08	39.27	40.81	39.53	41.49	38.79	40.53
FedProx [38]	46.89	53.47	46.41	51.86	43.44	48.73	34.98	40.16	35.03	40.95	34.32	39.83
MOON [40]	45.15	44.79	43.51	46.10	42.55	41.11	36.57	39.78	36.69	40.33	36.14	39.45
CCVR [50]	45.84	41.20	45.64	44.10	42.85	35.95	39.03	41.29	38.96	42.37	38.51	40.81
FedDisco [51]	43.11	41.89	43.37	43.94	40.42	36.91	38.11	40.87	38.22	41.45	37.54	40.57
FedSAM [52]	31.09	<u>71.31</u>	25.59	<u>75.25</u>	15.41	<u>65.69</u>	42.59	<u>44.61</u>	42.68	<u>45.64</u>	41.62	<u>44.29</u>
FedExp [53]	46.23	41.34	45.20	43.78	43.55	66.69	38.93	41.51	39.04	41.95	38.51	41.20
FedInit [54]	44.75	40.42	43.86	43.70	42.00	37.05	38.29	40.46	38.21	41.08	37.68	40.19
SymmetricCE [16]	<u>57.82</u>	50.69	<u>55.28</u>	52.28	<u>50.56</u>	45.71	**56.36**	39.25	**58.86**	40.52	**55.80**	38.65
FedLSR [55]	55.21	11.07	53.07	–	–	–	16.14	18.72	6.57	8.17	8.26	10.23
FedNoRo [15]	37.07	39.59	34.70	39.29	34.19	37.27	34.90	39.29	34.95	40.07	34.43	39.19
FedELC [56]	54.24	42.72	53.79	48.69	**50.70**	43.06	41.75	39.93	41.89	37.10	40.83	35.97
FedCorr [13]	49.53	42.33	48.60	43.98	45.22	39.84	36.75	39.55	40.05	40.25	42.67	38.48
FedClean [8]	54.28	40.77	52.69	44.66	47.77	42.16	39.57	38.88	40.22	36.71	39.83	36.79
FedRG	**62.70**	**74.55**	**58.83**	**80.62**	42.88	**68.83**	<u>51.95</u>	**56.49**	<u>54.31</u>	**60.80**	<u>51.25</u>	**55.46**

Recognition Accuracy (CRA), which measures how accurately the model distinguishes clean samples from noisy ones. Specifically, CRA is computed as the proportion of samples whose predicted clean/noisy status matches the ground truth, reflecting the model's ability to correctly identify both clean and noisy data instances.

Baselines: We compare the proposed FedRG with the representative FL methods: Representative FL methods: FedAvg [3], FedProx [38], MOON [40], CCVR [50], FedDisco [51], FedSAM [52], FedExp [53], FedInit [54].Representative noise-robust methods: Symmetric CE [44], FedCorr [13], FedClean [8], FedLSR [55], FedELC [56] and FedNoRo [15].

4.2. Implementation Details

Federated Learning Setup: Following previous works [50, 61], we conduct experiments under a strongly non-IID setting, where the client data is partitioned according to a Dirichlet distribution with concentration parameter $\alpha = 0.1$. Unless otherwise stated, we adopt $K = 10$ clients for all datasets. More details about the hyperparameters are provided in Appendix Sec.1.

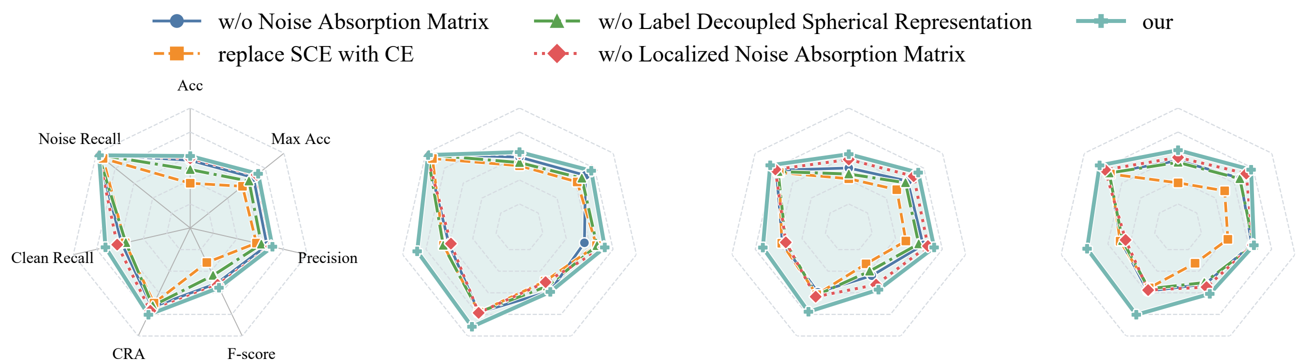
Federated Label Noise Setup: Following existing methods [10, 15, 56], we keep the noise rate fixed across all clients and then corrupt the local labels of each client using a transition matrix with noise level $\epsilon = 0.4$. In the *globalized* corruption setting, all clients share a common transition matrix, *i.e.*, $p(\tilde{y} = j \mid y = i)$. In the *localized* setting, each client maintains its own transition matrix that operates within its local label. For both settings, we consider two standard noise patterns: (i) *symmetric flipping* and (ii) *pairflip*.

4.3. Main Results

Tabs. 1 and 2 summarize the overall performance of different baselines under both symmetric and pairflip label-noise settings across CIFAR10, SVHN, and CIFAR100. Classical FL algorithms such as FedAvg, FedProx, and MOON degrade significantly under label corruption. Methods that explicitly incorporate regularization or noise modeling (e.g., SymmetricCE) achieve moderate improvements in symmetric corruption but show limited improvement under pairflip corruption.

In contrast, FedRGconsistently achieves the best or second-best performance across nearly all datasets, metrics, and noise patterns. For CIFAR10 (Tab 1), FedRGsurpasses all baselines under both localized and globalized noise, with remarkable gains in accuracy and F-score. For instance, FedRGimproves symmetric-globalized accuracy by a large margin over the second-best method. For SVHN (Tab. 2), FedRGdemonstrates substantial robustness, particularly under pairflip noise where the performance of most baselines collapses; FedRGmaintains high accuracy (80.62) and precision (68.83), indicating its strong capability to correct and filter corrupted labels. Similar trends are observed on CIFAR100, where FedRGoutperforms competing approaches by a clear margin despite the increased task difficulty and finer-grained label space.

Overall, these results validate the effectiveness of FedRGin handling heterogeneous and severe label noise in federated environments. Its consistent superiority across datasets and noise types suggests that the geometric-consistency-based noise identification and the personalized noise absorption matrix jointly suppress misleading gradients and mitigate the negative impact of noisy client labels.



(a) Symmetric with localized noise (b) Symmetric with globalized noise (c) Pairflip with localized noise (d) Pairflip with globalized noise

Figure 3. Comparison of the performance among different schemes under four heterogeneous noise scenarios. Each radar chart corresponds to a specific type of label noise: (a) symmetric with localized noise, (b) symmetric with globalized noise, (c) pairflip with localized noise, and (d) pairflip with globalized noise.

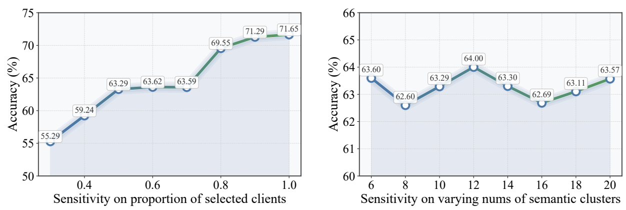


Figure 4. Comparison of model performance under different numbers of clients and semantic clusters.

4.4. Ablation Study

As illustrated in Fig. 3, we compare the FedRG with four degraded variants: ❶ w/o Noise Absorption Matrix, which removes the forward noise absorption matrix; ❷w/o Label Decoupled Spherical Representation, which eliminates the unsupervised SimCLR in stage1; ❸ replaces the SCE loss with a standard CE loss; and ❹ w/o Localized Noise Absorption Matrix, which aggregates the noise absorption matrix across rounds.

Removing the forward noise-absorption matrix causes a clear drop across almost all metrics, especially in Clean Recall and CRA. This confirms the importance of using the transition matrix to suppress the influence of noisy labels. Second, the removal of the label decoupled spherical representation stage also leads to noticeable degradation. In the early rounds, clients receive highly corrupted supervision.Thus, the unsupervised stages provides noise-free representation learning that significantly benefits downstream robustness. Third, replacing SCE with standard CE indicates that SCE is essential for mitigating gradient bias caused by incorrect labels and preventing overfitting to noise. Finally, disabling the aggregation of the noise absorption matrix mildly affects performance, particularly in the globalized noise setting, suggesting that global consolidation of noise patterns further enhances stability and allows clients to share consistent noise-handling information. Across all four scenarios, FedRGconsistently achieves the best performance on nearly all metrics. The consistent gaps between FedRG and its variants demonstrate that each component contributes meaningfully to the overall robustness to both localized and globalized patterns.

We further perform ablation studies about the client sampling rate at each round and the number of clusters G, the results are listed in Fig. 4. Based on Fig. 4 left, the performance of FedRG continues to improve with the increase in the number of participating clients in each round. The selection of the number of semantic clusters is also a key parameter for FedRG. The Fig. 4 right shows the performance of FedRG under different numbers of semantic clusters. In our other experiments, we set the number of labels to 10. In addition, we also test the impact of different numbers of semantic clusters ranging from 6 to 20 on the performance. It can be seen that the performance remains balanced overall and is robust to the number of semantic clusters.

5. Conclusion

We introduce FedRG to addresses the challenges of noisy labels under data heterogeneity by the representation geometry principle over traditional loss-based methods. FedRG robustly identifies noisy samples through geometric consistency and employs a personalized noise absorption matrix for optimization. Extensive experiments demonstrate that FedRG outperforms the state-of-the-art baselines.

Limitations and Future Directions: FedRG currently focuses on image classification, leaving graph and natural language domains unexplored. As the first work on OOD generalization on graphs [62] demonstrates, structural distribution shift poses unique challenges; future work could extend our geometric framework to such settings. Additionally, leveraging the first benchmark and library for curriculum learning [63, 64], incorporating sample difficulty into our clean/noisy partition could enable finer-grained robust training in heterogeneous FL scenarios.

Acknowledgement.

Tian Wen, Xuefeng Jiang and Yuwei Wang were supported by the National Key Research and Development Program of China (Grants No. 2023YFB2703700). Zhiqin Yang, Yonggang Zhang, and Bo Han were supported by NSFC General Program No. 62376235 and RGC GRF No. 12200725. Yonggang Zhang was also funded by Inno HK Generative AI R&D Center. Hao Peng was supported by the Local Science and Technology Development Fund of Hebei Province Guided by the Central Government of China, through grant 254Z9902G.

References

[1] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021. 1

[2] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 1

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *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*, pages 1273–1282. PMLR, 2017. 1, 2, 6, 7

[4] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3347–3366, 2021. 1

[5] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022. 1

[6] Yuan Wang, Huazhu Fu, Renuga Kanagavelu, Qingsong Wei, Yong Liu, and Rick Siow Mong Goh. An aggregation-free federated learning for tackling data heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26233–26242, 2024.

[7] Zhiqin Yang, Yonggang Zhang, Chenxin Li, Yiu-ming Cheung, Bo Han, and Yixuan Yuan. Fedgps: Statistical rectification against data heterogeneity in federated learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1

[8] Xiaoqian Jiang and Jing Zhang. Fedclean: A general robust label noise correction for federated learning. In *Forty-second International Conference on Machine Learning*, 2025. 1, 2, 3, 6, 7

[9] Junyi Li, Jian Pei, and Heng Huang. Communication-efficient robust federated learning with noisy labels. In Aidong Zhang and Huzefa Rangwala, editors, *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 914–924. ACM, 2022. 1

[10] Siqi Liang, Jintao Huang, Dun Zeng, Junyuan Hong, Jiayu Zhou, and Zenglin Xu. Fednoisy: Federated noisy label learning benchmark. *CoRR*, abs/2306.11650, 2023. 2, 3, 7

[11] Zhiqin Yang, Yonggang Zhang, Yu Zheng, Xinmei Tian, Hao Peng, Tongliang Liu, and Bo Han. Fedfed: Feature distillation against data heterogeneity in federated learning. *Advances in neural information processing systems*, 36:60397–60428, 2023. 2

[12] Miao Yang, Hua Qian, Ximin Wang, Yong Zhou, and Hongbin Zhu. Client selection for federated learning with label noise. *IEEE Trans. Veh. Technol.*, 71(2):2193–2197, 2022. 2

[13] Jingyi Xu, Zihan Chen, Tony Q. S. Quek, and Kai Fong Ernest Chong. 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*, pages 10174–10183. IEEE, 2022. 2, 3, 6, 7

[14] Xinyuan Ji, Zhaowei Zhu, Wei Xi, Olga Gadyatskaya, Zilong Song, Yong Cai, and Yang Liu. Fedfixer: Mitigating heterogeneous label noise in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12830–12838, 2024. 2

[15] 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 2023, 19th-25th August 2023, Macao, SAR, China*, pages 4424–4432. ijcai.org, 2023. 2, 3, 6, 7

[16] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vi-*

sion, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 322–330. IEEE, 2019. 6, 7

[17] Hidde Lycklama, Lukas Burkhalter, Alexander Viand, Nicolas Küchler, and Anwar Hithnawi. Rofl: Robustness of secure federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 453–476. IEEE, 2023.

[18] Bixiao Zeng, Xiaodong Yang, Yiqiang Chen, Zhiqi Shen, Hanchao Yu, and Yingwei Zhang. Fedes: federated early-stopping for hindering memorizing heterogeneous label noise. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5416–5424, 2024. 2

[19] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546, 2018. 2, 6

[20] Xian-Jin Gui, Wei Wang, and Zhang-Hao Tian. Towards understanding deep learning from noisy labels with small-loss criterion. In *IJCAI*, 2021.

[21] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*. 2

[22] Xiaochen Zhou and Xudong Wang. Federated label-noise learning with local diversity product regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17141–17149, 2024. 2

[23] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR, 2022. 2

[24] Yiqun Diao, Qinbin Li, and Bingsheng He. Towards addressing label skews in one-shot federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[25] Manyi Zhang, Xuyang Zhao, Jun Yao, Chun Yuan, and Weiran Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15890–15900, 2023. 2

[26] Jae Soon Baik, In Young Yoon, Kun Hoon Kim, and Jun Won Choi. Distribution-aware robust learning from long-tailed data with noisy labels. In *European Conference on Computer Vision*, pages 160–177. Springer, 2024. 2

[27] Shenwang Jiang, Jianan Li, Ying Wang, Bo Huang, Zhang Zhang, and Tingfa Xu. Delving into sample loss curve to embrace noisy and imbalanced data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7024–7032, 2022. 2

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020. 2

[29] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[30] Athanasios Psaltis, Anestis Kastellos, Charalampos Z Patrikakis, and Petros Daras. Fedlid: Self-supervised federated learning for leveraging limited image data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1039–1048, 2023. 2

[31] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 2, 4

[32] Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *International Conference on Machine Learning*, pages 154–162. PMLR, 2014. 2, 4

[33] Md Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, and Liming Chen. von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017. 2

[34] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37(3):50–60, 2020. 2

[35] Yuwei Wang, Runhan Li, Hao Tan, Xuefeng Jiang, Sheng Sun, Min Liu, Bo Gao, and Zhiyuan Wu. Federated skewed label learning with logits fusion. *CoRR*, abs/2311.08202, 2023. 2

[36] Bingjie Yan, Danmin Cao, Xinlong Jiang, Yiqiang Chen, Weiwei Dai, Fan Dong, Wuliang Huang, Teng Zhang, Chenlong Gao, Qian Chen, Zhen Yan, and Zhirui Wang. Fedeye: A scalable and flexible end-to-end federated learning platform for ophthalmology. *Patterns*, 5(2):100928, 2024. 2

[37] Haodong Zhao, Peng Peng, Chiyu Chen, Linqing Huang, and Gongshen Liu. Fedrs-bench: Realistic federated learning datasets and benchmarks in remote sensing. *arXiv preprint arXiv:2505.08325*, 2025. 2

[38] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In Inderjit S. Dhillon, Dimitris S. Papailiopoulos, and Vivienne Sze, editors, *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org, 2020. 2, 6, 7

[39] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020. 2

[40] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10713–10722. Computer Vision Foundation / IEEE, 2021. 2, 6, 7

[41] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022. 2

[42] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press, 2017. 2

[43] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8792–8802, 2018.

[44] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 322–330. IEEE, 2019. 2, 7

[45] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020. 3

[46] Geoffrey J. McLachlan and Suren I. Rathnayake. On the number of components in a gaussian mixture model. *WIREs Data Mining Knowl. Discov.*, 4(5):341–355, 2014. 3

[47] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[48] Nannan Wu, Zengqiang Yan, Nong Sang, Li Yu, and Chang Wen Chen. Fedpca: Noise-robust fair federated learning via performance-capacity analysis. *CoRR*, abs/2503.10567, 2025. 3

[49] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005. 4

[50] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021. 6, 7

[51] Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. Feddisco: Federated learning with discrepancy-aware collaboration. In *International Conference on Machine Learning*, pages 39879–39902. PMLR, 2023. 6, 7

[52] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18250–18280. PMLR, 2022. 6, 7

[53] Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedexp: Speeding up federated averaging via extrapolation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 6, 7

[54] Yan Sun, Li Shen, and Dacheng Tao. Understanding how consistency works in federated learning via stage-wise relaxed initialization. *Advances in Neural Information Processing Systems*, 36:80543–80574, 2023. 6, 7

[55] Xuefeng Jiang, Sheng Sun, Yuwei Wang, and Min Liu. Towards federated learning against noisy labels via local self-regularization. In Mohammad Al Hasan and Li Xiong, editors, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 862–873. ACM, 2022. 6, 7

[56] Xuefeng Jiang, Sheng Sun, Jia Li, Jingjing Xue, Runhan Li, Zhiyuan Wu, Gang Xu, Yuwei Wang, and Min Liu. 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*, pages 1015–1026, 2024. 6, 7

[57] Shikun Li, Xiaobo Xia, Jiankang Deng, Shiming Ge, and Tongliang Liu. Transferring annotator- and instance-dependent transition matrix for learning from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7377–7391, 2024. 6

[58] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 6

[59] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 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*, pages 770–778. IEEE Computer Society, 2016. 6

[61] Zhenheng Tang, Yonggang Zhang, Shaohuai Shi, Xin He, Bo Han, and Xiaowen Chu. Virtual homogeneity learning: Defending against data heterogeneity in federated learning. In *International Conference on Machine Learning*, pages 21111–21132. PMLR, 2022. 7

[62] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Ood-gnn: Out-of-distribution generalized graph neural network. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7328–7340, 2022. 8

[63] Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang, Zhixiao Xiong, Fangzhou Xiong, Peiyang Xu, Wenwu Zhu, et al. Curbench: curriculum learning benchmark. In *Forty-first International Conference on Machine Learning*, 2024. 8

[64] Yuwei Zhou, Hong Chen, Zirui Pan, Chuanhao Yan, Fanqi Lin, Xin Wang, and Wenwu Zhu. Curml: A curriculum machine learning library. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7359–7363, 2022. 8