This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Relaxed Contrastive Learning for Federated Learning

Seonguk Seo* [1] Jinkyu Kim* [1] Geeho Kim* [1] Bohyung Han [1,2]
[1]ECE & [2]IPAI, Seoul National University
{seonguk, jinkyu, snow1234, bhhan}@snu.ac.kr

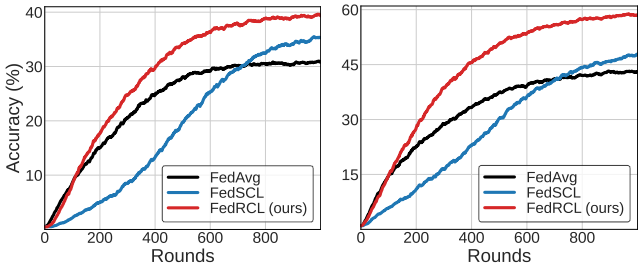
Abstract

We propose a novel contrastive learning framework to effectively address the challenges of data heterogeneity in federated learning. We first analyze the inconsistency of gradient updates across clients during local training and establish its dependence on the distribution of feature representations, leading to the derivation of the supervised contrastive learning (SCL) objective to mitigate local deviations. In addition, we show that a naïve integration of SCL into federated learning incurs representation collapse, resulting in slow convergence and limited performance gains. To address this issue, we introduce a relaxed contrastive learning loss that imposes a divergence penalty on excessively similar sample pairs within each class. This strategy prevents collapsed representations and enhances feature transferability, facilitating collaborative training and leading to significant performance improvements. Our framework outperforms all existing federated learning approaches by significant margins on the standard benchmarks, as demonstrated by extensive experimental results. The source code is available at our project page[1].

1. Introduction

Federated learning (FL) trains a shared model through the collaboration of distributed clients while safeguarding the privacy of local data by restricting their sharing and transfer. The primary challenge in this learning framework arises from the data heterogeneity across clients and the class imbalance in local data. These problems eventually lead to severe misalignments of the local optima of the client models, hindering the search for better global optima of the aggregated model and slowing down convergence.

To tackle these challenges, most existing approaches focus on minimizing the discrepancy between the global and local models by incorporating regularization techniques on either model parameters [1, 2, 12, 21, 44] or feature representations [15, 18, 19, 23, 40]. However, aligning the local models with the global model entails a trade-off as the global model is not necessarily optimal. Recently, there have been several attempts to analyze the inconsistent local training in a principled way [33, 42]. For example, Zhang *et al*. [42] investigate the label distribution skewness from a statistical perspective, and introduce a deviation bound for analyzing the inconsistency of gradient updates in local training.



(a) Tiny-ImageNet (b) CIFAR-100

Figure 1. Performance curves of our framework, dubbed as FedRCL, in comparison to other baselines on the Tiny-ImageNet and CIFAR-100 with non-*i.i.d.* setting ($\alpha = 0.1$). FedSCL incorporates the supervised contrastive learning objective into FedAvg, but it suffers from slow convergence and restrains performance enhancement. Our framework significantly improves both convergence speed and accuracy.

We reformulate the deviation bound proposed in [42] and establish its dependence on the distribution of feature representations. Subsequently, we derive that incorporating the supervised contrastive learning (SCL) objective enhances this bound, resulting in consistent local updates across heterogeneous clients. In other words, we show that employing SCL improves the convergence of federated learning by alleviating the variations of local models.

Although SCL is helpful for the optimization in federated learning, the empirical results show that a naïve integration of SCL suffers from slow convergence and limited performance gains as illustrated in Figure 1. Due to the limited and imbalanced training data in a local client, the intra-class attraction force in SCL hampers feature diversity and consequently weakens the transferability of neural

*indicates equal contribution.
[1]https://github.com/skynbe/FedRCL

networks to diverse tasks. Considering the consolidation principle of federated learning through the aggregation of heterogeneous local models, the lack of transferability impedes the collaborative training process.

To tackle this issue and enhance the transferability of models, we present a novel contrastive learning strategy for federated learning. Our approach imposes the penalty on the sample pairs within the same class that may exhibit excessively high similarity otherwise. Such a simple adaptive repulsion strategy effectively prevents the intra- and inter-class collapse of representations, enhancing the transferability across heterogeneous clients and leading to the discovery of better global optima. Furthermore, we expand the proposed approach to cover all intermediate levels of representations, promoting consistent local updates even further. The proposed approach demonstrates remarkable performance improvements in all datasets and settings consistently, surpassing existing baselines by significant margins. We present the effectiveness and robustness of the proposed method by thorough empirical analysis. Our main contributions are summarized as follows.

- By reformulating the deviation bound of local gradient update, we theoretically analyze that supervised contrastive learning mitigates inconsistent local updates across heterogeneous clients.
- We discover the feature collapse phenomenon caused by the standard SCL in federated learning, resulting in slow convergence and limited performance improvement.
- We propose a relaxed supervised contrastive loss, which adaptively imposes the divergence penalty on pairs of examples in the same class and prevents their representations from being learned to be indistinguishable.
- We demonstrate that our approach significantly outperforms existing federated learning algorithms on the standard benchmarks under various settings.

The rest of the paper is organized as follows. We review the prior works in Section 2 and discuss the preliminaries in Section 3. Section 4 presents the proposed approach in the context of federated learning and Section 5 validates its effectiveness empirically. Finally, we conclude our paper in Section 6.

2. Related Works

This section first overviews the existing FL algorithms, and discusses how contrastive learning has been explored in the context of FL.

2.1. Federated learning

McMahan *et al*. [22] propose a pioneer FL framework, FedAvg, which aggregates model updates from distributed clients to improve a global model without requiring the exchange of local data. However, it suffers from slow convergence and poor performance due to the heterogeneous nature of client data in practical scenarios [45]. To address the issue of heterogeneity in FL, numerous approaches have been proposed in two distinct directions, local training and global aggregation.

The major approaches in local training are imposing regularization constraints on model parameters or feature representations. Specifically, they incorporate proximal terms [21], introduce control variates [12, 20], or leverage primal-dual analysis [1, 44] to regularize model parameters, while adopting knowledge distillation [15, 18, 40], metric learning [19, 23, 46], logit calibration [42], feature decorrelation [33], or data augmentation [37, 41] for effective representation learning. Our framework also belongs to representation learning, where it particularly focuses on gradient deviations in local training and transferability of trained models across heterogeneous local clients.

Besides the local training methods, server-side optimization techniques have been explored to expedite convergence using momentum [9, 14, 27] or decrease the communication cost by quantization [6, 24, 28, 35]. These server-side works are orthogonal to our client-side approach and are easily combined with the proposed algorithm.

2.2. Contrastive learning in FL

Recent works have explored the integration of contrastive learning techniques [3, 7, 25] into federated learning to prevent local client drift and assist local training. FedEMA [46] adopts self-supervised contrastive learning to deal with unlabeled data collected from edge devices. MOON [19] introduces a model-contrastive loss, which aims to align the current local model with the global model, while pushing the current model away from the local model of the previous round. FedProc [23] employs a contrastive loss to align local features with the global prototypes to reduce the representation gap, where the global class prototypes are distributed from the server. FedBR [5] conducts contrastive learning to align local and global feature spaces using local data and globally shared proxy data to reduce bias in local training. In contrast to prior approaches, our framework does not require additional communication overhead for contrastive learning and does not rely on global models or prototypes to mitigate the deviations in local training.

3. Preliminaries

Before discussing the proposed approach, we briefly describe the main idea and formulation of federated learning and supervised contrastive learning.

3.1. Problem setup

Suppose that there are N clients, $\{C_1, ..., C_N\} = \mathcal{C}$. Each client C_i has a dataset $\mathcal{D}_i$, which comprises a set of pairs

of an example and its class label. The goal of federated learning is to optimize a global model parametrized by $\theta = [\phi; \psi]$, corresponding to a feature extractor, ϕ, and a classifier, ψ, that minimizes the average losses over all clients as

$$\arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i(\theta), \tag{1}$$

where $\mathcal{L}_i(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_i}\left[\ell(\mathbf{x}, y; \theta)\right]$ is the empirical loss in C_i, given by the expected loss over all samples in $\mathcal{D}_i$. Note that data distributions in individual clients may be heterogeneous, and privacy concerns strictly prohibit transfering training data across clients. We employ FedAvg [22] as a baseline algorithm. In the tth communication round, a central server sends a global model θ^{t-1} to the active client set $\mathcal{C}_t \subseteq \mathcal{C}$. Each client $C_i \in \mathcal{C}_t$ initializes its parameter $\theta^t_{i,0}$ to θ^{t-1}, and performs K iterations for optimization using its local data. The server collects the resulting local models $\theta^t_{i,K}$ and computes the global model θ^t for the next round of training by simply averaging the local model parameters. This training process is repeated until the global model θ^t converges.

3.2. Supervised contrastive learning

Supervised contrastive learning (SCL) [13] is a variant of self-supervised contrastive learning [3, 25], where, given the ith example and its ground-truth label denoted by $(\mathbf{x}_i, y_i)$, the supervised contrastive loss $\mathcal{L}_{\text{SCL}}$ is defined as

$$\mathcal{L}_{\text{SCL}}(\mathbf{x}_i, y_i) = -\sum_{\substack{j\neq i,\\ y_j=y_i}} \log \frac{\exp\left(\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle/\tau\right)}{\sum_{k\neq i}\exp\left(\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_k)\rangle/\tau\right)}, \tag{2}$$

where $\phi(\cdot)$ denotes the feature representation of an input example, $\langle\cdot,\cdot\rangle$ indicates the cosine similarity function, and τ is a temperature. To boost its effectiveness, hard example mining is usually adopted to construct both positive and negative pairs. Eq. (2) is also expressed as follows:

$$\begin{aligned}\mathcal{L}_{\text{SCL}}(\mathbf{x}_i, y_i) = &\sum_{y_j=y_i, j\neq i}\Big\{-\left(\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle/\tau\right)\\ &+\log\Big(\sum_{k\neq i}\exp\left(\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_k)\rangle/\tau\right)\Big)\Big\}.\end{aligned} \tag{3}$$

This loss function encourages feature representations from the same class to be similar while pushing features from different classes apart.

4. Relaxed Supervised Contrastive Learning

This section begins by analyzing the local deviations in federated learning with heterogeneous clients, and presents that supervised contrastive learning (SCL) mitigates the deviations (Section 4.1). Then, we identify the challenges in employing SCL in the FL context (Section 4.2) and discuss our solution to address the challenges (Section 4.3 and 4.4).

4.1. Benefit of SCL for local training

One of the main challenges in federated learning is inconsistent local updates caused by the heterogeneity of local client data. Zhang *et al.* [42] present that existing FL methods based on softmax cross-entropy result in biased local models, and introduce a deviation bound to measure the deviation of the gradient update during the local training. To analyze this further, we revisit the deviation bound and formulate a sample-wise deviation bound considering all classes, which is formally defined below.

Definition 1 (Sample-wise deviation bound) *Let $\mathbf{x} \in \mathcal{O}_r$ denote a training example with ground-truth class label r. The sample-wise deviation bound is defined as*

$$D(\mathbf{x}) = \frac{\left(1 - P_r^{(r)}\right)\Phi_r|\mathcal{O}_r|S_r(\mathbf{x})}{\sum_{j\neq r} P_r^{(j)}\Phi_j|\mathcal{O}_j|S_j(\mathbf{x})}, \tag{4}$$

where $P_z^{(y)} = \frac{1}{|\mathcal{O}_y|}\sum_{i\in\mathcal{O}_y} p_z(\mathbf{x}_i)$ means the average prediction score for class z, estimated with the examples that belong to class y, $\Phi_y = \frac{1}{|\mathcal{O}_y|}\sum_{i\in O_y}\|\phi(\mathbf{x}_i)\|_2$ is the average feature norm of the examples in class y, and $S_y(\mathbf{x}) = \frac{1}{|\mathcal{O}_y|}\sum_{i\in O_y}\langle\phi(\mathbf{x}), \phi(\mathbf{x}_i)\rangle$ denotes the average feature similarity with respect to an example $\mathbf{x}$.

Proposition 1 *If $D(\mathbf{x}) \ll 1$, the local updates of the parameters in classification layer, $\{\Delta\psi_y\}_{y\in\mathcal{Y}}$, are prone to deviate from the desirable direction,* i.e., *$\Delta\psi_r\phi(\mathbf{x}) < 0$ and $\Delta\psi_j\phi(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{O}_r$ and $\forall j \neq r$.*

The proof of Proposition 1 is provided in Section A of the supplementary document. Eq. (4) indicates that the deviation bound of an example depends on the distribution of feature representations with respect to the example, $S_r(\mathbf{x})$ and $S_j(\mathbf{x})$. This proposition means that lower values of $D(\mathbf{x})$ incur inconsistent local training.

Proposition 1 states that it is possible to prevent the local gradient deviation of each example by increasing $D(\mathbf{x})$. If $\frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r} S_j(\mathbf{x}) - S_r(\mathbf{x}) \leq 0$, then the lower bound of $D(\mathbf{x})$ in (4) becomes $\frac{\left(1-P_r^{(r)}\right)\Phi_r|\mathcal{O}_r|}{|\mathcal{Y}|-1}\min_{j\neq r}\left\{\frac{1}{P_r^{(j)}\Phi_j|\mathcal{O}_j|}\right\}$. Thus, we formulate the surrogate objective to minimize $\max\left(0, \frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r} S_j(\mathbf{x}) - S_r(\mathbf{x})\right)$, which is highly correlated to the increase of the lower bound. By using a smooth

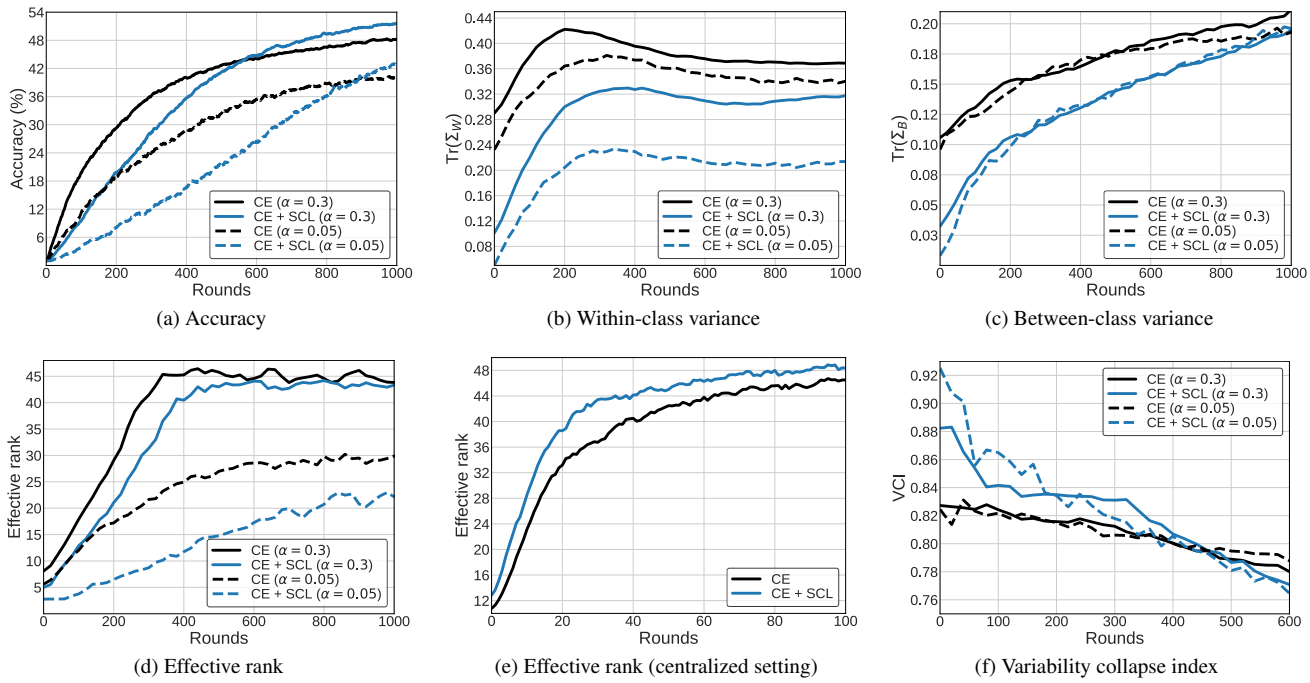


Figure 2. Effects of employing supervised contrastive loss on the CIFAR-100 under non-*i.i.d.* settings. Black and blue lines denote models trained with $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SCL}}$, respectively. Dotted and solid lines indicate different data heterogeneity with Dirichlet parameters $\alpha \in \{0.05, 0.3\}$.

approximation to the maximum function with the *LogSumExp* operator, we derive its upper bound as follows

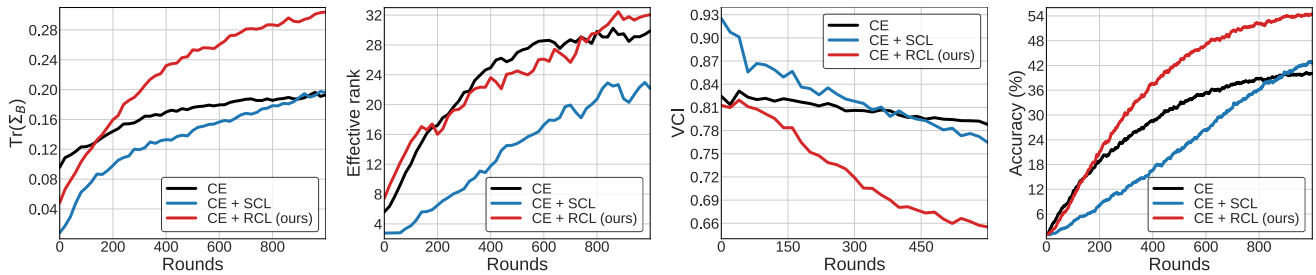
$$\max\left(0, \sum_{j \neq r} \frac{S_j(\mathbf{x})}{|\mathcal{Y}| - 1} - S_r(\mathbf{x})\right)$$
$$\leq \log\left(\exp(0) + \exp\left(\sum_{j \neq r} \frac{S_j(\mathbf{x})}{|\mathcal{Y}| - 1} - S_r(\mathbf{x})\right)\right)$$
$$\leq -\frac{1}{|\mathcal{O}_r| - 1} \sum_{\mathbf{x}_i \in \mathcal{O}_r \setminus \mathbf{x}} \log\left(\frac{\exp(\langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle)}{\sum_{\mathbf{x}_k \neq \mathbf{x}} \exp(\langle \phi(\mathbf{x}), \phi(\mathbf{x}_k) \rangle)}\right).$$

Please refer to Section B for further details. This derivation demonstrates how the optimization of $\mathcal{L}_{\text{SCL}}$ contributes to mitigating local gradient deviations.

4.2. Representation collapse in FL with SCL

Based on our analysis in Section 4.1, we empirically validate the effectiveness of SCL in federated learning under data heterogeneity with Dirichlet parameters $\alpha \in \{0.05, 0.3\}$ in Figure 2a. We train the ResNet-18 model using the loss function $\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{SCL}}$ at each local client on the CIFAR-100 dataset, using 5% participation rate out of 100 distributed clients, where $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss. As depicted in the figure, while SCL eventually achieves improved performance by reducing local deviations over the baseline methods, it is accompanied by a noticeable lag in the convergence speed during at early stage of training. We conjecture that, due to limited and skewed local training data, SCL leads to excessively compact representations of the examples in the same classes, hindering effective knowledge transfer across clients in federated learning.

To delve into these phenomena, we first compute the within-class and between-class covariance matrices of the feature embeddings provided by a local model, denoted by $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$, respectively. Figure 2b and 2c plot the trace of the two matrices. SCL effectively reduces the within-class variance compared to the baseline model only with the cross-entropy loss, due to the attraction term between samples from the same class. However, it is noteworthy that SCL also yields a lower between-class variance than the baseline, especially at the early stages of training, despite the repulsion term between the examples in different classes. Since the attraction and repulsion forces interact in contrastive learning, the excessive representation similarity between positive pairs weakens the repulsion force between negative pairs. In other words, the collapse of intra-class representations negatively affects the separation between inter-class examples, leading to an overall reduction in the diversity of feature representations. To evaluate this feature collapse quantitatively, we observe the effective rank [30] of the covariance matrix of all feature embeddings given by a local model, which estimates the actual dimensionality of the learned feature manifold of training data. Formally, the



(a) Between-class variance (b) Effective rank (c) Variability collapse index (d) Accuracy

Figure 3. Results of our relaxed contrastive learning (RCL) approach on the CIFAR-100 dataset under a non-*i.i.d.* setting ($\alpha = 0.05$). RCL outperform SCL in all metrics.

effective rank is defined as follows:

Definition 2 (Effective rank) *Consider a matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *with its singular values* $\{\sigma_1, ..., \sigma_Q\}$*, where* $Q = \min(m, n)$*, and let* $\mathrm{p}_k = \sigma_k / \sum_{i=1}^{Q} |\sigma_i|$*. Then, the effective rank of matrix* $\mathbf{A}$ *is defined as* $\exp(H(\mathrm{p}_1, ..., \mathrm{p}_Q)) = \exp(-\sum_{k=1}^{Q} \mathrm{p}_k \log \mathrm{p}_k)$*, where* $H(\cdot)$ *is the Shannon entropy.*

Figure 2d illustrates the impact of SCL on the effective rank in the CIFAR-100 test set. It supports that SCL diminishes the effective rank when compared to the baseline methods, particularly during the early stage of training, and leads to overall representation collapses. Interestingly, SCL does not exhibit dimensional collapse in the centralized setting[2] as in Figure 2e, which implies that limited and skewed local training data incurs the problem in SCL.

These collapsed representations exacerbate the transferability of neural networks across heterogeneous tasks and clients. Previous studies [4, 32, 38] have emphasized the close relationship between feature diversity and transferability, highlighting that representation collapses of trained models hamper maintaining crucial information beneficial for knowledge transfer to downstream tasks. To quantitatively analyze this, we employ the variability collapse index [38], $\mathrm{VCI} = 1 - \frac{\mathrm{Tr}[\Sigma_T^{\dagger} \Sigma_B]}{\mathrm{rank}(\Sigma_B)}$, where Σ_T and Σ_B denote the total covariance and between-class covariance matrices for a given feature matrix. It provides a robust measurement of transferability in terms of optimal linear probing loss, where lower values denote better transferability. As observed in Figure 2f, SCL yields higher VCI values at the early stage, indicating low transferability even in comparison to the baselines. Given that federated learning can be regarded as a continual fine-tuning process across heterogeneous local tasks, the lack of transferability impedes collaborative training, resulting in slow convergence and limited performance gain. We will discuss strategies for addressing these challenges in the following subsection.

[2]We trained a ResNet-18 model with a single client using the whole CIFAR-100 training set.

Algorithm 1 FedRCL

Input: initial model θ^0, # of communication rounds T, # of local iterations K, # of layers L
for each round $t = 1, \ldots, T$ **do**
  Sample a subset of clients $\mathcal{C}_t \subseteq \mathcal{C}$
  Server sends θ^{t-1} to all active clients $C_i \in \mathcal{C}_t$
  for each $C_i \in \mathcal{C}_t$, **in parallel do**
    $\theta_{i,0}^t \leftarrow \theta^{t-1}$
    for $k = 1, \ldots, K$ **do**
      for each $(\mathbf{x}, y)$ in a batch **do**
        $\mathcal{L}_{\mathrm{RCL}} \leftarrow \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}_{\mathrm{RCL}}(\mathbf{x}, y; \phi_l)$
        $\mathcal{L}(\theta_{i,k-1}^t) \leftarrow \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{RCL}}$
        $\theta_{i,k}^t \leftarrow \theta_{i,k-1}^t - \eta \nabla \mathcal{L}(\theta_{i,k-1}^t)$
      end for
    end for
    Client sends $\theta_{i,K}^t$ back to the server
  end for
  In server:
    $\theta^t = \frac{1}{|\mathcal{C}_t|} \sum_{C_i \in \mathcal{C}_t} \theta_{i,K}^t$
end for

4.3. Relaxed contrastive loss for FL

To address the representation collapse issue identified in Section 4.2, we propose a novel federated learning approach with an advanced contrastive learning strategy, referred to as Federated Relaxed Contrastive Learning (FedRCL). The proposed algorithm adopts the relaxed contrastive loss $\mathcal{L}_{\mathrm{RCL}}$, imposing the feature divergence on intra-class samples as

$$\mathcal{L}_{\mathrm{RCL}}(\mathbf{x}_i, y_i; \phi) = \sum_{\substack{j \neq i, \\ y_j = y_i}} \Big\{ -\log \frac{\exp\left(\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle / \tau\right)}{\sum_{k \neq i} \exp\left(\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_k) \rangle / \tau\right)} + \beta \cdot \mathbb{1}_{\{\mathbf{x}_j \in \mathcal{P}(\mathbf{x}_i)\}} \log\left(\exp\left(\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle / \tau\right)\right) \Big\} \quad (5)$$

where $\mathcal{P}(\mathbf{x}) = \{\mathbf{x}' | y_{\mathbf{x}'} = y_{\mathbf{x}}, \langle \phi(\mathbf{x}'), \phi(\mathbf{x}) \rangle > \lambda\}$ represents a set of intra-class samples more similar to the anchor $\mathbf{x}$ than the threshold λ and β is a hyperparameter for

Table 1. Results from 5% participation rate over 100 distributed clients on the CIFAR-10, CIFAR-100, and Tiny-ImageNet for the different levels of Dirichlet parameter (α). Accuracies at the target round are based on the exponential moving average results with parameter 0.9.

Dataset	Method	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.3$		$\alpha = 0.6$		*i.i.d.*	
		500R	1000R	500R	1000R	500R	1000R	500R	1000R	500R	1000R
CIFAR-10	FedAvg [22]	51.47	63.42	58.80	70.82	75.63	83.18	80.93	85.52	84.67	88.19
	FedAvg + FitNet [29]	51.34	63.25	58.67	71.09	74.87	83.03	79.14	84.84	84.20	87.67
	FedProx [21]	48.61	59.58	56.22	68.87	70.30	80.46	76.06	83.48	84.14	87.66
	MOON [19]	49.68	61.73	69.16	77.12	83.32	86.30	84.95	87.99	88.24	89.66
	FedMLB [15]	32.81	49.16	52.01	72.31	74.98	84.08	77.84	85.96	86.84	89.93
	FedLC [42]	54.30	65.62	62.39	72.52	78.37	84.79	81.17	86.02	84.57	88.41
	FedNTD [18]	52.33	63.36	62.23	73.54	76.05	83.78	81.20	86.46	85.98	89.44
	FedProc [23]	25.61	47.77	33.28	62.56	63.03	80.93	69.41	84.57	78.30	87.66
	FedDecorr [33]	53.04	66.62	63.74	75.35	76.62	83.40	81.39	85.28	85.41	88.16
	FedRCL (ours)	**64.44**	**76.74**	**74.82**	**82.72**	**84.01**	**88.44**	**86.00**	**89.45**	**89.70**	**91.90**
CIFAR-100	FedAvg [22]	31.20	39.86	36.65	43.04	41.70	47.47	43.23	49.29	43.52	48.12
	FedAvg + FitNet [29]	31.09	38.35	36.48	43.25	42.96	48.59	44.20	49.82	44.61	49.33
	FedProx [21]	30.27	39.44	35.78	43.11	42.24	48.19	43.21	48.48	45.20	49.37
	MOON [19]	34.28	40.64	42.91	50.31	53.15	58.37	55.76	61.42	58.50	64.73
	FedMLB [15]	30.89	43.89	38.64	48.94	47.39	54.58	49.36	56.70	50.12	56.40
	FedLC [42]	34.24	40.84	39.80	44.40	42.74	47.23	44.24	48.89	44.06	47.63
	FedNTD [18]	33.10	41.75	35.84	42.86	43.22	49.29	44.26	50.32	44.93	50.15
	FedProc [23]	18.41	38.56	25.19	43.73	32.66	49.68	36.09	49.89	40.76	52.94
	FedDecorr [33]	33.31	41.73	38.88	43.89	43.52	49.17	44.01	49.08	45.46	49.30
	FedRCL (ours)	**43.71**	**54.63**	**49.82**	**58.23**	**57.89**	**63.46**	**58.71**	**64.06**	**60.25**	**64.81**
Tiny-ImageNet	FedAvg [22]	22.49	25.90	26.62	29.71	31.80	33.58	33.91	35.01	35.62	37.02
	FedAvg + FitNet [29]	22.82	26.95	27.37	30.51	32.96	33.95	33.46	34.70	35.79	37.31
	FedProx [21]	22.91	27.02	27.31	30.93	32.35	34.34	34.33	35.53	35.94	36.11
	MOON [19]	23.30	26.34	30.31	32.03	36.97	39.32	38.98	42.07	41.88	45.62
	FedMLB [15]	19.31	26.88	29.31	34.41	37.20	40.16	39.34	42.15	40.69	42.98
	FedLC [42]	26.30	28.28	30.63	32.25	35.03	35.95	35.38	36.48	36.57	37.75
	FedNTD [18]	22.83	28.96	28.86	33.74	33.91	37.33	36.47	39.43	37.77	40.85
	FedProc [23]	10.74	22.74	14.02	27.43	16.62	32.43	19.64	32.60	21.59	35.43
	FedDecorr [33]	22.55	26.18	28.15	30.74	33.40	34.86	33.31	34.90	35.02	35.82
	FedRCL (ours)	**27.21**	**34.60**	**34.30**	**39.36**	**40.25**	**44.95**	**43.20**	**46.70**	**45.01**	**47.25**

the divergence term. The second term of Eq. (5) serves to prevent within-class representation collapses, which also promotes the separation of the examples between different classes. This ultimately enhances overall feature diversity and transferability, which is crucial in the context of federated learning with non-*i.i.d.* settings. As illustrated in Figure 3, FedRCL facilitates inter-class separation, mitigates dimensional collapse, and improves the transferability of trained models, resulting in early convergence and significant performance improvement.

4.4. Multi-level contrastive training

Existing contrastive learning approaches [3, 13, 25] concentrate on aligning the feature representations of the last layer, resulting in predominant model updates in deeper layers while having limited influence on lower-layer parameters. To mitigate this issue, we expand the proposed contrastive learning approach to encompass feature representations in earlier layers. Let $\phi_l(\mathbf{x})$ denotes the lth level feature representation of sample $\mathbf{x}$. Then, we construct $\mathcal{L}_{\text{RCL}}$ by aggregating $\frac{1}{L}\sum_{l=1}^{L}\mathcal{L}_{\text{RCL}}(\mathbf{x}, y; \phi_l)$, where L is the number of layers. The comprehensive algorithm of our framework is presented in Algorithm 1.

4.5. Discussion

FedRCL has something common with existing methods incorporating contrastive loss for local updates, but it has clear differences and advantages over them. While most existing works [5, 19, 23] employ contrastive learning to regulate local training towards the global model for consistent local updates, this constraint often leads to suboptimal solutions as the global model is not fully optimized. Our approach is free from this issue, because FedRCL mitigates local deviations by itself so it does not align with the global model explicitly. Some algorithms require proxies for contrastive learning such as global prototypes [23] or globally shared data [5], which rely on extra communication overhead and full client participation. Note that transferring such prototypes or data is vulnerable and incurs privacy concerns. In contrast, FedRCL does not involve any

Table 2. Results from 2% participation rate over 100 and 500 clients on three benchmarks. The Dirichlet parameter is commonly set to 0.3.

	CIFAR-10				CIFAR-100				Tiny-ImageNet			
	100 clients		500 clients		100 clients		500 clients		100 clients		500 clients	
Method	500R	1000R	500R	1000R	500R	1000R	500R	1000R	500R	1000R	500R	1000R
FedAvg [22]	65.92	78.13	59.88	72.12	38.19	44.62	29.01	37.86	28.63	34.62	21.00	27.37
FedAvg + FitNet [29]	66.88	79.22	57.29	70.94	36.89	46.69	28.52	36.41	27.80	34.88	20.17	27.10
FedProx [21]	65.78	75.82	60.23	72.78	36.69	45.16	28.44	35.45	27.45	32.91	22.34	29.04
MOON [19]	71.52	75.42	69.15	78.06	39.91	46.51	33.51	42.41	27.26	32.25	26.69	31.81
FedMLB [15]	65.85	79.45	58.68	71.38	40.90	53.34	32.03	42.61	31.17	38.09	28.39	33.67
FedLC [42]	72.90	80.90	60.16	71.39	39.70	42.10	29.58	36.78	30.94	35.59	22.14	26.83
FedNTD [18]	69.11	80.43	60.65	73.20	38.13	48.03	28.95	36.31	28.39	36.41	24.67	32.16
FedProc [23]	49.71	73.54	50.91	70.10	24.20	44.52	23.74	36.90	12.69	28.84	15.00	23.74
FedDecorr [33]	71.29	78.99	60.01	72.38	39.42	48.45	30.56	38.20	27.93	33.51	24.34	30.28
FedRCL (ours)	**75.94**	**84.67**	**72.93**	**81.71**	**50.83**	**59.07**	**37.23**	**46.98**	**32.09**	**40.87**	**30.44**	**36.44**

additional overhead and consistently demonstrates strong performance improvement even with an extremely low participation rate.

5. Experiment

5.1. Experimental setup

Datasets and baselines We employ three standard benchmarks for experiments: CIFAR-10, CIFAR-100 [16], and Tiny-ImageNet [17], covering various levels of data heterogeneity and participation rates. We generate *i.i.d.* datasets by randomly assigning training examples to each client without replacement. For non-*i.i.d.* cases, we simulate data heterogeneity by sampling label ratios from a Dirichlet distribution with a symmetric parameter $\alpha \in \{0.05, 0.1, 0.3, 0.6\}$ following [9]. The participation ratio is 5% out of 100 distributed clients unless stated otherwise. Following existing literature, each client holds an equal number of examples. For evaluation, we use the complete test set for each dataset and measure the accuracy achieved at the 500th and 1,000th rounds. We compare our method, dubbed as FedRCL, with several state-of-the-art federated learning techniques, which include FedAvg [22], FedAvg + FitNet [29], FedProx [21], MOON [19], FedMLB [15], FedLC [42], FedNTD [18], FedProc [23], and FedDecorr [33].

Implementation details We adopt a ResNet-18 as the backbone network, where we replace the batch normalization with the group normalization [36] as suggested in [8]. We trained the model from scratch, using the SGD optimizer with a learning rate of 0.1, an exponential decay parameter of 0.998, a weight decay of 0.001, and no momentum, following prior works [1, 15, 39]. The number of local training epochs is set to 5 and the batch size is adjusted to ensure a total of 10 local iterations at each local epoch throughout all experiments. We apply contrastive learning to conv1, conv2_x, conv3_x, conv4_x, and conv5_x layers. Other hyperparameter settings are as follows for all experiments unless specified otherwise: $\lambda = 0.7$, $\beta = 1$, and $\tau = 0.05$. We used the PyTorch framework [26] for implementation and executed on NVIDIA A5000 GPUs. Please refer to the supplementary document for further details about our implementation.

Table 3. Ablative results of contrastive training in the non-*i.i.d.* settings on the CIFAR-100 dataset.

	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.3$	
	500R	1000R	500R	1000R	500R	1000R
Baseline	31.20	39.86	36.65	43.04	41.70	47.47
FedSCL	21.22	42.93	30.93	48.09	41.54	51.70
FedCL	35.29	41.45	40.39	45.91	45.99	50.16
FedRCL (ours)	**43.71**	**54.63**	**49.82**	**58.23**	**57.89**	**63.46**

5.2. Results

We compare the proposed method, FedRCL, with numerous client-side federated learning baselines [15, 18, 19, 21–23, 29, 33, 42] on the CIFAR and Tiny-ImageNet datasets. Table 1 demonstrates that our framework outperforms all other existing algorithms by large margins on all datasets and experiment settings. Among the baselines, MOON presents meaningful performance improvement, but its gains are marginal under severe data heterogeneity, *e.g.*, $\alpha = 0.05$. FedLC employs adaptive label margin to mitigate local deviations, but its impact on performance is limited. FedDecorr exhibits minor improvement but degraded performance in some settings. This implies that the blind mitigation of dimensional collapse is not necessarily helpful for FL. Compared to other works, our algorithm achieves significant performance improvements in all datasets, regardless of the level of data heterogeneity.

5.3. Analysis

Low participation rate and large-scale clients We validate our framework in more challenging scenarios with lower client participation rates and a larger number of distributed clients. Table 2 presents the robust performance improvement of FedRCL on three benchmarks, where a par-

Table 4. Ablative results of multi-level contrastive training in various non-*i.i.d.* settings on the CIFAR-100 dataset.

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.6$
Baseline	39.86	43.04	47.47	49.29
Last-layer only	46.36	52.94	57.09	58.20
Multi-layers (ours)	**54.63**	**58.23**	**63.46**	**64.06**

ticipation rate is 0.02 and the number of clients is set to one of $\{100, 500\}$. All methods suffer from performance degradation, compared with the results in Table 1, due to the reduced client data, increased data disparity, and lower participation rate. Particularly, FedProc experiences a significant performance drop, because it relies on the global class prototypes aggregated from participating clients at each round, which may not be accurate in extremely low participation settings. MOON exhibits performance degradation compared to FedAvg in some challenging configurations, partly because it utilizes outdated previous local models due to the sparse participation of local clients. Despite these challenges, FedRCL consistently demonstrates promising performance on all the tested datasets.

Contrastive learning strategies Table 3 compares the effectiveness of various contrastive learning strategies, all employing multi-level contrastive training for fair comparisons. FedSCL is an ablative model of our framework, which incorporates a naïve supervised contrastive loss into the FedAvg baseline. While FedSCL improves upon the baseline in general, its gains are moderate and even negative in the early stage of training. In contrast, our full framework consistently enhances performance throughout the entire learning process, as also observed in Figure 1. We also employ another variant, denoted as FedCL, which adopts a self-supervised contrastive loss [25], but its benefits are not salient. This is partly because the objective in [25] does not directly align with the reduction of local deviations.

Multi-level contrastive learning Table 4 presents the ablative results of FedRCL on CIFAR-100, where the proposed contrastive learning is applied only to the last-layer feature outputs. The results show that FedRCL benefits from the contrastive learning on intermediate representations.

Sensitivity of the divergence penalty We study the impact of β in Eq. (5) on the performance of FedRCL under a non-*i.i.d.* setting with $\alpha = 0.1$. Figure 4 illustrates that a large β leads to early convergence and the improvements are consistent over a wide range of β, although its excessively high values marginally degrades the final performances.

Combination with server-side optimization methods Our approach is orthogonal to server-side algorithms, which allows seamless combinations of FedRCL and the server-side techniques such as FedAvgM [9], FedADAM [27], and FedACG [14]. Table 5 presents the consistent and promising performance gains by the combinations,

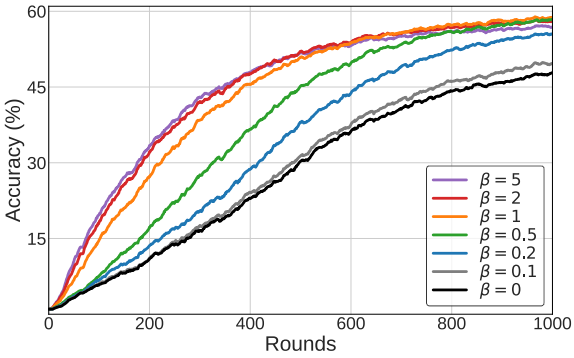


Figure 4. Ablative results by varying the weight of the divergence penalty (β), which exhibit stability across its wide range.

Table 5. Integration of FedRCL into various server-side federated learning approaches under a non-*i.i.d.* setting ($\alpha = 0.3$).

Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
FedAvgM [10]	85.48	53.29	38.51
FedAvgM + FedRCL	**88.51**	**64.61**	**47.23**
FedADAM [27]	81.82	52.81	39.74
FedADAM + FedRCL	**85.69**	**57.84**	**41.57**
FedACG [14]	89.10	62.51	46.31
FedACG + FedRCL	**89.67**	**66.38**	**47.97**

6. Conclusion

We presented a novel federated learning approach to address the challenges of data heterogeneity effectively. We initiated our investigation by analyzing gradient deviations at each local model and showed that the SCL objective mitigates the local deviations, but it entails representation collapses and limited transferability. To tackle this issue, we proposed a federated relaxed contrastive learning framework that successfully prevents representation collapses, which is further enhanced by encompassing all levels of intermediate feature representations. We demonstrated the superiority and robustness of our framework through extensive experiments and analyses.

Acknowledgement This work was supported by the National Research Foundation of Korea grants [No. 2022R1A2C3012210; No. 2022R1A5A7083908; No. 2021M3A9E4080782] and the IITP grants [2021-0-02068; 2021-0-01343] funded by the Korea government (MSIT).

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *ICLR*, 2021. 1, 2, 7

[2] Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, and Afshin Rostamizadeh. Federated learning via posterior averaging: A new perspective and practical algorithms. In *ICLR*, 2021. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 3, 6

[4] Quan Cui, Bingchen Zhao, Zhao-Min Chen, Borui Zhao, Renjie Song, Boyan Zhou, Jiajun Liang, and Osamu Yoshie. Discriminability-transferability trade-off: an information-theoretic perspective. In *ECCV*, 2022. 5

[5] Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedbr: Improving federated learning on heterogeneous data via local learning bias reduction. In *ICML*, 2023. 2, 6

[6] Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *AISTATS*, 2021. 2

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[8] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *ICML*, 2020. 7

[9] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 2, 7, 8, 13

[10] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. In *ECCV*, 2020. 8, 14

[11] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 13

[12] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. In *ICML*, 2020. 1, 2

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 2020. 3, 6

[14] Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with accelerated client gradient. In *CVPR*, 2024. 2, 8, 13, 14

[15] Jinkyu Kim, Geeho Kim, and Bohyung Han. Multi-level branched regularization for federated learning. In *ICML*, 2022. 1, 2, 6, 7, 14

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[17] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 7

[18] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning. In *NeurIPS*, 2022. 1, 2, 6, 7, 14

[19] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, 2021. 1, 2, 6, 7, 14

[20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smithy. Feddane: A federated newton-type method. In *ACSCC*, 2019. 2

[21] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020. 1, 2, 6, 7, 14

[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017. 2, 3, 6, 7, 14

[23] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143:93–104, 2023. 1, 2, 6, 7, 14

[24] Hyeon-Woo Nam, Moon Ye-Bin, and Tae-Hyun Oh. Fedpara: Low-rank hadamard product for communication-efficient federated learning. In *ICLR*, 2022. 2

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3, 6, 8

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 7

[27] Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv

Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *ICLR*, 2021. 2, 8, 13, 14

[28] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *AISTATS*, 2020. 2

[29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 6, 7, 14

[30] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *EUSIPCO*, 2007. 4

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 13

[32] Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, and Diane Larlus. Improving the generalization of supervised models. *arXiv preprint arXiv:2206.15369*, 2022. 5

[33] Yujun Shi, Jian Liang, Wenqing Zhang, Vincent YF Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *ICLR*, 2023. 1, 2, 6, 7, 13, 14, 15

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014. 13

[35] Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In *ICML*, 2022. 2

[36] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 7

[37] Chencheng Xu, Zhiwei Hong, Minlie Huang, and Tao Jiang. Acceleration of federated learning with alleviated forgetting in local training. In *ICLR*, 2022. 2

[38] Jing Xu and Haoxiong Liu. Quantifying the variability collapse of neural networks. In *ICML*, 2023. 5

[39] Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021. 7

[40] Dezhong Yao, Wanning Pan, Yutong Dai, Yao Wan, Xiaofeng Ding, Hai Jin, Zheng Xu, and Lichao Sun. Local-global knowledge distillation in heterogeneous federated learning with non-iid data. *arXiv preprint arXiv:2107.00051*, 2021. 1, 2

[41] Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. In *ICLR*, 2021. 2

[42] 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 *ICML*, 2022. 1, 2, 3, 6, 7, 11, 13, 14, 15

[43] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 13

[44] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. In *arXiv preprint arXiv:2005.11418*, 2020. 1, 2

[45] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 2

[46] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. In *ICLR*, 2022. 2