

WHOM TO TRUST? ADAPTIVE COLLABORATION IN PERSONALIZED FEDERATED LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Data heterogeneity poses a fundamental challenge in federated learning (FL), especially when clients differ not only in distribution but also in the reliability of their predictions across individual examples. While personalized FL (PFL) aims to address this, we observe that many PFL methods fail to outperform two necessary baselines, local training and centralized training. This suggests that meaningful personalization only emerges in a narrow regime, where global models are insufficient, but collaboration across clients still holds value. Our empirical findings point to two key ingredients for success in this regime: adaptivity in collaboration and fine-grained trust, at the level of individual examples. We show that these properties can be achieved within federated semi-supervised learning, where clients exchange predictions over a shared unlabeled dataset. This enables each client to align with public consensus when it is helpful, and disregard it when it is not, without sharing model parameters or raw data. As a concrete realization of this idea, we develop FEDMOSAIC, a personalized co-training method where clients reweight their loss and their contribution to pseudo-labels based on per-example agreement and confidence. FEDMOSAIC outperforms strong FL and PFL baselines across a range of non-IID settings, and we prove convergence under standard smoothness, bounded-variance, and drift assumptions. In contrast to many of these baselines, it also outperforms local and centralized training. These results clarify when federated personalization can be effective, and how fine-grained, trust-aware collaboration enables it.

1 INTRODUCTION

Federated learning (FL) enables collaborative machine learning across distributed data sources without direct data sharing. Classical methods such as FedAvg (McMahan et al., 2017), aim to train a single global model across all clients. This approach can succeed when data distributions are sufficiently similar, but collapses under strong distributional shifts. In highly heterogeneous settings, the promise of collaboration breaks down: models trained jointly may perform worse than models trained independently.

Personalized Federated Learning (PFL) addresses this challenge by shifting the goal. Rather than optimizing a shared global model, the goal is to use collaboration to improve each client’s personalized model. For example, Tab. 1 shows that in heterogeneous regimes both FL and even centralized training perform worse than local training, i.e., clients learning independently without any communication. This underlines the requirement for PFL, but also highlights an often-overlooked baseline: when no method outperforms local training, collaboration is not just ineffective—it is detrimental. Yet many PFL methods fail to beat this baseline (cf. Tab. 1), casting doubt on their utility.

Table 1: **Average test Accuracy on DomainNet and Office-10 dataset** (details in sec.4). Most personalized FL methods fail to surpass local training baseline. FEDMOSAIC exceeds both core baselines through adaptive, example-level collaboration. Color Map: **baselines**, **worse than baselines**, **worse than local training**, **better than baselines**.

| | Method | DomainNet | Office |
|-----|------------------|---------------------|---------------------|
| | Centralized | 66.24 (0.4) | 40.92 (0.6) |
| FL | FedAvg | 31.00 (0.8) | 37.25 (0.8) |
| | FedProx | 55.23 (0.1) | 58.39 (0.3) |
| | Per-FedAvg | 72.48 (0.4) | 71.92 (0.5) |
| PFL | pFedMe | 75.21 (0.5) | 74.83 (0.7) |
| | APFL | 80.59 (0.3) | 80.91 (0.1) |
| | FedPHP | 78.25 (0.6) | 76.36 (0.4) |
| | Local Training | 84.64 (0.1) | 86.79 (0.4) |
| | FEDMOSAIC | 87.44 (0.02) | 89.06 (0.01) |

This widespread failure to measure true collaborative gain arises because "personalization" is often treated as a vague remedy for heterogeneity without a clear underlying principle. We argue that progress requires a new foundation. Personalization shouldn't be a default modification to an existing FL algorithm; it should emerge from a principled understanding of what each client needs and how collaboration can help. A meaningful PFL solution must adapt the degree and nature of collaboration based on client context. It must also account for heterogeneity not just between clients, but at the level of individual examples. Clients may align on some concepts (e.g., identifying cats) and diverge on others (e.g., identifying specific dog breeds), and collaboration should reflect this granularity.

In formal terms, PFL aims to minimize the sum of local risks across m clients with heterogeneous data distribution \mathcal{D}_i and personalized models h_1, \dots, h_m :

$$\min_{h_1, \dots, h_m} \sum_{i=1}^m \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\mathcal{L}(h_i(x), y)] .$$

In this setting, local model may outperform global or centralized models, making strong local and centralized baselines essential. The key trade-off between the massive data access of a centralized model versus the specialization of a local one, is the central tension PFL must navigate in an adaptive and data-specific way.

While federated learning can adapt by weighing parameters according to similarity (Huang et al., 2021; Zhang et al., 2021), data-specific collaborations require a shift in mechanism. Rather than aggregating model parameters, we propose to use federated semi-supervised learning (Bistriz et al., 2020; Abourayya et al., 2025) where clients share predictions on a public dataset. Collaboration is achieved by enforcing consensus between clients. We propose to adapt this consensus mechanism so that clients can contribute only on examples where they have expertise and can selectively trust others based on their demonstrated competence. Two clients familiar with cats can confidently collaborate on a new cat photo, while a client that has only seen cars should not influence the labeling of cat images. This form of selective, example-level trust is fundamentally difficult to achieve through parameter averaging alone.

In this work, we demonstrate this principle in practice. We propose a personalized Federated Co-Training approach (FEDMOSAIC) that enables adaptive, fine-grained collaboration through two key mechanisms: a dynamic weighting strategy allowing clients to balance global and local signals in each communication round, and an expertise-aware consensus mechanism that weights peer contributions by their competence on different data regions. Both mechanisms operate on predictions over a public dataset, enabling personalization that is responsive to the data's true structure.

While FEDMOSAIC achieves state-of-the-art empirical performance across benchmarks, its main contribution is conceptual. It redefines personalization as a question of collaborative structure, not just algorithm design. Our results show that principled, example-level collaboration can unlock the full potential of personalized federated learning.

2 RELATED WORK

Federated Learning (FL) aims to train models collaboratively across decentralized clients without compromising data privacy. However, heterogeneous data distributions across clients (non-IID settings) present a persistent challenge that degrades performance. Approaches addressing heterogeneity broadly fall into two categories: traditional FL and personalized FL (PFL) methods. We review these groups in relation to our method, FEDMOSAIC.

Traditional Federated Learning: Traditional federated learning methods typically learn a single global model. FEDAVG (McMahan et al., 2017) averages local models but struggles under non-IID data due to client drift. Subsequent methods attempt to correct this: SCAFFOLD (Karimireddy et al., 2020) uses control variates to correct the local updates, FedProx (Li et al., 2020) adds a proximal term to each client's loss function to stabilize training, and FedDyn (Acar et al., 2021) introduces dynamic regularization. Others use representation alignment, such as MOON (Li et al., 2021a), which applies a contrastive loss to align local and global features. These methods implicitly assume a global model can suffice, which may fail under strong heterogeneity. Moreover, parameter sharing can pose privacy risks (Zhu et al., 2019; Abourayya et al., 2025).

Personalized Federated learning (PFL): Personalized Federated learning methods tailor models to individual clients, addressing non-IID challenges through different strategies.

Meta-learning and Regularization-Based Methods optimize a shared initialization or constrain local updates. E.g., Per-FedAvg (Fallah et al., 2020) learns a shared initialization, while Ditto (Li et al., 2021b) regularizes local updates toward a global reference. PFedMe (T Dinh et al., 2020) applies bi-level optimization to decouple personalization from global learning. **Personalized Aggregation strategies** dynamically aggregate models based on client similarity or adaptive weighting. APFL (Deng et al., 2020) introduces an adaptive mixture of global and local models, allowing clients to interpolate between shared and personalized parameters based on their data distribution. FedAMP (Huang et al., 2021) uses attention to weight client contributions based on similarity. Other methods select collaborators (e.g., FedFomo (Zhang et al., 2021), FedPHP (Li et al., 2021d)) or apply layer-wise attention (FedALA (Zhang et al., 2023c)). **Model Splitting Architectures** partition models into shared and personalized components. FedPer (Arivazhagan et al., 2019) keeps shared base layers and personalizes top layers. FedRep shares a backbone but personalizes the head. (Collins et al., 2021) shares a backbone but personalizes the head. FedBN (Li et al., 2021c) personalizes batch normalization layers to tackle feature shift. Other recent methods such as FedAS (Yang et al., 2024), GPFL (Zhang et al., 2023b), and FedBABU (Oh et al., 2021) disentangle or freeze specific parts of the model to balance generalization and personalization. PFedHN (Shamsian et al., 2021) uses a hypernetwork that generates personalized model parameters conditioned on client identity. **Knowledge Distillation Approaches** transfer knowledge from global or peer models to personalized local models. FedProto (Tan et al., 2022) aligns class-wise feature prototypes across clients, FedPAC (Xu et al., 2023) uses contrastive learning to distill knowledge into personalized models, and FedKD (Wu et al., 2022) reduces communication cost by distilling knowledge from a teacher ensemble to lightweight client models. FedMatch (Chen et al., 2021) uses consistency regularization to unlabeled and noisy data, FedDF (Lin et al., 2020) aggregates predictions via ensemble distillation, and FedNoisy (Liang et al., 2023) focuses on robust aggregation in the presence of noisy labels or adversarial participants. PerFed-CKT (Cho et al., 2021) enhances personalization by clustering clients with similar data distributions and facilitating knowledge transfer through logits instead of model parameters. Jeong & Kountouris (2023) proposes a fully decentralized PFL framework where clients share distilled knowledge with neighboring clients, enabling personalization without a central server. FedD2S (Atapour et al., 2024) introduces a data-free federated knowledge distillation approach that employs a deep-to-shallow layer-dropping mechanism.

Despite this progress, existing PFL methods often share several limitations: (i) *Static collaboration*: Most PFL methods rely on fixed rules (e.g., aggregation weights or model splits), lacking adaptivity to client-specific or example-level variation. (ii) *Privacy risks*: Sharing model parameters, gradients, or even soft labels may expose sensitive information. (iii) *Limited generality*: Many methods are tailored to specific heterogeneity types (e.g., label skew in case of FedMix, or feature shift in case of FedBN). (iv) *Communication / computational overhead*: Some require complex multi-model training or costly synchronization. To overcome these limitations, we argue that PFL methods should use some form of dynamic modulation and per-example trust weighting.

3 PERSONALIZED FEDERATED CO-TRAINING: ADAPTIVE AND EXPERT-AWARE COLLABORATION

We now introduce Personalized Federated Co-Training (FEDMOAIC), a concrete realization of the principle that effective personalization arises from adaptive, data-specific collaboration. Our method builds upon the framework of federated co-training (Abourayya et al., 2025), a privacy-preserving paradigm where clients collaborate by sharing hard predictions on a shared, unlabeled public dataset, U (we analyze the impact of this dataset’s size and distribution in sec.4). This process creates a consensus pseudo-labeled dataset, which clients use to augment their local training.

While this approach avoids sharing sensitive model parameters and soft labels, it introduces two critical challenges for personalization:

1. **When to trust the global signal?** A client’s local data may conflict with the global consensus. Blindly trusting pseudo-labels can harm a model that is already well-specialized.

2. **Whose predictions to trust?** Clients possess varying levels of expertise across the data space. A naive consensus that treats all clients equally will be corrupted by noisy or misaligned predictions.

FEDMOSAIC addresses these challenges directly with two core mechanisms: (1) dynamic loss weighting, which allows each client to adaptively decide when to trust the global signal, and (2) confidence-based aggregation, which intelligently decides whose predictions to trust.

Dynamic Loss Weighting: Deciding When to Trust: To allow clients to autonomously balance global collaboration with local specialization, we introduce a dynamic weight λ_i^t , into the local objective. At each round t , client i minimizes the combined loss:

$$\mathcal{L}_i^t(h) = \mathcal{L}(h, D_i) + \lambda_i^t \cdot \mathcal{L}(h, P_t)$$

where D_i is the client’s private data and P_t is the pseudo-labeled public dataset. The weight λ_i^t modulates the influence of the global signal. Our choice of the function for computing λ_i^t was driven by the need for a smooth, bounded, and interpretable mechanism. We define it as:

$$\lambda_i^t = \exp \left(- \frac{\mathcal{L}(h_{t-1}^i, P_t) - \mathcal{L}(h_{t-1}^i, D_i)}{\mathcal{L}(h_{t-1}^i, D_i)} \right)$$

This exponential form satisfies several desirable properties. It ensures positivity ($\lambda_i^t > 0$), avoids discontinuities, and smoothly adjusts the client’s trust based on the relative performance of its model on global versus local data. The behavior is highly intuitive:

- **Conflict** ($\mathcal{L}_{\text{global}} \gg \mathcal{L}_{\text{local}}$): If the consensus pseudo-labels are harmful, the global loss term increases, causing $\lambda_i^t \rightarrow 0$ and prompting the client to rely on its local data.
- **Alignment** ($\mathcal{L}_{\text{global}} \approx \mathcal{L}_{\text{local}}$): If the consensus is helpful and aligns with local data, $\lambda_i^t \approx 1$ achieving a balance between personalization and collaboration.
- **Enhancement** ($\mathcal{L}_{\text{global}} < \mathcal{L}_{\text{local}}$): If the consensus provides a cleaner signal than the noisy local data, $\lambda_i^t > 1$, encouraging the client to trust the collaborative signal more heavily.

Confidence-Based Aggregation: Deciding Whose to Trust: To address the varying expertise of clients, we replace the standard uniform aggregation of predictions with a confidence-based consensus. Instead of just sharing hard labels, each client i also communicates a confidence vector $E_t^i \in (0, \infty)^{|U|}$, where $E_t^i[j]$ quantifies its estimated expertise on its prediction for example $x_j \in U$. The server then computes a weighted score matrix S_t by aggregating the one-hot predictions L_t^i from each client, weighted by their corresponding expertise:

$$S_t = \sum_{i=1}^m \text{diag}(E_t^i) \cdot L_t^i \in \mathbb{R}^{|U| \times C}$$

The final consensus pseudo-label for each example is determined by the highest aggregate score:

$$L_t[j] = \arg \max_{c \in [C]} S_t[j, c], \quad \forall j \in \{1, \dots, |U|\}$$

This mechanism allows clients who are more confident or reliable about specific data regions to have a greater influence on the consensus, effectively reducing the impact of noise from non-expert clients. We explore two practical instantiations for the confidence scores E_t^i : a class-frequency-based heuristic and an uncertainty-based score derived from the model’s predictive entropy. The full procedure is detailed in Algorithm 1.

Communication. In each communication round (every b local steps), client i sends a one-hot matrix $L_t^i \in \{0, 1\}^{|U| \times C}$ and expertise vector $E_t^i \in \mathbb{R}^{|U|}$; thus it adds exactly one scalar per public example compared to federated co-training (Abourayya et al., 2025). Encoding L_t^i by class indices (majority vote depends only on $\arg \max$) uses $\lceil \log_2 C \rceil$ bits per example instead of C bits, and quantizing expertise to b_E bits gives a per-round uplink budget $B_{\text{FEDMOSAIC}} = |U| (\lceil \log_2 C \rceil + b_E)$ bits. By contrast, parameter sharing (e.g., FEDAVG) uploads $32P$ bits for a model with P parameters. For example, as in our Fashion-MNIST experiments with $|U| = 10^4$ and $C = 10$, choosing $b_E = 8$ gives $B_{\text{FEDMOSAIC}} = 10^4(4 + 8) = 1.2 \times 10^5$ bits (≈ 15 KB) per client and round; parameter sharing instead communicates ≈ 2.6 MB, so FEDMOSAIC reduces communication by a factor of ≈ 177 .

Algorithm 1: Federated Co-Training with Adaptivity and Specialization (FEDMOSAIC)

Input: communication period b , m clients with local datasets D^1, \dots, D^m and learning algorithms $\mathcal{A}^1, \dots, \mathcal{A}^m$, unlabeled public dataset U , total rounds T

Output: final models h_T^1, \dots, h_T^m

```

1 initialize local models  $h_0^1, \dots, h_0^m$ ,  $P \leftarrow \emptyset$ 
2 Locally at client  $i$  at time  $t$  do
3   compute local loss  $\ell_{\text{priv}} = \mathcal{L}(h_{t-1}^i, D^i)$ 
4   compute pseudo-label loss  $\ell_{\text{pseudo}} = \mathcal{L}(h_{t-1}^i, P)$ 
5   compute adaptive weight  $\lambda_t^i = \exp\left(-\frac{\ell_{\text{pseudo}} - \ell_{\text{priv}}}{\ell_{\text{priv}}}\right)$ 
6   compute loss  $\ell = \ell_{\text{priv}} + \lambda_t^i \ell_{\text{pseudo}}$ 
7   update  $h_t^i \leftarrow \mathcal{A}^i(\ell, h_{t-1}^i)$ 
8   if  $t \% b = b - 1$  then
9     construct prediction matrix  $L_t^i \in \{0, 1\}^{|U| \times C}$ 
10    construct expertise vector  $E_t^i \in (0, \infty)^{|U|}$ 
11    send  $(L_t^i, E_t^i)$  to server and receive  $L_t$ 
12     $P \leftarrow (U, L_t)$ 
13  end
14 At server at time  $t$  do
15  receive  $(L_t^1, E_t^1), \dots, (L_t^m, E_t^m)$  from clients
16  compute weighted score matrix  $S_t = \sum_{i=1}^m \text{diag}(E_t^i) \cdot L_t^i$ 
17  set pseudo-labels  $L_t[j] = \arg \max_{c \in [C]} S_t[j, c]$  for all  $j \in \{1, \dots, |U|\}$ 
18  send  $L_t$  to all clients

```

Convergence under dynamic pseudo-labels: To provide theoretical support, we analyze the convergence behavior of FEDMOSAIC under standard assumptions in stochastic optimization. Our goal is to characterize the rate at which each client’s objective approaches a stationary point, despite the dynamic pseudo-labeling and the heterogeneity of local objectives.

We assume standard conditions, including smoothness of the loss functions, bounded gradient variance, and bounded drift of pseudo-labels across rounds. These assumptions reflect the structure of FEDMOSAIC, where local objectives are updated periodically but converge due to the stabilization of pseudo-labels as shown by [Abourayya et al. \(2025\)](#).

Assumptions 1. The following conditions hold for each client $i \in [m]$ at round t :

1. Each loss function $\mathcal{L}_i^{\text{local}}$ and $\mathcal{L}_i^{\text{global}, t}$ is $L(1+e)^{-1}$ -smooth.
2. The gradient estimator g_i^t is unbiased and has bounded variance:
$$\mathbb{E}[g_i^t] = \nabla \mathcal{L}_i^t(\theta_t), \quad \mathbb{E}[\|g_i^t - \nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \sigma^2.$$
3. The global loss has bounded gradients: $\|\nabla \mathcal{L}_i^{\text{global}, t}(\theta)\| \leq G$ for all θ and t .
4. The objective drift is bounded: $|\mathcal{L}_i^{t+1}(\theta) - \mathcal{L}_i^t(\theta)| \leq \delta$, $\forall \theta$.
5. The per-sample gradient variance is bounded:

$$\mathbb{E}_{x \in D_i} \left[\left\| \nabla_{\theta} \ell(\theta, x, \hat{y}^t) - \nabla \mathcal{L}_i^{\text{local}, t} \right\|^2 \right] \leq \bar{\sigma}^2, \quad \mathbb{E}_{x \in U} \left[\left\| \nabla_{\theta} \ell(\theta, x, \hat{y}^t) - \nabla \mathcal{L}_i^{\text{global}, t} \right\|^2 \right] \leq \bar{\sigma}^2$$

Under these conditions, we establish that FEDMOSAIC converges to an approximate stationary point. Specifically, after T communication rounds, the average squared gradient norm decreases at a rate of $\mathcal{O}(1/T)$ plus additive terms accounting for local and global variance and pseudo-label drift.

Table 2: Average test accuracy (%) under pathological and practical Non-IID Settings for $m = 15$ clients. Color Map: **baselines**, **worse than both baselines**, **worse than local training**, **better than both baselines**.

| | Method | Pathological non-IID | | Practical non-IID | |
|-----|------------------|----------------------|---------------------|---------------------|---------------------|
| | | Fashion-MNIST | CIFAR-10 | Fashion-MNIST | CIFAR-10 |
| | Centralized | 99.28 (0.1) | 87.90 (0.1) | 99.28 (0.03) | 87.90 (0.04) |
| | Local training | 99.32 (0.02) | 88.01 (0.01) | 98.23 (0.01) | 83.91 (0.2) |
| FL | FedAvg | 76.72 (0.1) | 64.42 (0.2) | 83.71 (0.2) | 70.28 (0.4) |
| | FedProx | 77.88 (0.3) | 70.25 (0.2) | 84.14 (0.3) | 73.35 (0.4) |
| | FedCT | 78.15 (0.01) | 73.91 (0.02) | 85.27 (0.01) | 74.39 (0.01) |
| | FedBN | 78.04 (0.3) | 81.35 (0.5) | 85.39 (0.3) | 80.41 (0.7) |
| | Per-FedAvg | 98.63 (0.02) | 87.20 (0.01) | 97.11 (0.01) | 81.37 (0.2) |
| PFL | Ditto | 99.37 (0.01) | 87.94 (0.01) | 98.39 (0.02) | 83.89 (0.04) |
| | pFedMe | 74.80 (0.4) | 81.47 (0.3) | 80.01 (0.1) | 81.61 (0.4) |
| | APFL | 99.26 (0.04) | 87.98 (0.01) | 97.96 (0.03) | 83.81 (0.2) |
| | FedPHP | 99.30 (0.01) | 87.90 (0.01) | 98.40 (0.01) | 83.75 (0.03) |
| | PerFed-CKT | 99.34 (0.01) | 87.95 (0.01) | 98.20 (0.01) | 83.87 (0.03) |
| | FEDMOSAIC | 99.40 (0.01) | 88.03 (0.01) | 98.43 (0.01) | 86.15 (0.01) |

Proposition 1 (Convergence of FEDMOSAIC). *Let each client’s objective at round t be*

$$\mathcal{L}_i^t(\theta) = \mathcal{L}_i^{\text{local}}(\theta) + \lambda_i^t \mathcal{L}_i^{\text{global},t}(\theta), \text{ where } \lambda_i^t = \exp\left(-\frac{\mathcal{L}_i^{\text{global}}(\theta_t) - \mathcal{L}_i^{\text{local},t}(\theta_t)}{\mathcal{L}_i^{\text{local},t}(\theta_t)}\right),$$

and $\mathcal{L}_i^{\text{global},t}$ may change at each round due to pseudo-label updates. Under Assumptions 1-5, for a fixed step size $0 < \eta \leq (2L)^{-1}$ and $\min_i |D_i| = d$, after T rounds of FEDMOSAIC, it holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \mathcal{L}_i^t(\theta_t)\|^2] \leq \frac{4L(\mathcal{L}_i^0 - \mathcal{L}_i^*)}{T} + \frac{\bar{\sigma}^2}{2Ld} + \frac{e^2 \bar{\sigma}^2}{2L|U|} + 2\delta.$$

The proof is provided in Appendix A. Abourayya et al. (2025) show that under the assumption of increasing local accuracy, pseudo-labels stabilize after some round t_0 , so the assumption of a bounded change in the client objective is realistic. In fact, the global loss term effectively becomes stationary under these assumptions quickly and the expected drift becomes negligibly small as t increases.

Client-Level Privacy: In each round t , client i communicates a hard-label matrix $L_t^i \in \{0, 1\}^{|U| \times C}$ (one-hot predictions on U) and an *expertise* vector $E_t^i \in \mathbb{R}^{|U|}$ (one scalar per $u \in U$). Compared to Abourayya et al. (2025), which releases only L_t^i , the present protocol adds exactly one real value per unlabeled example. We apply the XOR mechanism to L_t^i . For this, Abourayya et al. (2025) showed that for on-average replace-one stable learning algorithms the sensitivity s^* of L_t^i is bounded, yielding a per-round ε_L -DP guarantee at the client level. For the expertise scores E_t^i we apply the Gaussian mechanism (Dwork et al., 2014) with variance σ^2 . Since the expertise scores are in $[0, 1]$ for class frequencies and in $[0, \log C]$ for predictive entropy, the (per-coordinate) sensitivity of E_t^i is bounded, which yields (ε_E, δ) -DP with

$$\varepsilon_E = \frac{c\sqrt{|U|}}{\sigma} \sqrt{2 \ln(1.25/\delta)},$$

where $c = 1$ for class frequencies and $c = \log C$ for predictive entropy. Combined, these two mechanisms on L_t^i and E_t^i yield $(\varepsilon_L + \varepsilon_E, \delta)$ -DP for FEDMOSAIC in each round.

4 EMPIRICAL EVALUATION

In this section, we evaluate FEDMOSAIC¹ against a suite of strong baselines in three challenging heterogeneity scenarios: (1) label skew, (2) feature shift, and (3) a hybrid setting combining both. We evaluate our method against FL (FedAvg, FedProx, FedCT, FedBN), state-of-the-art PFL methods (Per-FedAvg, Ditto, pFedMe, APFL, FedPHP, PerFed-CKT), and crucial local training and centralized baselines, which are essential for measuring true collaborative benefit. Centralized training refers

Table 3: Average test accuracy (%) on the Office-10 and DomainNet datasets in feature shift scenarios. For Office-10: A, C, D, W = Amazon, Caltech, DSLR, WebCam. For DomainNet: C, I, P, Q, R, S = Clipart, Infograph, Painting, Quickdraw, Real, Sketch. Color Map: see Table 2.

| | Method | Office-10 | | | | DomainNet | | | | | | |
|------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | | A | C | D | W | C | I | P | Q | R | S | |
| FL | Centralized | 74.03 (0.1) | 58.24 (0.2) | 79.12 (0.2) | 78.52 (0.01) | 70.53 (0.4) | 30.59 (0.3) | 61.87 (0.2) | 71.50 (0.1) | 70.17 (0.4) | 64.62 (0.3) | |
| | Local training | 71.36 (0.02) | 38.67 (0.3) | 81.25 (0.1) | 76.27 (0.2) | 65.31 (0.5) | 38.25 (0.7) | 66.52 (0.3) | 78.43 (0.3) | 71.04 (0.2) | 70.53 (0.6) | |
| | FedAvg | 71.88 (0.1) | 48.44 (0.1) | 40.63 (0.2) | 54.24 (0.6) | 55.71 (0.2) | 28.42 (0.5) | 40.25 (0.3) | 52.64 (0.2) | 54.15 (0.1) | 56.12 (0.2) | |
| | FedProx | 73.44 (0.2) | 52.00 (0.2) | 68.75 (0.4) | 79.66 (0.4) | 59.41 (0.2) | 35.74 (0.4) | 48.82 (0.4) | 55.37 (0.1) | 56.82 (0.5) | 59.17 (0.2) | |
| | FedCT | 73.96 (0.1) | 57.21 (0.2) | 68.73 (0.01) | 70.31 (0.02) | 61.53 (0.3) | 35.19 (0.01) | 64.73 (0.03) | 60.82 (0.01) | 71.85 (0.02) | 69.25 (0.01) | |
| | FedBN | 75.39 (0.01) | 58.13 (0.01) | 78.54 (0.2) | 78.23 (0.8) | 69.45 (0.3) | 38.01 (0.1) | 68.12 (0.2) | 79.21 (0.2) | 76.20 (0.1) | 69.23 (0.1) | |
| | PFL | Per-FedAvg | 73.04 (0.1) | 51.81 (0.5) | 69.22 (0.3) | 77.58 (0.01) | 68.42 (0.01) | 36.21 (0.2) | 60.49 (0.2) | 72.63 (0.1) | 70.84 (0.3) | 68.16 (0.3) |
| | | Ditto | 75.30 (0.01) | 57.91 (0.3) | 78.39 (0.02) | 78.39 (0.1) | 70.97 (0.01) | 39.13 (0.01) | 67.31 (0.02) | 80.33 (0.03) | 77.35 (0.01) | 73.14 (0.03) |
| | | pFedMe | 70.83 (0.3) | 49.78 (0.1) | 75.00 (0.03) | 64.41 (0.01) | 67.21 (0.1) | 37.42 (0.3) | 65.17 (0.2) | 75.24 (0.2) | 74.19 (0.1) | 68.93 (0.3) |
| APFL | | 71.30 (0.01) | 39.05 (0.06) | 50.85 (0.2) | 69.63 (0.1) | 68.73 (0.1) | 38.05 (0.3) | 67.39 (0.3) | 79.14 (0.01) | 77.42 (0.1) | 71.85 (0.2) | |
| FedPHP | | 70.63 (0.5) | 40.13 (0.04) | 51.78 (0.01) | 72.74 (0.02) | 65.29 (0.4) | 36.32 (0.3) | 66.01 (0.5) | 77.03 (0.2) | 75.28 (0.6) | 70.11 (0.1) | |
| PerFed-CKT | | 71.26 (0.1) | 46.80 (0.3) | 74.22 (0.2) | 73.50 (0.02) | 67.49 (0.2) | 37.41 (0.1) | 62.83 (0.5) | 72.45 (0.1) | 65.39 (0.2) | 62.59 (0.1) | |
| FEDMOSAIC | | 80.21 (0.01) | 60.00 (0.02) | 81.25 (0.02) | 83.05 (0.1) | 71.36 (0.1) | 41.59 (0.2) | 69.38 (0.4) | 84.27 (0.1) | 79.25 (0.3) | 75.03 (0.2) | |

to applying the local training algorithm on the pooled data from all clients, as if it were stored in a single location. Local training refers to each client training a model independently using only its own local data, without any collaboration.¹

Experimental Setup: A core component of our method is the shared, unlabeled public dataset U . Following standard practice in semi-supervised learning, for each experiment this dataset is a small, class-balanced sample from the original training set, omitting its labels. This ensures that U is drawn IID from the global training distribution and is disjoint from every client dataset D_i ($U \cap D_i = \emptyset$); since the D_i are non-IID, U 's distribution differs from each D_i . This way, U provides a comprehensive view of the label space, even when clients' private data is highly skewed.

We set the size of U to: CIFAR-10—3,000 samples; Fashion-MNIST—2,250 samples; DomainNet—300 samples; and Office-10—80 samples. A comprehensive ablation study detailing the impact of the public dataset's size and distribution as well as an investigation of individual clients' losses, is provided in the Appendix B.

Label Skew: We first evaluate FEDMOSAIC under label distribution skew, a common protocol where clients see only subsets of the available classes. We test on two variants: a "pathological" setting where each of the 15 clients on Fashion-MNIST and CIFAR-10 holds data from only 2 classes, and a more practical setting where label proportions are drawn from a Dirichlet distribution. These settings are widely adopted in the literature (T Dinh et al., 2020; Fallah et al., 2020; Zhang et al., 2023a;d;b). For these experiments, we use the class-frequency-based confidence score, a natural fit for scenarios dominated by class imbalance.

As shown in Table 2, FEDMOSAIC achieves top performance across all settings. In the pathological case on CIFAR-10, it scores 0.8803, surpassing all PFL methods and, crucially, the strong local training baseline (0.8801). This result is significant: it demonstrates that FEDMOSAIC's adaptive collaboration successfully extracts useful signals from peers without being corrupted by their extreme data skew, achieving a better outcome than local training. Performance trends are similar in the practical scenario, confirming the method's robustness to varying degrees of label imbalance.

¹Code to reproduce all experimental results: <https://anonymous.4open.science/r/FEDMOSAIC/README.md>

Table 4: Average test accuracy (in %) on the DomainNet and Office-10 dataset in hybrid settings for $m = 30$ clients on DomainNet and $m = 20$ on Office-10. Color map: see Table 2.

| | Method | DomainNet | DomainNet (ViT) | Office-10 |
|-----|----------------|--------------|-----------------|--------------|
| | Centralized | 66.24 (0.4) | 68.25 (0.2) | 40.92 (0.6) |
| | Local training | 84.64 (0.1) | 84.92 (0.3) | 86.79 (0.4) |
| FL | FedAvg | 31.00 (0.8) | 33.28 (0.5) | 37.25 (0.8) |
| | FedProx | 55.23 (0.1) | 57.18 (0.3) | 58.39 (0.3) |
| | FedCT | 56.38 (0.01) | 67.52 (0.02) | 59.42 (0.02) |
| | FedBN | 71.54 (0.3) | 70.39 (0.4) | 75.48 (0.3) |
| PFL | Per-FedAvg | 72.48 (0.4) | 73.19 (0.3) | 71.92 (0.5) |
| | Ditto | 81.47 (0.01) | 83.82 (0.02) | 80.63 (0.01) |
| | pFedMe | 75.21 (0.5) | 76.81 (0.8) | 74.83 (0.7) |
| | APFL | 80.59 (0.3) | 83.27 (0.5) | 80.91 (0.1) |
| | FedPHP | 78.25 (0.6) | 77.31 (0.7) | 76.36 (0.4) |
| | PerFed-CKT | 79.24 (0.4) | 80.16 (0.2) | 82.49 (0.1) |
| | FEDMOSAIC (W) | 87.44 (0.02) | 88.52 (0.2) | 89.06 (0.01) |
| | FEDMOSAIC (U) | 88.36 (0.01) | 87.35 (0.1) | 89.43 (0.03) |

Feature Shift: To evaluate robustness to heterogeneous input distributions, we test on feature shift scenarios using the Office-10 and DomainNet datasets. Here, each domain (e.g., "Webcam," "Sketch") acts as a client, sharing a common label space but having a unique data style. Table 3 shows that FEDMOSAIC consistently sets the state-of-the-art on all domains. On the complex DomainNet benchmark, it achieves the highest accuracy across all six domains, outperforming specialized methods like Ditto and FedBN. This demonstrates that the dynamic weighting and confidence-based aggregation are not limited to label skew; they effectively manage domain-specific features, allowing clients to learn from each other while preserving their specialized knowledge.

Hybrid Distribution (Label Skew + Feature Shift): We now consider the most challenging scenario: a hybrid of label skew and feature shift. To simulate this, we partition each domain in DomainNet and Office-10 into 5 clients, each assigned only 2 of the 10 classes. This results in 30 highly heterogeneous clients for DomainNet and 20 for Office-10. In this demanding setup, we evaluate both our confidence mechanisms: the class-frequency heuristic (FEDMOSAIC-W) and the uncertainty-based score (FEDMOSAIC-U).

The results in Table 4 confirm the superiority of our approach. With both AlexNet and ViT architectures, FEDMOSAIC variants significantly outperform all baselines. On Office-10, for instance, FEDMOSAIC-U achieves 0.8943 accuracy, a remarkable improvement over the next best baseline, Ditto (0.8063). One can note that centralized training is worse than local training due to the highly heterogeneous setting, meaning that a single global model cannot fit all clients effectively.

Interestingly, both the simple class-frequency heuristic and the more complex uncertainty-based score yield similarly strong results. This suggests that in settings with extreme label skew, class frequency serves as a powerful and efficient proxy for model expertise.

Taken together, these results validate that FEDMOSAIC's principled approach to adaptive, expert-aware collaboration enables it to deliver state-of-the-art performance, consistently outperforming strong baselines in diverse and realistic non-IID settings.

The Effect of the Unlabeled Dataset: FEDMOSAIC relies heavily on a shared unlabeled dataset $|U|$. To understand how sensitive FEDMOSAIC is to the characteristics of this dataset, we conducted a study on the effect of the size and distribution of this dataset. We simulated varying degrees of skew by sampling $|U|$ (with a fixed size of 3,000) using a Dirichlet distribution. We tested concentration parameters $\alpha = \{1, 0.7, 0.5, 0.3, 0.1\}$, where $\alpha = 1$ corresponds to a perfectly IID distribution and lower values induce increasingly severe skew. As shown in Fig. 1 and Fig. 2, performance degrades as the public dataset becomes more skewed, especially at low α (e.g., 0.3, 0.1) where some classes are missing. However, a key finding is that FEDMOSAIC never performs worse than the local baseline. This highlights the robustness of the adaptive aggregation scheme: when the global signal is unhelpful,

the dynamic weight λ steers clients toward local training, acting as a fail-safe. More details are provided in App. B.

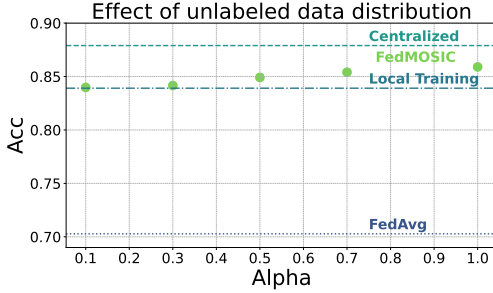


Figure 1: Average test accuracy of FEDMOSAIC on CIFAR-10 under different distribution of U .

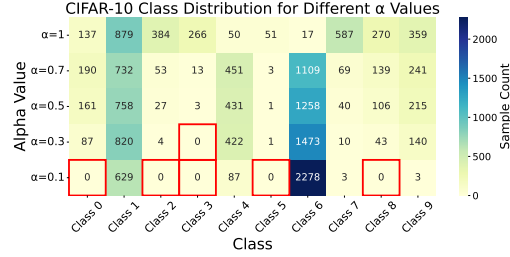


Figure 2: Class distribution of U under different values of alpha.

5 DISCUSSION AND CONCLUSION

Personalized Federated Learning (PFL) aims to address data heterogeneity by tailoring models to client-specific distributions. Yet, as we have demonstrated, many existing approaches fall short of their promise, often failing to outperform even local training or centralized baselines. This raises fundamental concerns about the core premise of collaboration in personalized federated learning.

We argue that meaningful personalization in federated learning requires more than per-client modeling: it must involve adaptive, data-specific collaboration. In particular, effective PFL methods should support example-level decision-making, allowing clients to modulate the degree and direction of collaboration based on local context and per-sample reliability. Without this level of adaptivity, personalization risks becoming a superficial modification of global training.

FEDMOSAIC is one concrete instantiation of this principle. It enables example-level collaboration through dynamic loss weighting and confidence-based aggregation over a shared unlabeled dataset. Unlike prior methods that personalize only at the client level, FEDMOSAIC allows each client to adapt both how much and whom to trust, based on the alignment between public and private data.

Empirical results across a diverse set of non-IID scenarios support the effectiveness of this approach. In the hybrid scenario, which combines label skew and feature shift, FEDMOSAIC outperforms all competitors and baselines by a wide margin. In the feature shift scenarios, it again surpasses all methods across most domains, often with substantial gains. In the label skew setting, FEDMOSAIC consistently achieves the best performance for the pathological non-IID scenario, though with very narrow margins, in particular with respect to local training. In the practical non-IID scenario with milder heterogeneity, centralized training performs best, as expected. Yet, traditional federated learning methods fall short, being outperformed by several PFL approaches, including FEDMOSAIC.

These results illustrate both the strengths and limitations of personalized FL. One limitation is that, particularly in the label skew setting, the advantage over strong local baselines can be modest. Such scenarios, especially the pathological non-IID one, raise the question of whether collaboration is truly justified, and whether evaluation setups that favor strong local baselines but show weak global benefit are well-posed. We therefore emphasize the need for more meaningful benchmarks: scenarios where collaboration has a clear potential upside, and where the evaluation criteria capture the practical value of federated interaction, not just statistical differences. That said, FEDMOSAIC demonstrates that adaptive and data-aware collaboration is both feasible and effective. Across our experiments, it outperforms both local and centralized baselines in most settings, supporting its robustness and practical utility.

While FEDMOSAIC represents a principled and practically validated advance in personalized federated learning, it also opens new directions for future work. A key limitation is the assumption of a public unlabeled dataset. Although such datasets exist in many domains, e.g., healthcare, vision, and language, it remains an open question how to extend this paradigm when such data are limited or unavailable. Developing mechanisms for privacy-preserving dataset synthesis, or leveraging foundation models for public data distillation, could further broaden the applicability of our framework.

REFERENCES

- Amr Abourayya, Jens Kleesiek, Kanishka Rao, Erman Ayday, Bharat Rao, Geoff Webb, and Michael Kamp. Little is enough: Boosting privacy by sharing only hard labels in federated semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
- Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.
- Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.
- Kawa Atapour, S Jamal Seyedmohammadi, Jamshid Abouei, Arash Mohammadi, and Konstantinos N Plataniotis. Fedd2s: Personalized data-free federated knowledge distillation. *arXiv preprint arXiv:2402.10846*, 2024.
- Ilai Bistriz, Ariana Mann, and Nicholas Bambos. Distributed distillation for on-device learning. *Advances in Neural Information Processing Systems*, 33:22593–22604, 2020.
- Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- Ann Cavoukian et al. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada*, 5(2009):12, 2009.
- Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. Fedmatch: Federated learning over heterogeneous question answering data. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 181–190, 2021.
- Yae Jee Cho, Jianyu Wang, Tarun Chiruvolu, and Gauri Joshi. Personalized federated learning for heterogeneous clients with clustered knowledge transfer. *arXiv preprint arXiv:2109.08119*, 2021.
- Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.
- Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.
- Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4):211–407, 2014.
- Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.
- Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7865–7873, 2021.
- Eunjeong Jeong and Marios Kountouris. Personalized decentralized federated learning with knowledge distillation. In *ICC 2023-IEEE International Conference on Communications*, pp. 1982–1987. IEEE, 2023.
- Donglin Jiang, Chen Shan, and Zhihui Zhang. Federated learning algorithm based on knowledge distillation. In *2020 International conference on artificial intelligence and computer engineering (ICAICE)*, pp. 163–167. IEEE, 2020.
- Michael Kamp. *Black-Box Parallelization for Machine Learning*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, Universitäts-und Landesbibliothek Bonn, 2019.
- Michael Kamp, Sebastian Bothe, Mario Boley, and Michael Mock. Communication-efficient distributed online learning with kernels. In *ECMLPKDD*, pp. 805–819. Springer, 2016.

- 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*, pp. 5132–5143. PMLR, 2020.
- Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021a.
- Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.
- Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021b.
- Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021c.
- Xin-Chun Li, De-Chuan Zhan, Yunfeng Shao, Bingshuai Li, and Shaoming Song. Fedphp: Federated personalization with inherited private models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 587–602. Springer, 2021d.
- Siqi Liang, Jintao Huang, Junyuan Hong, Dun Zeng, Jiayu Zhou, and Zenglin Xu. Fednoisy: Federated noisy label learning benchmark. *arXiv preprint arXiv:2306.11650*, 2023.
- Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.
- Osman Mian, David Kaltenpoth, Michael Kamp, and Jilles Vreeken. Nothing but regrets — privacy-preserving federated causal discovery. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 8263–8278. PMLR, 2023.
- Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021.
- Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International conference on machine learning*, pp. 9489–9502. PMLR, 2021.
- Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.
- Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8432–8440, 2022.
- Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022.
- Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. *arXiv preprint arXiv:2306.11867*, 2023.
- Xiyuan Yang, Wenke Huang, and Mang Ye. Fedas: Bridging inconsistency in personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11986–11995, 2024.
- Jianqing Zhang, Yang Hua, Jian Cao, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Eliminating domain bias for federated learning in representation space. In *NeurIPS*, 2023a.

- Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, Jian Cao, and Haibing Guan. Gpfl: Simultaneously learning global and personalized feature information for personalized federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5041–5051, 2023b.
- Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11237–11244, 2023c.
- Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedcp: Separating feature information for personalized federated learning via conditional policy. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3249–3261, 2023d.
- Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2021.
- Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.