Pattern Recognition Letters 203 (2026) 105–110

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Beyond data dependency: FedPET enables robust federated learning via data-free dual-teacher knowledge distillation

Bo Wang [a], Yiming Yan [a], Zhaoning Wang [a], Jialin Liu [a], Wei Wang [b,*]

[a] School of Information and Communication Engineering, Dalian University of Technology, No. 2 Linggong Road, Ganjingzi District, Dalian, 116024, Liaoning, China
[b] New Laboratory of Pattern Recognition (NLPR), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), No. 95 Zhongguancun East Road, Haidian, 100190, Beijing, China

ARTICLE INFO

Editor: Yonghuai Liu

*Keywords:*
Data-free knowledge distillation
Federated learning
Communication efficiency
Model compression

ABSTRACT

Federated Learning (FL) is a distributed paradigm for privacy-preserving machine learning, where clients jointly build deep models by exchanging local updates instead of raw data. Despite its advantages, FL typically requires many communication rounds. When the models contain large parameter sets, this leads to high communication overhead and burden on the system. To overcome these challenges, we introduce a **federated data-free knowledge distillation** framework specifically tailored for image classification tasks, which utilizes a **P**re-trained model and forms an **E**nsemble of client models serving as **T**eachers during the distillation stage (FedPET). To further alleviate the problem of low-quality synthetic samples in data-free settings, we design the Training Sample Generation (TSG) module that enhances sample quality through channel feature exchange by exploiting the spatial and channel characteristics of visual data, thereby boosting student model training. Comprehensive experiments on multiple image classification benchmarks verify that FedPET consistently delivers substantial gains over state-of-the-art (SOTA) approaches. The results confirm the effectiveness of our method in reducing communication costs while sustaining student model performance in visual federated learning scenarios, highlighting its potential for real-world deployment in federated knowledge distillation scenarios.

## 1. Introduction

Ensuring data security and safeguarding privacy have become paramount concerns in modern technology development. Traditional machine learning paradigms rely on centralized data collection, which poses significant risks to user data privacy, security, and compliance. Moreover, in real-world distributed scenarios (e.g., terminal devices, medical systems), data is often scattered across clients (i.e., data silos), making centralized training models that aggregate multi-source data prohibited. Thus, alternative approaches balancing data utility, security, and privacy compliance are urgently needed [1].

Federated Learning (FL) emerges as a promising solution, enabling model training across decentralized data sources without transferring raw data to a central server [2]. FL operates by training local models on clients, iteratively uploading updates to the server for global model aggregation, and then distributing the refined global model back to clients for further optimization–this cycle continues until convergence.

However, frequent exchange of model updates between the server and clients in FL leads to high communication overhead, especially for large-scale deep learning models with massive parameters. For decentralized clients with limited bandwidth and throughput, such transmission costs are impractical [3].

Efforts to enhance FL communication efficiency include gradient compression and knowledge distillation. Gradient compression reduces update size but suffers from severe performance degradation at high compression ratios and struggles with decentralized data heterogeneity [4]. Knowledge distillation transmits local model predictions on shared public data to cut overhead, but sharing even anonymized data is unfeasible in privacy-sensitive scenarios (e.g., medical records, personalized recommendations). Thus, developing communication-efficient FL techniques independent of additional data remains a critical unsolved problem.

To address these challenges, we propose FedPET, a federated model compression approach designed for visual applications, integrating data-free knowledge distillation and ensemble distillation. FedPET uses a pre-trained model and an ensemble of client models as teachers to guide distillation: first, a pre-trained model on the server constrains random noise and labels to train an image generator, with intermediate features

* Corresponding author.
*E-mail addresses:* bowang@dlut.edu.cn (B. Wang), yiming@mail.dlut.edu.cn (Y. Yan), hailap@mail.dlut.edu.cn (Z. Wang), flinx_ly@mail.dlut.edu.cn (J. Liu), wwang@nlpr.ia.ac.cn (W. Wang).

https://doi.org/10.1016/j.patrec.2026.03.006
Received 5 November 2025; Received in revised form 21 January 2026; Accepted 5 March 2026
Available online 7 March 2026

stored in the Training Sample Generation (TSG) module's feature pool; after clients upload local models, the TSG module samples features to generate diverse synthetic data; finally, the pre-trained model and client ensemble model jointly train the client-aggregated student model using synthetic data, achieving model compression while maintaining accuracy compared to state-of-the-art methods.

Specifically, our contributions can be summarized in three aspects:

- To address the balance between model accuracy and compression rate in data-free distillation within federated learning, we innovatively use pretrained models and ensemble models as the teacher for data-free knowledge distillation, ensuring both accuracy and compression.
- To improve the diversity of generated samples and the effectiveness of model distillation, we use a channel feature exchange method optimized for convolutional feature maps. It randomly swaps half of the feature channels between two sampled features from a feature pool, generating mixed features that are passed through the deeper layers of the generator to produce diverse synthetic images as training data.
- We conduct extensive experiments to validate our method. Compared to traditional federated learning methods such as FedAvg and federated data-free distillation, our method demonstrates significant improvements in performance by at least 1.6%.

## 2. Related work

### 2.1. Federated learning

Extensive research focuses on federated learning aggregation algorithms to address communication overhead and data heterogeneity. FedAvg [2] periodically aggregates local models on a central server for global updates, serving as the baseline for FL optimization. FedProx [5] introduces a proximal term to constrain local updates, keeping them close to the global model to mitigate heterogeneity. SCAFFOLD [6] adopts variance reduction techniques with control variables to correct client drift caused by Non-IID data. FedDyn [7] employs linear and quadratic penalty terms to dynamically align global and local optimization objectives. MOON [8] further corrects local training via model-level contrastive learning, enhancing adaptation to heterogeneous data.

### 2.2. Data-free knowledge distillation

Data-Free Knowledge Distillation (DFKD) transfers knowledge from a pre-trained teacher to a lightweight student without raw data. DFKD [9], the earliest attempt, reconstructs training data using the teacher model and its metadata. ZSKD [10] models teacher outputs as a Dirichlet distribution to constrain synthetic data generation. DeepInversion [11] improves generation quality by regularizing via batch normalization statistics of the teacher. Additionally, generative models create surrogate data, supporting data-free distillation across scenarios.

### 2.3. Knowledge distillation in federated learning

DFKD aligns with FL's privacy-preserving nature, driving federated distillation research. FedDF [12] uses ensemble distillation, optimizing the global model with averaged client logits. FedGen [13] transfers local model parameters to the server, training a conditional generator via DFKD for local model refinement. DENSE [14] introduces a framework for one-shot federated learning (FL) that trains a generator to approximate the prediction distribution of an ensemble of client models, enabling the distillation of a global model without requiring multiple rounds of communication. However, its generation strategy may encounter difficulties in fully capturing the intricate characteristics of non-independent and identically distributed (Non-IID) data. To mitigate the issue of catastrophic forgetting of the generator across iterations, DFRD [15] introduces an Exponential Moving Average (EMA) copy of the generator and proposes dynamic weighting strategies to extract local knowledge more accurately. While these methods mitigate communication overheads, they primarily rely on recovering information from local models, often overlooking the value of pre-trained prior knowledge, and the diversity of their generated synthetic samples remains limited.

## 3. Proposed method

First, we define key assumptions: Let $D = \{\mathcal{X} \in R^{c\times h\times w}, \mathcal{Y} = 1, 2, \dots, K\}$ denote the training dataset, where $x^i \in \mathcal{X}$ are training images and $y^i \in \mathcal{Y}$ their corresponding labels. Let $f_P(x;\theta_P)$ be the teacher network pre-trained on $D$. The goal of traditional data-free knowledge distillation is to learn a lightweight student classification network $f_S(x;\theta_S)$ that mimics $f_P(x;\theta_P)$'s classification capabilities without using $D$.

### 3.1. Framework overview

The FedPET process is illustrated in Fig. 1. Based on conventional data-free knowledge distillation, we generate synthetic images in the federated learning framework to train the student network.

Specifically, FedPET has three core stages: 1) Optimize the generative network with the pretrained teacher; 2) Generate training images via Channel Feature Exchange (CFE); 3) Conduct knowledge distillation on the client-aggregated model using pretrained and client ensemble models as dual teachers. The CFE module enhances the generator's ability to produce diverse synthetic images, improving distillation effectiveness. Dual teachers enable the student network to learn visual representations from the pretrained model and extract client knowledge by mimicking the ensemble model's logits, thus boosting the aggregated student model's performance.

### 3.2. Generator optimization and data generation

This stage aims to train a high-performance generator that captures the pretrained model's training space (stored on the server) to enhance generated data diversity. We use a generative network $G$ to produce synthetic data and labels $\{\hat{x}, y\}$, aligning their distribution with the pretrained model's real training dataset $D = \{\mathcal{X}, \mathcal{Y}\}$. At the start of each global epoch, we initialize a mini-batch of Gaussian random noise $z$ and uniform random class labels $y \in \{1, 2, \dots, K\}$ as input to $G$, which outputs $\hat{x} = G(z)$.

To refine this output, we apply cross-entropy loss $\mathcal{L}_{CEg}$ via the pretrained model to constrain $\hat{x}$ and $y$, ensuring synthetic data is accurately classified into target categories under the teacher's guidance.

Moreover, we incorporate Batch Normalization (BN) regularization loss $\mathcal{L}_{BNg}$ [11,16] (common in data-free distillation) to align feature mean/variance in BN layers with the teacher's running mean/variance. Training $G$ with both $\mathcal{L}_{CEg}$ and $\mathcal{L}_{BNg}$ aligns synthetic data with the distribution of clients' original data theoretically. The overall loss for optimizing $G$ is:

$$\mathcal{L}_G = \lambda_{CEg} \cdot \mathcal{L}_{CEg} + \lambda_{BNg} \cdot \mathcal{L}_{BNg} \tag{1}$$

where $\lambda_{CEg}$ and $\lambda_{BNg}$ are weight coefficients balancing the two loss terms.

To further improve generated data quality, we design a feature pool storing hidden representations of synthetic images during $G$'s optimization. With channel feature exchange (CFE), these hidden features enhance training sample diversity. The synthetic sample generation process is shown in Fig. 2.

To train $f_S(\cdot)$ with diverse synthetic data, we first sample features from the feature pool. Let $F_a^i$ and $F_b^i$ denote features from different synthetic images (generated by $G$) extracted from layer $i$. After sampling, we perform channel-wise feature exchange by randomly swapping half the feature channels between $F_a^i$ and $F_b^i$, producing two mixed features:

$$F_{ab}^i = cfe(F_a^i, F_b^i) \tag{2a}$$

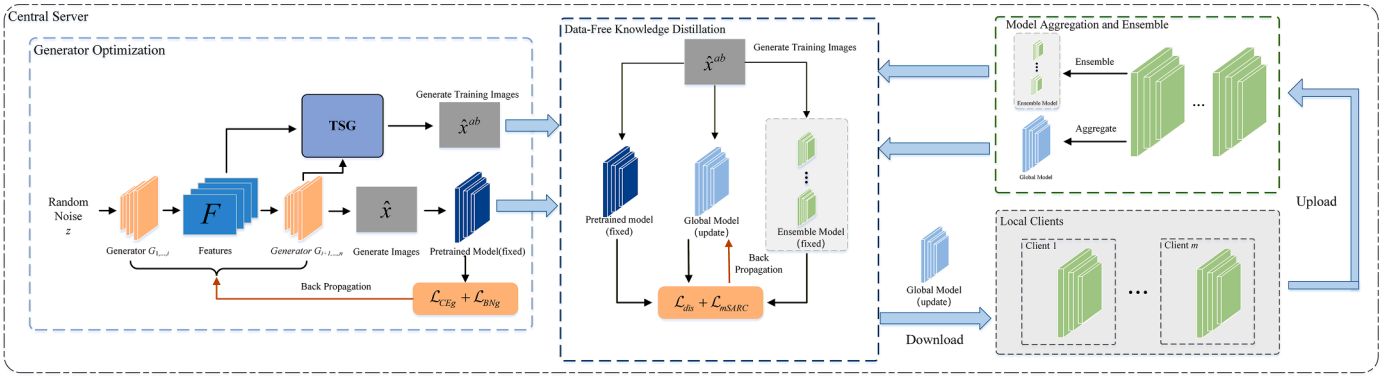


**Fig. 1.** Full pipeline of FedPET. FedPET runs on the server and includes two phases: generator optimization and data-free knowledge distillation. $\mathcal{L}_{CEg}$, $\mathcal{L}_{BNg}$ are conditional generator losses; $\mathcal{L}_{mSARC}$, $\mathcal{L}_{dis}$ are global model losses. $G_{1,\dots,i}$ represent layers 1 to $n$ of the generative network $G$.

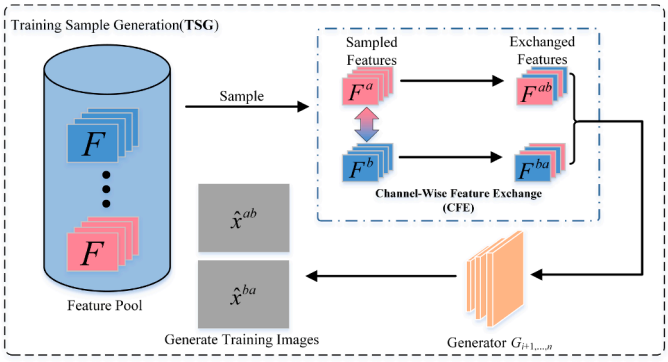


**Fig. 2.** Framework of the Training Sample Generation (TSG) module using CFE. $G_{i+1,\dots,k}$ represent layers $i+1$ to $n$ of the generative network $G$.

$$F_{ba}^{i} = cfe(F_{b}^{i}, F_{a}^{i}) \tag{2b}$$

We then feed $F_{ab}^{i}$ and $F_{ba}^{i}$ into $G$'s deeper layers (from $i+1$ to $n$), resulting in two distinct synthetic images $\hat{x}_{ab}^{i}$ and $\hat{x}_{ba}^{i}$. The CFE method broadens feature variations to enhance generated image diversity compared to traditional models.

### 3.3. Multi-teacher knowledge distillation based on pretrained model and ensemble model

Methods like FedAvg (direct model aggregation) fail to fully exploit clients' unique data characteristics under heterogeneous data. We use an ensemble model with pretrained models to train the global model via generator-based data-free distillation. Specifically, we distill knowledge from the ensemble model into the client-aggregated global model by minimizing prediction discrepancy between dual teachers (ensemble + pretrained models) and the student (global model) on synthetic data. The distillation process is detailed below:

First, we compute the ensemble model's logits for synthetic data:

$$f_E(\hat{x}, \{\theta^k\}_{k=1}^{m}) \leftarrow \frac{1}{m}\sum\nolimits_{k\in C} f^k(\hat{x}, \theta^k) \tag{3}$$

Let $C$ be the client set; $f_E(\hat{x}, \{\theta^k\}_{k=1}^{m})$ is the ensemble model's average logits for input $\hat{x}$. $\theta^k$ denotes the $k$-th client's model parameter, and $f^k(\hat{x}, \theta^k)$ is the $k$-th client's prediction function outputting logits for $\hat{x}$.

Next, we extract knowledge from the ensemble and pretrained models into the global model by minimizing the following objective function, which combines their logits via weighted average:

$$\mathcal{L}_{dis}(\hat{x}; \theta_S) = KL\Big(\big(\alpha_E f_E(\hat{x}, \{\theta^k\}_{k=1}^{m}) + \alpha_P f_P(\hat{x})\big), f_S(\hat{x}; \theta_S)\Big) \tag{4}$$

where $\alpha_E$ and $\alpha_P$ are weight coefficients for the ensemble and pretrained teacher models respectively; $f_P(\hat{x})$ is the pretrained model's logits for $\hat{x}$.

Relying solely on KL divergence may make student and teacher focus on different features despite same labels. We introduce a multi-scale Spatial Activation Region Consistency (mSARC) constraint [17] to guide the student network $f_S(\cdot)$'s intermediate layers to focus on spatial regions aligned with the pretrained teacher $f_P(\cdot)$. Concretely, we enforce consistency between Class Activation Maps (CAMs) [18] of the $i_k$-th layer (student) and $j_k$-th layer (teacher), defined as:

$$\mathcal{L}_{mSARC} = \mathrm{E}_{\hat{x}} \sum_{k=1}^{t} \left\| \mathrm{CAM}_{i_k}(S, \hat{x}) - \mathrm{CAM}_{j_k}(T, \hat{x}) \right\|_2^2 \tag{5}$$

where $k = 1, 2, \cdots, t$, and $i_k$, $j_k$ denote layer indices.

Finally, the overall multi-teacher knowledge distillation loss $\mathcal{L}_{MTKD}$ is:

$$\mathcal{L}_{MTKD} = \lambda_{KL} \cdot \mathcal{L}_{dis} + \lambda_{mSARC} \cdot \mathcal{L}_{mSARC} \tag{6}$$

where $\lambda_{KL}$ and $\lambda_{mSARC}$ are weight coefficients balancing the KL divergence and mSARC loss terms.

## 4. Results

In this section, we provide an empirical evaluation of FedPET. Section 4.1 outlines the experimental setup. Section 4.2 presents comparisons between FedPET and representative state-of-the-art FL methods as well as federated data-free knowledge distillation approaches. Section 4.3 examines communication efficiency improvements. Section 4.4 reports ablation studies to assess the contribution of individual components and the impact of key parameters.

### 4.1. Experimental settings

#### 4.1.1. Dataset

To thoroughly evaluate the performance of FedPET, we adopt the CIFAR-10 [19] and CIFAR-100 [19] datasets, which are widely applied in FL studies and are considered challenging benchmarks under federated settings. CIFAR-10 and CIFAR-100 each contain 60,000 color images with a resolution of $32 \times 32$, comprising 10 and 100 categories, respectively.

To model client-level data heterogeneity, following prior studies [20–22], we partition the training data according to a Dirichlet distribution $Dir(\omega)$, which produces Non-IID local datasets across clients. Here, $\omega$ controls the imbalance degree: smaller values of $\omega$ correspond to higher data heterogeneity.

#### 4.1.2. Baselines

We evaluate FedPET against several representative federated learning approaches, including FedAvg, FedProx, SCAFFOLD, FedDyn, and

**Table 1**
Test accuracy comparison on CIFAR-10 and CIFAR-100.

| Method | CIFAR-10 | | | CIFAR-100 | | |
|---|---|---|---|---|---|---|
| | $\omega = 1$ | $\omega = 0.6$ | $\omega = 0.3$ | $\omega = 1$ | $\omega = 0.6$ | $\omega = 0.3$ |
| FedAvg | 83.81 | 82.02 | 79.43 | 50.13 | 50.67 | 50.17 |
| FedProx | 84.10 | 82.36 | 80.12 | 51.25 | 50.94 | 50.82 |
| FedDyn | 85.19 | 82.87 | 80.15 | 53.27 | 51.68 | 50.51 |
| MOON | 84.34 | 82.67 | 80.97 | 52.51 | 52.55 | 51.88 |
| SCAFFOLD | 85.99 | 84.55 | 82.14 | 53.32 | 53.91 | 54.36 |
| FedGen | 83.91 | 82.23 | 79.72 | 50.38 | 50.71 | 50.08 |
| FedDF | 84.47 | 82.92 | 80.97 | 52.12 | 51.36 | 51.26 |
| FedFTG | 87.34 | 86.06 | 84.38 | 56.94 | 56.49 | 55.96 |
| FedPET(ours) | **92.54** | **91.21** | **88.98** | **68.77** | **68.23** | **67.59** |

MOON, along with data-free knowledge distillation methods such as FedGen, FedDF, FedFTG, DENSE, and DFRD. Since FedDF does not explicitly define the procedure for generator training, we adopted the same strategy as FedGen. For DENSE and DFRD, we strictly followed the configurations reported in DFRD to ensure a fair and consistent comparison.

#### 4.1.3. Network architecture

For both the CIFAR-10 and CIFAR-100 datasets, we utilize ResNet18 [23] as the backbone architecture for the student model, while ResNet34 [23] serves as the backbone for the pretrained teacher model. Additionally, to investigate the impact of our method on improving communication efficiency, we further employed VGG [24] and Wide ResNet [25] as the backbone architectures for the student and teacher models.

#### 4.1.4. Hyperparameters

In our approach, the number of local training epochs is set to $E = 5$, and the number of active clients is fixed at $K = 10$. The global training is conducted for 200 rounds, where the learning rates of the local model and the generator are configured as 0.01 and 0.001, respectively. The temperature parameter for knowledge distillation is set to 30. Unless otherwise specified, the weight coefficients of the pretrained teacher $\alpha_P$ and the ensemble teacher $\alpha_E$ in Equation 1.6 are assigned values of 3.2 and 0.1, respectively.

### 4.2. Performance comparison

#### 4.2.1. Comparative analysis with mainstream methods

We evaluated accuracy of all baselines on CIFAR-10 and CIFAR-100 with $\omega$ set to 1, 0.6 and 0.3. Detailed results are in Table 1. FedPET achieves the best performance across all IID and Non-IID scenarios. It outperforms mainstream methods by at least 4% in highly heterogeneous settings with $\omega$ equal to 0.3 or 0.6. FedPET maintains performance improvements under IID conditions. Its higher accuracy over FedGen and FedDF comes from the CFE module, which enhances synthetic sample diversity and improves knowledge transfer.

#### 4.2.2. Comparative analysis with state-of-the-art methods

We compared FedPET with SOTA federated data-free knowledge distillation methods DENSE and DFRD on CIFAR-100. FedPET's parameters match those in DFRD for fairness. Table 2 presents detailed results. FedPET consistently outperforms DENSE and DFRD by 1.6% in test accuracy across various $\omega$ values. It shows strong robustness in high heterogeneity scenarios with $\omega$ equal to 0.01 or 0.1, where traditional methods struggle.

The observed performance improvement can be attributed primarily to the enhanced diversity of synthetic samples and the synergistic effect of the proposed dual-teacher strategy. In contrast to DENSE, which relies on a basic generator, FedPET integrates the TSG module to generate synthetic images with richer semantic variations by mixing feature channels, thereby forcing the student model to learn more robust decision boundaries. Furthermore, while DFRD mitigates generator forgetting via EMA, its knowledge base is confined to client models. In comparison, FedPET leverages both the domain-specific knowledge derived from the client ensemble and the general visual features from the server-side pre-trained teacher. This combination effectively compensates for the fragmented knowledge of client models under extreme Non-IID settings, achieving higher accuracy than DFRD.

**Table 2**
Test accuracy comparison of SOTA federated data-free knowledge distillation methods on CIFAR-100.

| Method | CIFAR-100 | | | Runtime Ratio |
|---|---|---|---|---|
| | $\omega = 1$ | $\omega = 0.1$ | $\omega = 0.01$ | |
| DENSE | 58.43 | 50.84 | 40.24 | **0.06** |
| DFRD | 61.69 | 54.05 | 43.47 | **0.05** |
| FedPET(ours) | **63.29** | **55.82** | **49.56** | 1 |

#### 4.2.3. Computational complexity analysis

We further analyze the computational complexity of FedPET. Theoretically, client-side complexity is minimal since clients are required to train only a lightweight student network, which is essential for resource-limited edge devices. The primary computational burden is shifted to the server, involving the optimization of the generator and the multi-teacher distillation process.

We evaluate efficiency using the Runtime Ratio, defined as the runtime of the compared method divided by that of FedPET over the same number of global epochs. As shown in Table 2, FedPET demonstrates greater computational complexity compared to DENSE and DFRD. This increased training time is a deliberate trade-off for enhanced performance and robustness, directly attributed to the additional computations required by the TSG module and the multi-teacher distillation mechanism. Given that servers typically possess abundant computing resources than clients, and considering the significant reduction in communication overhead, we believe that the server-side computational cost is justified by the substantial accuracy gains achieved in heterogeneous federated scenarios.

### 4.3. Communication efficiency improvement

We analyzed compression performance of distilled models on CIFAR-10 and CIFAR-100 to verify FedPET's ability to reduce model transmission volume. Compression ratio is defined as Model Storage Size Reduction divided by Original Model Storage Size, with the pretrained teacher model as the benchmark. Detailed results are in Table 3.

FedPET achieves a minimum compression ratio of 47.39% relative to the pretrained teacher model, nearly halving model storage size. Compression is accompanied by a maximum accuracy decrease of 6.29%, showing a good balance between compression efficiency and performance retention. ResNet-based architectures perform notably: the student model reduces storage by over 40% while maintaining comparable classification accuracy, making it suitable for bandwidth-limited scenarios.

### 4.4. Ablation study

#### 4.4.1. Impacts of weight coefficient $\alpha_P$ and $\alpha_E$

To assess how the pretrained teacher model and client ensemble teacher model affect the student model's training effectiveness, we conducted ablation experiments on CIFAR-10. We varied the weight ratio between the pretrained model and client ensemble model, denoted as the teacher weight ratio $\rho$. $\alpha_P$ represents the weight coefficient of the pretrained teacher model, and $\alpha_E$ represents that of the ensemble teacher model, as defined in Eq. (4). Specific results are shown in Tables 4, 5 and Fig. 3. Fig. 3 visualizes the accuracy differences between different $\rho$ values and the baseline ($\rho = 1$), clarifying the relationship between the pretrained teacher model's weight ratio and the student model's accuracy.

**Table 3**
Model compression ratio and accuracy comparison under different architectures.

| Teacher Model | Student Model | Teacher Model Storage(Mb) | Student Model Storage(Mb) | Compression Ratio(%) | Model Accuracy Decrease(%) |
|---|---|---|---|---|---|
| ResNet-34 | ResNet-18 | 81.50 | 42.88 | 47.39 | 3.16 |
| WRN-40 | WRN-16 | 94.87 | 88.58 | 69.80 | 6.29 |
| VGG-11 | VGG-8 | 92.24 | 87.58 | 57.60 | 4.66 |

**Table 4**
Test accuracy (%) for different teacher weight ratio ($\rho \geq 1$).

| Weight Coefficient $\alpha_P$ | Weight Coefficient $\alpha_E$ | Teacher Weight Ratio $\rho$ | Global Model Accuracy(%) |
|---|---|---|---|
| 6.4 | 0.1 | 64 | 92.34 |
| 3.2 | 0.1 | 32 | 92.54 |
| 1.6 | 0.1 | 16 | 92.5 |
| 0.8 | 0.1 | 8 | 92.49 |
| 0.4 | 0.1 | 4 | 90.33 |
| 0.2 | 0.1 | 2 | 90.09 |
| 0.1 | 0.1 | 1 | 89.97 |

**Table 5**
Test accuracy (%) for different teacher weight ratio ($\rho \leq 1$).

| Weight Coefficient $\alpha_P$ | Weight Coefficient $\alpha_E$ | Teacher Weight Ratio $\rho$ | Global Model Accuracy(%) |
|---|---|---|---|
| 0.1 | 0.1 | 1 | 89.97 |
| 0.1 | 0.2 | 1/2 | 89.89 |
| 0.1 | 0.4 | 1/4 | 89.75 |
| 0.1 | 0.8 | 1/8 | 88.95 |
| 0.1 | 1.6 | 1/16 | 88.92 |
| 0.1 | 3.2 | 1/32 | 88.96 |
| 0.1 | 6.4 | 1/64 | 85.61 |

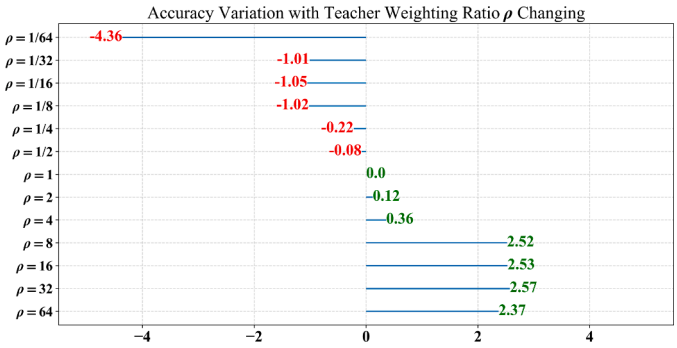


**Fig. 3.** Accuracy variations with teacher weight ratio $\rho$. $\rho = 1$ is the baseline. The global model has the lowest accuracy at $\rho = \frac{1}{64}$ and highest accuracy at $\rho = 32$.

With $\rho = 1$ as the baseline, we calculated accuracy differences between different $\rho$ values and the baseline. As the pretrained teacher model's weight ratio $\rho$ increases, the student model's accuracy improves consistently. This trend continues until $\rho$ reaches 32:1, where the student model achieves peak performance. This shows the pretrained teacher model's importance in transferring robust and generalized knowledge to the student model. When $\rho$ further increases to 64:1, the student model's accuracy decreases slightly. This decline results from the client ensemble teacher model's disproportionately small contribution. The ensemble model's limited influence reduces effective transfer of local knowledge learned by clients, which could otherwise complement the global knowledge from the pretrained teacher.

In contrast, as the client ensemble teacher model's weight ratio increases, the student model's overall accuracy shows a slight downward trend. It reaches the lowest point when $\rho$ is 1:64. These results indicate the client ensemble teacher model represents incomplete knowledge learned by local clients. It serves as a repository for local client knowledge in the federated learning knowledge distillation framework. It mainly strengthens the student model's ability to learn from knowledge shared among clients, but its contribution to improving the student model's performance is slightly inferior to that of the pretrained teacher model.

The results confirm the pretrained teacher model's dominant role in enhancing the student model's performance. As a global knowledge source, the pretrained model's high-quality representations help the student model achieve superior accuracy, especially in scenarios with heterogeneous client data distributions. This also shows proper allocation of teacher model weights is crucial for optimizing the student model's performance.

**Table 6**
Impact of different key modules on model accuracy.

| Key Module/Loss Function | Associated Process | Global Model Accuracy |
|---|---|---|
| Full Training | | 92.54 |
| $-\mathcal{L}_{CEg}$ | Sample Generation | 92.14 |
| $-\mathcal{L}_{BNg}$ | Sample Generation | 84.92 |
| $-\mathcal{L}_{mSARC}$ | Knowledge Distillation | 91.59 |
| −Pretrained Model | Knowledge Distillation | 91.05 |
| −Ensemble Model | Knowledge Distillation | 92.13 |

#### 4.4.2. *Necessity of each component for FedPET*

We conducted comprehensive ablation experiments on CIFAR-10 to verify effectiveness of key modules and loss functions in FedPET. Table 6 presents student model test accuracy after systematically excluding specific components. The analysis clarifies their impact on performance and stability.

Results show removing any key module leads to noticeable accuracy drop. Excluding the mSARC constraint reduces accuracy by over 0.95%, demonstrating its significant role in guiding the student model to mimic the teacher model's spatial activation patterns. Absence of the CFE module causes marked decline in both performance and stability, emphasizing its importance in enhancing generated sample diversity.

The cross-entropy loss $\mathcal{L}_{CEg}$ aligns generated synthetic data with teacher predictions. Its removal reduces accuracy substantially, highlighting its role in ensuring the generator produces meaningful training data. Batch Normalization regularization loss $\mathcal{L}_{BNg}$ aligns synthetic data distribution with real data distribution. It is indispensable as its absence leads to less stable performance and significant accuracy decline, underlining its necessity for maintaining consistency in generator outputs.

Removing the pretrained teacher model causes a sharp accuracy drop, confirming its knowledge serves as a robust foundation for initializing the student model especially in heterogeneous data scenarios. Excluding the ensemble teacher model which aggregates knowledge from all clients slightly decreases global model performance, indicating its utility in capturing diverse client knowledge and balancing local variations.

Notably, the accuracy decrease caused by removing $\mathcal{L}_{CEg}$ is similar to that caused by removing the ensemble model. Our analysis shows that among parameters yielding the best results, the ensemble model's low weight parameter makes its distillation effect on the student model less pronounced, leading to this outcome.

## 5. Conclusion

This paper introduces FedPET, a novel federated data-free knowledge distillation method designed to tackle the challenges of model compression and communication efficiency in federated learning environments. FedPET leverages a multi-teacher knowledge transfer approach based on synthetic data. In each global epoch, a pretrained teacher model on the central server is first used to optimize an image generator, enhancing the diversity of generated images through the TSG module. After aggregating and ensembling local client models, data-free knowledge distillation is performed, utilizing both the pretrained model and the clients' ensemble model as teachers. Extensive experiments demonstrate that FedPET significantly outperforms state-of-the-art methods.

Despite its promising performance, FedPET has several limitations that should be discussed. First, with respect to **modality specificity**, while the dual-teacher framework is conceptually generic, the current TSG module depends on Channel Feature Exchange (CFE), which is specifically tailored for the spatial and channel characteristics of visual data. Extending FedPET to non-visual domains would require adapting the generation strategy. Second, in terms of **computational overhead**, the iterative optimization of the generator and the distillation process on the server incur higher computational costs compared to simple aggregation methods. While FedPET can theoretically extend to larger datasets, server-side training runtime would increase dramatically, resulting in a trade-off for reduced communication bandwidth. Finally, the method's effectiveness depends partially on the **domain relevance** of the pre-trained teacher model. A significant domain gap between the pre-trained model and the client data may limit the effectiveness of knowledge initialization. Future work will focus on addressing these limitations and exploring more efficient generation methods.

## CRediT authorship contribution statement

**Bo Wang:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization; **Yiming Yan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization; **Zhaoning Wang:** Writing – review & editing, Methodology, Formal analysis, Conceptualization; **Jialin Liu:** Writing – review & editing, Methodology, Conceptualization; **Wei Wang:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

## Data availability

Data will be made available on request.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

This work is supported by the Application Fundamental Research Project of Liaoning Province (No. 2025JH2/101330123), and the National Natural Science Foundation of China (No. 62372452).

## References

[1] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210.
[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.
[3] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: a survey, Sci. China Technol. Sci. 63 (10) (2020) 1872–1897.
[4] D. Li, J. Wang, FedMD: Heterogenous federated learning via model distillation, NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality, (2019), 1–8. arXiv:1910.03581
[5] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Syst. 2 (2020) 429–450.
[6] S.P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, A.T. Suresh, SCAFFOLD: stochastic controlled averaging for federated learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 5132–5143.
[7] D.A.E. Acar, Y. Zhao, R.M. Navarro, M. Mattina, P.N. Whatmough, V. Saligrama, Federated learning based on dynamic regularization, International Conference on Learning Representations (ICLR), (2021), 1 – 43. arXiv:2111.04263
[8] Q. Li, B. He, D. Song, Model-contrastive federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10713–10722.
[9] R.G. Lopes, S. Fenu, T. Starner, Data-free knowledge distillation for deep neural networks, NeurIPS Workshop on Learning with Limited Data, (2017), 1–8. arXiv:1710.07535
[10] G.K. Nayak, K.R. Mopuri, V. Shaj, V.B. Radhakrishnan, A. Chakraborty, Zero-shot knowledge distillation in deep networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 4743–4751.
[11] H. Yin, P. Molchanov, J.M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N.K. Jha, J. Kautz, Dreaming to distill: data-free knowledge transfer via deepinversion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8715–8724.
[12] A. Afonin, S.P. Karimireddy, Towards model agnostic federated learning using knowledge distillation, International Conference on Learning Representations (ICLR), (2021), 1–23. arXiv:2110.15210
[13] Z. Zhu, J. Hong, J. Zhou, Data-free knowledge distillation for heterogeneous federated learning, in: International Conference on Machine Learning, PMLR, 2021, pp. 12878–12889.
[14] J. Zhang, C. Chen, B. Li, L. Lyu, S. Wu, S. Ding, C. Shen, C. Wu, DENSE: data-free one-shot federated learning, Adv. Neural Inf. Process. Syst. 35 (2022) 21414–21428.
[15] S. Wang, Y. Fu, X. Li, Y. Lan, M. Gao, et al., DFRD: Data-free robustness distillation for heterogeneous federated learning, Adv. Neural Inf. Process. Syst. 36 (2023) 17854–17866.
[16] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 843–852.
[17] S. Yu, J. Chen, H. Han, S. Jiang, Data-free knowledge distillation via feature exchange and activation region constraint, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24266–24275.
[18] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[19] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, University of Toronto, Toronto, ON, Canada, (2009). https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
[20] K. Luo, X. Li, Y. Lan, M. Gao, GradMA: a gradient-memory-based accelerated federated learning with alleviated catastrophic forgetting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3708–3717.
[21] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, Y. Khazaeni, Bayesian nonparametric federated learning of neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 7252–7261.
[22] H. Wang, M. Yurochkin, Y.K. Sun, D. Papailiopoulos, Y. Khazaeni, Federated learning with matched averaging, newblock International Conference on Learning Representations (ICLR), (2020), 1–16. arXiv:2002.06440
[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[24] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations (ICLR), (2015), 1–14. arXiv:1409.1556
[25] S. Zagoruyko, N. Komodakis, Wide residual networks, Proceedings of the British Machine Vision Conference (BMVC), (2016), 1–15. arXiv:1605.07146