electronics

MDPI

Article

FedSKF: Selective Knowledge Fusion via Optimal Transport in Federated Class Incremental Learning

Minghui Zhou * and Xiangfeng Wang

School of Computer Science and Technology, East China Normal University, Shanghai 200062, China; xfwang@cs.ecnu.edu.cn
* Correspondence: 51215901041@stu.ecnu.edu.cn

Abstract: Federated learning has been a hot topic in the field of artificial intelligence in recent years due to its distributed nature and emphasis on privacy protection. To better align with real-world scenarios, federated class incremental learning (FCIL) has emerged as a new research trend, but it faces challenges such as heterogeneous data, catastrophic forgetting, and inter-client interference. However, most existing methods enhance model performance at the expense of privacy, such as uploading prototypes or samples, which violates the basic principle of only transmitting models in federated learning. This paper presents a novel selective knowledge fusion (FedSKF) model to address data heterogeneity and inter-client interference without sacrificing any privacy. Specifically, this paper introduces a PIT (projection in turn) module on the server side to indirectly recover client data distribution information through optimal transport. Subsequently, to reduce inter-client interference, knowledge of the global model is selectively absorbed via knowledge distillation and an incomplete synchronization classifier at the client side, namely an SKS (selective knowledge synchronization) module. Furthermore, to mitigate global catastrophic forgetting, a global forgetting loss is proposed to distill knowledge from the old global model. Our framework can easily integrate various CIL methods, allowing it to adapt to application scenarios with varying privacy requirements. We conducted extensive experiments on CIFAR100 and Tiny-ImageNet datasets, and the performance of our method surpasses existing works.

Keywords: federated learning; artificial intelligence; knowledge distillation; deep learning

Citation: Zhou, M.; Wang, X. FedSKF: Selective Knowledge Fusion via Optimal Transport in Federated Class Incremental Learning. *Electronics* **2024**, *13*, 1772. https://doi.org/10.3390/electronics13091772

Academic Editors: Najlae Idrissi, Yassine Sadqi, Abdul Wahid and Gurjot Singh Gaba

Received: 22 March 2024
Revised: 30 April 2024
Accepted: 2 May 2024
Published: 4 May 2024

1. Introduction

Federated learning (FL) has become popular in the field of artificial intelligence due to the growing focus on privacy protection, and it allows multiple distributed devices or data centers to cooperatively train a robust global model without transmitting samples [1]. The standard FL typically assumes a fixed class set for all clients, which does not align with real-world scenarios. For instance, when a new virus emerges and spreads globally, numerous medical data centers participating in federated learning will update the class set. Recognizing the need to adapt to such changes, there is a universal call for the integration of continual learning into federated learning, which is referred to as federated continual learning (FCL). As a subdivision of FCL, federated class incremental learning (FCIL) has gained increasing attention in recent years, resulting in a significant amount of excellent work [2–7]. FCIL presents a more formidable challenge as it must simultaneously address three key issues: catastrophic forgetting, data heterogeneity, and inter-client interference [5]. Moreover, FCIL also involves more complex privacy protection considerations [7].

Catastrophic forgetting is a challenge inherent in class incremental learning (CIL) [8], where the model tends to forget the knowledge of the old classes when learning new classes, leading to a significant decline in overall performance [9]. Over the years of exploration, techniques in CIL such as knowledge distillation [10], rehearsal [10,11], and class prototypes [12] have been proven to be effective in mitigating catastrophic forgetting.

In FCIL, local catastrophic forgetting is propagated to the global model through aggregation, leading to global catastrophic forgetting. On the basis of using local old model knowledge distillation to weaken local catastrophic forgetting, Ref. [13] introduces global model knowledge distillation as a strategy to alleviate global catastrophic forgetting. Ref. [14] uploads a small number of samples to the server in order to create a compact global dataset, while effective, this approach clearly compromises privacy. Ref. [4] proposed a compromise solution by uploading perturbed class prototypes to the proxy server to build additional global knowledge. Simultaneously, the clients utilize the real data rehearsal method to mitigate forgetting. This approach achieves excellent performance but comes with increased communication costs and longer training times. A more interesting solution is generative data replay [7,15], which synthesizes global dataset through GAN [16] models to avoid uploading additional information about samples. Ref. [7] pre-trains the global model on the synthesized global data to transfer knowledge between tasks. Unfortunately, the performance of this method is notably poor.

Data heterogeneity is an inherent problem in FL, stemming from factors such as non-independent and identical data distributions (non-IID), missing classes, and different data sources among clients [17]. Due to the non-IID nature of client datasets, local models initialized with the same parameters will converge into divergent models [3]. This divergence slows down the convergence speed and diminishes the performance of the global model. Furthermore, in FCIL, the classes of each task among different clients may not align perfectly, which leads to varying local models across clients. As these local models are aggregated, they interfere with each other, resulting in bad performance of the global model. This phenomenon is known as inter-client interference [5].

In FCIL, common methods to address data heterogeneity and inter-client interference include global dataset [2,4,6,7] and parameter decomposition [5,18]. Creating a global dataset or knowledge base can not only alleviate forgetting but is also beneficial for addressing data heterogeneity. The global dataset or knowledge base is typically class-balanced and contains the most representative class knowledge. Ref. [4] employs global prototypes to select the best old model. Ref. [6] pre-trains the global model on global fractal images. However, these approaches increase the communication burden. Ref. [2] constructs a global dataset from unlabeled public datasets and uses it to distill knowledge. However, it requires additional prior knowledge and may not be ideal for all types of data. Parameter decomposition [5,18] is another feasible and more secure approach to solve the above problems. In FedWeIT [5], network weights are divided into global federated parameters and sparse task-specific parameters. Each client receives selective knowledge from other clients by re-weighting its task-specific parameters. This method not only effectively mitigates inter-client interference but also addresses catastrophic forgetting. It is important to notice that the setting of FedWeIT does not fully adhere to traditional continuous learning. Specifically, there are overlapping tasks in the FedWeIT setting. Old tasks may appear in other clients during subsequent training, which reduces the challenge of catastrophic forgetting. In this paper, we concentrate on a *non-overlapping* tasks scenario, which presents greater challenges in terms of catastrophic forgetting and closely aligns with the original setting of CL.

The privacy challenge in FCIL is more intricate than in traditional federated learning. This complexity arises from the need to not only safeguard privacy beyond the client but also carefully consider whether CIL methods within the client might compromise privacy [7]. From the perspective of FL, privacy protection is its fundamental objective, and this principle should also hold true for FCIL. Therefore, we advocate for privacy protection in FCIL to be as stringent as in FL, which means that only the model should be transmitted, without any additional information about data, such as class prototypes or samples. However, certain existing approaches have made compromises on privacy in order to enhance model performance [4,6,14]. While these compromises might be viable under relaxed privacy constraints, they are not worthy of recommendation. From the perspective of CIL, privacy protection is not always imperative and varies depending on

the application scenarios. For instance, real data rehearsal involves storing small batches of real samples for extended periods, which may not be acceptable in privacy-sensitive domains like the medical field [7,15]. In order to deal with this problem, we propose the exploration of a more general and adaptable algorithmic framework that can accommodate different levels of privacy requirements and seamlessly integrate with various CL methods.

Recalling the original intention of federated learning [19], the exchange of models without global datasets is crucial for maximizing privacy security. Exploring ways to fully leverage the model to mitigate both global catastrophic forgetting and data heterogeneity remains a valuable and ongoing research challenge. Model fusion via optimal transport (OT), which is based on mathematics and fuses models without accessing training data, has captured our attention. As far as we know, there has not been any work about OT in FCIL. Yet, there is already some related work in FL. Refs. [20–22] successfully achieve one-time knowledge transfer between non-IID neural networks using optimal transport. However, it is worth noting that while optimal transport has shown promise in addressing data heterogeneity in FL, it may not effectively tackle the challenges of class missing and global catastrophic forgetting in FCIL scenarios. We noticed that, in recent years, pre-trained transformer [23] has gained substantial of attention in the computer vision field and has achieved state-of-the-art (SOTA) performance on public datasets in FCIL [24]. However, transformers need substantial training data, which easily leads to over-fitting when dealing with small datasets [25]. Additionally, their interpretability is poor, which limits application fields. On the contrary, while the performance of convolutional neural networks may be slightly lower than pre-trained ViT on public datasets, they still hold broader applicability in real-world scenarios.

Based on the aforementioned analysis, we consider the following question: *How can three key issues—catastrophic forgetting, data heterogeneity, and inter-client interference—be effectively mitigated in FCIL without transmitting any additional data except the model?*

In this paper, we propose a method called FedSKF (**s**elective **k**nowledge **f**usion via optimal transport in **fed**erated class incremental learning) within a non-overlapping tasks setting, which effectively addresses three key issues in FCIL and can flexibly integrate CL methods to deal with different privacy protection scenarios. Specifically, we introduce a selective knowledge fusion mechanism to mitigate inter-client interference and data heterogeneity. This method includes two modules: the selective knowledge synchronization (SKS) module at the client and the projection in turn (PIT) module at the server. The SKS module selectively assimilates knowledge through personalized classifiers and global classification consistency loss. the PIT module sequentially projects the average model into the feature space of each local model via optimal transport to fuse heterogeneous knowledge. This process enhances the performance of the global model and facilitates model convergence. To alleviate global catastrophic forgetting, we propose global forgetting loss, which treats the old global model as a secondary distillation teacher. These approaches operate directly on the model or its logit output without depending on specific CIL methods, thereby ensuring the flexibility of our framework. Throughout the process, only the model parameters are transmitted, maintaining a communication burden consistent with traditional FL. Overall, our main contributions are as follows:

- A selective knowledge fusion mechanism with global classification consistency loss is proposed in this paper which can effectively alleviate data heterogeneity and inter-client interference in FCIL.
- A novel model aggregation strategy called PIT (projection in turn) is proposed to improve the performance of the global model. This paper is the first to introduce optimal transport into FCIL in order to mitigate data heterogeneity.
- Global forgetting loss is designed to reduce global catastrophic forgetting, which makes use of the potential knowledge in the old global model.
- A universal and flexible FCIL framework is proposed in this paper which can integrate various CIL methods to cater to different application scenarios with varying privacy

levels. Furthermore, we conducted numerous experiments on CIFAR100 and Tiny-ImageNet datasets.

2. Formulation

Before formally defining the FCIL problem, let us revisit the setting of CIL from the client's perspective. Traditionally, a client learns T tasks $\{\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, ..., \mathcal{T}^{(T)}\}$ in sequence, where $\mathcal{T}^{(t)}$ represents the t-th task and is trained on a labeled dataset $\mathcal{D}^{(t)} = \{x_i^{(t)}, y_i^{(t)}\}_{i=1}^{N_t}$, where N_t is the size of dataset $\mathcal{D}^{(t)}$. The entire class set is denoted as $\mathcal{C}$, and $\mathcal{T}^{(t)}$ contains non-overlapping subsets of classes $\mathcal{C}^{(t)} \in \mathcal{C}$, which means $\mathcal{C}^{(a)} \cap \mathcal{C}^{(b)} = \varnothing$ for any $a \neq b$. During the t-th task training phase, the data from previous tasks are unavailable. In rehearsal-based methods, a small number of representative instances are stored as exemplar set $\mathcal{E}$; accordingly, $\mathcal{D}^{(t)}$ is updated to $\mathcal{D}^{(t)} \cup \mathcal{E}$. The goal of CIL is to fix a model $\boldsymbol{\theta}^{(t)}$, which not only acquires knowledge from the current task $\mathcal{T}^{(t)}$ but also retains knowledge from historical tasks $\{\mathcal{T}^{(1)}, ..., \mathcal{T}^{(t-1)}\}$.

Now, let us extend CIL to the scene of federal class-incremental learning, which involves a global server and multiple clients. We denote the number of clients as K. Each client $c_i \in \{c_1, ..., c_K\}$ trains the t-th task $\mathcal{T}_i^{(t)}$ on local dataset $\mathcal{D}_i^{(t)} \in \mathcal{D}^{(t)}$. For any two clients c_a and c_b, at the t-th task, $\mathcal{D}_a^{(t)} \cap \mathcal{D}_b^{(t)} = \varnothing$. It is worth mentioning that for each class, the data from different clients are assumed to be non-IID [3]. After local training, each client uploads its model $\boldsymbol{\theta}_i^{(t)}$ to the server $\mathcal{S}$, which then aggregates them into a global model $\boldsymbol{\theta}_g^{(t)}$. None of the clients or servers are allowed to obtain raw data from each other. The ultimate objective of FCIL is to aggregate a robust global model $\boldsymbol{\theta}_g$, which grasps all the knowledge of all clients, including previous knowledge. Formally, the objective function is defined as:

$$\boldsymbol{\theta}_g^* = \arg\min_{\boldsymbol{\theta}_g} \sum_{k=1}^{K} \sum_{t=1}^{T} \mathcal{L}(\mathcal{T}_k^{(t)}; \boldsymbol{\theta}_k^{(t)}; \boldsymbol{\theta}_g), \tag{1}$$

where $\mathcal{L}$ is the loss function used to measure classification error.

3. Methodology

The overview of our method is depicted in Figure 1. Our framework achieves selective knowledge fusion through two modules: selective knowledge synchronization (SKS) and projection in turn (PIT). In the SKS module, each local model is segmented into a general feature extractor and a personalized classifier. In each communication round, the parameters of the global feature extractor overwrite those of the local feature extractor. However, the classifier parameters are not directly synchronized. Instead, they are constrained by global classification consistency loss, which guides local classifiers to selectively assimilate knowledge from the global model. In the PIT module, we use optimal transport to fuse models on the basis of FedAvg. Specifically, PIT selects a model from among local models as the base model. Then, it projects the average of the local models into the feature space of the base model, effectively projecting the abstract global knowledge into a specific feature space. To ensure model convergence, PIT iteratively selects each client model as the base model. Additionally, to address catastrophic forgetting, we employ three forgetting losses to distill knowledge from the old model. The local forgetting loss and task-semantic distillation loss both distill knowledge from the old local model, working together to mitigate local catastrophic forgetting. Thanks to the personalized classifier in SKS, there exist knowledge discrepancies between the old global model and the old local model. Therefore, we propose global forgetting loss, which helps alleviate global catastrophic forgetting by distilling knowledge from the old global model. Each module is described in detail below.

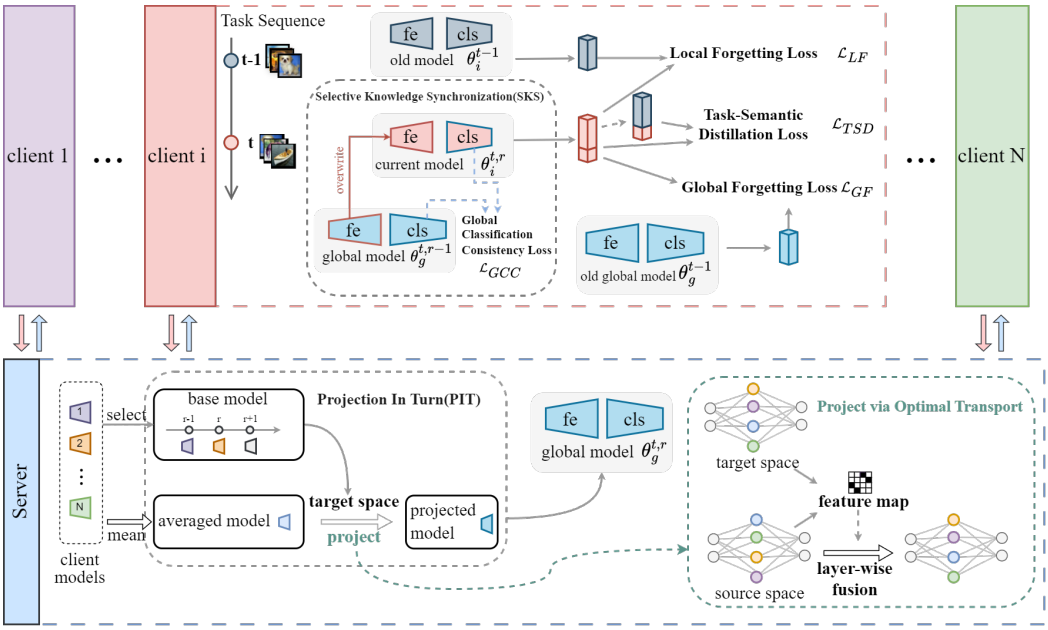


Figure 1. The presentation of our proposed FedSKF model. Each client selectively absorbs knowledge from the global model through the SKS module and distills knowledge from both the old local model and the old global model using three distinct loss functions. After local training, each client sends the local model to the server. The server then performs model fusion through the PIT module to generate a new global model, which is subsequently sent back to each client. The SKS module and PIT module work together to complete the selective knowledge fusion mechanism (Note: 'fe' is short for feature extractor, and 'cls' is short for classifier.).

3.1. Server Side: Projection in Turn via Optimal Transport

Optimal transport (OT) [26] is utilized for one-time knowledge transfer in FL. It treats the parameters of two models as two feature spaces and computes a feature map between them, thereby projecting the models from the source space to the target space, i.e., achieving the soft alignment of neural networks. Several works [20–22] have demonstrated the effectiveness of OT in knowledge fusion and mitigating data heterogeneity.

In contrast to existing works, we utilize OT not simply for a one-time knowledge transfer from local models to the global model, but rather to leverage its knowledge fusion capability for model projection in the reverse direction. Our proposed method transfers knowledge from the global model to a local model. Data heterogeneity leads to model heterogeneity in FL [3]. While this is typically viewed as a challenge, it also contains hidden information: model parameters indirectly encode client data distribution information. In FedAvg [19], the averaging operation quickly imbues the global model with certain global distribution insights. However, relying solely on the averaging model to access global knowledge makes it challenging for clients to accurately learn the data distribution of other clients, which hinders information transmission efficiency and accuracy. Based on the aforementioned analysis, we speculate that enriching the global model with more accurate data distribution information will improve its performance. Therefore, we employ OT to re-project the averaged model space back into the original client space, thus indirectly recovering the client's original data distribution information. The experimental results confirm the effectiveness of this idea.

We design a module named PIT (projection in turn) for model aggregation. In each aggregation step, PIT selects a model from the numerous client models as the base model and computes the average model. Subsequently, we follow the approach outlined in [21] to leverage OT for computing the feature map and performing layer-wise model fusion. In this process, the average model is projected into the feature space of the base model, leading to the creation of a new global model. For fairness and model convergence, we select the model of each client as the base model in turn, which ensures that the knowledge

from each client is evenly spread among all clients. Existing work [21] executes OT between the global model and each client model individually. This results in OT running k times, leading to high computational complexity. In contrast, our proposed method executes OT only once in each communication round, significantly reducing computational complexity and enhancing the scalability.

Let $\mathrm{OT}(\cdot)$ denote the model fusion function and $\boldsymbol{\theta}_{base}^{(t,r)}$ denote the base model in the r-th aggregation step for the t-th task. Then, the global model can be computed using the following formula:

$$\boldsymbol{\theta}_{g}^{(t,r)} = OT(\boldsymbol{\theta}_{avg}^{(t,r)}, \boldsymbol{\theta}_{base}^{(t,r)}), \tag{2}$$

where $\boldsymbol{\theta}_{avg}^{(t,r)} = \frac{1}{K}\sum_{i=1}^{K}\boldsymbol{\theta}_{i}^{(t,r-1)}$. Then, the data distribution knowledge of client i will be transmitted to all clients through the global model. Each client selectively utilizes the knowledge of the global model via the SKS module.

3.2. Client Side: Selective Knowledge Synchronization

In traditional centralized deep learning, the model can theoretically capture all features of the entire dataset, which represents the upper limit of performance in FL [19]. In the FCIL setting, even though the client class set may not perfectly align, the idealized upper limit still aims to extract all features across clients. To address inter-client interference, selectively absorbing knowledge from the global model is a natural approach. Differing from existing works that allocate task-specific parameters for each task [5], we propose concentrating inter-client differences within the classifier. We allow the neural network to autonomously determine what knowledge to learn through a loose consistency constraint. Additionally, we encourage clients to collaboratively train a robust feature extractor.

We design the SKS module to realize the above idea. Specifically, we segment the client model into a feature extractor and a classifier, each with distinct updating strategies. For the feature extractor, we apply the conventional FL process, during which global feature extractor parameters overwrite local parameters. For local classifiers, parameter alignment occurs only once in the training process of a task, that is, during local model updating after the first communication. Throughout subsequent training processes of this task, the local classifier no longer maintains the same parameters as the global classifier, but selectively absorbs knowledge via a global classification consistency (GCC) loss, which controls the distance between the local and global classifier. For the neural network $\boldsymbol{\theta}$, its number of layers is denoted as $|\boldsymbol{\theta}|$. Furthermore, its parameter set is denoted as $\mathcal{P}(\boldsymbol{\theta}) = \{W_i, B_i\}_{i\in\{1,\ldots,|\theta|\}}$, which uniformly represents the parameters of all layers in the $\boldsymbol{\theta}$, including weights and bias. Although the parameters differ, the global classifier shares the same structure as the local classifier, ensuring consistency in their parameter sets. Then, the calculation formula for the GCC loss can be expressed as follows:

$$\mathcal{L}_{GCC}^{(t,r)} = \left\|\boldsymbol{\theta}_{local\_cls}^{(t,r)} - \boldsymbol{\theta}_{global\_cls}^{(t,r)}\right\|^2 = \sum_{p\in\mathcal{P}_g}\left\|\mathcal{P}_g[p] - \mathcal{P}_c[p]\right\|^2, \tag{3}$$

where $\boldsymbol{\theta}_{local\_cls}^{(t,r)}$ and $\boldsymbol{\theta}_{global\_cls}^{(t,r)}$ represent the classifiers of the local model and global model, respectively, $\mathcal{P}_g = \mathcal{P}(\boldsymbol{\theta}_{global\_cls}^{(t,r)})$, and $\mathcal{P}_c = \mathcal{P}(\boldsymbol{\theta}_{local\_cls}^{(t,r)})$.

In other words, this paper introduces a personalized classifier. However, unlike the personalization layer in FL [27,28], our goal is not to directly enhance client model performance (which is not guaranteed here), but rather to alleviate inter-client interference. Furthermore, unlike typical personalization layers in FL that often do not contribute to model aggregation, our personalized classifier actively participates in each round of model aggregation without being overwritten by the global model. The form of GCC may appear similar to the proximal term of FedProx [29], but they are actually quite distinct from each other. In FedProx, the local model aims to be consistent with the global model, with the proximal term restricting updates to the local model. In contrast, at the beginning of local

training in our proposed method, the local classifier is distant from the global model, and GCC is intended to guide the local model towards convergence with the global model. Furthermore, unlike FedProx, which imposes constraints upon the entire model, we allow the feature extractor to update freely and only apply consistency constraints to the classifier.

By refraining from forcibly overwriting the classifier parameters, our approach can effectively mitigate the impact of data heterogeneity and inter-client interference during training.

3.3. Objective Function

The effectiveness of knowledge distillation (KD) [30] in mitigating catastrophic forgetting has been demonstrated in numerous studies [2,4,11]. It is often used together with the rehearsal-based method in CIL. In this paper, KD serves as the primary approach to mitigate forgetting, and we introduce multiple distillation losses to extract multidimensional historical knowledge. The utilization of samples rehearsal is contingent upon the privacy demands of the application scenario. When privacy and storage permissions allow, we endorse this approach, as it significantly boosts model effectiveness. Desensitized prototypes can substitute for samples in scenarios with stringent privacy requirements. For privacy-sensitive settings, we solely rely on knowledge distillation to reinforce the retention of old knowledge. The loss mentioned in this subsection will be calculated in exactly the same way at each client, so the subscript representing the client's serial number is omitted in the following loss function.

The training data of the t-th task is denoted as $\mathcal{D}^{(t)}$. Each client trains the local model $\boldsymbol{\theta}^{(t,r)}$ at the r-th round of communication for task t on $\mathcal{D}^{(t)}$. The classification loss is denoted as $\mathcal{L}_{cls}^{(t,r)}$:

$$\mathcal{L}_{cls}^{(t,r)} = CE(\mathcal{D}^{(t)}; \boldsymbol{\theta}^{(t,r)}), \tag{4}$$

where $CE(\cdot)$ is the cross-entropy loss function. We denote the data stored for the previous task as $\mathcal{M}$, such as exemplars or prototypes. We uniformly express the sum of the loss on the $\mathcal{M}$ and the classification loss on the $\mathcal{D}^{(t)}$ as $\mathcal{L}_{data}^{(t,r)}$, which represents the knowledge gained directly or indirectly from the samples:

$$\mathcal{L}_{data}^{(t,t)} = \mathcal{L}_{m}^{(t,r)}(\mathcal{M}; \boldsymbol{\theta}^{(t,r)}; \boldsymbol{\theta}^{(t-1)}) + \mathcal{L}_{cls}^{(t,r)}, \tag{5}$$

where $\mathcal{L}_{m}^{(t,r)}(\cdot)$ is the supplementary loss of memory $\mathcal{M}$.

It does not matter which method is used, real samples or generated samples or prototypes; all of them can adapt to our framework by simply putting the corresponding loss in $\mathcal{L}_{m}^{(t,r)}$.

Local Catastrophic Forgetting. In order to alleviate local catastrophic forgetting, we distill knowledge from the old local model $\boldsymbol{\theta}^{(t-1)}$ at the logits-level by minimizing the local forgetting (LF) loss:

$$\mathcal{L}_{LF}^{(t,r)} = -\sum_{i=1}^{\left|\mathcal{C}^{(1:t-1)}\right|} \sigma(\hat{z}_i/\pi)(log\sigma(z_i/\pi)), \tag{6}$$

where $\hat{z}$ is the logit output of the old local model $\boldsymbol{\theta}^{(t-1)}$, z is the output logit of the current model $\boldsymbol{\theta}^{(t,r)}$, σ is a softmax function, π is the distillation temperature, and $\left|\mathcal{C}^{(1:t-1)}\right|$ is the total number of classes for tasks 1 to $t-1$. Moreover, the logits output of $\boldsymbol{\theta}^{(t,r)}$ indicates the semantic similarity relations of classes [4], providing insights into the class-level relationship between old and new tasks. Inspired by the work of [4], we introduce task-semantic distillation (TSD) loss to absorb more knowledge from the old local model $\boldsymbol{\theta}^{(t-1)}$. For a batch dataset $\{X_b^{(t)}, Y_b^{(t)}\} \in \mathcal{D}^{(t)}$, we denote the logit output of the current model $\boldsymbol{\theta}^{(t,r)}$ as $Z(X_b^{(t)}, \boldsymbol{\theta}^{(t,r)})$. Similarly, we obtain the corresponding logit output $Z(X_b^{(t)}, \boldsymbol{\theta}^{(t-1)})$ of the old local model $\boldsymbol{\theta}^{(t-1)}$. As shown in Figure 1, we recombine two logit vectors by retaining

the value of $Z(X_b^{(t)}, \boldsymbol{\theta}^{(t-1)})$ for old classes $\mathcal{C}^{(1:t-1)} = \{\mathcal{C}^{(i)}\}_{i=1}^{t-1}$ and using the output of $Z(X_b^{(t)}, \boldsymbol{\theta}^{(t,r)})$ for new classes $\mathcal{C}^t$, resulting in a new logical vector $\mathbb{Z}_b^{(t)}$:

$$\mathbb{Z}_b^{(t)} = \begin{cases} Z(X_b^{(t)}, \boldsymbol{\theta}^{(t-1,r)})_i, & i \in \mathcal{C}^{(1:t-1)} \\ Z(X_b^{(t)}, \boldsymbol{\theta}^{(t,r)})_i, & i \in \mathcal{C}^t \end{cases}. \tag{7}$$

Then, for $Y_b^{(t)}$, we calculate its one-hot encoding labels $\mathbb{Y}_b^{(t)}$. The TSD loss can be calculated by the following formula:

$$\mathcal{L}_{TSD}^{(t,r)} = \mathcal{H}(\mathbb{Z}_b^{(t)}, \mathbb{Y}_b^{(t)}), \tag{8}$$

where $\mathcal{H}(\cdot)$ is a metric function, which is set as binary cross-entropy in this paper. LF loss $\mathcal{L}_{LF}^{(t,r)}$ and TSD loss $\mathcal{L}_{TSD}^{(t,r)}$ jointly complete the knowledge extraction from the old local model $\boldsymbol{\theta}^{(t-1)}$.

Global Catastrophic Forgetting. Inspired by the distillation of local models, we recognize that old knowledge may be embedded within the old global model $\boldsymbol{\theta}_g^{(t-1)}$. With our hierarchical architecture, the classifiers of the local old model $\boldsymbol{\theta}_{local\_cls}^{(t-1)}$ and the global old model $\boldsymbol{\theta}_{global\_cls}^{(t-1)}$ are independent of each other, implying that the classifier of the global model holds the global old knowledge. Therefore, we follow the example of local distillation and define the global forgetting loss as

$$\mathcal{L}_{GF}^{(t,r)} = -\sum_{i=1}^{|\mathcal{C}^{(1:t-1)}|} \sigma(\tilde{z}_i/\pi)(log\sigma(z_i/\pi)), \tag{9}$$

where $\tilde{z}$ is the output logit of the old model $\boldsymbol{\theta}_g^{(t-1)}$.

Thus, we obtain the complete loss function for the client as follows:

$$\mathcal{L}^{(t,r)} = \mathcal{L}_{data}^{(t,r)} + \alpha \cdot \mathcal{L}_{GCC}^{(t,r)} + \beta \cdot (\mathcal{L}_{LF}^{(t,r)} + \mathcal{L}_{TSD}^{(t,r)}) + \gamma \cdot \mathcal{L}_{GF}^{(t,r)}, \tag{10}$$

where α is the hyperparameter to ensure global classification consistency, and β and γ are used to control the distillation weight of the old local model and the old global model, respectively.

4. Experiments Results

4.1. Dataset

We utilize the two datasets, namely CIFAR100 [31] and Tiny-ImageNet [32], to evaluate the performance of our method in two data partition settings. The CIFAR100 dataset is sourced from the "80 Million Tiny Images" collection, containing 100 classes. Each class consists of 600 color images with dimensions of 32×32 pixels, of which 500 images per class are allocated for training and 100 for testing. The Tiny-ImageNet contains 200 classes, and each class has 500 training images and 50 test images. Specifically, we divided the CIFAR100 dataset into two data partition settings, consistent with [7]. The first setting involves 10-task continuous learning, with each task containing 10 classes, which is denoted as CIFAR100_10_10. The second setting consists of 5-task continuous learning, with 20 classes per task, which is denoted as CIFAR100_5_20. Similarly, Tiny-ImageNet is divided into 5 tasks, each with 40 classes.

To simulate a non-IID setting, we utilize the Dirichlet distribution, a commonly employed method in FCIL [3,7]. We adjust the hyperparameter λ to control the degree of non-IID. When $\lambda > 0$, the smaller the λ, the higher the degree of data heterogeneity. The exception is that when $\lambda = 0$, it means that the data between clients are independent and identically distributed, i.e., IID. It is noteworthy that when $\lambda > 0$, each client may include majority, minority, or missing classes. As an example, Figure 2 shows the data statistics results of the first task on the CIFAR100_10_10 setting with $\lambda = 0.5$. It can be seen that

the data distribution among different clients is irregular and exhibits significant variations, which is consistent with the real world.

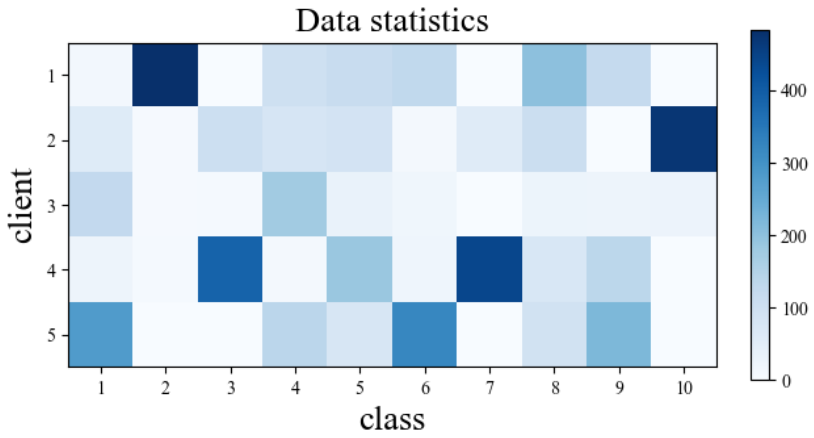


Figure 2. The data statistics results of the first task within the CIFAR100_10_10 setting with $\lambda = 0.5$.

4.2. Implementation Details

We implemented all experiments using PyTorch and ran them on a 40GB NVIDIA PCIE A100 GPU. Following the settings in [7], we set the number of clients to 5, with each task containing 100 rounds of communication, and the clients trained 5 epochs locally.

We ran the experiment in both IID ($\lambda = 0$) and non-IID ($\lambda = 0.5$, abbreviated as NIID) scenarios. We trained models using an SGD optimizer with the learning rate to 0.05. The distillation temperature was set to $\pi = 2$, and the weight of the distillation loss from the old local model was set to $\beta = 3$, as commonly used in other works [7,10,11]. By using grid search, our proposed method achieves the best performance with the hyperparameters $\alpha = 2$ and $\gamma = 1$. Additionally, we implemented three client-side privacy scenarios, and each scenario had its own baselines:

- **Privacy—sensitive** scenarios, in which only old models can be stored. We selected five methods as baselines: (1) FedLwf, the integration of CL classical algorithm LWF [10] and the FL classical algorithm FedAvg [19]; (2) FedLwf, the integration of LWF and another classical FL algorithm, FedProx [29]; (3) FedLwf2T [13], a simple method based on FedLwf in FCIL that uses the global model as the secondary teacher; (4) FedWeIT [5], a regularization-based method in FCIL that maximizes the knowledge transfer between clients; (5) TARGET [7], a generative-based method in FCIL that synthesizes global samples through a GAN model and distills knowledge from old local models without exemplar.
- **Privacy—moderate** scenarios, which allow prototypes to be stored without any concerns. We selected the integration of an excellent prototype-based method PASS [12] and FedAvg, i.e., FedPASS, as a baseline.
- **Privacy—relaxed** scenarios, in which real samples can be stored. We selected FedICaRL (the integration of iCaRL [11] and FedAvg) as a baseline.

In order to make a fair comparison, we employed ResNet18 as the backbone for the CIFAR100 dataset, with a classifier consisting of a single, fully connected layer. Moreover, we further utilized VGG16 [33] with three fully connected layers to conduct experiments on the Tiny-ImageNet dataset. However, only the last fully connected layer is considered as the classifier in our proposed method. Specifically, for rehearsal-based methods, we set the exemplar size of each class to 20, following the classic work [4,11]. We ran our experiments with three random seeds {2021, 2022, 2023} and reported the averaged results.

4.3. Evaluation Metric

In this paper, the average accuracy $\bar{A}$ and the average forgetting $\bar{F}$ are utilized to evaluate the model's performance, which are the commonly used evaluation metrics for CL [9,12] and FCIL [4,18,24]. For the t-th task of the client, its accuracy on the previous task i is $a_{i,t}$, then the accuracy of the current task $A^t = \frac{1}{t}\sum_{i=1}^{t} a_{i,t}$. Thus, the average accuracy of T tasks can be calculated by the following formula:

$$\bar{A} = \frac{1}{T}\sum_{t=1}^{T} A^t. \tag{11}$$

The average forgetting $\bar{F}$ is used to measure the forgetting degree of the model. A lower $\bar{F}$ indicates better model performance. The calculation formula is as follows:

$$\bar{F} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{t}\sum_{i=1}^{t-1}(a_{i,t-1} - a_{i,t}). \tag{12}$$

4.4. Results and Analysis

To evaluate the universality of our proposed method, we separately integrated LWF [10], PASS [12], and iCaRL [11] into FedSKF, denoted as ours(Lwf), ours(PASS), and ours(iCaRL), to address three privacy scenarios, respectively. Furthermore, we compared them with baselines in diverse settings. Additionally, we present the results of ablation experiments to analyze hyperparameters and the function of each module in FedSKF.

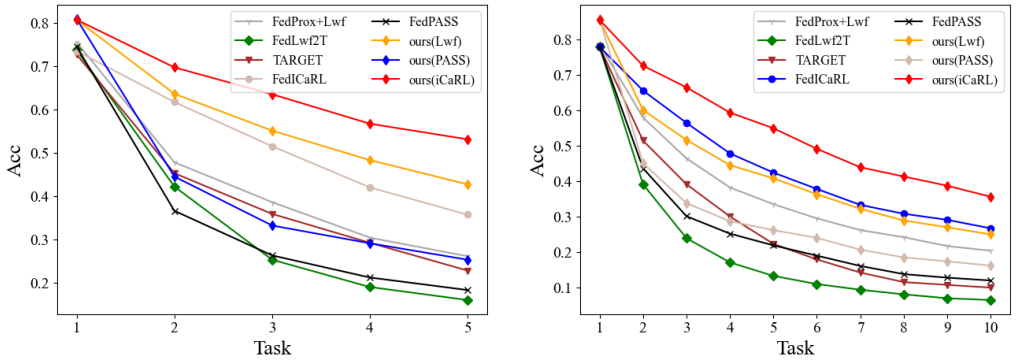
Experiments on CIFAR-100: We conducted numerous experiments on two data partition settings, namely CIFAR100_10_10 and CIFAR100_5_20, considering both IID and non-IID settings. The results of comparative experiments are presented in Table 1. On the one hand, in a vertical comparison of average accuracy and average forgetting, it can be seen that our method is obviously superior to the baselines in the same scenario. It is worth mentioning that the comparison across privacy scenarios may not ensure fairness. For example, when exemplar is permitted, FedICaRL exhibits competitive performance, surpassing the SOTA method in privacy—sensitive scenarios. On the other hand, when comparing different settings horizontally, it becomes evident that as the degree of data heterogeneity or the number of tasks increases, the model's performance decreases. However, our method exhibits the smallest decline in performance. Specifically, in the $T = 10$ setting, when λ increases from 0 to 0.5, the average accuracy of our method decreases by only 1.43%, whereas the second-best method, FedICaRL, experiences a 3.65% reduction. This indicates that our method effectively mitigates the challenges posed by data heterogeneity. Furthermore, let us go deep into each privacy scenarios. In privacy—sensitive settings, our method outperforms existing methods by 5.52% to 15.8%. It is important to note that for three models, FedLwf, FedWeIT, and TARGET, we reference the experimental results from the literature [7]. Since we maintain consistent implementation details, the comparison is still fair. Similarly, compared to FedPASS, PASS integrated in FedSKF shows better performance. Moreover, the same enhancement in performance is also observed in privacy-relaxed scenarios, even though FedICaRL is already the second-best method. These results highlight the universality and effectiveness of FedSKF in integrating existing continuous learning methods.

A more intuitive test accuracy comparison curve is shown in Figure 3, which depicts the accuracy of different models after completing each task in the non-IID setting. For instance, for task 3, the accuracy value represents the average test accuracy of the classes belonging to the first three tasks. Despite varying task numbers, our method consistently achieves the highest performance, demonstrating its strong resistance to catastrophic forgetting. Figure 4 depicts the fluctuation in average accuracy across different methods as data heterogeneity shifts in the CIFAR100_5_20 setting. The Dirichlet parameter γ is set to $\{0, 0.1, 0.5, 1.0\}$. As λ decreases from 1 to 0.1, our proposed method demonstrates notably

higher accuracy and less performance degradation compared to existing methods across three privacy scenarios.

Table 1. Comparative experiment results on CIFAR100 within both IID and non-IID settings. $\bar{A}$ refers to the average accuracy (%) and $\bar{F}$ to the average forgetting (%). The higher the value of $\bar{A}$, the lower the value of $\bar{F}$, indicating better model performance. The best results are in bold.

Data Heterogeneity		**IID ($\lambda = 0$)**				**NIID ($\lambda = 0.5$)**			
Tasks		**T = 5**		**T = 10**		**T = 5**		**T = 10**	
Scenario	**Method**	$\bar{A}$	$\bar{F}$	$\bar{A}$	$\bar{F}$	$\bar{A}$	$\bar{F}$	$\bar{A}$	$\bar{F}$
Privacy-sensitive	FedLwf	30.61	45	23.27	37	27.59	44	17.98	45
	FedProx+Lwf	46.12	22	38.71	19	43.78	22	37.65	19
	FedLwf2T	37.25	24	25.43	**16**	35.34	24	21.32	**16**
	FedWeIT	28.45	52	20.39	43	24.57	54	15.45	48
	TARGET	36.31	22	24.76	26	33.33	27	20.71	29
Privacy-moderate	FedPASS	39.28	20	31.52	20	35.48	21	27.23	21
Privacy-relaxed	FedICaRL	58.72	22	48.41	25	52.98	21	44.76	20
Ours	ours(Lwf)	61.15	20	48.06	21	58.18	19	43.17	20
	ours(PASS)	45.21	19	34.52	20	42.6	20	31.62	17
	ours(iCaRL)	**66.52**	**16**	**56.14**	24	**64.84**	**15**	**54.71**	21



(**a**) CIFAR100_5_20-NIID ($\lambda = 0.5$) (**b**) CIFAR100_10_10-NIID ($\lambda = 0.5$)

Figure 3. Comparison curve of the average accuracy of models for CIFAR100_5_20-NIID (**a**) and CIFAR100_10_10-NIID (**b**).

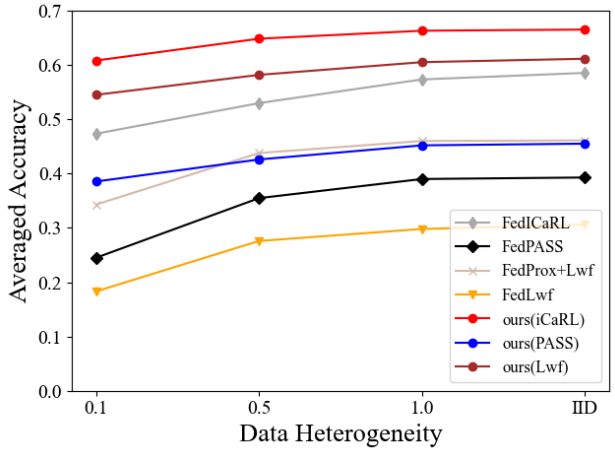
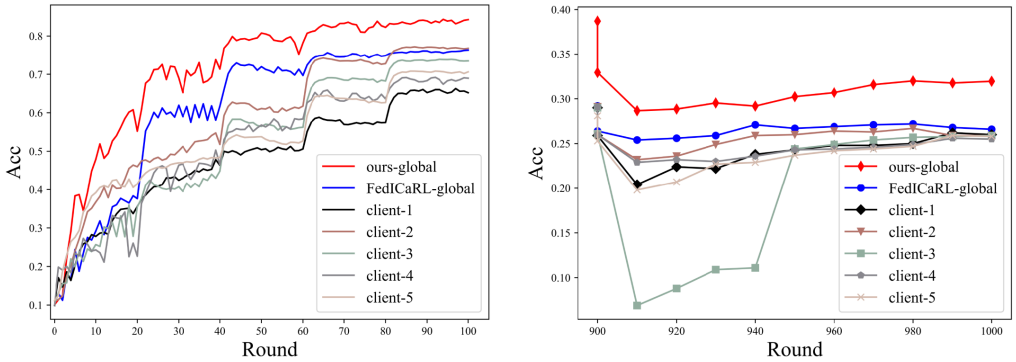


Figure 4. Comparison curve of the average accuracy of models for various degrees of data heterogeneity with CIFAR100_5_20.

Figure 5 illustrates the performance of ours(iCaRL) within the CIFAR100_10_10-NIID setting in more detail. Specifically, Figure 5a shows the test accuracy during the first task. Since the first task is unaffected by catastrophic forgetting, it can better demonstrate the model's ability to resist inter-client interference. Compared to the suboptimal method FedICaRL, the global model of our method can swiftly and accurately acquire knowledge. Notably, when the model converges, the performance of FedAvg-based methods can hardly surpass the local optimal model. In contrast, our method efficiently aggregates the heterogeneous models into the global model with outstanding performance and ensures that the global model is optimal. This indicates that our model effectively resists inter-client interference and highlights the effectiveness of our aggregation algorithm, i.e., PIT module. Figure 5b depicts the training process of the last task, which is most affected by catastrophic forgetting. A significant performance improvement appears in the later stages of training of our method, highlighting the effectiveness of knowledge distillation from the old global model, i.e., global forgetting loss.

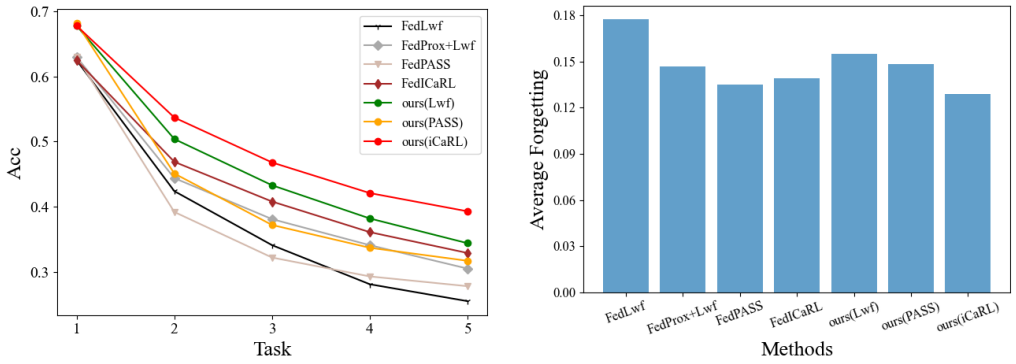


(**a**) The test accuracy of the first task. (**b**) The test accuracy of the last task.

Figure 5. The detailed test accuracy of ours (iCaRL) in the CIFAR100_10_10-NIID setting.

Experiments on Tiny-ImageNet: To demonstrate the universality and effectiveness of our proposed method across various datasets and models, we conducted experiments using VGG16 [33] as the backbone on the more challenging Tiny-ImageNet dataset with $\lambda = 0.1$. The results are similar to those observed on CIFAR-100. Figure 6 presents the test accuracy across 5 tasks and the average forgetting of different methods. It is evident that, among all the compared methods in three privacy scenarios, our proposed method outperforms existing methods in more challenging settings. Ours(iCaRL) achieves the highest average accuracy $\bar{A} = 50.12\%$ and the lowest average forgetting $\bar{F} = 13\%$. This underscores the universality and effectiveness of our proposed method across different models and datasets.

Ablation study: To better illustrate the necessity of each module in our method, we ran ablation experiments in the CIFAR100_5_20-NIID setting via ours(iCaRL), and the results are presented in Table 2. When the PIT module is removed, the test accuracy of the first task and the average accuracy of our proposed method is significantly reduced. Similarly, the removal of SKS also leads to a decrease in the average accuracy of the model and an increase in the average forgetting. This indicates the pivotal role of these two modules in mitigating data heterogeneity and inter-client interference. The impact of the global forgetting loss is observed in the detailed data of each task; its removal affects the ability to address catastrophic forgetting, resulting in a 6.1% reduction in the accuracy of the global model at the last task.



(**a**) The test accuracy of tasks. (**b**) The average forgetting of all tasks.

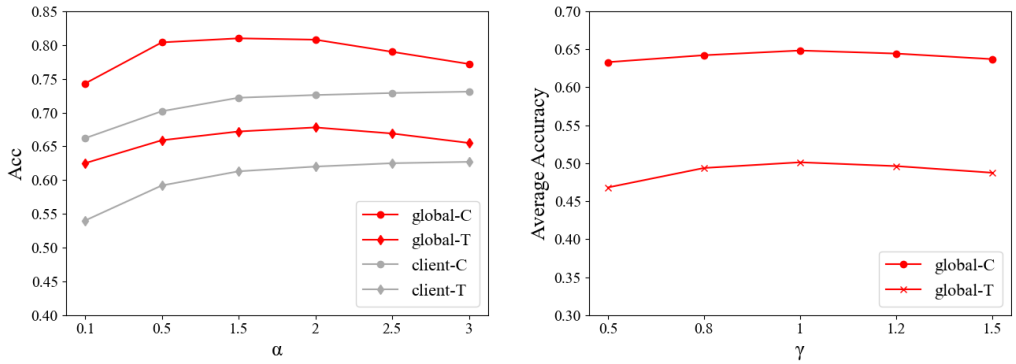
Figure 6. The test accuracy and the average forgetting on Tiny-ImageNet_5_40 with $\lambda = 0.1$.

Table 2. Ablation experimental results in the CIFAR100_5_20-NIID setting. The test accuracy for each task is listed. $\bar{A}$ refers to average accuracy (%) and $\bar{F}$ to average forgetting (%). The best results are in bold.

Method	**Tasks**					$\bar{A}$	$\bar{F}$
	1	**2**	**3**	**4**	**5**		
Ours	**0.808**	**0.698**	**0.636**	**0.568**	**0.532**	**64.84**	**15**
W/o SKS	0.771	0.639	0.576	0.503	0.463	59.04	16
W/o PIT	0.762	0.663	0.556	0.451	0.364	55.92	24
W/o GF	0.808	0.687	0.609	0.542	0.471	62.34	19

We also meticulously designed ablation experiments to investigate the impact of hyperparameters in Equation (10). Herein, α regulates the importance of the GCC loss, influencing the similarity between the global model and local models. The hyperparameter β, inherited from traditional continual learning, influences the model's resistance to local catastrophic forgetting, while γ strikes a balance between global and local catastrophic forgetting. Based on existing works [7,11], we set β to 3 and then employed grid search to determine the optimal values of α and γ. We conducted experiments in two settings: (1) CIFAR100_5_20, denoted as C, with $\lambda = 0.5$ and ResNet18 as the model backbone, and (2) Tiny-ImageNet_5_40, denoted as T, with $\lambda = 0.1$ and VGG16 as the backbone. To prevent catastrophic forgetting from biasing the results, we measured the influence of α using the test accuracy of the first task. As depicted in Figure 7a, when $\alpha \in [0.5, 2]$, global models achieved optimal or superior performance. While our method is sensitive to α, its optimal value remains relatively stable. We recommend using a relatively higher α for settings with high data heterogeneity. However, setting the α value too high is not desirable, as this is not conducive in mitigating the interference caused by bad clients on other clients, thereby impeding the overall performance of the global model. Additionally, the client accuracy is shown in Figure 7a, which represents the average value of the test accuracy of all client models. We observed that higher values of α led to closer alignment between client models and the global model. Conversely, lower α values resulted in a significant disparity between the local and global models, indicating overfitting of local models. Contrary to intuition, within the [0.5, 2] interval, the global model exhibits better generalization, while local models show slight overfitting. We theorize that mild overfitting enhances other clients' understanding of the real data distribution, with its negative effects mitigated through a balance between optimal transport and the SKS module. This observation echoes the original intention of designing the PIT module, confirming its effectiveness. After fixing $\alpha = 2$, the average accuracy of all tasks is used to determine the best value of γ. Figure 7b

illustrates that when $\gamma = 1$, the balance between local and global catastrophic forgetting is optimal.



(**a**) The test accuracy of the first task for various α. (**b**) The average accuracy of all tasks for various γ.

Figure 7. The ablation results of hyperparameters α and γ.

5. Conclusions

In conclusion, this paper introduces a novel selective knowledge fusion (FedSKF) model to tackle three key challenges in FCIL. FedSKF is the first work to introduce optimal transport into FCIL to mitigate data heterogeneity. Furthermore, we propose a selective knowledge fusion mechanism to alleviate inter-client interference through optimal transport and global classification consistency loss. Unlike most existing methods, our framework is universal and not strongly coupled with CIL methods, enabling the integration of different CIL methods to address various privacy scenarios. Numerous experiments demonstrate the effectiveness and superiority of our proposed FedSKF model compared to existing works.

Author Contributions: Conceptualization, M.Z.; methodology, M.Z.; software, M.Z.; validation, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, M.Z. and X.W.; supervision, X.W.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Science and Technology Commission of Shanghai Municipality under grant number 22511106000 and 22511106004, and National Natural Science Foundation of China under grant number 62231019.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; Gao, Y. A survey on federated learning. *Knowl.-Based Syst.* **2021**, *216*, 106775. [CrossRef]
2. Ma, Y.; Xie, Z.; Wang, J.; Chen, K.; Shou, L. Continual Federated Learning Based on Knowledge Distillation. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, Vienna, Austria, 23–29 July 2022; pp. 2182–2188.
3. Criado, M.F.; Casado, F.E.; Iglesias, R.; Regueiro, C.V.; Barro, S. Non-IID data and Continual Learning processes in Federated Learning: A long road ahead. *Inf. Fusion* **2022**, *88*, 263–280. [CrossRef]
4. Dong, J.; Wang, L.; Fang, Z.; Sun, G.; Xu, S.; Wang, X.; Zhu, Q. Federated Class-Incremental Learning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022.
5. Yoon, J.; Jeong, W.; Lee, G.; Yang, E.; Hwang, S.J. Federated Continual Learning with Weighted Inter-client Transfer. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Volume 139, pp. 12073–12086.
6. Shenaj, D.; Toldo, M.; Rigon, A.; Zanuttigh, P. Asynchronous Federated Continual Learning. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023.

7. Zhang, J.; Chen, C.; Zhuang, W.; Lyu, L. TARGET: Federated Class-Continual Learning via Exemplar-Free Distillation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023.
8. Chaudhry, A.; Dokania, P.K.; Ajanthan, T.; Torr, P.H. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 532–547.
9. Wang, L.; Zhang, X.; Su, H.; Zhu, J. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *1*, 1–20. [CrossRef] [PubMed]
10. Li, Z.; Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2935–2947. [CrossRef] [PubMed]
11. Rebuffi, S.A.; Kolesnikov, A.; Sperl, G.; Lampert, C.H. iCaRL: Incremental Classifier and Representation Learning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5533–5542.
12. Zhu, F.; Zhang, X.Y.; Wang, C.; Yin, F.; Liu, C.L. Prototype Augmentation and Self-Supervision for Incremental Learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 5867–5876.
13. Usmanova, A.; Portet, F.; Lalanda, P.; Vega, G. A distillation-based approach integrating continual learning and federated learning for pervasive services. In Proceedings of the 3rd Workshop on Continual and Multimodal Learning for Internet of Things—Co-located with IJCAI 2021, Montreal, QC, Canada, 19–27 August 2021; pp. 1–7.
14. Wang, Z.; Zhang, Y.; Xu, X.; Fu, Z.; Yang, H.; Du, W. Federated probability memory recall for federated continual learning. *Inf. Sci.* **2023**, *629*, 551–565. [CrossRef]
15. Qi, D.; Zhao, H.; Li, S. Better Generative Replay for Continual Federated Learning. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023; pp. 1–17.
16. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Networks. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2014; Volume 27, pp. 2672–2680.
17. Mendieta, M.; Yang, T.; Wang, P.; Lee, M.; Ding, Z.; Chen, C. Local learning matters: Rethinking data heterogeneity in federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8397–8406.
18. Zhang, Z.; Guo, B.; Sun, W.; Liu, Y.; Yu, Z. Cross-FCL: Toward a Cross-Edge Federated Continual Learning Framework in Mobile Edge Computing Systems. *IEEE Trans. Mob. Comput.* **2024**, *23*, 313–326. [CrossRef]
19. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
20. Farnia, F.; Reisizadeh, A.; Pedarsani, R.; Jadbabaie, A. An Optimal Transport Approach to Personalized Federated Learning. *IEEE J. Sel. Areas Inf. Theory* **2022**, *3*, 162–171. [CrossRef]
21. Singh, S.P.; Jaggi, M. Model fusion via optimal transport. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22045–22055.
22. Chiang, Y.H.; Terai, K.; Chiang, T.W.; Lin, H.; Ji, Y.; Lui, J.C.S. Optimal Transport-Based One-Shot Federated Learning for Artificial Intelligence of Things. *IEEE Internet Things J.* **2024**, *11*, 2166–2180. [CrossRef]
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021.
24. Bagwe, G.; Yuan, X.; Pan, M.; Zhang, L. Fed-CPrompt: Contrastive Prompt for Rehearsal-Free Federated Continual Learning. Available online: https://openreview.net/pdf?id=xEyopZpViw (accessed on 1 March 2024).
25. Lin, M.; Chen, M.; Zhang, Y.; Shen, C.; Ji, R.; Cao, L. Super vision transformer. *Int. J. Comput. Vis.* **2023**, *131*, 3136–3151. [CrossRef]
26. Kantorovitch, L. On the translocation of masses. *Manag. Sci.* **1958**, *5*, 1–4. [CrossRef]
27. Kulkarni, V.; Kulkarni, M.; Pant, A. Survey of Personalization Techniques for Federated Learning. In Proceedings of the 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, UK, 27–28 July 2020; pp. 794–797.
28. Tan, A.Z.; Yu, H.; Cui, L.; Yang, Q. Towards Personalized Federated Learning. *IEEE Trans. Neural Networks Learn. Syst.* **2023**, *34*, 9587–9603. [CrossRef] [PubMed]
29. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
30. Hinton, G.E.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:abs/1503.02531.
31. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images; Technical Report. 2009. Available online: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (accessed on 1 March 2024).

32. Le, Y.; Yang, X. Tiny imagenet visual recognition challenge. *CS 231N* **2015**, *7*, 3.
33. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–14.