

AFCL: Achieving Spatio-Temporal Invariance to Data Heterogeneity in Federated Continual Learning

Jianheng Tang¹, Jingyu He², Kejia Fan², Run He³, Jingchao Wang¹, Anfeng Liu², Houbing Herbert Song⁴, Leye Wang¹, Zhanxing Zhu⁵, Huiping Zhuang^{† 3}, Yunhuai Liu^{† 1,6}

¹Peking University; ²Central South University;

³South China University of Technology; ⁴University of Maryland, Baltimore County;

⁵University of Southampton; ⁶Key Lab of High Confidence Software Technologies (PKU)

Abstract

Federated Continual Learning (FCL) enables distributed clients to collaboratively train a global model from online task streams in dynamic real-world scenarios. However, existing FCL methods face challenges of both spatial data heterogeneity among distributed clients and temporal data heterogeneity across online tasks. Such data heterogeneity significantly degrades the model performance with severe spatial-temporal catastrophic forgetting of local and past knowledge. In this paper, we identify that the root cause of this issue lies in the inherent vulnerability and sensitivity of gradients to Non-IID data. To fundamentally address this issue, we propose a gradient-free method, named Analytic Federated Continual Learning (AFCL), by deriving analytical (i.e., closed-form) solutions from frozen extracted features. In local training, our AFCL enables single-epoch learning with only a lightweight forward-propagation process for each client. In global aggregation, the server can recursively and efficiently update the global model with single-round aggregation. Theoretical analyses validate that our AFCL achieves spatio-temporal invariance of Non-IID data. Extensive experiments show the consistent superiority of our AFCL regarding accuracy and efficiency over state-of-the-art baselines across various benchmark datasets and settings.

1. Introduction

Federated Learning (FL) is a widely adopted paradigm that enables distributed clients to collaboratively train machine learning models while preserving the privacy of their respective data [9, 11, 13]. In traditional FL, the data heterogeneity (i.e., Non-IID issue) across distributed clients represents one of the most critical challenges, leading to a degradation in the performance of the aggregated global model on local tasks [48, 59, 72]. While many studies have been proposed

to address this issue [3, 57, 67], they are typically based on an unrealistic static assumption that the domains and categories of all clients' data remain unchanged [62, 64, 66].

Federated Continual Learning (FCL) breaks these static limitations by allowing clients to continuously acquire new knowledge from online task streams [35, 36, 70]. Although FCL expands the applicability of FL in dynamic real-world scenarios, it introduces a more challenging issue: spatio-temporal data heterogeneity [60, 63, 65]. This heterogeneity arises not only across different clients (i.e., spatial level) but also in different tasks of the same client (i.e., temporal level), forming a complex interplay, as shown in Figure 1 (a).

The direct negative impact of spatial-temporal data heterogeneity is spatial-temporal catastrophic forgetting [63, 65], as illustrated in Figure 1(b). Catastrophic forgetting is originally a term from Continual Learning (CL), used to describe the phenomenon where deep models, after being trained on new data samples, tend to rapidly forget previously acquired knowledge [10, 39]. In FCL, the local models face temporal forgetting as new data continue to arrive over time for both existing clients and incoming clients [6, 40]. Meanwhile, the Non-IID data across clients can cause the aggregated global model to forget each client's local knowledge in a spatial sense [5, 53]. Moreover, as the local clients employ the aggregated global model as the basis to further learn the next task, the spatial and temporal forgetting interplay with each other, making directly applying CL to FL infeasible [6, 65].

Spatio-temporal data heterogeneity can severely disrupt the stability of gradient-based updates to affect the model parameters [63, 65]. In this paper, we identify that the root cause of spatial-temporal catastrophic forgetting lies in the inherent sensitivity of gradients to Non-IID data. Specifically, as data distributions evolve over time and space, gradient-based updates tend to overwrite well-learned model parameters from previous tasks, leading to spatio-temporal catastrophic forgetting. Thus, our key insight is to fundamentally address spatio-temporal catastrophic forgetting in FCL by

[†]Corresponding Authors

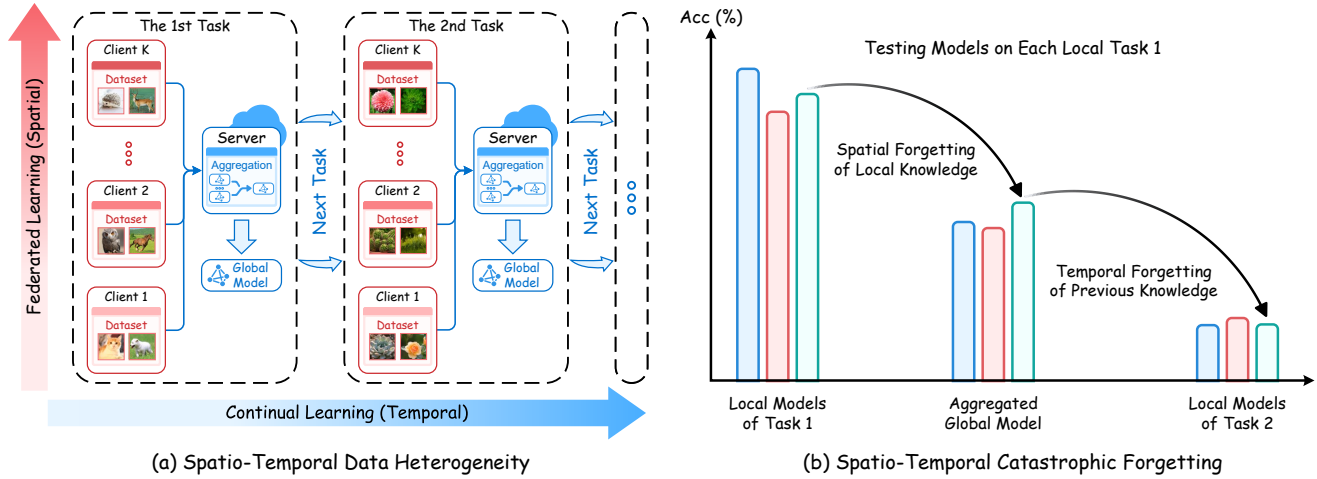


Figure 1. The spatial-temporal data heterogeneity and spatial-temporal catastrophic forgetting in FCL.

avoiding the primary culprit of gradients. Recent studies have shown the effectiveness of frozen pre-trained models in handling spatial-temporal heterogeneity among the fields of FL [4, 47, 52], CL [10, 42, 44, 45, 68], and FCL [49, 63, 65]. These findings validate our insight, but no existing work has fundamentally addressed the heterogeneity issue in FCL.

In this paper, we first posit that the ultimate goal for bridging this gap in FCL is achieving spatio-temporal invariance to data heterogeneity, i.e., regardless of how heterogeneous the data are distributed across clients and tasks in FCL, the aggregated model remains invariant. To fulfill this goal and fundamentally address the heterogeneity in FCL, we propose a gradient-free method, named Analytic Federated Continual Learning (AFCL). In our AFCL, each client shares a frozen foundation model as a gradient-free feature extractor. Then, we further develop a gradient-free classifier with analytical (i.e., closed-form) solutions, allowing each client to embed local knowledge via lightweight and forward-only training. Thus, by aggregating each client’s local knowledge in a one-shot manner, the server can recursively construct and update a meticulously designed *Global Knowledge Matrix*, from which the final global classifier can be derived. To enable our AFCL to support the challenging class-continual scenario, where arbitrary class-level heterogeneity exists in the clients’ online data increments, we design a known-unknown class splitting mechanism for flexible and interpretable adaptation.

Our main contributions are summarized as follows:

- Conceptually, identifying the root cause of spatio-temporal catastrophic forgetting, we posit the ultimate goal of spatio-temporal invariance in FCL, and fulfill it for the first time.
- Technically, we devise an efficient gradient-free protocol for our AFCL, involving only single-epoch local training and single-round aggregation for each client. In this way, our AFCL represents the first gradient-free method in FCL.

- Theoretically, we validate our AFCL’s ideal property of *spatio-temporal invariance to Non-IID data* based on its closed-form solution. Moreover, we also show our AFCL’s *interpretability* via the lens of knowledge aggregation.
- Experimentally, our AFCL delivers consistent superiority across a wide range of datasets and settings, outperforming competitive baselines by up to 25.07% in accuracy, with more than 90% better runtime efficiency.

2. Related Work

2.1. Federated Continual Learning

As a prevalent distributed learning paradigm, FL has made significant progress, particularly in tackling spatial data heterogeneity across clients [48, 59, 72]. Traditional FL studies are typically based on an unrealistic static assumption that the domains and categories of training data across all clients remain unchanged, yet failing to handle more realistic and dynamic scenarios where each client’s data continuously evolve with online task stream [3, 32, 67]. In this context, by introducing the concept of CL to FL, FCL has emerged as a prevalent topic for distributed knowledge acquisition from the clients’ online and heterogeneous data [41, 62, 64, 66].

Recent researchers have highlighted that the interaction between spatial and temporal heterogeneity in FCL has been largely overlooked in existing studies [1, 60], resulting in the challenge of spatio-temporal catastrophic forgetting [6, 53]. To date, while a few studies have attempted to address this issue from the temporal-spatial perspective, none of them have truly realized gradients as the root cause of forgetting, preventing them from fundamentally resolving spatio-temporal catastrophic forgetting [51, 60, 63, 65, 69]. To the best of our knowledge, we are the first to posit and successfully achieve the ideal pursuit of spatio-temporal invariance in FCL.

2.2. Analytic Learning

Analytic learning is a representative gradient-free technique, originally developed to address the common gradient-related issues, e.g., vanishing and exploding gradients [25, 73]. Due to its flexible usage of matrix inversion, analytic learning is also referred to as pseudoinverse learning [2, 14, 61]. Its core idea is to directly derive analytical solutions via least squares [2, 52, 54]. With the wide adoption of the block-wise recursive Moore-Penrose inverse, AL has shown power in many applications, particularly in CL [18, 19, 26, 30, 52].

Nonetheless, previous research on analytic learning has primarily focused on its partial property from analytical solutions (e.g., the temporal invariance for online CL [10, 25, 73]). In contrast, we first introduce the concept of analytic learning into an under-explored field of FCL, extending its invariance benefit to *a holistic spatial-temporal perspective*. In proposing our AFCL, we carefully design a known-unknown class splitting mechanism to address the unique class-continual challenges in FCL, where arbitrary class-level (along with domain-level) heterogeneity may exist in the clients' online data increments across both temporal and spatial dimensions.

3. Our Proposed AFCL

3.1. System Model

In FCL, the server typically needs to cope with two forms of heterogeneous incremental data from clients: (1) new data continuously collected by existing clients [6, 60, 63, 64], and (2) new data introduced by newly joined clients [1, 7, 53, 65]. To uniformly and flexibly address the two types of temporal increments, our proposed AFCL adopts a unified manner by treating all the new data increments as if they were brought by newly joined clients. Specifically, each batch of new data continuously collected by an existing client is treated as if it were introduced by a newly joined *virtual client*.

Thus, the entire system can be modeled as consisting of a central server $\mathcal{S}$ and a set of K (virtual) clients, denoted by $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$. The k -th client's dataset is denoted as $\mathcal{D}_k \sim \{\mathcal{X}_k, \mathcal{Y}_k\}$, where $\mathcal{X}_k \in \mathbb{R}^{N_k \times \omega \times h \times c}$ (containing N_k images with the shape of $\omega \times h \times c$) and $\mathcal{Y}_k \in \mathbb{R}^{N_k}$ (containing N_k examples) correspond to the stacked input and label data for the k -th client, respectively. In FCL, each client is considered online and incremental, and no single virtual client can access the data of any other owing to the spatial-temporal data isolation, even if both virtual clients originate from the same physical entity's varying tasks.

We further consider a challenging class-continual setting in FCL, where each client's data increments simultaneously consist of both previously encountered classes and entirely new classes, exhibiting arbitrary class-level heterogeneity. In practice, the server is often unable to pre-determine both the total number and specific types of classes that may appear

during training. Fortunately, each client can declare the set of classes contained in its current local data during registration, enabling the server to dynamically manage the class space (e.g., the one-hot encoding mapping). Notably, since image recognition is the most common task in FCL [32, 33, 36, 60, 63, 65, 66], we also follow this default convention in our system model, leaving more extensions of our AFCL to other tasks for a brief discussion in Section 5.

3.2. Motivation and Overview

In FCL, facing spatio-temporal data heterogeneity, the inaccessibility of local and historical data causes gradient-based updates to induce conflicts among clients and with prior updates, manifesting as spatio-temporal catastrophic forgetting. Thus, we identify that the root cause of spatio-temporal catastrophic forgetting lies in the vulnerability and sensitivity of gradients to Non-IID data. Motivated by this insight, our AFCL is designed to fundamentally address spatio-temporal catastrophic forgetting by avoiding the reliance on gradients.

The models within our AFCL mainly consist of two key components: a pre-trained feature extractor and a gradient-free analytic classifier. For the first component, employing a pre-trained model for feature extraction has been widely recognized and adopted across the fields of FL [4, 47, 52], CL [42, 44, 71], and FCL [49, 63, 65]. Thus, we follow this practice in our AFCL. This pre-trained backbone can be flexibly determined by downloading a public model from the Internet, according to the specific requirements of FCL (e.g., different sizes or performance levels). Then, we further develop an analytic classifier for the extracted features, with its final global parameters $\mathbf{W}_K$ optimized under the empirical risk minimization objective:

$$\mathbf{W}_K = \underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{Y}_{1:K} - \mathbf{F}_{1:K} \mathbf{W}\|_F^2 + \gamma \|\mathbf{W}\|_F^2, \quad (1)$$

where, $\mathbf{F}_{1:K}$ is the stacked features and $\mathbf{Y}_{1:K}$ is the stacked labels with global one-hot encoding, both assembled from the full datasets $\mathcal{D}_{1:K}$ of all K clients. The term $\gamma \|\mathbf{W}\|_F^2$ is used for regularization, with adjustable γ for flexibility. Notably, we use Mean Squared Error (MSE) instead of Cross-Entropy (CE) as the loss function for classification to enable the closed-form solutions for our AFCL. In fact, recent work has broadly shown that MSE can yield performance comparable to or even better than CE [21, 22], with better robustness to combat catastrophic forgetting in CL [12, 23, 25].

The framework of our AFCL is illustrated in Figure 2. Specifically, each virtual client first registers with the server based on the set of classes contained in its current local data. After that, the server dynamically manages the global class space and publishes the setting information, including the pre-trained backbone and the one-hot encoding functions with known-unknown class splitting. Then, each client performs feature extraction and label encoding, followed by localized analytic computation within a single epoch through

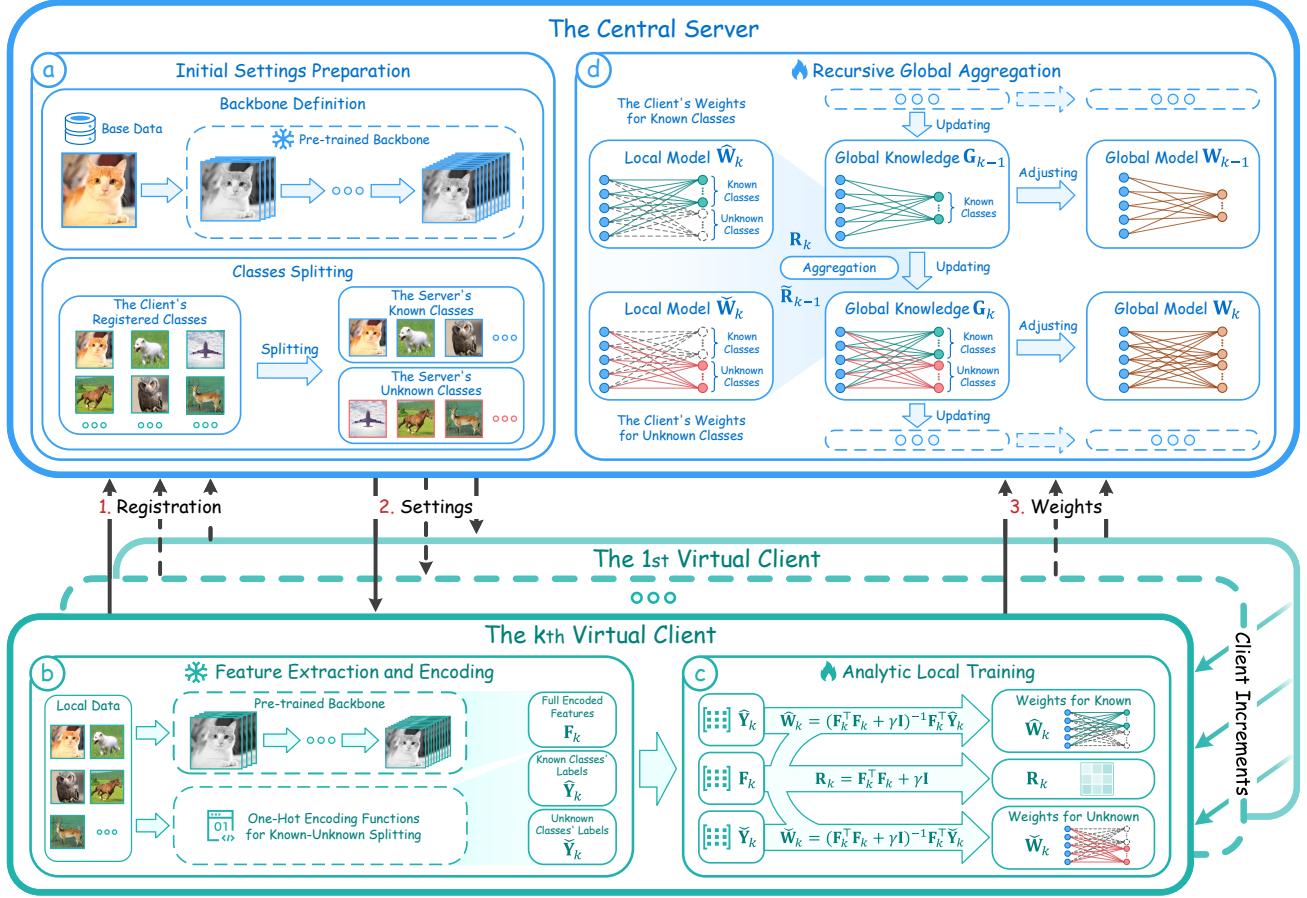


Figure 2. The framework of our proposed AFCL.

least-squares. Finally, each client uploads the local models, and the server recursively aggregates them to construct and update a meticulously designed *Global Knowledge Matrix*, from which the optimal global model for objective (1) can be ultimately derived. Our AFCL is gradient-free and forward-only, making it highly efficient for edge environments.

3.3. Known-Unknown Class Splitting

Here, considering the challenging class-continual setting of FCL, we propose a known-unknown class split mechanism to enable dynamic adjustments of the one-hot mapping and model structure for effective incorporating the new classes. Specifically, in practice, the clients' online data stream may introduce new classes that have not been encountered by previous clients. Thus, faced with the continuously joining clients, we need to not only update knowledge for previously encountered classes but also dynamically expand the model structure to accommodate knowledge for newly emerging classes. To achieve this goal, each newly participating client needs to transmit the set of classes contained in its current local data to the server during registration. The server then designates previously encountered classes as "known classes"

and the newly introduced classes by this new client as "unknown classes". Therefore, the server generates two one-hot encoding functions for known classes and unknown classes, which will assist the client in local training to obtain two local models for known and unknown classes, respectively. We show how it influences the computation in Sections 3.4–3.5.

When the k -th client connects to the server, we denote the sets of known and unknown classes as $\hat{S}_k$ and $\tilde{S}_k$. Upon receiving the k -th client's class information, the server needs to generate the one-hot encoding functions for known and unknown classes, denoted as $\text{Onehot}_k(\cdot)$ and $\text{Onehot}_k^*(\cdot)$, respectively. Specifically, $\text{Onehot}_k^*(\cdot)$ is directly generated based on the set of unknown classes $\tilde{S}_k$. As a special case, when $k = 1$, there is no previous knowledge, and all classes introduced by the first client are deemed unknown. In other cases, $\text{Onehot}_k(\cdot)$ is obtained via the recursive update in (2). The server then transmits both $\text{Onehot}_k(\cdot)$ and $\text{Onehot}_k^*(\cdot)$ to the k -th client as part of the setting information.

$$\text{Onehot}_k(s) = \begin{cases} [\text{Onehot}_{k-1}(s) & \mathbf{0}], & \text{if } s \in \hat{S}_{k-1}, \\ [\mathbf{0} & \text{Onehot}_{k-1}^*(s)], & \text{if } s \in \tilde{S}_{k-1}. \end{cases} \quad (2)$$

3.4. Client-Side: Analytic Local Training

After obtaining the setting information from the server, each client then locally extracts features from its local data using the pre-trained backbone and maps the labels using the one-hot encoding functions, as shown in Figure 2(b). Specifically, the k -th client performs a forward propagation on its raw local data $\mathcal{X}_k$ to extract the features $\mathbf{F}_k$:

$$\mathbf{F}_k = \text{Backbone}(\mathcal{X}_k, \Theta_B). \quad (3)$$

Here, $\mathbf{F}_k \in \mathbb{R}^{N_k \times l_B}$ demonstrates the features extracted by the backbone, with its dimension l_B determined by the backbone's output. Meanwhile, the client also employs the obtained one-hot encoding functions $\text{Onehot}_k(\cdot)$ and $\text{Onehot}_k^*(\cdot)$ to map labels from its local data $\mathcal{Y}_k$, as follows:

$$\hat{\mathbf{Y}}_k = \text{Onehot}_k(\mathcal{Y}_k), \quad \check{\mathbf{Y}}_k = \text{Onehot}_k^*(\mathcal{Y}_k). \quad (4)$$

Here, $\hat{\mathbf{Y}}_k \in \mathbb{R}^{N_k \times d_{k-1}}$ and $\check{\mathbf{Y}}_k \in \mathbb{R}^{N_k \times (d_k - d_{k-1})}$ represent the label matrices for known and unknown classes, respectively, where d_k and d_{k-1} represent the total number of classes registered by the server across the datasets from the first k clients and the first $(k-1)$ clients, respectively.

Upon obtaining $\mathbf{F}_k$, $\hat{\mathbf{Y}}_k$, and $\check{\mathbf{Y}}_k$, the client subsequently performs localized analytic computation, as shown in Figure 2(c). Since the client's dataset may contain previously unseen classes for the server, structural variations may occur in the aggregated global models. To address this issue, in our AFCL, each client needs to train two separate local analytic models $\hat{\mathbf{W}}_k \in \mathbb{R}^{l_B \times d_{k-1}}$ and $\check{\mathbf{W}}_k \in \mathbb{R}^{l_B \times (d_k - d_{k-1})}$ for the known and unknown classes, respectively. Specifically, the training of these local models is reformulated as two regularized optimization problems in the form of (1), which can be analytically solved as (5) and (6) through least-squares, with the detailed derivations provided in Section 3.6.

$$\hat{\mathbf{W}}_k = (\mathbf{F}_k^\top \mathbf{F}_k + \gamma \mathbf{I})^{-1} \mathbf{F}_k^\top \hat{\mathbf{Y}}_k, \quad (5)$$

$$\check{\mathbf{W}}_k = (\mathbf{F}_k^\top \mathbf{F}_k + \gamma \mathbf{I})^{-1} \mathbf{F}_k^\top \check{\mathbf{Y}}_k. \quad (6)$$

Additionally, we define the *Regularized Gram Matrix* for the k -th client as $\mathbf{R}_k \in \mathbb{R}^{l_B \times l_B}$, which is computed by (7).

$$\mathbf{R}_k = \mathbf{F}_k^\top \mathbf{F}_k + \gamma \mathbf{I}. \quad (7)$$

This *Regularized Gram Matrix* is subsequently utilized to assist the server in performing global model aggregation. Upon finishing all the gradient-free and forward-only computations as mentioned above, the k -th client then uploads its two constructed local models (i.e., $\hat{\mathbf{W}}_k$ and $\check{\mathbf{W}}_k$) and the *Regularized Gram Matrix* (i.e., $\mathbf{R}_k$), as its compressed local knowledge, to the server for global knowledge aggregation.

3.5. Server-Side: Recursive Global Aggregation

After the k -th client completes the analytic local training, the server subsequently performs recursive global aggregation based on the received information from this client (i.e., $\hat{\mathbf{W}}_k$, $\check{\mathbf{W}}_k$, and $\mathbf{R}_k$). Specifically, by aggregating each client's local knowledge, the server recursively constructs and updates a *Global Knowledge Matrix* $\mathbf{G}_k \in \mathbb{R}^{l_B \times d_k}$, from which the global model $\mathbf{W}_k \in \mathbb{R}^{l_B \times d_k}$ can be derived. Notably, this aggregation process needs only a single round for each client, thus avoiding the multi-round communication overhead.

Considering the class-continual setting of FCL, the newly joined (virtual) clients may dynamically introduce brand new classes that the server has never encountered. Thus, such previously unknown classes should alter the structure of the analytic model, changing its output dimension from d_{k-1} to d_k . To dynamically accommodate the structural increments caused by previously unknown classes while simultaneously adjust the parameters of the remaining structure based on new data from known classes, we design a unique updating formula for the *Global Knowledge Matrix*, as shown in (8). In this way, $\mathbf{G}_k$ can capture and maintain the accumulated global knowledge from the beginning up to the k -th round.

$$\mathbf{G}_k = \underbrace{[\mathbf{A}_k \mathbf{G}_{k-1} + \mathbf{B}_k \hat{\mathbf{W}}_k]}_{(a)} \underbrace{[\mathbf{B}_k \check{\mathbf{W}}_k]}_{(b)}, \quad k \in (1, K], \quad (8)$$

where

$$\begin{cases} \mathbf{A}_k = \mathbf{I} - (\tilde{\mathbf{R}}_{k-1})^{-1} \mathbf{R}_k (\mathbf{I} - (\tilde{\mathbf{R}}_k)^{-1} \mathbf{R}_k), \\ \mathbf{B}_k = \mathbf{I} - (\mathbf{R}_k)^{-1} \tilde{\mathbf{R}}_{k-1} (\mathbf{I} - (\tilde{\mathbf{R}}_k)^{-1} \tilde{\mathbf{R}}_{k-1}), \\ \tilde{\mathbf{R}}_k = \sum_{i=1}^k \mathbf{R}_i, \quad \mathbf{R}_i = \mathbf{F}_i^\top \mathbf{F}_i + \gamma \mathbf{I}. \end{cases} \quad (9)$$

In the first round, since all the classes are previously unknown, there is no knowledge for the introduced known classes, and thus $\mathbf{G}_1 = \tilde{\mathbf{W}}_1$ serves as a special case. The recursive formula (8) for $\mathbf{G}_k$ expresses the knowledge update in an explicit and interpretable manner. Specifically, the former term in (8) represents the knowledge for known classes, which is a combination of the previously stored knowledge $\mathbf{A}_k \mathbf{G}_{k-1}$ and the newly acquired knowledge $\mathbf{B}_k \hat{\mathbf{W}}_k$. The latter term $\mathbf{B}_k \check{\mathbf{W}}_k$, which is *concatenated into* $\mathbf{G}_k$, reflects the acquired knowledge for the previously unknown classes.

Notably, the server does not need to store all historical *Regularized Gram Matrices*, as $\tilde{\mathbf{R}}_k$ can also be updated recursively. As the clients continually join, the *Global Knowledge Matrix* is dynamically updated to accommodate the incremental data, while preserving the knowledge and memory of all previously learned data. Via the recursive computation for $\mathbf{G}_k$, we can further derive its closed-form general expression in Lemma 3 within Appendix A.2. The explicit and closed-form formulation for $\mathbf{G}_k$ also provides superior interpretability for the AFCL's knowledge update process.

Notably, the *Global Knowledge Matrix* is a dedicated unit for storing globally obtained knowledge in our AFCL, rather

than the final global model itself. After updating the *Global Knowledge Matrix* $\mathbf{G}_k$, the server needs to further compute the corresponding global analytical model $\mathbf{W}_k$ by (10).

$$\mathbf{W}_k = \mathbf{G}_k + [(k-1)\gamma(\tilde{\mathbf{R}}_k - (k-1)\gamma\mathbf{I})^{-1}]\mathbf{G}_k. \quad (10)$$

In this way, the server continuously and recursively executes the above process until all the K clients' local models have been aggregated into the global model, as illustrated in Figure 2(d). Notably, the final global model $\mathbf{W}_K$ obtained by our AFCL is equivalent to empirical risk minimization in (1) with the full datasets $\mathcal{D}_{1:K}$ from all K clients.

3.6. Theoretical Analyses

To deliver our AFCL's validity, we first confirm its derived closed-form solutions in Lemma 1 and Theorem 1. Then, we further deduce our AFCL's ideal property of *spatio-temporal invariance to Non-IID data*, as displayed in Theorem 2.

Lemma 1: For any optimization problem of the form:

$$\underset{\mathbf{W}}{\operatorname{argmin}} \quad \|\mathbf{Y} - \mathbf{F}\mathbf{W}\|_{\mathbf{F}}^2 + \gamma\|\mathbf{W}\|_{\mathbf{F}}^2, \quad (11)$$

it admits a unique closed-form solution given by:

$$\mathbf{W} = (\mathbf{F}^\top \mathbf{F} + \gamma\mathbf{I})^{-1} \mathbf{F}^\top \mathbf{Y}. \quad (12)$$

Proof. The derivation is based on the least-squares method with ℓ_2 regularization. See Appendix A.1 for details.

According to Lemma 1, we can verify that the local models of the k -th client for the known and unknown classes (i.e., $\hat{\mathbf{W}}_k$ and $\tilde{\mathbf{W}}_k$) obtained via (5) and (6) are the corresponding optimal local models. In addition, we can derive that the global optimization objective with the empirical risk minimization over the datasets $\mathcal{D}_{1:k}$ from the first k clients similarly admits the following closed-form solution:

$$\mathbf{W}_k = (\mathbf{F}_{1:k}^\top \mathbf{F}_{1:k} + \gamma\mathbf{I})^{-1} (\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k}). \quad (13)$$

Based on this conclusion, we further reveal that the global model derived from our AFCL is exactly equivalent to the optimal solution (13) for empirical risk minimization over the full datasets $\mathcal{D}_{1:k}$ from the first k clients, as follows.

Theorem 1: In our AFCL, the computation results for $\mathbf{W}_k$ by (10) is exactly equivalent to (13), which corresponds to the optimal solution obtained via centralized joint learning over the full datasets $\mathcal{D}_{1:k}$ from the first k clients.

Proof. The proof involves two primary steps: (1) inductively establishing the closed-form expression for $\mathbf{G}_k$ utilizing the *Woodbury Matrix Identity*, and (2) substituting this expression for $\mathbf{G}_k$ into (10). See Appendix A.2 for details.

Thus, we can go on to establish our AFCL's ideal property of *spatio-temporal invariance to Non-IID data*, as follows.

Theorem 2: The final global model obtained by our AFCL is independent of the spatio-temporal data heterogeneity and client registration order, being identical to that obtained by centralized joint learning over the full dataset $\mathcal{D}_{1:K}$.

Proof. The proof leverages Theorem 1 and the orthogonality of permutation matrices. See Appendix A.3 for details.

4. Experiments

4.1. Experimental Setup

Datasets & Settings. To evaluate the performance of our AFCL, we conduct extensive experiments on three benchmark vision datasets: CIFAR-100 [24], Tiny-ImageNet [27], and ImageNet-R [20]. Moreover, to align with existing studies in FCL, we partition the training process into multiple tasks, each involving collaborative model training among clients. Specifically, we use the Si-Blurry setting [46] to partition each dataset into a total of 5 or 10 tasks to simulate varying temporal data heterogeneity. Then, we allocate each task's dataset among 5 clients using the Dirichlet distribution with different parameters α to simulate varying spatial data heterogeneity [38]. More details are given in Appendix B.1.

Baselines & Metrics. In our experiments, we compare our AFCL against ten classical baselines, encompassing three types: FL, FL+CL, and FCL. The FL type includes classic methods that fine-tune across online tasks: FedAvg [43], FedProx [31], and FedDyn [8]. The FL+CL type includes early methods that use CL ideas with FL: FedIcaRL [50], FedLwF [37], and FedEwc [29]. The FCL type includes recent state-of-the-art methods: FedMGP [64], LGA [7], FedCBC [63], and TagFCL [55]. For a fair comparison, we use ResNet-18 [17] as the unified backbone across all methods. We employ average accuracy and knowledge retention as performance metrics, and use the cumulative runtime for efficiency. More details are given in Appendix B.2.

4.2. Experimental Evaluations

Invariance Analyses. First, we compare the performance of our AFCL with baselines under various spatio-temporal heterogeneity settings in Table 1. As shown in the results, our AFCL consistently outperforms the state-of-the-art baselines by up to 0.21%–25.07%. Notably, as the temporal heterogeneity intensifies (i.e., T rises from 5 to 10), the performance of all baselines exhibits a significant decline. A similar degradation is observed in the baselines as the spatial heterogeneity intensifies (e.g., α decreases from 0.2 to 0.1 on CIFAR-100). These results validate our previous analyses regarding the spatio-temporal catastrophic forgetting of these gradient-based methods. Conversely, benefiting from the ideal property of spatio-temporal invariance, our AFCL always maintains invariant and superior performance across all settings on each dataset. Moreover, it is foreseeable that if spatio-temporal data heterogeneity further increases in practice (e.g., the number of tasks $T \gg 10$, the number of clients $K \gg 5$), the advantage of our AFCL can be even more pronounced. Notably, while analytic-learning-based methods' advantages have been widely validated in prior studies [18, 19, 23, 45, 68, 73], our AFCL's advance is markedly greater. This is presumably due to the fact that prior studies only contended with heterogeneity in a sin-

Table 1. Performance comparison of average accuracy among our AFCL and other baselines. The parameters T and α control the levels of temporal and spatial heterogeneity in FCL. The **bold** and underlined results indicate the best and second-best performance, respectively.

Method		CIFAR-100				Tiny-ImageNet				ImageNet-R			
		$T = 5$		$T = 10$		$T = 5$		$T = 10$		$T = 5$		$T = 10$	
		$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 1.0$
FedAvg [43]	FL	4.96%	6.35%	4.15%	4.98%	1.85%	2.24%	1.32%	1.47%	1.41%	1.73%	1.47%	1.65%
FedProx [31]		10.01%	12.56%	7.92%	10.45%	6.73%	7.85%	4.83%	6.27%	9.79%	12.46%	6.58%	8.65%
FedDyn [8]		11.50%	13.64%	9.01%	10.72%	6.65%	14.66%	5.37%	11.49%	13.41%	14.81%	11.95%	13.48%
FedIcaRL [50]	FL+CL	4.05%	5.52%	3.64%	3.93%	2.05%	2.41%	1.84%	1.86%	2.03%	2.22%	1.57%	1.78%
FedLwF [37]		8.37%	11.11%	4.18%	5.63%	3.13%	3.62%	2.70%	2.21%	3.13%	3.19%	2.17%	2.48%
FedEwc [29]		12.64%	14.28%	7.92%	11.48%	3.34%	4.56%	2.63%	3.39%	4.27%	4.71%	3.21%	3.30%
FedMGP [64]	FCL	23.87%	26.75%	15.21%	16.17%	12.53%	14.82%	9.60%	10.10%	18.23%	18.69%	8.75%	9.33%
LGA [7]		39.09%	40.21%	31.78%	<u>38.56%</u>	36.70%	38.47%	27.36%	29.33%	23.19%	29.75%	19.18%	21.59%
FedCBC [63]		43.34%	44.93%	27.32%	29.02%	31.13%	33.87%	22.48%	24.62%	23.51%	23.79%	15.72%	14.14%
TagFCL [55]		<u>45.10%</u>	<u>46.64%</u>	<u>36.66%</u>	36.13%	<u>40.94%</u>	<u>44.98%</u>	<u>29.60%</u>	<u>32.56%</u>	<u>34.44%</u>	<u>38.01%</u>	<u>24.76%</u>	<u>28.37%</u>
AFCL (Ours)		58.56%	58.56%	58.56%	58.56%	54.67%	54.67%	54.67%	54.67%	38.22%	38.22%	38.22%	38.22%
Advance $\uparrow$		13.46%	11.92%	21.90%	20.00%	13.73%	9.69%	25.07%	22.17%	3.78%	0.21%	13.46%	9.85%

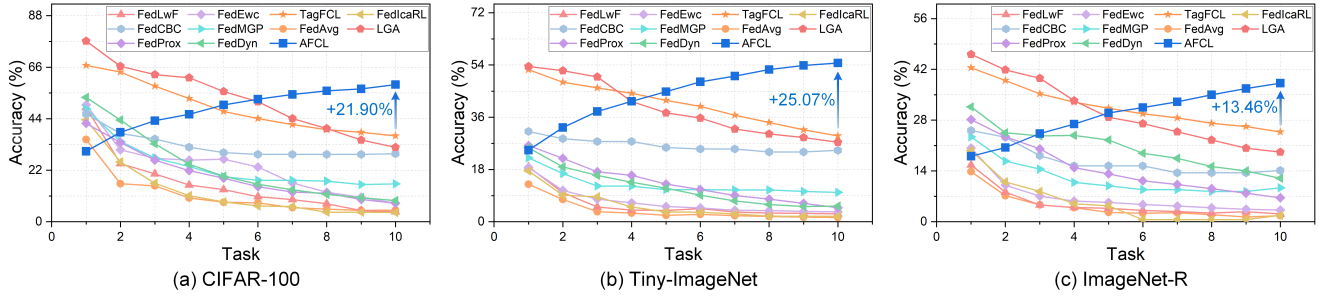


Figure 3. Average accuracy of our AFCL and the baselines throughout task progression.

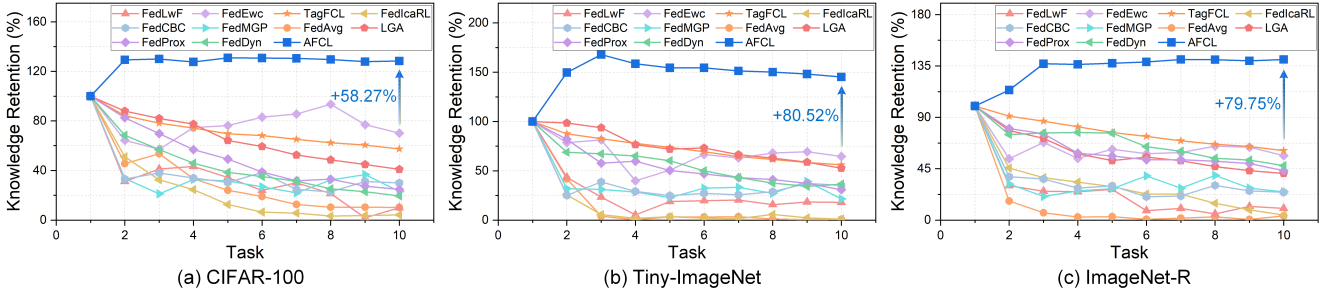


Figure 4. Average knowledge retention of our AFCL and the baselines throughout task progression.

gle dimension (either temporal or spatial), while our FCL scenario involves the more challenging dual heterogeneity.

Forgetting Analyses. Second, to further explore how model performance evolves as tasks progress, we examine the variations in average accuracy and knowledge retention across all methods, as shown in Figures 3–4. Specifically, the baselines all suffer from severe spatio-temporal forgetting, resulting in a marked decrease in average accuracy and knowledge retention as tasks progress. In contrast, as our AFCL continually learns from new tasks, its accuracy shows an initial increase followed by stabilization, while maintaining

the highest knowledge retention throughout the entire process. Notably, the knowledge retention of AFCL can exceed 100%, indicating its continual acquisition of generalizable knowledge from incoming data. In contrast, the knowledge retention of all baselines remains below 100%. These results show AFCL’s gradient-free power for combating forgetting.

Efficiency Analyses. To clearly illustrate the advantage of our AFCL, we compared its cumulative runtime in Figure 5 against two representative baselines: FedAvg, which offers the highest efficiency, and TagFCL, which yields the best performance. Note that the y-axis in Figure 5 is plotted on

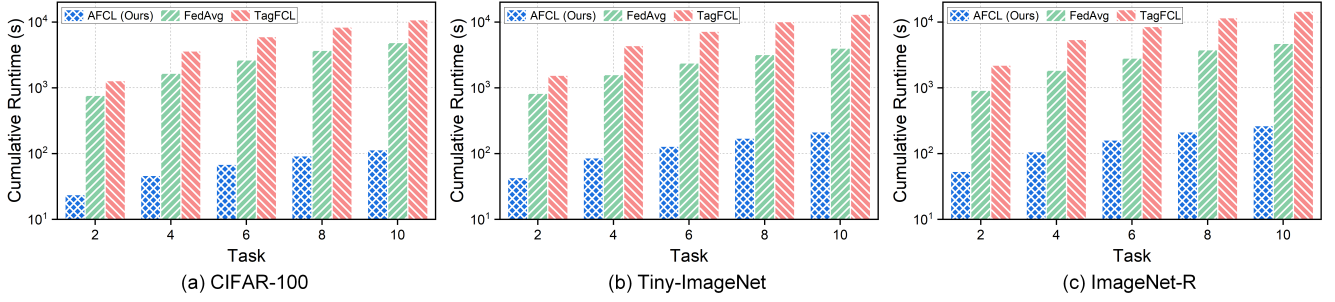


Figure 5. Cumulative runtime of our AFCL and the baselines throughout task progression.

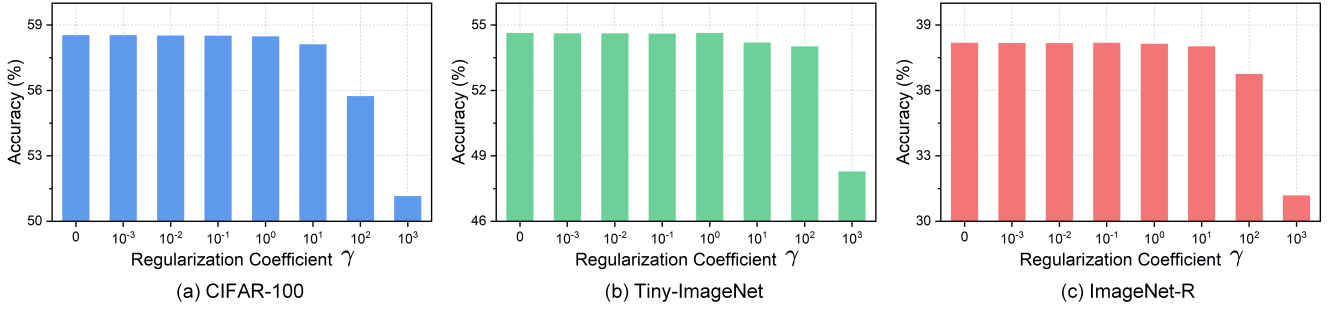


Figure 6. Sensitivity analyses for the regularization coefficient γ .

a logarithmic scale. Specifically, even compared with the most efficient baseline, FedAvg, AFCL still reduces runtime by over 94% across all three datasets, corresponding to an approximately $\times 18$ to $\times 43$ speedup. Furthermore, compared with the best-performing baseline, TagFCL, the efficiency advantage of our AFCL becomes even more pronounced, with runtime reductions exceeding 98% and speedups of roughly $\times 55$ to $\times 95$ across the three datasets.

Sensitivity Analyses. Finally, we show sensitivity analyses for our AFCL to its hyperparameter, i.e., the regularization coefficient γ , as shown in Figure 6. We can observe that the performance of our AFCL remains relatively stable across a wide range of the regularization coefficient γ , indicating that it can achieve strong performance without extensive hyperparameter tuning. Performance degradation occurs only when γ becomes excessively large, which leads to underfitting.

5. Conclusion and Discussion

In this paper, identifying the sensitivity of gradients as the root cause of spatio-temporal catastrophic forgetting in FCL, we propose our AFCL to fundamentally address this issue in a gradient-free manner. To the best of our knowledge, we are the first to posit and successfully achieve the ideal pursuit of spatio-temporal invariance in FCL. Through meticulous technical design, our AFCL can accommodate a wide variety of challenging settings (e.g., class-continual scenarios) and conditions (e.g., stragglers and dropouts) in FCL. Moreover, our AFCL also exhibits superior interpretability with

great efficiency, making it highly suitable for edge environments in FCL. Thus, our AFCL promotes a new direction for achieving high-reliability and low-cost deployment in FCL.

Since the adoption of pre-trained models is a common practice among FL, CL, and FCL, the backbone in our AFCL can be easily downloaded/distilled from a public model, or trained/fine-tuned centrally on public datasets, without risking private data leakage. Unlike many recent FCL methods that are strictly limited to heavy ViT backbones [41, 51, 65], our AFCL supports arbitrary backbone architectures (e.g., lightweight ResNet), offering better adaptability for FCL. In addition, while our AFCL focuses on the most common task in FCL, its potential is not limited. Due to the broad adaptability of analytic learning, our AFCL can be extended to many other tasks in future work, such as graph learning [15], semantic segmentation [30], and multimodal learning [58]. Moreover, the latest technical advancements about analytic learning can also be incorporated into our AFCL for further boosting its performance in future work [18, 23, 45, 68].

Acknowledgment

This work is supported partly by the National Natural Science Foundation of China (NSFC) under grant 62576013, the National Key Research and Development of China under grant 2024YFC2607404, the Jiangsu Provincial Key Research and Development Program under Grants BE2022065-1, BE2022065-3, and the Ningxia Domain-Specific Large Model Health Industry R&D Program 2024JBGS001.

References

- [1] Sara Babakniya, Zalan Fabian, Chaoyang He, Mahdi Soltanolkotabi, and Salman Avestimehr. A data-free approach to mitigate catastrophic forgetting in federated class incremental learning for vision tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3
- [2] Zhenjiao Cai, Sulan Zhang, Ping Guo, Jifu Zhang, and Lihua Hu. A progressive stacking pseudoinverse learning framework via active learning in random subspaces. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(5):2822–2832, 2024. 3
- [3] Huancheng Chen and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. *Advances in Neural Information Processing Systems*, 37:65525–65561, 2024. 1, 2
- [4] Hong-You Chen, Cheng-Hao Tu, Ziwei Li, Han Wei Shen, and Wei-Lun Chao. On the importance and applicability of pre-training for federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3
- [5] Marcos F Criado, Fernando E Casado, Roberto Iglesias, Carlos V Regueiro, and Senén Barro. Non-iid data and continual learning processes in federated learning: A long road ahead. *Information Fusion*, 88:263–280, 2022. 1
- [6] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10164–10173, 2022. 1, 2, 3
- [7] Jiahua Dong, Hongliu Li, Yang Cong, Gan Sun, Yulun Zhang, and Luc Van Gool. No one left behind: Real-world federated class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2054–2070, 2024. 3, 6, 7
- [8] Alp Emre Durmus, Zhao Yue, Matas Ramon, Mattina Matthew, Whatmough Paul, and Saligrama Venkatesh. Federated learning based on dynamic regularization. In *International conference on learning representations*, 2021. 6, 7
- [9] Kejia Fan, Jialin Guo, Runsheng Li, Yuanye Li, Anfeng Liu, Jianheng Tang, Tian Wang, Mianxiong Dong, and Houbing Song. RMDF-CV: A reliable multi-source data fusion scheme with cross validation for quality service construction in mobile crowd sensing. *IEEE Transactions on Services Computing*, 18(1):399–413, 2024. 1
- [10] Kejia Fan, Yajiang Huang, Jingyu He, Feijiang Han, Jianheng Tang, Huiping Zhuang, Anfeng Liu, Tian Wang, Mianxiong Dong, Houbing Herbert Song, et al. CALM: A ubiquitous crowdsourced analytic learning mechanism for continual service construction with data privacy preservation. *ACM UbiComp 2025 / ACM IMWUT*, 9(2):1–30, 2025. 1, 2, 3
- [11] Tao Fan, Hanlin Gu, Xuemei Cao, Chee Seng Chan, Qian Chen, et al. Ten challenging problems in federated foundation models, 2025. 1
- [12] Zijian Gao, Wangwang Jia, Xingxing Zhang, Dulan Zhou, Kele Xu, Feng Dawei, Yong Dou, Xinjun Mao, and Huaimin Wang. Knowledge memorization and rumination for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3
- [13] Jialin Guo, Jianheng Tang, Anfeng Liu, Neal N. Xiong, Jie Wu, Xiaomin Ouyang, and Jun Huang. COOL: A cloud-fog federated learning system with multimodal isolated client data in edge networks. *IEEE Transactions on Mobile Computing*, 25(4):5199–5216, 2026. 1
- [14] Ping Guo, Michael R Lyu, and NE Mastorakis. Pseudoinverse learning algorithm for feedforward neural networks. *Advances in Neural Networks and Applications*, 1(321-326), 2001. 3
- [15] Zihao Guo, Qingyun Sun, Ziwei Zhang, Haonan Yuan, Huiping Zhuang, Xingcheng Fu, and Jianxin Li. GraphKeeper: Graph domain-incremental learning via knowledge disentanglement and preservation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 8
- [16] William W Hager. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989. 2, 3
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6
- [18] Run He, Di Fang, Yicheng Xu, Yawen Cui, Ming Li, Cen Chen, Ziqian Zeng, and Huiping Zhuang. Semantic shift estimation via dual-projection and classifier reconstruction for exemplar-free class-incremental learning. In *Forty-second International Conference on Machine Learning*, 2025. 3, 6, 8
- [19] Run He, Kai Tong, Di Fang, Han Sun, Ziqian Zeng, Haoran Li, Tianyi Chen, and Huiping Zhuang. AFL: A single-round analytic approach for federated learning with pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 6
- [20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 6
- [21] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2021. 3
- [22] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*, 2017. 3
- [23] Kai Jiang, Zhengyan Shi, Dell Zhang, Hongyuan Zhang, and Xuelong Li. Mixture of noise for pre-trained model-based class-incremental learning. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 3, 6, 8
- [24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [25] Songning Lai, Mingqian Liao, Zhangyi Hu, Jiayu Yang, Wenshuo Chen, Hongru Xiao, Jianheng Tang, Haicheng Liao, and

- Yutao Yue. Learning new concepts, remembering the old: Continual learning for multimodal concept bottleneck models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12314–12322, 2025. 3
- [26] Songning Lai, Mingqian Liao, Zhangyi Hu, Jiayu Yang, Wenshuo Chen, et al. Learning new concepts, remembering the old: Continual learning for multimodal concept bottleneck models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12314–12322, 2025. 3
- [27] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6
- [28] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning. In *Advances in Neural Information Processing Systems*, pages 38461–38474. Curran Associates, Inc., 2022. 6
- [29] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6, 7
- [30] Jiaxu Li, Rui Li, Jianyu Qi, Songning Lai, Linpu Lv, Kejia Fan, Jianheng Tang, Yutao Yue, Dongzhan Zhou, Yunhui Liu, and Huiping Zhuang. CFSSeg: Closed-form solution for class-incremental semantic segmentation of 2D images and 3D point clouds. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 247–256, 2025. 3, 8
- [31] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 6, 7
- [32] Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12820–12829, 2024. 2, 3
- [33] Yichen Li, Wenchao Xu, Yining Qi, Haozhao Wang, Ruixuan Li, and Song Guo. SR-FDIL: Synergistic replay for federated domain-incremental learning. *IEEE Transactions on Parallel and Distributed Systems*, 35(11):1879–1890, 2024. 3
- [34] Yichen Li, Haozhao Wang, Yining Qi, Wei Liu, and Ruixuan Li. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2025. 6
- [35] Yichen Li, Haozhao Wang, Wenchao Xu, Tianzhe Xiao, Hong Liu, Minzhu Tu, Yuying Wang, Xin Yang, Rui Zhang, Shui Yu, et al. Unleashing the power of continual learning on non-centralized devices: A survey. *IEEE Communications Surveys & Tutorials*, 2025. 1
- [36] Yichen Li, Yuying Wang, Haozhao Wang, Yining Qi, Tianzhe Xiao, and Ruixuan Li. FedSSI: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In *International Conference on Machine Learning*, pages 36151–36169. PMLR, 2025. 1, 3
- [37] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. 6, 7
- [38] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems*, pages 2351–2363. Curran Associates, Inc., 2020. 6
- [39] Yaoyao Liu, Yingying Li, Bernt Schiele, and Qianru Sun. Online hyperparameter optimization for class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8906–8913, 2023. 1
- [40] Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In *IJCAI*, pages 2182–2188, 2022. 1
- [41] Muhammad Anwar Masum, Mahardhika Pratama, Lin Liu, H Habibullah, and Ryszard Kowalczyk. Federated few-shot class-incremental learning. In *International Conference on Representation Learning (ICLR)*, pages 99479–99508, 2025. 2, 8
- [42] Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36:12022–12053, 2023. 2, 3
- [43] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 6, 7
- [44] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):1–50, 2023. 2, 3
- [45] Saleh Momeni, Changnan Xiao, and Bing Liu. AnaCP: Toward upper-bound continual learning via analytic contrastive projection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 6, 8
- [46] Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, and Gyeong-Moon Park. Online class incremental learning on stochastic blurry task boundary via mask and visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11731–11741, 2023. 6
- [47] John Nguyen, Jianyu Wang, Kshitiz Malik, Maziar Sanjabi, and Michael Rabbat. Where to begin? on the impact of pre-training and initialization in federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3
- [48] Hongyi Peng, Han Yu, Xiaoli Tang, and Xiaoxiao Li. FedCal: achieving local and global calibration in federated learning via aggregated parameterized scaler. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 1, 2
- [49] Hongming Piao, Yichen Wu, Dapeng Wu, and Ying Wei. Federated continual learning via prompt-based dual knowledge transfer. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3
- [50] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Confer-*

- ence on Computer Vision and Pattern Recognition (CVPR), 2017. 6, 7
- [51] Xuankun Rong, Jianshu Zhang, Kun He, and Mang Ye. CAN: Leveraging clients as navigators for generative replay in federated continual learning. In *Forty-second International Conference on Machine Learning (ICML)*, 2025. 2, 8
- [52] Jianheng Tang, Yajiang Huang, Kejia Fan, Feijiang Han, Jiaxu Li, Jinfeng Xu, Run He, Anfeng Liu, Houbing Herbert Song, Huiping Zhuang, and Yunhuai Liu. DeepAFL: Deep analytic federated learning. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. 2, 3
- [53] Minh-Tuan Tran, Trung Le, Xuan-May Le, Mehrtash Harandi, and Dinh Phung. Text-enhanced data-free approach for federated class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23870–23880, 2024. 1, 2, 3, 6
- [54] Jue Wang, Ping Guo, and Yanjun Li. Densepilae: a feature reuse pseudoinverse learning algorithm for deep stacked autoencoder. *Complex & Intelligent Systems*, pages 1–11, 2022. 3
- [55] Qiang Wang, Bingyan Liu, and Yawen Li. Traceable federated continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12872–12881, 2024. 6, 7
- [56] Yuan Wang, Huazhu Fu, Renuga Kanagavelu, Qingsong Wei, Yong Liu, and Rick Siow Mong Goh. An aggregation-free federated learning for tackling data heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26233–26242, 2024. 6
- [57] Canran Xiao et al. Confusion-resistant federated learning via diffusion-based data harmonization on non-iid data. *Advances in Neural Information Processing Systems*, 37:137495–137520, 2024. 1
- [58] Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8
- [59] Kunda Yan, Sen Cui, Abudukelimu Wuerkaixi, Jingfeng Zhang, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Balancing similarity and complementarity for federated learning. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 1, 2
- [60] Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, and Tianrui Li. Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):3832–3850, 2024. 1, 2, 3
- [61] Qian Yin, Bingxin Xu, Kaiyan Zhou, and Ping Guo. Bayesian pseudoinverse learners: From uncertainty to deterministic learning. *IEEE Transactions on Cybernetics*, 52(11):12205–12216, 2022. 3
- [62] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, pages 12073–12086. PMLR, 2021. 1, 2
- [63] Hao Yu, Xin Yang, Xin Gao, Yihui Feng, Hao Wang, Yan Kang, and Tianrui Li. Overcoming spatial-temporal catastrophic forgetting for federated class-incremental learning. *Proceedings of the 32nd ACM International Conference on Multimedia*, page 5280–5288, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 3, 6, 7
- [64] Hao Yu, Xin Yang, Xin Gao, Yan Kang, Hao Wang, Junbo Zhang, and Tianrui Li. Personalized federated continual learning via multi-granularity prompt. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 4023–4034, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 3, 6, 7
- [65] Hao Yu, Xin Yang, Le Zhang, Hanlin Gu, Tianrui Li, Lixin Fan, and Qiang Yang. Handling spatial-temporal data heterogeneity for federated continual learning via tail anchor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3, 8
- [66] Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. Target: Federated class-continual learning via exemplar-free distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4782–4793, 2023. 1, 2, 3, 6
- [67] Jianqiao Zhang, Caifeng Shan, and Jungong Han. FedGMKD: An efficient prototype federated learning framework through knowledge distillation and discrepancy-aware aggregation. *Advances in Neural Information Processing Systems*, 37:118326–118356, 2024. 1, 2
- [68] Xiang Zhang, Run He, Chen Jiao, Di Fang, Ming Li, Ziqian Zeng, Cen Chen, and Huiping Zhuang. L3A: Label-augmented analytic adaptation for multi-label class incremental learning. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 74938–74949. PMLR, 2025. 2, 6, 8
- [69] Yifei Zhang, Hao Zhu, Alysa Ziyang Tan, Dianzhi Yu, Longtao Huang, and Han Yu. pFedMxF: Personalized federated class-incremental learning with mixture of frequency aggregation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30640–30650, 2025. 2
- [70] Zhouyangzi Zhang, Bin Guo, Wen Sun, Yan Liu, and Zhiwen Yu. Cross-FCL: Toward a cross-edge federated continual learning framework in mobile edge computing systems. *IEEE Transactions on Mobile Computing*, 23(1):313–326, 2024. 1
- [71] Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8363–8371. International Joint Conferences on Artificial Intelligence Organization, 2024. Survey Track. 3
- [72] Yajie Zhou, Xiaoyi Pang, Zhibo Wang, Jiahui Hu, Peng Sun, and Kui Ren. Towards efficient asynchronous federated learning in heterogeneous edge environments. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, pages 2448–2457, 2024. 1, 2
- [73] Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. *Advances in Neural Information Processing Systems*, 35:11602–11614, 2022. 3, 6