sensors

MDPI

Article

Clustered Federated Spatio-Temporal Graph Attention Networks for Skeleton-Based Action Recognition

Tao Yu [1], Sandro Pinto [1], Tiago Gomes [1], Adriano Tavares [1] and Hao Xu [2,*]

[1] Centro Algoritmi, University do Minho, 4800-058 Guimarães, Portugal; id8670@alunos.uminho.pt (T.Y.); sandro.pinto@dei.uminho.pt (S.P.); mr.gomes@dei.uminho.pt (T.G.); atavares@dei.uminho.pt (A.T.)
[2] College of Computer Science and Technology, Jilin University, Changchun 130012, China
* Correspondence: xuhao@jlu.edu.cn

Abstract

Federated learning (FL) for skeleton-based action recognition remains underexplored, particularly under strong client heterogeneity where regular FedAvg tends to cause client drift and unstable convergence. We introduce Clustered Federated Spatio-Temporal Graph Attention Networks (CF-STGAT), a clustered FL framework that leverages attention-derived spatio-temporal statistics from local STGAT models to dynamically group clients and perform attention-weighted inter-cluster fusion that gently align cluster models. Concretely, the server periodically extracts multi-head parameter-based attention descriptors, normalizes and projects them via PCA, and applies K-means to form clusters; a global reference is then computed by attention–similarity weighting and used to regularize each cluster model with a lightweight fusion step. On NTU RGB+D 60/120(NTU 60/120), CF-STGAT consistently outperforms strong FL baselines with the STGAT backbone, yielding absolute top-1 gains of +0.84/+4.09 (NTU 60, X-Sub/X-Setup) and +7.98/+4.18 (NTU 120, X-Sub/X-Setup) over FedAvg, alongside smoother per-client trajectories and lower terminal test loss. Ablations indicate that attention-guided clustering and inter-cluster fusion are complementary: clustering reduces within-group variance whereas fusion limits cross-cluster divergence. The approach keeps local training unchanged and adds only server-side statistics and clustering.

Keywords: skeleton-based action recognition; spatio-temporal graph attention; clustered federated learning; inter-cluster regularization

check for updates

Academic Editor: Stefan Poslad

Received: 21 October 2025
Revised: 17 November 2025
Accepted: 27 November 2025
Published: 29 November 2025

Citation: Yu, T.; Pinto, S.; Gomes, T.; Tavares, A.; Xu, H. Clustered Federated Spatio-Temporal Graph Attention Networks for Skeleton-Based Action Recognition. *Sensors* **2025**, *25*, 7277. https://doi.org/10.3390/s25237277

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Skeleton-based human action recognition (HAR) is an active area in computer vision with applications in intelligent surveillance and robotic perception [1,2]. Given a sequence of 3D skeletal joints, the goal is to predict the action label by jointly modeling spatio-temporal configurations and dynamics. Early CNN/RNN approaches often treated skeletons as images or plain sequences and thus underutilized the explicit skeletal graph structure. Graph neural networks (GNNs) address this by operating on graph-structured data; the landmark ST-GCN [3] constructs a spatio-temporal skeleton graph and achieves strong centralized performance. Also, recent work [4] has demonstrated the effectiveness of spatial–temporal graph convolutional networks for skeleton-based activity recognition in practical applications. Subsequent advances enrich spatial topology modeling and temporal expressiveness through two-stream or channel-wise variants and multi-scale graph designs [5–7]. To capture long-range interactions across distant joints and timesteps,

the Spatio-Temporal Graph Attention Network (STGAT) provides interpretable attention patterns across joints and time, yielding compact structural descriptors [8]. However, most methods are developed and evaluated in centralized settings and do not address privacy constraints.

In practice, skeleton data are compartmentalized between devices or institutions, and privacy rules impede centralization. Federated learning (FL) enables collaborative training without sharing raw data, thereby preserving privacy and supporting personalization [9]. A baseline method is established by integrating STGAT with the standard FedAvg algorithm, referred to as Fed-STGAT (see Figure 1). This configuration preserves data privacy by sharing model parameters, thereby keeping the raw skeleton data streams on the client devices. Nevertheless, non-IID heterogeneity which span subjects, viewpoints, and recording setups induces client drift and unstable convergence when updates are naively averaged. In skeleton HAR the effect is exacerbated: the same action may manifest distinct dynamics across subjects and capture setups, inflating variance and degrading global averaging. Existing FL-HAR and clustered FL partially alleviate these issues. Prior efforts with regularized or contrasting objectives (e.g., FedProx [10], MOON [11]) constrain local training, while clustered FL (e.g., FedCLAR [12]) groups similar clients before aggregation to reduce cross-client variance and stabilize optimization. However, few methods explicitly leverage attention statistics a privacy-friendly, structure-aware signals for clustering and coordination. Compared to sample-level features, an attention descriptor is a low-dimensional statistic derived from the trained model that summarizes cross-joint and cross-time importance patterns and aligns naturally with the skeleton graphs.

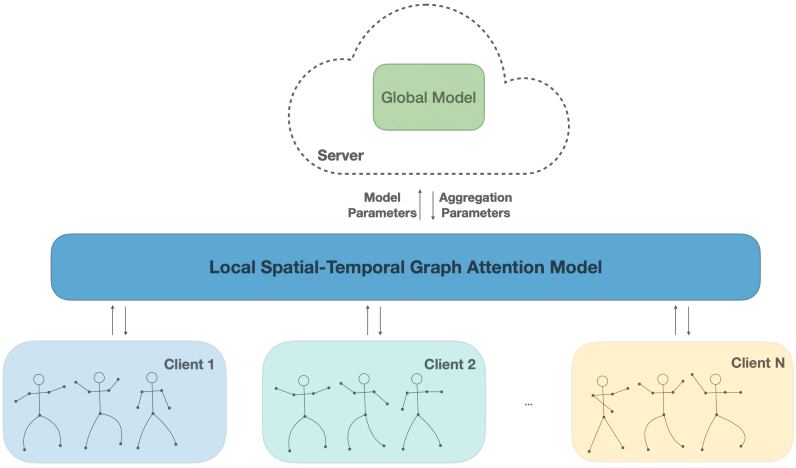


Figure 1. Overview of Fed-STGAT. Each client trains a local STGAT on its private skeleton sequences and periodically uploads model updates. The server aggregates the received updates to obtain the global model and broadcasts it back to clients for the next round. Raw data never leaves the clients, preserving data privacy.

We therefore propose Clustered Federated Spatio-Temporal Graph Attention Networks (CF-STGAT). Clients train STGAT locally; the server periodically extracts lightweight multi-head attention statistics and performs K-means to obtain several structure-aware, homogeneous clusters. Beyond standard within-cluster aggregation, we compute a global reference by attention–similarity weighting (via cosine similarity between cluster-level attention centroids and a global prototype) and apply a lightweight fusion to the intra-cluster aggregates. This design preserves personalization while mitigating cross-cluster drift and adds only minimal server-side statistical and clustering over-head.

The contributions of this paper are summarized as follows:

- We introduce an attention–descriptor-driven federated clustering paradigm for skeleton HAR, turning spatio-temporal attention into a privacy-friendly client representation for dynamic cluster assignment.
- We propose an attention–similarity-weighted cross-cluster fusion that balances personalization and consistency via a lightweight update, without changing the client-side training pipeline.
- We perform comprehensive experiments on NTU 60/120 with X-Sub/X-Setup, benchmarking against FedAvg, FedProx, MOON, and clustered baselines FedCLAR. CF-STGAT achieves consistent accuracy gains and smoother convergence.

2. Related Work

2.1. Spatio-Temporal Model for Skeleton-Based Action Recognition

Modeling spatio-temporal dependencies is central to skeleton-based action recognition. Beyond early CNN/TCN pipelines [13–15], recent 3D-CNNs designs revisit skeleton encoding with task-specific representations, for example, PoseConv3D [16] encodes poses as 3D heatmap volumes and achieves strong cross-dataset robustness. Graph-based methods continue to advance topology modeling: InfoGCN [17] integrates an information-bottleneck objective with attention-based graph convolution to infer context-dependent topology, HD-GCN hierarchically decomposes skeleton graphs to capture semantically meaningful edges [18], ref. [19] reviewed skeleton graph neural network-based human action recognition, and BlockGCN [20] redefines topology awareness via block-wise graph convolutions with improved efficiency. Transformer-style models further extend receptive fields, FG-STFormer [21] balances focal/local and global attention, STAR-Transformer employs cross-attention to fuse spatio-temporal cues [22], and SkateFormer partitions joints and frames to efficiently model skeletal–temporal relations [23]. While attention improves global dependency modeling, its computational cost can be high, efficient self-attention and local dynamic STGAT reduce over-head while maintaining accuracy [8,24]. Recent surveys synthesize the field's shift from RNN/LSTM to GCN/GAT/Transformer families and highlight the need for structure-aware, efficient, and privacy-preserving learning, which is the target setting of our framework CF-STGAT [25,26].

2.2. Federated Learning for Skeleton-Based Action Recognition

Federated learning (FL) enables privacy-preserving collaboration by keeping raw data on device and exchanging only model updates, thereby reducing both data-exposure risk and the costs of centralized collection; in practice, it is often paired with secure aggregation and lightweight communication protocols, making FedAvg [27] a widely adopted, communication-efficient baseline for deep models across modalities [28–30]. To address non-IID data, regularized and contrastive formulations improve robustness to statistical heterogeneity, FedProx [10] augments the local objective with a proximal term to constrain client drift under variable local steps, while MOON [11] introduces a model-contrastive objective that aligns client and global representations to reduce drift. Variance-reduction and normalization strategies such as SCAFFOLD and FedNova further correct client-specific bias and the objective inconsistency induced by unbalanced local updates [31,32], yet these methods remain structure-agnostic to skeleton topology. Furthermore, clustered and personalized schemes address client heterogeneity that is typical in skeleton HAR, where agglomerative clustered FL groups clients by update similarity to stabilize training [12,33], while structured FL via clustered additive modeling provides a principled formulation to reconcile cluster-wise personalization with global coordination [34]; however, they rely on generic similarity measures that do not explicitly reflect spatio-temporal skeleton dependencies. When skeleton streams and their topologies are explicitly modeled as graphs,

federated graph learning (FGL) [35] offers task-tailored aggregation and topology-aware updates, recent systems and algorithms demonstrate gains by combining global coordination with local graph-specific objectives [36,37]. Related cross-silo variants such as vertical FL and split learning are also relevant when different parties hold complementary features or wish to offload parts of the network; recent surveys and studies document privacy risks and corresponding defenses, as well as practical split feature-partitioned training pipelines that can be combined with FL for spatio-temporal data [38–40]. In contrast, we adopt FedAvg as the baseline privacy boundary and inject structure awareness through leveraging parameter-based attention features from STGAT to drive clustering and attention–similarity-weighted fusion, achieving cluster-wise personalization while gently regularizing cross-cluster drift.

3. Methods

3.1. Preliminary

We adopt STGAT as the local model due to its strong empirical performance in capturing short-term features for skeleton sequences via spatio-temporal patterns. Below, we formalize the graph construction and summarize the network components relevant to our method.

3.1.1. Spatio-Temporal Graph Building

Formally, the human skeleton graph is always represented as $G = (\mathcal{V}, A)$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes skeleton joints and $A \in \mathbb{R}^{N\times N}$ is the adjacency matrix. $A_{i,j} \neq 0$ indicates a connection between joint v_i and v_j, otherwise $A_{i,j} = 0$ without connections. Regarding the input, the action sequences are characterized by a set of node features, represented as a feature tensor $X \in \mathbb{R}^{C\times T\times N}$ where each joint is described by a C dimensional feature vector across T frames. To enhance the capture of local cross-spacetime joint relationships without necessitating multiple transmissions, the construction of a spatio-temporal graph is proposed in which joints within a localized cross-spacetime neighborhood are represented as nodes. These nodes are connected by edges not only to other nodes within their original spatio graph but also to their counterparts in adjacent temporal frames. For a given time point t and window size $\tau \in \mathbb{N}^*$, the concatenation of features within the time window can be expressed as $X_\tau^t = \left[X_{:,t-\frac{\tau}{2},:}; X_{:,t-\frac{\tau}{2}+1,:}; \ldots; X_{:,t+\frac{\tau}{2},:}\right] \in \mathbb{R}^{C\times\tau\times N}$. Then, a time-aware adjacency is defined as $\tilde{A}_\tau^t = \left[\tilde{A}_1^t; \ldots; \tilde{A}_\tau^t\right] \in \mathbb{R}^{\tau\times N\times N}$.

3.1.2. STGAT Network

As is widely recognized, for a graph G, with the node feature matrix X serving as the input, the formulation for a new feature matrix H derived from a layer within a graph convolutional network (GCN) at each timestep can be described by

$$H = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}XW\right), \tag{1}$$

The degree matrix $\tilde{D}$ corresponds to $\tilde{A}$ and is a diagonal matrix whose elements represent the sum of the connections, including self-loops, that incorporate its own features into the calculation. Also, W typically represents the trainable weight matrix, and σ denotes a non-linear activation function. In parallel to the development of the spatio-temporal graph, by integrating $\tilde{A}_\tau^t$ into graph operations, the output for each timestep input X_τ^t can be derived as

$$H_\tau^t = \sigma\left(\tilde{A}_\tau^t X_\tau^t W\right), \tag{2}$$

Furthermore, STGAT utilizes multi-head attention modules to learn various kinds of edge weights. Specifically, independent self-attention modules S are employed to discern diverse graph structures. Consequently, the aggregated output from all heads across various spatio-temporal graphs can be expressed as

$$Y^t = \sigma\left(\frac{1}{S}\sum_{s=1}^{S} \hat{A}^t_{\tau,s} X^t_\tau W_s\right), \tag{3}$$

where $\hat{A}^t_{\tau,s}$ denotes the unique spatio-temporal graph computed at timestamp t, with a temporal length of τ for the s_{th} head.

3.2. Fed-STGAT

As illustrated in Figure 1, the structure comprises N distributed clients and a central aggregation server. Each client trains a local STGAT on private data and uploads model parameters Θ after local updates. The server aggregates the received parameters and broadcasts the global model for the next round, repeating until convergence or a fixed communication budget.

3.2.1. Local Updates

For each client $i \in N$, at the beginning of each communication round, the local STGAT model is initialized with the global model parameters Θ^t received from the server. The model is then trained on the client's private dataset using multiple steps of SGD to perform parameter updates. The detailed procedure is as follows:

$$\Theta_i^{t+1} = \Theta^t - \eta \cdot \nabla_\Theta \mathcal{L}(\Theta^t; D_i). \tag{4}$$

where η is the learning rate and $\nabla_\Theta \mathcal{L}(\Theta^t; D_i)$ is the set of gradient updates of the local dataset on client i. Upon completion of the local training, each client uploads its updated parameters to the central server for aggregation.

3.2.2. Aggregation

The server performs global model aggregation when receiving the locally updated model parameters $\{\Theta_i^t\}_{i=1}^N$ from N participating clients. The global model is updated on FedAvg as follows:

$$\Theta^{t+1} = \sum_{i=1}^{N} \frac{n_i}{n} \Theta_i^t, \tag{5}$$

where n_i denotes the number of the local data samples on client i, and n is the total number of samples from all selected clients. The updated global parameters are redistributed to clients for the next round.

3.2.3. Limitation of Fed-STGAT

Under non-IID partitions, FedAvg can induce client drift and unstable convergence. The lack of structure-aware aggregation and personalization further limits cross-subject generalization. Thus, we introduce dynamic clustering in Section 3.3 and a light fusion step in Section 3.4 to mitigate drift and reduce inter-client variance.

3.3. Dynamic Cluster Adjustment

The approach extends FedAvg by grouping clients dynamically into clusters and training cluster-specific models. To obtain a structure-aware similarity, we extract attention-layer parameters from each client's STGAT as privacy-preserving features that reflect spatio-temporal patterns. Every I_c rounds, the server reclusters participating clients based on these

features. Before clustering, we normalize the feature vector of each client $\mathbf{a}_i = [a_{i1}, \ldots, a_{im}]$ by Z-score, then we fit a PCA model on the normalized attention vectors to reduce the dimensionality from m to d automatically by retaining 90% of the cumulative explained variance. This yields low-dimensional descriptors that preserve most of the information in the original attention parameters while reducing noise and improving the efficiency and stability of the subsequent clustering step. Having assembled the matrix $\widetilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_n]$, the empirical covariance matrix is computed as $\Sigma = \frac{1}{n}\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}}$. Solving the eigenproblem $\Sigma \mathbf{U}_d = \mathbf{U}_d \Lambda_d$ yields the top-deigenvectors $\mathbf{U}_d = [\mathbf{u}_1, \ldots, \mathbf{u}_d]$ associated with the largest eigenvalues in Λ_d. Each normalized vector $\tilde{\mathbf{a}}_i$ is then projected as

$$\mathbf{z}_i = \mathbf{U}_d^\top \tilde{\mathbf{a}}_i, \tag{6}$$

yielding the final d-dimensional clustering feature for client i. If the current round is one of the designated clustering intervals I_c, the server uses the collected feature vectors from all participating clients to recompute the client clusters. If a client is assigned to another cluster, it means the characteristics of its model have shifted or the clustering decision boundary has changed.

We adopt a K-means approach for the clustering algorithm due to its simplicity and efficiency in the federated setting, partitioning the clients into C clusters by minimizing the within-cluster variance in the feature space. As shown in Figure 2, the outcome is a set of clusters $C_1, C_2, \ldots, C_C$, where clients within each cluster have high internal similarity in their attention features $\mathbf{z}_i$. After clustering, each cluster C_k is associated with its own sub-model. Clients do not remain tied to a single cluster for the entire training; they are allowed to change clusters if their attention-based representation shifts or the clustering boundaries are refined over time.

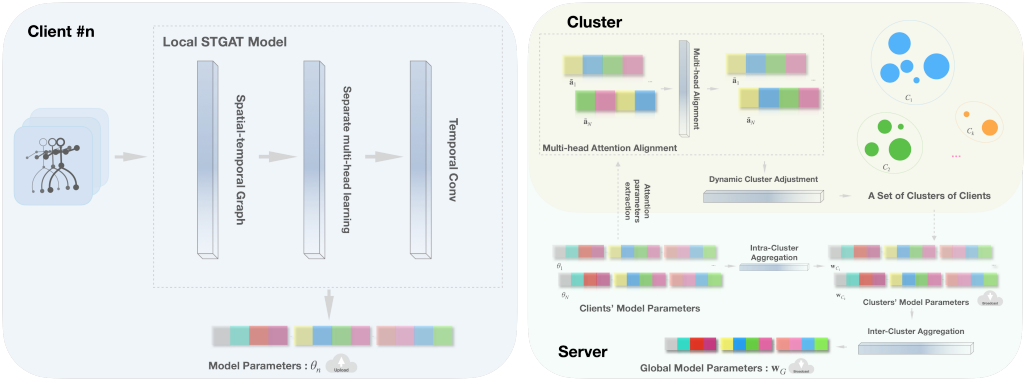


Figure 2. CF-STGAT pipeline. (**Left**) Each client trains a local STGAT and uploads model parameters $\boldsymbol{\theta}_n$ after E local steps. (**Right, top**) The server extracts and normalizes attention statistics and performs multi-head attention alignment; every I_c rounds, a dynamic clustering partitions clients into $\{C_1, \ldots, C_C\}$. (**Right, bottom**) Inter-cluster aggregation forms an attention-weighted global reference $\boldsymbol{w}_G$ and updates each cluster model by $\boldsymbol{w}_{C_k} \leftarrow (1-\gamma)\boldsymbol{w}_{C_k} + \gamma\, \boldsymbol{w}_G$ before the next round. Notation: $\gamma \in [0, 1]$ fusion coefficient; E local steps; I_c reclustering interval.

Complexity note. Let N denote the number of participating clients in a clustering step and m the dimensionality of the raw attention features. Computing the covariance matrix and its eigendecomposition for PCA on the $N \times m$ matrix $\widetilde{\mathbf{A}}$ has complexity $\mathcal{O}(Nm^2 + m^3)$. Since m is a small, fixed hyperparameter in our setup, this cost is effectively linear in N. Performing K-means in the reduced d-dimensional space has complexity $\mathcal{O}(NCId)$, where C is the number of clusters and I is the number of iterations. In practice, m, d, C, and I are all small constants, and clustering is only performed every I_c rounds on the subset of active clients, so the server-side over-head is negligible compared to local STGAT training and communication.

3.4. Cluster Federated Aggregation

While clustered federated learning provides personalization by clustering models to groups of similar clients, it is important to prevent these cluster-specific models from diverging too far from each other, which could harm overall generalization. Therefore, for Cluster C_k, the server forms a size-weighted average $\bar{\mathbf{w}}_{C_k}$ of the received client models; for inter-cluster regularization aggregation, to maintain global coherence between clusters and prevent cluster models from diverging excessively.

We construct a global attention prototype $\mathbf{a}_G$ from cluster attention centroids and compute cosine similarities $s_k = cosine(\mathbf{a}_k, \mathbf{a}_G)$ between cluster C_k. Attention-aware weights

$$\alpha_k = \frac{N_k\, s_k}{\sum_j N_j\, s_j}, \quad \sum_k \alpha_k = 1. \tag{7}$$

which yield a reference model:

$$\mathbf{w}_G^{(t)} = \sum_k \alpha_k \bar{\mathbf{w}}_{C_k}^{(t)}. \tag{8}$$

Each cluster model is then updated by a light fusion:

$$\bar{\mathbf{w}}_{C_k}^{(t+1)} \leftarrow (1-\gamma)\, \bar{\mathbf{w}}_{C_k}^{(t)} + \gamma\, \mathbf{w}_G^{(t)}, \tag{9}$$

with γ controlling inter-cluster regularization. Instead of treating the fusion coefficient γ as a fixed hyperparameter, we adapt it automatically at each communication round based on the discrepancy between the previous global model and the current cluster models. In practice, this makes γ increase when cluster models deviate more from the global model (encouraging stronger fusion) and decrease when the models are already similar (allowing more specialization). This step preserves local training pipelines while aligning clusters through periodic server-side coordination to effectively balance model personalization with overall consistency in a federated learning setting.

The above formulation can be understood as follows. The server clusters clients through similar attention-based descriptors which form the uploaded spatio-temporal structure, so that each cluster-wise model is trained on a group of clients that share similar dependency patterns. The fusion regularization term softly couples these cluster-specific models, allowing them to share transferable knowledge while still adapting to the particular characteristics of each cluster. In this way, the method balances global sharing and local specialization in a structure-aware manner.

4. Experimental Results

4.1. Datasets

NTU 60 [41]. A widely adopted benchmark for 3D skeleton-based human action recognition. Captured with Microsoft Kinect v2 sensors in a controlled laboratory setting, it comprises 56,880 samples across 60 action classes performed by 40 subjects. Each sample encodes a sequence of 25 skeletal joints, with each joint represented by its (x, y, z) coordinates in 3D space. The dataset offers two standardized evaluation protocols: cross-subject (X-Sub), which trains on half of the subjects and tests on the other half, and cross-view (X-View). We follow the standard X-Sub split; for X-View, we instead use a setup-based split for stress-testing distribution shift.

NTU 120 [42]. An extended version of NTU 60 that includes 120 action classes, each performed twice by 60 subjects (40 male, 20 female) across four distinct recording setups. NTU 120 retains the same data modalities as NTU 60, offering synchronized RGB and depth streams along with the core 3D skeleton data (25 joints per frame). NTU 120 introduces the Cross-Setup (X-Setup) protocol: even-ID setups for training and odd-ID setups for testing.

Each setup features a distinct location and background and, as in NTU 60, it employs three horizontally spaced cameras to capture each action from three different side-view angles simultaneously.

We also adopt a federated-by-dataset setting, where federated clients are constructed from the underlying dataset structure. Concretely, for both NTU 60 and NTU 120, each $(\text{subject}, \text{setup})$ pair thus defines a federated client, inducing a naturally non-IID label and setup distribution across clients. The statistics of federated data partitions are shown in Table 1. The central server evaluates on a global test set defined over subject–setup pairs. For NTU 60, in addition to the standard X-Sub, we define a cross-setup protocol (NTU 60 X-Setup) to stress-test distribution shifts across recording setups: following the rule of NTU 120 X-Setup, sequences from even-numbered setups are used for training and those from odd-numbered setups for testing.

Table 1. Statistics of the federated data partitions on NTU 60 and NTU 120.

Dataset	Protocol	Classes	Train Clients	Test Clients
NTU 60	X-Sub	60	20	20
NTU 60	X-Setup	60	8	9
NTU 120	X-Sub	120	53	53
NTU 120	X-Setup	120	16	16

4.2. The Details of Implementation

All experiments are conducted using the PyTorch 1.12.1 on a single NVIDIA RTX 4090 GPU. The backbone network is based on STGAT, consisting of eight graph attention layers, each with eight attention heads. We use batch size 32 and weight decay 5×10^{-4}. From each sequence, we first sample 150 frames and then randomly crop to 128 frames for both training and testing to match sequence length. Within FL, each client performs $E = 1$ local epoch per communication cycle. The initial learning rate is set to 0.1 and is reduced by a factor of 10 at communication rounds $\{60, 90, 120\}$, respectively. The total number of communication rounds is fixed at $R = 200$. For clustering, the interval is $I_c = 5$ communication rounds. We set $C = 4$ on NTU 60 and $C = 6$ on NTU 120. The fusion regularization coefficient γ is selected from the set $\{0.1, 0.2, 0.5\}$ based on a small grid search.

4.3. Experiments Results and Discussion

4.3.1. General Results

Table 2 presents a comparative performance analysis on NTU 60/120 under the X-Sub and X-Setup protocols with STGAT as the backbone. (1) CF-STGAT achieves the best overall performance, surpassing FedAvg by +0.84 (NTU 60, X-Sub), +4.09 (NTU 60, X-Setup), +7.98 (NTU 120, X-Sub), and +4.18 (NTU 120, X-Setup), yielding an average gain of +4.27 percentage points (pp). The largest single-cell improvement occurs on NTU 120 (X-Sub) (+7.98). (2) CF-STGAT also outperforms FedCLAR across all settings, with gains of +0.68, +3.31, +4.78, and +2.06 pp on the same four protocol–dataset combinations (mean +2.71 pp). Notably, the margin is largest on NTU 120 (X-Sub) (+4.78). (3) CF-STGAT Improvements over FedAvg are larger on NTU 120 (mean +6.08 pp) than on NTU 60 (mean +2.47 pp), suggesting greater benefits under the more heterogeneous setting. Overall, relative to optimizer-based baselines (FedAvg, FedProx, MOON) and clustered FL (FedCLAR), CF-STGAT attains the highest accuracy across all protocols, with robust gains on NTU 120.

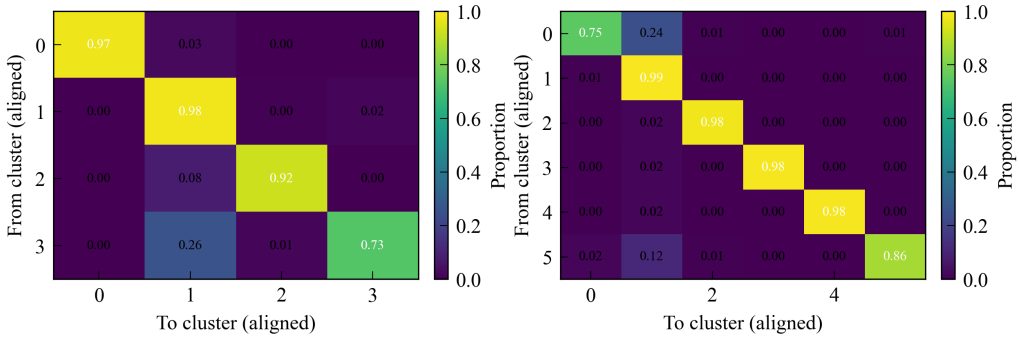
Table 2. Overall comparison on NTU RGB+D 60/120 under X-Sub and X-Setup. We report top-1 accuracy (%) on each protocol. Bold denotes the best. Values in parentheses indicate the absolute margin vs. FedAvg (baseline row marked with *).

Model Backbone: STGAT	**NTU 60**		**NTU 120**	
	X-Sub (%)	**X-Setup (%)**	**X-Sub (%)**	**X-Setup (%)**
Fed-STGAT *	89.72	86.46	72.35	78.71
FedProx	89.31 (−0.41)	87.49 (+1.03)	75.36 (+3.01)	80.65 (+1.94)
MOON	88.01 (−1.71)	86.96 (+0.50)	73.31 (1.04)	79.86 (+1.15)
FedCLAR	89.88 (+0.16)	87.24 (+0.78)	75.55 (+3.20)	80.83 (+2.12)
CF-STGAT (our)	**90.56** (+0.84)	**90.55** (+4.09)	**80.33** (+7.98)	**82.89** (+4.18)

4.3.2. Clustering Stability Analysis

To analyze how client cluster assignments evolve over communication rounds, we measure round-to-round agreement between consecutive reclustering checkpoints using the Adjusted Rand Index (ARI) and report a churn rate (the fraction of clients that change clusters between two checkpoints). Across NTU 60 X-Sub configurations, the reclustering dynamics stabilize after a short warm-up. As shown in Table 3, the $I_c = 5$, $C = 4$ variant yields the strongest consistency (ARI = 0.846 ± 0.213) and the lowest churn (0.040 ± 0.056), outperforming $C = 3$ (ARI = 0.674 ± 0.326; churn = 0.064 ± 0.077) and $I_c = 3$ (ARI = 0.750 ± 0.236; churn = 0.064 ± 0.064). This indicates that too small C conflates heterogeneous clients, producing fuzzy boundaries. Conversely, when C is too large, semantically similar clients are over-partitioned, leading to cluster fragmentation and elevated churn. Also, overly frequent reclustering (smaller I_c) can introduce transient boundary oscillations, even if the fraction of migrating clients is moderate, the induced pairwise disagreements depress ARI.

In addition, we visualize the row-normalized average transition matrix to reveal where changes concentrate. The transition matrices in Figure 3 are strongly diagonal on both datasets, confirming persistent cluster identities. On NTU 120, clusters 1–4 exhibit near-perfect self-transition ($\geq$0.98), with localized drift from $0 \to 1$ (0.24) and $5 \to 1$ (0.12), that means a single fuzzy boundary while the rest remain stable. On NTU 60, overall stability is also high, but the boundary between clusters shows more activity: $3 \to 1 \approx 0.26$ and $2 \to 1 \approx 0.08$, consistent with the slightly lower ARI mean and the small but non-zero churn pulses. Overall, the table and heatmaps show that instability is concentrated at a few adjacent-cluster boundaries rather than reflecting global oscillations.



(**a**) Average Transition | NTU 60 (X-Sub) (**b**) Average Transition | NTU 120 (X-Sub)

Figure 3. Row-normalized average transition matrices across consecutive checkpoints. Values are averaged over all checkpoint pairs after Hungarian label alignment.

Table 3. Round-to-round clustering stability on NUT 60 (X-Sub). Values are mean ± std over consecutive checkpoint pairs.

Setup	ARI	Churn
$I_c = 5, C = 4$	0.846 ± 0.213	0.040 ± 0.056
$I_c = 5, C = 3$	0.674 ± 0.326	0.064 ± 0.077
$I_c = 3, C = 4$	0.750 ± 0.236	0.064 ± 0.064

4.3.3. Effectiveness of Clustering

As shown in Figure 4, in the absence of clustering (treating all clients as a single group), the baseline method FedAvg and its regularized extensions FedProx and MOON exhibit slow and uneven convergence under non-IID partitions. Training a single global model with FedAvg on highly non-IID clients can induce large variance in the aggregated update, since local gradients are optimized for very different data distributions. This variance manifests as oscillatory behavior of the global model and unstable performance across communication rounds. Although MOON accelerates the initial rise, the jagged behavior remains, indicating that optimizer-level regularization alone is insufficient to eliminate client drift. Consistently, the test loss shown in Figure 5 decreases more slowly and stabilizes at a higher asymptote, particularly on NTU 120, where label and domain shifts are stronger. By contrast, our clustering mechanism groups clients with similar attention-based descriptors, i.e., similar spatio-temporal dependency patterns over the skeleton graph, so that each cluster-wise model is trained on a more homogeneous subset of clients. Within a cluster, local updates are better aligned and the variance of the aggregated update is reduced, which mitigates client drift and yields smoother training trajectories. The resulting cluster-specific models thus exhibit more stable convergence behavior than a single global model trained on all heterogeneous clients.

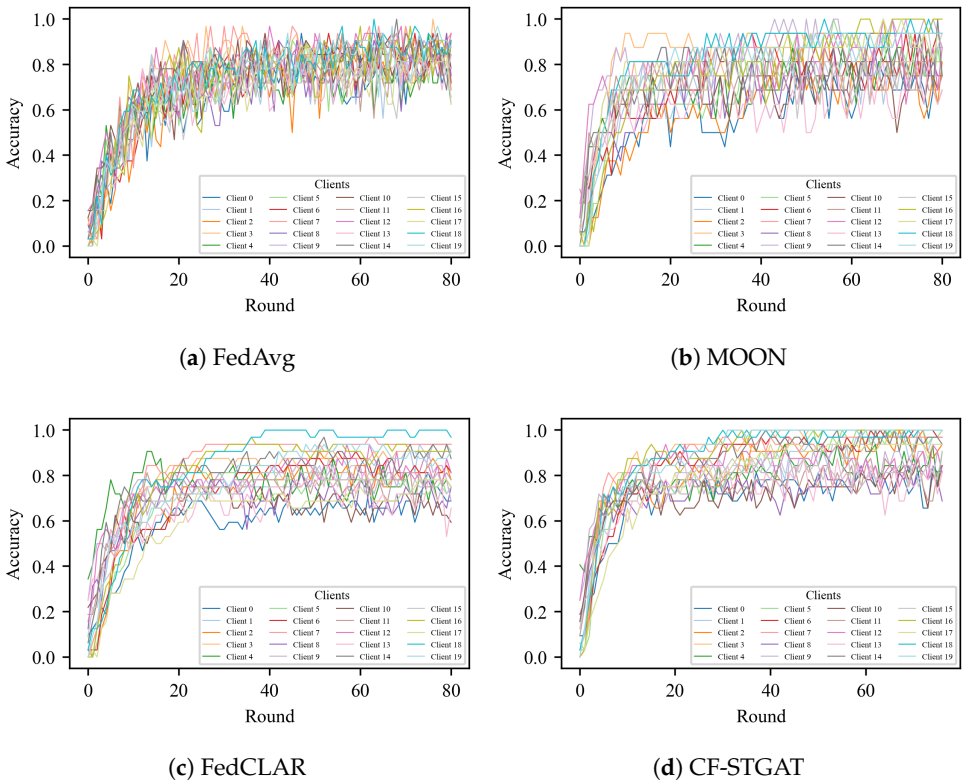
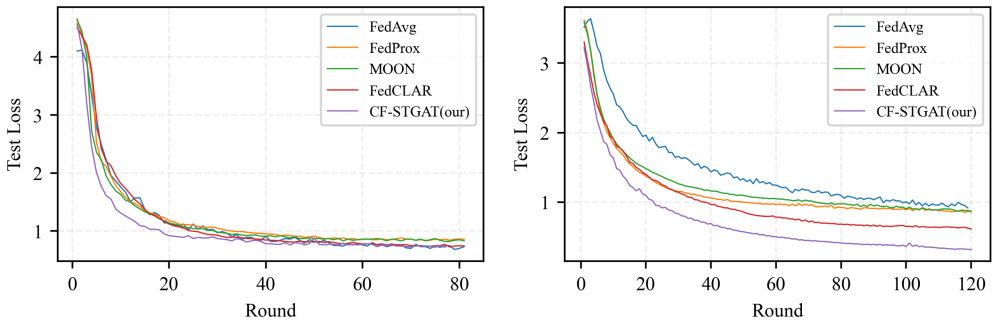


Figure 4. Per-client training accuracy vs. communication rounds on NTU RGB+D 60 (X-Sub). (**a**) FedAvg. (**b**) MOON. (**c**) FedCLAR (clustered-only). (**d**) CF-STGAT (clustering + global fusion). Each curve denotes individual clients; tighter bundles indicate lower dispersion.



(**a**) NTU 60 (X-Sub) (**b**) NTU 120 (X-Sub)

Figure 5. Test loss vs. communication rounds. (**a**) NTU 60 (X-Sub): CF-STGAT(ours) remains lowest throughout. (**b**) NTU 120 (X-Sub): separation is pronounced. CF-STGAT(ours) attains the fastest decay and lowest asymptote (lower is better).

4.3.4. Effectiveness of Fusion

With the clustered-only method FedCLAR, most clients indeed benefit from more homogeneous aggregation; however, the persistent low-accuracy tail and residual jaggedness indicate that inter-cluster drift is not fully controlled. Compared with our approach, this limitation is reflected in a higher late-stage test loss (Figure 5). By contrast, CF-STGAT, clustering augmented with global fusion, produces tighter and smoother client trajectories, elevates the plateau height, and compresses the long tail (Figure 4). On the test set, fusion yields faster loss decay and a lower terminal loss, with the largest margin observed on NTU 120. Overall, these results show that our method complements clustering by gently yet effectively aligning clusters, thereby delivering additional accuracy and stability beyond clustered-only federated learning.

4.3.5. Discussion and Future Work

A limitation of the present study is that we do not include explicit empirical comparisons with some recent personalized federated learning methods, such as FedPer [43] and FedRep [44]. These approaches adopt a complementary perspective: they typically share a global feature extractor while learning client-specific heads or representations, thereby personalizing predictions at the client level without explicitly modeling relationships among clients. In contrast, our method focuses on structure-aware clustering of clients based on attention-derived descriptors and trains cluster-specific STGAT models to capture shared spatio-temporal patterns within each group. This design is orthogonal to head- or representation-based personalization and could, in principle, be combined with FedPer/FedRep-style techniques by applying them within each discovered cluster. We leave a systematic empirical comparison and such hybrid combinations as important directions for future work. From a privacy perspective, the server only observes compressed attention-based descriptors aggregated over each client, rather than raw data or per-sample activations; while this reduces the risk of direct data leakage, our method is complementary to secure aggregation and differential-privacy mechanisms, and a formal privacy analysis of such descriptors is left for future work.

5. Conclusions

In summary, this work introduced CF-STGAT, a clustered federated framework for skeleton-based action recognition that integrates multi-head spatio-temporal attention with dynamic client clustering and a lightweight, attention-weighted inter-cluster fusion. The approach preserves cluster-level personalization while providing gentle cross-cluster align-

ment. Empirically, CF-STGAT delivers higher accuracy, smoother and faster convergence, and consistently lower terminal test loss, with the advantages becoming more pronounced under stronger heterogeneity. The method leaves client-side training unchanged and adds only modest server-side statistics and clustering.

Author Contributions: Conceptualization and study design: all authors. Methodology, implementation, experiments, and data analysis: T.Y. Supervision and project oversight: S.P., T.G., A.T., and H.X. Writing—original draft: T.Y. Writing—review and editing: all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (NSFC): "Research on Understanding Ancient Characters Based on Multi-modal Large Models" (No. 62476111), the Department of Science and Technology of Jilin Province: "Research and Application of Key Technologies for Cognitive Intelligence-Oriented Smart Education" (No. 20230201086GX), the Science and Technology Bureau of Changchun: "AI + Education: Large Model for Educational Evaluation" (No. 24GNYZ11), and the Industry University Research Innovation Fund of the Ministry of Education project "Research and Application of an Integrated Teaching Model for Human centered Artificial Intelligence" (Grant No. 2022XF017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in [NTU RGB+D] at [https://doi.org/10.1109/CVPR.2016.115], reference number [41].

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

References

1. Dai, C.; Lu, S.; Liu, C.; Guo, B. A light-weight skeleton human action recognition model with knowledge distillation for edge intelligent surveillance applications. *Appl. Soft Comput.* **2024**, *151*, 111166. [CrossRef]
2. Terreran, M.; Barcellona, L.; Ghidoni, S. A general skeleton-based action and gesture recognition framework for human–robot collaboration. *Robot. Auton. Syst.* **2023**, *170*, 104523. [CrossRef]
3. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018 ; Volume 32.
4. Xiao, L.; Yang, X.; Peng, T.; Li, H.; Guo, R. Skeleton-Based Activity Recognition for Process-Based Quality Control of Concealed Work via Spatial–Temporal Graph Convolutional Networks. *Sensors* **2024**, *24*, 1220. [CrossRef] [PubMed]
5. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
6. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13359–13368.
7. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 143–152.
8. Hu, L.; Liu, S.; Feng, W. Skeleton-based action recognition with local dynamic spatial–temporal aggregation. *Expert Syst. Appl.* **2023**, *232*, 120683. [CrossRef]
9. Li, C.; Niu, D.; Jiang, B.; Zuo, X.; Yang, J. Meta-har: Federated representation learning for human activity recognition. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 912–922.
10. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. In Proceedings of the Machine Learning and Systems, Austin, TX, USA, 2–4 March 2020; Volume 2, pp. 429–450.
11. Li, Q.; He, B.; Song, D. Model-contrastive federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10713–10722.

12. Presotto, R.; Civitarese, G.; Bettini, C. Fedclar: Federated clustering for personalized sensor-based human activity recognition. In Proceedings of the 2022 IEEE International Conference on Pervasive Computing and Communications (PerCom), Pisa, Italy, 21–25 March 2022; IEEE: New York, NY, USA, 2022; pp. 227–236.
13. Kim, T.S.; Reiter, A. Interpretable 3d human action analysis with temporal convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 1623–1631.
14. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165.
15. Nguyen, H.C.; Nguyen, T.H.; Scherer, R.; Le, V.H. Deep learning for human activity recognition on 3D human skeleton: Survey and comparative study. *Sensors* **2023**, *23*, 5121. [CrossRef] [PubMed]
16. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
17. Chi, H.g.; Ha, M.H.; Chi, S.; Lee, S.W.; Huang, Q.; Ramani, K. Infogcn: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20186–20196.
18. Lee, J.; Lee, M.; Lee, D.; Lee, S. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 10444–10453.
19. Feng, M.; Meunier, J. Skeleton graph-neural-network-based human action recognition: A survey. *Sensors* **2022**, *22*, 2091. [CrossRef] [PubMed]
20. Zhou, Y.; Yan, X.; Cheng, Z.Q.; Yan, Y.; Dai, Q.; Hua, X.S. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2049–2058.
21. Gao, Z.; Wang, P.; Lv, P.; Jiang, X.; Liu, Q.; Wang, P.; Xu, M.; Li, W. Focal and global spatial-temporal transformer for skeleton-based action recognition. In Proceedings of the Asian Conference on Computer Vision, Macau, China, 4–8 December 2022; pp. 382–398.
22. Ahn, D.; Kim, S.; Hong, H.; Ko, B.C. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 3330–3339.
23. Do, J.; Kim, M. Skateformer: Skeletal-temporal transformer for human action recognition. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Cham, Switzerland, 2024; pp. 401–420.
24. Qin, X.; Cai, R.; Yu, J.; He, C.; Zhang, X. An efficient self-attention network for skeleton-based action recognition. *Sci. Rep.* **2022**, *12*, 4111. [CrossRef] [PubMed]
25. Xin, W.; Liu, R.; Liu, Y.; Chen, Y.; Yu, W.; Miao, Q. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing* **2023**, *537*, 164–186. [CrossRef]
26. Ren, B.; Liu, M.; Ding, R.; Liu, H. A survey on 3d skeleton-based action recognition using learning method. *Cyborg Bionic Syst.* **2024**, *5*, 0100. [CrossRef] [PubMed]
27. 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.
28. Li, Q.; Wen, Z.; Wu, Z.; Hu, S.; Wang, N.; Li, Y.; Liu, X.; He, B. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3347–3366. [CrossRef]
29. Liu, B.; Lv, N.; Guo, Y.; Li, Y. Recent advances on federated learning: A systematic survey. *Neurocomputing* **2024**, *597*, 128019. [CrossRef]
30. Zhao, J.; Bagchi, S.; Avestimehr, S.; Chan, K.; Chaterji, S.; Dimitriadis, D.; Li, J.; Li, N.; Nourian, A.; Roth, H. The federation strikes back: A survey of federated learning privacy attacks, defenses, applications, and policy landscape. *ACM Comput. Surv.* **2025**, *57*, 1–37. [CrossRef]
31. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. Scaffold: Stochastic controlled averaging for federated learning. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 5132–5143.
32. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the objective inconsistency problem in heterogeneous federated optimization. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver BC Canada, 6–12 December 2020; Volume 33, pp. 7611–7623.
33. Mehta, M.; Shao, C. A greedy agglomerative framework for clustered federated learning. *IEEE Trans. Ind. Informatics* **2023**, *19*, 11856–11867. [CrossRef]

34. Ma, J.; Zhou, T.; Long, G.; Jiang, J.; Zhang, C. Structured federated learning through clustered additive modeling. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 43097–43107.
35. Chen, C.; Xu, Z.; Hu, W.; Zheng, Z.; Zhang, J. FedGL: Federated graph learning framework with global self-supervision. *Inf. Sci.* **2024**, *657*, 119976. [CrossRef]
36. Li, X.; Wu, Z.; Zhang, W.; Zhu, Y.; Li, R.H.; Wang, G. FedGTA: Topology-Aware Averaging for Federated Graph Learning. *VLDB Endow.* **2023**, *17*, 41–50. [CrossRef]
37. Huang, W.; Wan, G.; Ye, M.; Du, B. Federated graph semantic and structural learning. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 3830–3838.
38. Liu, Y.; Lou, Y.; Liu, Y.; Cao, Y.; Wang, H. Label leakage in vertical federated learning: A survey. In Proceedings of the IJCAI, Jeju, Republic of Korea, 3–9 August 2024.
39. Li, Z.; Yan, C.; Zhang, X.; Gharibi, G.; Yin, Z.; Jiang, X.; Malin, B.A. Split learning for distributed collaborative training of deep learning models in health informatics. *AMIA Annu. Symp. Proc.* **2024**, *2023*, 1047–1056. [PubMed]
40. Ye, M.; Shen, W.; Du, B.; Snezhko, E.; Kovalev, V.; Yuen, P.C. Vertical federated learning for effectiveness, security, applicability: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–32. [CrossRef]
41. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
42. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2684–2701. [CrossRef] [PubMed]
43. Arivazhagan, M.G.; Aggarwal, V.; Singh, A.K.; Choudhary, S. Federated learning with personalization layers. *arXiv Prepr.* **2019**, arXiv:1912.00818.
44. Collins, L.; Hassani, H.; Mokhtari, A.; Shakkottai, S. Exploiting shared representations for personalized federated learning. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 2089–2099.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.