Personalized Federated Learning for Spatio-Temporal Forecasting: A Dual Semantic Alignment-Based Contrastive Approach

Qingxiang Liu[1, 2], **Sheng Sun**[1], **Yuxuan Liang**[3], **Min Liu**[1, 4*], **Jingjing Xue**[1, 2]

[1]Institute of Computing Technology Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]The Hong Kong University of Science and Technology (Guangzhou)
[4]Zhongguancun Laboratory

Abstract

The existing federated learning (FL) methods for spatio-temporal forecasting fail to capture the inherent spatio-temporal heterogeneity, which calls for personalized FL (PFL) methods to model the spatio-temporally variant representations. While contrastive learning is promising in tackling spatio-temporal heterogeneity, the existing methods are noneffective in distinguishing positive and negative pairs and can hardly apply to PFL paradigm. To tackle this limitation, we propose a novel PFL method, named **F**ederated d**U**al s**E**mantic a**L**ignment-based contra**S**tive learning (FUELS), which can adaptively align positive and negative pairs based on semantic similarity, thereby injecting precise spatio-temporal heterogeneity into the latent representation space by auxiliary contrastive tasks. From temporal perspective, a hard negative filtering module is introduced to dynamically align heterogeneous temporal representations for the supplemented intra-client contrastive task. From spatial perspective, we design lightweight-but-efficient prototypes as client-level semantic representations, based on which the server evaluates spatial similarity and yields client-customized global prototypes for the supplemented inter-client contrastive task. Extensive experiments demonstrate that FUELS outperforms state-of-the-art methods, with impressive communication cost reduction.

1 Introduction

Spatio-temporal forecasting aims to predict the future trends with historical records from distributed devices. Given the masses of generated data, a better way is decentralized processing, thus significantly decreasing transmission latency (Meng, Rambhatla, and Liu 2021). Federated Learning (FL) enables distributed devices (termed *clients*) to collaboratively optimize a shared model without the disclosure of local data, which can release the concerns on privacy leakage and simultaneously achieve comparable performance with centralized learning methods (McMahan et al. 2017). Therefore, FL generates great promise for spatio-temporal forecasting tasks and a multitude of relevant methods have been proposed (Li, Li, and Wang 2021; Li and Wang 2022). Since the server can only have access to local model parameters, these FL methods cannot exploit a joint spatio-temporal correlation-capturing module like centralized methods (Li et al. 2018; Li and Zhu 2021). Generally, most of them decouple the spatio-temporal correlation, and evaluate spatial correlation in the server by Graph Neural Networks (Zhao et al. 2020; Lin et al. 2021) and temporal correlation at clients by Recurrent Neural Networks (Zhu and Wang 2021; Mahdy et al. 2020).

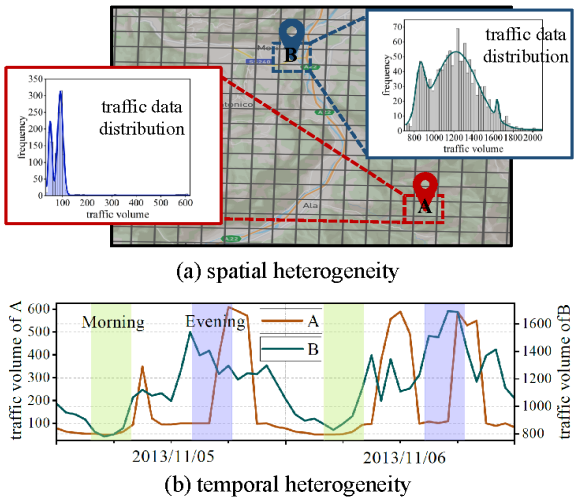

Figure 1: Spatio-temporal heterogeneity inside traffic flows.

Nevertheless, these FL forecasting methods cannot capture the inherent spatio-temporal heterogeneity by adopting shared model parameters across clients. As shown in Fig. 1, in wireless traffic prediction (Zhang et al. 2021), traffic patterns of two base stations (BSs) are spatio-variant (*traffic distribution of two clients varies*) and time-variant (*traffic patterns in the morning and evening are diverse*). The spatio-temporal heterogeneous traffic patterns call for personalized federated learning (PFL) methods to tackle with differentiated representations in the latent space from spatio-temporal level. However, the existing PFL methods mainly focus on increasing the generalization ability of local models by sharing diverse knowledge across clients (Karimireddy et al. 2020; Tan et al. 2022a; Chen et al. 2023), which can hardly work on coupled spatio-temporal heterogeneity.

Contrastive learning has emerged as an effective technique which shows great promise in improving PFL paradigm to address spatio-temporal heterogeneity. Generally, con-

*Min Liu is the corresponding author. E-mail: liumin@ict.ac.cn

trastive learning can minimize the agreement between disparate semantic representations (*negative pairs*), which can increase representations' differentiation, thereby injecting spatio-temporal heterogeneity into the latent representation space. Since there are no explicit semantic labels in spatio-temporal forecasting, it is a challenge to determine positive and negative pairs. In (Ji et al. 2023), for spatial heterogeneity, learnable embeddings are introduced for clustering regions' representations, which can hardly apply to PFL paradigm, due to the exposure of raw data and frequent communication in the process of parameter optimization. For temporal heterogeneity, region-level and city-level representations of the same time stamps make up positive pairs, otherwise negative pairs, which ignores semantic similarity among representations of different time stamps (termed *hard negatives*). Hence, **it remains a question to exactly determine negative and positive pairs when adopting contrastive learning in PFL paradigm for tackling spatio-temporal heterogeneity.**

To this end, we propose a novel PFL method, named **F**ederated d**U**al s**E**mantic a**L**ignment-based contra**S**tive learning (FUELS), which can dynamically establish positive and negative pairs based on spatio-temporal representations' similarity, hence guaranteeing the validity of spatio-temporal heterogeneity modeling. *From temporal perspective*, we propose a hard negative filtering module for each client, which can adaptively align true negative pairs for intra-client contrastive task so as to inject temporal heterogeneity into temporal representations. *From spatial perspective*, we define prototypes as client-level semantic representations which directly serve as the communication carrier, thus enabling sharing noise-robust knowledge across clients in a communication-efficient manner. We propose a Jensen Shannon Divergence (JSD)-based aggregation mechanism, which firstly aligns homogeneous and heterogeneous prototypes from clients and then yields client-customized global positive and negative prototypes for inter-client contrastive task to increase the spatial differentiation of local models. We summarize the key contributions as follows.

- To the best of our knowledge, this is the first PFL method towards spatio-temporal heterogeneity, where local training is enhanced by two well-crafted contrastive loss items to increase prediction models' ability of discerning spatial and temporal heterogeneity.
- We dynamically evaluate the heterogeneous degree among temporal representations and design a hard negative filtering module. Furthermore, we propose a JSD-based aggregation mechanism to generate particular global positive and negative prototypes for clients, striking a balance between sharing semantic knowledge and evaluating spatial heterogeneity.
- We validate the effectiveness and efficiency of the proposed FUELS from both theoretical analysis and extensive experiments.

2 Related Work

Federated Learning for Spatio-Temporal Forecasting. Due to the effectiveness of FL, there have been many works focusing on incorporating FL into spatio-temporal forecasting tasks (Zhang et al. 2021; Perifanis et al. 2023; Zhang, Zhang, and Shihada 2022). These FL methods aim to evaluate the decoupled spatial and temporal correlation at the server and at the clients respectively by split learning (Meng, Rambhatla, and Liu 2021), clustering (Liu et al. 2020; Zhang et al. 2021), online learning (Liu et al. 2023), and so on. However, all of these methods ignore the spatio-temporal heterogeneity and yield prediction results for different time stamps or different clients with the shared parameter space.

Personalized Federated Learning. PFL tackles with the problem of inference performance decline aroused by statistics heterogeneity across clients. The existing PFL methods can be divided into 2 categories, i.e., clients training a global model or training personalized local models. Methods in the first category aim to increase the generalization of the global model by designing client-selection strategy (Yang et al. 2021; Li et al. 2021) or adding proximal item to original functions (Karimireddy et al. 2020; Li, He, and Song 2021) In the second category, researchers modify the conventional FL procedure by network splitting (Arivazhagan et al. 2019; Bui et al. 2019; Liang et al. 2020), multitask learning (Smith et al. 2017; Ghosh et al. 2020; Xie et al. 2021), knowledge distillation (Li and Wang 2019; Zhu, Hong, and Zhou 2021; Lin et al. 2020) and so on. However, all of these works fail to solve the decoupled spatio-temporal heterogeneity.

Contrastive Learning. In contrastive learning, embeddings from similar samples are pulled closer and those from different ones are pushed away by constraining the loss function (Chen et al. 2020; He et al. 2020). Due to its effectiveness, contrastive learning has been applied to many scenarios, i.e., graph learning (Li et al. 2022; You et al. 2020; Zhu et al. 2021), traffic flow prediction (Ji et al. 2023; Yue et al. 2022; Woo et al. 2022), *etc*. Furthermore, some researches have focused on introducing contrastive learning into PFL paradigm mainly for handling statistics heterogeneity (Li, He, and Song 2021; Tan et al. 2022c; Yu et al. 2022; Mu et al. 2023). However, all of these methods focus on improving the performance on image classification. How to tackle with the spatio-temporal heterogeneity in PFL paradigm by contrastive learning remains an open problem.

3 Problem Formulation

Without loss of generality, we formulate the spatio-temporal forecasting problem for wireless traffic prediction, which is also applicable to other scenarios. Given N BSs as clients in the FL paradigm, the n-th BS has its traffic observations $\mathcal{V}_n = \left\{ v_n^k \,|k \in [1, K] \right\}$, where $v_n^k \in \mathbb{R}$ denotes the detected traffic volume of client n at the k-th time stamp. Based on sliding window mechanism (Zhang et al. 2021), $\mathcal{V}_n$ can be divided into input-output pairs (samples), which is denoted as $\mathcal{D}_n = \left\{ \left(\boldsymbol{x}_n^k; y_n^k \right) \right\}$. Thereinto, $y_n^k = v_n^k$ denotes the value to be predicted. $\boldsymbol{x}_n^k = \left(\boldsymbol{cv}_n^k, \boldsymbol{pv}_n^k \right)$, where $\boldsymbol{cv}_n^k \in \mathbb{R}^c, \boldsymbol{pv}_n^k \in \mathbb{R}^q$, and $\boldsymbol{x}_n^k \in \mathbb{R}^{c+q}$. $\boldsymbol{cv}_n^k = \left(v_n^{k-c}, \cdots, v_n^{k-2}, v_n^{k-1} \right)$ and $\boldsymbol{pv}_n^k = (v_n^{k-qp}, \cdots, v_n^{k-2p}, v_n^{k-p})$ denote two historical traffic sequences to evaluate the closeness and periodicity of traffic data. c, q, and p represent the size of close window, the size of periodic window, and the periodicity of traffic volume.

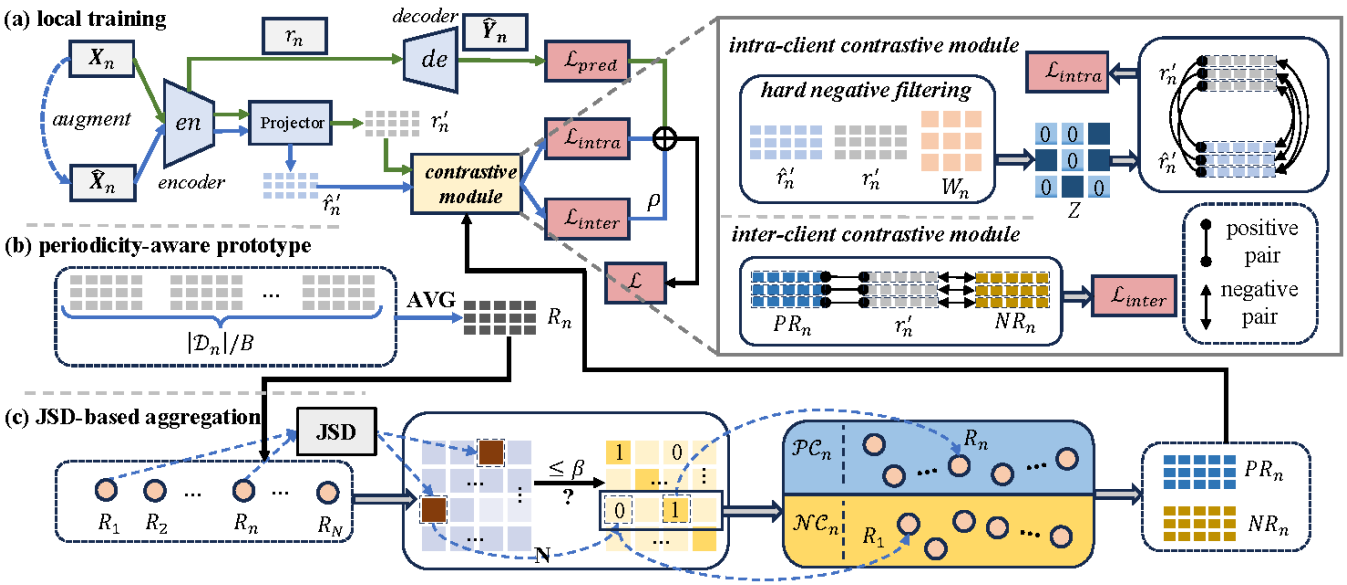

Figure 2: An overview of the proposed FUELS. (a) Each client performs local training by the supplemented inter- and intra-client contrastive loss items for spatio-temporal heterogeneity. (b) The designed periodicity-aware prototype works as the communication carrier. (c) The JSD-based aggregation generates client-customized global prototypes.

Since BSs are deployed at different locations, the detected traffic data across different BSs have diverse distributions. Therefore, in PFL paradigm, each BS tends to learn its own prediction model for performance improvement. The objective can be formulated as

$$\min_{\{w_n\}} \frac{1}{N} \sum_{n=1}^{N} \frac{|\mathcal{D}_n|}{D} \sum_{(\boldsymbol{x}_n^k, y_n^k) \in \mathcal{D}_n} \mathcal{L}\left(f\left(w_n; \boldsymbol{x}_n^k\right), y_n^k\right). \quad (1)$$

where $f(\cdot)$ represents prediction model. $\mathcal{L}(\cdot)$ denotes the loss function. $D = \sum_{n=1}^{N} |\mathcal{D}_n|$ denotes the total number of samples over all N clients. w_n denotes the model parameters of client n.

4 Methodology

We firstly provide an overarching depiction of FUELS, as is shown in Fig. 2. In macroscopic view, local training is enhanced by two designed contrastive loss items (Fig. 2 (a)). From temporal perspective, we propose a hard negative filtering module for true negative alignment. From spatial perspective, we employ prototypes as communication carrier (Fig. 2 (b)), and generate specific global prototypes (Fig. 2 (c)) for spatial heterogeneity evaluation. We elaborate the technical details in the following parts.

4.1 Prediction Model Architecture

(i) encoder $en(\delta_n; \boldsymbol{X}_n) : \mathbb{R}^{B\times(c+q)} \to \mathbb{R}^{B\times d_r}$, parameterized by δ_n, maps $\boldsymbol{X}_n \in \mathbb{R}^{B\times(c+q)}$ into a latent space with d_r dimensions, where $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ denotes a batch of training samples from $\mathcal{D}_n$ with batch size B. Actually, $\boldsymbol{X}_n = (\boldsymbol{V}_n^c, \boldsymbol{V}_n^p)$, where $\boldsymbol{V}_n^c \in \mathbb{R}^{B\times c}$ and $\boldsymbol{V}_n^p \in \mathbb{R}^{B\times q}$. we adopt two widely used Gated Recurrent Unit (GRU) models to evaluate the closeness and periodicity from $\boldsymbol{V}_n^c$ and $\boldsymbol{V}_n^p$, denoted as closeness-GRU ($\mathrm{GRU_c}$) and periodicity-GRU ($\mathrm{GRU_p}$) respectively. $\mathrm{GRU_c}(\delta_n^c; \boldsymbol{V}_n^c)$: $\mathbb{R}^{B\times c} \to \mathbb{R}^{B\times \frac{d_r}{2}}$, is parameterized by δ_n^c. $\mathrm{GRU_p}(\delta_n^p; \boldsymbol{V}_n^p)$: $\mathbb{R}^{B\times q} \to \mathbb{R}^{B\times \frac{d_r}{2}}$, is parameterized by δ_n^p. We denote the representation of $\boldsymbol{X}_n$ output by the encoder en as $r_n \in \mathbb{R}^{B\times d_r}$, which can be formulated as $r_n = en(\delta_n; \boldsymbol{X}_n) = \textbf{concat}\left[\mathrm{GRU_c}(\delta_n^c; \boldsymbol{V}_n^c); \mathrm{GRU_p}(\delta_n^p; \boldsymbol{V}_n^p)\right]$.

(ii) decoder $de(\phi_n; \cdot) : \mathbb{R}^{B\times d_r} \to \mathbb{R}^{B\times 1}$, a linear layer parameterized by ϕ_n, outputs the final predicted result $\hat{\boldsymbol{Y}}_n$ from the representation, i.e., $\hat{\boldsymbol{Y}}_n = de(\phi_n; r_n)$.

(iii) projector $pr(\theta_n; \cdot) : \mathbb{R}^{B\times d_r} \to \mathbb{R}^{B\times d_p}$, a linear layer parameterized by θ_n, transformed the representations to the targeted prototype dimension d_p. We denote $r'_n = pr(\theta_n; r_n)$.

4.2 Intra-Client Contrastive Task for Temporal Heterogeneity

In (Liu et al. 2022), hard negatives are filtered out based on traffic closeness but those out of the preset closeness scope can still form negative pairs, which may perturb the latent semantic space. Therefore, we propose a hard negative filtering module to align representations by semantic similarity and then design a intra-client contrastive task to maximize the divergence of negative pairs, thereby injecting precise temporal heterogeneity into temporal representations.

Firstly, we adopt the temporal shifting manner in (Liu et al. 2022) to generate the augmented dataset for client n, which is denoted as $\hat{\mathcal{D}}_n$. For $\boldsymbol{X}_n$, the corresponding augmented batch is denoted as $\hat{\boldsymbol{X}}_n$. We denote $\hat{r}'_n = pr(en(\hat{\boldsymbol{X}}_n))$ and $\hat{r}'_n \in \mathbb{R}^{B\times d_p}$. The procedure of hard negative filtering is formulated as

$$SM = \exp\left(\mathrm{sim}\left(r'_n, \hat{r}'_n\right)/\tau\right); Z = ReLu\left(SM \odot W_n\right),$$

where $SM \in \mathbb{R}^{B\times B}$ denotes the similarity matrix, $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity and τ denotes the temperature factor. $W_n \in \mathbb{R}^{B\times B}$ denotes the learnable filtering matrix and $\odot$ denotes the Hadamard product. Let $r'_{n,i} \in \mathbb{R}^{d_p}$ denote the i-th row of r'_n, i.e., the temporal representation of the i-th time stamp of $\boldsymbol{X}_n$. If $Z_{b,i} = 0(i \neq b)$, we treat $\hat{r}'_{n,i}$ as a hard negative of $r'_{n,b}$ (The augmented representation at time step i is a hard negative of the raw representation at time step b.). If $Z_{b,i} > 0$, $\hat{r}'_{n,i}$ is seen as a true negative of $r'_{n,b}$ and i is added into $\mathcal{TN}_b$. Ideally, $Z_{b,b}(1 \leq b \leq B)$ should be equal to 0, as we can also filter out the positive pairs (detailed results in Section 6.3).

By this procedure, (raw and augmented) representations with similar semantics at different or same time stamps can be filtered out. The subsequent contrastive task will repel the representations with differing semantics (true negatives) so as to validly inject temporal heterogeneity into representations. We can obtain the intra-client contrastive loss item as

$$\mathcal{L}_{intra} = \frac{1}{B}\sum_{b=1}^{B} -\log \frac{SM_{b,b}}{SM_{b,b} + \sum_{i\in\mathcal{TN}_b} SM_{b,i}}, \quad (2)$$

where $SM_{b,b}$ represents the similarity of positive pair and $SM_{b,i}$ denotes that of negative pair. Under the constraint of $\mathcal{L}_{intra}$, the prediction model will generate traffic representations with great temporal distinguishability.

4.3 Inter-Client Contrastive Task for Spatial Heterogeneity

Local Prototype Definition. In conventional FL paradigm, model parameters serve as communication carriers which are not compact and hardly preserve well the spatial heterogeneity. Inspired by prototype learning, where prototypes aggregate the representations of samples with the same labels and thus can carry the class-specific semantic knowledge, we can extract the client-level prototype to carry the client-specific knowledge.

We start with a concatenation-based prototype, which is formulated as

$$R_n = \textbf{concat}\left[r'_n \,|r'_n = pr(en\left(\boldsymbol{X}_n\right)), (\boldsymbol{X}_n, \boldsymbol{Y}_n) \in \mathcal{D}_n\right],$$

where R_n denotes the local prototype of client n. $\mathcal{D}_n$ can be divided into $|\mathcal{D}_n|/B$ batches, with batch size equal to B and hence $R_n \in \mathbb{R}^{|\mathcal{D}_n|\times d_p}$. Each client should transmit $|\mathcal{D}_n| \times d_p$ parameters to the server at each training round, which potentially incurs more communication overhead than conventional FL methods with $\mathcal{D}_n$ increasing.

Given the periodicity of traffic data, if we set the batch size B equal to the traffic periodicity p, representations of different batches will have similar distribution, as they all carry the traffic knowledge within a periodicity. Therefore, we can fuse these representations from different batches to obtain the **periodicity-aware prototype** as

$$R_n = \textbf{AVG}\left[r'_n \,|r'_n = pr(en\left(\boldsymbol{X}_n\right)), \forall\, (\boldsymbol{X}_n, \boldsymbol{Y}_n) \in \mathcal{D}_n\right],$$

where **AVG** represents the averaging operation. Therefore, $R_n \in \mathbb{R}^{B\times d_p}$. When employing periodicity-aware prototype as communication carrier, the number of communication parameters keeps stable regardless of local dataset size, and is significantly fewer than the parameter size of local prediction model (see Section 6.1). Furthermore, compared with the aforementioned concatenation-based prototype, the periodicity-aware prototype is less affected by traffic noise and thus contains more available client-specific knowledge. *Therefore, in FUELS, we adopt the periodicity-aware prototype as communication carrier, unless otherwise stated.*

JSD-Based Aggregation. The diversity of traffic data across clients results in the distribution difference over $R_n(1 \leq n \leq N)$, which can be evaluated by JSD at the server. Let $JS(R_n||R_m) \in \mathbb{R}$ denote the JSD value between R_n and R_m. According to the calculation logic, if the heterogeneity between R_n and R_m is stronger, $JS(R_n||R_m)$ will be higher. Given a threshold β, if $JS(R_n||R_m) \leq \beta$, the server considers client n and m share similar knowledge and have homogeneous traffic data. Then, R_m will be added into $\mathcal{PC}_n$, which denotes the set of positive prototypes for client n. Otherwise, R_m will be put into $\mathcal{NC}_n$, which denotes the set of negative prototypes for client n. Finally, the server performs customized aggregation for client n as

$$PR_n = \frac{1}{|\mathcal{PC}_n|}\sum_{R_m\in\mathcal{PC}_n} R_m; NR_n = \frac{1}{|\mathcal{NC}_n|}\sum_{R_m\in\mathcal{NC}_n} R_m,$$

where $PR_n \in \mathbb{R}^{B\times d_p}$ and $NR_n \in \mathbb{R}^{B\times d_p}$ represent the global positive and negative prototypes for client n respectively.

It is worth noting that the computation complexity of JSD values is $\mathcal{O}\left(\frac{N^2-N}{2}\right)$, which can hardly apply to forecasting tasks with masses of clients. The server can randomly select several clients as participants each round, with the selection ratio is α. Therefore, the JSD values of those clients which are not selected can be reused. The computation complexity can be reduced by $\mathcal{O}\left(\frac{N^2(1-\alpha)^2-N(1-\alpha)}{2}\right)$. The reduction enlarges significantly with N increasing.

Inter-Client Contrastive Loss. After aggregation, the customized global positive and negative prototypes are distributed to the corresponding clients. *Each client can further enforce the spatial discrimination of its local prediction model by approaching the positive prototype and keeping away from the negative prototype in the training process.* We define the inter-client contrastive loss item as

$$\begin{aligned}
\mathcal{L}_{inter} &= \frac{1}{B}\sum_{b=1}^{B} -\log\frac{loss_{pos}}{loss_{pos}+loss_{neg}},\\
loss_{pos} &= \exp\left(\mathrm{sim}\left(r_{n,b}, PR_{n,b}\right)/\tau\right),\\
loss_{neg} &= \exp\left(\mathrm{sim}\left(r_{n,b}, NR_{n,b}\right)/\tau\right),
\end{aligned} \quad (3)$$

where $PR_{n,b}$ and $NR_{n,b} \in \mathbb{R}^{d_p}$ denote the b-th row of PR_n and NR_n respectively. Under the constraint of $\mathcal{L}_{inter}$, the local prediction models can be empowered with spatial personalization.

4.4 Local Training and Inference

In the training process, client n inputs the representation r_n from the encoder en to the decoder de for generating the predicted values. Then, it calculates the prediction loss as

$$\mathcal{L}_{pred}\left(\delta_n;\phi_n;\boldsymbol{Y}_n,\hat{\boldsymbol{Y}}_n\right) = \frac{1}{B}\left\|\boldsymbol{Y}_n - \hat{\boldsymbol{Y}}_n\right\|^2. \quad (4)$$

Algorithm 1: Traning process of FUELS

Input: $\mathcal{D}_n$, $\mathcal{D}'_n$, $n = 1, \cdots, N$, ρ, β, τ, α.

SERVEREXECUTE:
Initialize $\{PR_n\}_{n=1}^N$ and $\{NR_n\}_{n=1}^N$.
Initialize the set of local prototypes $\mathcal{R}$.
for $t = 1, 2, \cdots, T$ **do**
  Randomly select $N\alpha$ clients.
  for *each selected client* n ***in parallel*** **do**
    $R_n \leftarrow$ `ClientExecute` (n, PR_n, NR_n)
    Update R_n in $\mathcal{R}$.
  Yield $\mathcal{PC}_n$ and $\mathcal{RC}_n, n \in [N]$ based on JSD values.
  $PR_n = \frac{1}{|\mathcal{PC}_n|} \sum_{R_m \in \mathcal{PC}_n} R_m$,
  $NR_n = \frac{1}{|\mathcal{NC}_n|} \sum_{R_m \in \mathcal{NC}_n} R_m, \forall n \in [N]$.
Function `ClientExecute`(*n, PR_n, NR_n*)**:**
  for *each epoch* **do**
    Initialize the set of representations $\mathcal{R}_n = \emptyset$.
    for $(\boldsymbol{X}_n, \boldsymbol{Y}_n, \hat{\boldsymbol{X}}_n)$ **do**
      $r_n \leftarrow en(\boldsymbol{X}_n); r'_n \leftarrow pr(r_n); \hat{r}'_n \leftarrow pr(en(\hat{\boldsymbol{X}}_n))$
      $\mathcal{R}_n \leftarrow \mathcal{R}_n + \{r'_n\}$
      Calculate $\mathcal{L}_{intra}$, $\mathcal{L}_{inter}$, and $\mathcal{L}_{pred}$ via Eq. (2), (3), and (4).
      $\mathcal{L} \leftarrow \mathcal{L}_{intra} + \mathcal{L}_{inter} + \rho\mathcal{L}_{pred}$
      Update δ_n, ϕ_n, and θ_n via gradient descent.
    $R_n \leftarrow$ **AVG**$(\mathcal{R}_n)$.
  return R_n

Therefore, the local loss function $\mathcal{L}$ in Eq. (1) is defined as a combination of $\mathcal{L}_{intra}$, $\mathcal{L}_{inter}$, and $\mathcal{L}_{pred}$, which is formulated as

$$\begin{aligned}&\mathcal{L}\left(\delta_n, \phi_n, \theta_n, W_n; \boldsymbol{X}_n, \boldsymbol{Y}_n, \hat{\boldsymbol{X}}_n, PR_n, NR_n\right) \\ =&\mathcal{L}_{pred}\left(\delta_n, \phi_n; \boldsymbol{X}_n, \hat{\boldsymbol{Y}}_n\right) \textit{//supervised loss} \\ +&\mathcal{L}_{intra}\left(\delta_n, \theta_n, W_n; \boldsymbol{X}_n, \hat{\boldsymbol{X}}_n\right) \textit{//intra-client loss} \\ +&\rho\mathcal{L}_{inter}\left(\delta_n, \theta_n; \boldsymbol{X}_n, PR_n, NR_n\right), \textit{//inter-client loss}\end{aligned} \tag{5}$$

where ρ denotes the additive weight of $\mathcal{L}_{inter}$. Then, client n performs gradient descent to update local parameters. The training process of FUELS is elaborated in Algorithm 1. After the training process, the local encoders are of great personalization to generate spatio-temporal heterogeneous representations. Therefore, in the inference process, each client just needs to input the test samples into the encoder and then decoder to obtain prediction results in an end-to-end manner.

5 Model Analysis

5.1 Generalization Analysis

We provide insights into the generalization bound of FUELS. The detailed proof and derivations are presented in Appendix 1.1. For ease of notation, we use the shorthand $\mathcal{L}(w_n) := \mathcal{L}\left(\delta_n, \phi_n, \theta_n, W_n; \boldsymbol{X}_n, \boldsymbol{Y}_n, \hat{\boldsymbol{X}}_n, PR_n, NR_n\right)$.

Assumption 5.1. (*Bounded Maximum*) The Loss function $\mathcal{L}(\cdot)$ has an upper bound, i.e., $\max \mathcal{L}(\cdot) \leq \mathrm{C}, \mathrm{C} < \infty$.

Theorem 5.2. *(Generalization Bounded) Let $w_n^*, n \in [1, N]$ denote the optimal model parameters for client n by FUELS. Denote the prediction model f as a hypothesis from $\mathcal{F}$ and d as the VC-dimension of $\mathcal{F}$. With the probability at least 1-κ:*

$$\begin{aligned}&\max_{(w_1, \cdots, w_N)} \left[\sum_{n=1}^N \frac{|\mathcal{D}_n|}{D}\mathcal{L}(w_n) - \sum_{n=1}^N \frac{|\mathcal{D}_n|}{D}\mathcal{L}(w_n^*)\right] \\ &\leq \sqrt{\frac{2d}{D}\log\frac{eD}{d}} + \sqrt{\frac{\mathrm{C}^2D^2}{2}\log\frac{1}{\kappa}},\end{aligned} \tag{6}$$

where e denotes the Euler's number. Theorem 5.2 indicates that the performance gap between FUELS and the optimal parameters is related to the VC-dimension of $\mathcal{F}$, which can be narrowed by carefully-selected prediction networks.

5.2 Convergence Analysis

We provide the convergence analysis of FUELS and the detailed proof is presented in Appendix 1.2. We begin with the commonly-used assumptions in FL works (Kairouz et al. 2021).

Assumption 5.3. (*Bounded Expectation of Gradients*) The expectation of gradient of loss function $\mathcal{L}(\cdot)$ is uniformly bounded, i.e., $\mathbb{E}(||\nabla\mathcal{L}(\cdot)||) \leq G$.

Assumption 5.4. (*Lipschitz Smooth*) The loss function $\mathcal{L}$ is L_1-smooth, i.e., $\mathcal{L}(w) - \mathcal{L}(w') \leq \langle\nabla\mathcal{L}(w'), w - w'\rangle + L_1||w - w'||^2, \forall w, w', \exists L_1 > 0$.

Assumption 5.5. (*Lipschitz Continuity*) Suppose $h : \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \rightarrow \mathcal{A}_T$ is Lipschitz Continuous in $\mathcal{A}_j$, i.e., $\exists L_h, \forall a_j, \hat{a}_j \in \mathcal{A}_j, \|h(a_1, \cdots, a_j, \cdots) - h(a_1, \cdots, \hat{a}_j, \cdots)\| \leq L_h \|a_j - \hat{a}_j\|$.

Let $w_{n,t}^i$ denote the local model parameters of client n at the i-th iteration in the t-th FL round, where $0 \leq t \leq T-1$ and $0 \leq i \leq I-1$. T and I denote the total number of FL rounds and local iterations respectively.

Theorem 5.6. *(Convergence Rate) Let Assumption 5.3 to 5.5 hold. Let $\mathcal{L}\left(w_{n,0}^0\right) - \mathcal{L}(w_n^*) = \Lambda$. If clients adopt stochastic gradient descent method to optimize local prediction models with the learning rate equal to η, for any client, given $\xi > 0$, after*

$$T = \frac{\Lambda}{\xi I(\eta - L_1\eta^2) - \rho\eta L_h N\alpha I^2 G} \tag{7}$$

FL rounds, we can obtain

$$\frac{1}{TI}\sum_{t=0}^{T-1}\sum_{i=0}^{I-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(w_{n,t}^i\right)\right\|^2\right] < \xi \tag{8}$$

with

$$\eta < \frac{\xi - \rho L_h N\alpha IG}{L_1\xi}. \tag{9}$$

Theorem 5.6 provides the convergence rate of FUELS. By adopting the learning rate η computed via Eq. (9), after T FL rounds (calculating T via Eq. (7)), the expectation of model updates will not exceed the given arbitrary value ξ.

Dataset	Net			Electricity			METR-LA			PEMS-BAY		
Metric	MSE	MAE	Comms	MSE	MAE	Comms	MSE	MAE	Comms	MSE	MAE	Comms
FedAvg	2.02	0.60	100737	0.34	0.40	100737	0.22	0.24	100737	0.05	0.12	100737
FedProx	1.76	0.61	100737	0.39	0.45	100737	0.21	0.23	100737	0.06	0.13	100737
FedRep	2.50	0.69	100608	0.27	0.35	100608	0.29	0.28	100608	0.07	0.15	100608
pFedMe	1.62	0.62	100737	0.77	0.71	100737	0.26	0.27	100737	0.10	0.16	100737
PerFedAvg	1.65	0.64	100737	0.29	0.39	100737	0.53	0.44	100737	0.14	0.21	100737
FedProto	2.56	0.90	6144	0.70	0.67	6144	0.88	0.58	73728	0.64	0.54	73728
FedPCL	4.99	1.34	106881	0.98	0.82	106881	1.27	0.71	174465	0.97	0.67	174465
CNFGNN	2.28	0.87	72065	1.01	0.49	72065	0.91	0.57	72065	0.79	0.59	72065
FCGCN	2.57	0.92	/	4.20	1.30	/	2.17	0.96	/	0.82	0.51	/
FedGRU	2.19	0.91	/	1.04	0.61	/	0.98	0.78	/	0.76	0.56	/
FUELS	**0.96**	**0.51**	**768**	**0.24**	**0.34**	**768**	**0.19**	**0.23**	**4608**	**0.04**	**0.11**	**36864**

Table 1: Performance comparisons. "Comms" refers to the number of parameters each client sends to the server per round.

5.3 Model Complexity

The number of parameters each client uploads is $\mathcal{O}(Bd_p)$, which is much more lightweight than model parameters (detailed results in Section 6.1). The computation cost at clients in FL training round can be summarized as $\mathcal{O}(I(3FE + 2FF + 2FD))$, where FE, FF, and FD represent the computation cost of encoder en, W_n, and decoder de respectively for a batch in the forward propagation. Given the light weight of W_n, most of additional computation cost results from encoding the augmented data. In Appendix 1.3, we further provide detailed comparison of computation and communication complexity.

6 Experiments

Baselines: The baselines cover a wide range of relevant methods, which can be classified into 3 categories: **TYPE 1** (*personalized federated learning methods*): FedAvg (McMahan et al. 2017), FedProx (Li et al. 2020), FedRep (Collins et al. 2021), PerFedAvg (Fallah, Mokhtari, and Ozdaglar 2020), and pFedMe (T Dinh, Tran, and Nguyen 2020); **TYPE 2** (*federated prototype learning methods*): FedProto (Tan et al. 2022b) and FedPCL (Deng and Yang 2023); **TYPE 3** (*federated learning methods for spatio-temporal forecasting*): CNFGNN (Meng, Rambhatla, and Liu 2021), FedGRU (Liu et al. 2020), and FCGCN (Xia, Jin, and Chen 2022).

Settings. Our evaluations are conducted on four widely-used benchmark datasets: Net (Zhang et al. 2021), Electricity (Liu et al. 2024), METR-LA (Meng, Rambhatla, and Liu 2021) and PEMS-BAY (Meng, Rambhatla, and Liu 2021). Detailed introduction of datasets and experimental implementations are provided in Appendix 2. We use Mean Square Error (MSE) and Mean Absolute Error (MAE) as metrics.

6.1 Main Results

The evaluation results of the baselines and our proposed method are presented in Table 1. In all methods except FCGCN and FedGRU, each device serves as a FL participant client. Therefore, we assess the "Comms" of these methods. We have the key observations that FUELS shows promising performance by consistently outperforming all the baselines. Moreover, the communication cost of clients in FUELS is significantly lower than that in the baselines. Specifically, compared with the second communication-efficient method FedProto, FUELS provides **87.5%**, **87.5%**, **99.94%**, and **50%** communication overhead reduction on four datasets respectively. We attribute such superiority to the designed contrastive tasks and novel communication carrier.

Effectiveness. Fig. 3 (a) and (b) show the predicted values of four methods on Net dataset and the cumulative distribution functions (CDFs) of MSEs over all clients. While the predicted values of these methods have similar prediction performance at the smooth traffic sequences, FUELS can generate more accurate results in fluctuating traffic sequences, which further underscores the effectiveness of FUELS in tackling heterogeneous temporal patterns. The area under the CDF curve of FUELS is larger than those of the other three methods, which indicates that clients' prediction MSEs in FUELS distribute around lower values. Specifically, 90% of clients have prediction MSEs lower than 2, while the cases for FedProx, pFedMe and PerFedAvg are 71%, 73% and 75% respectively.

Efficiency. Fig. 3 (c) shows the training MSE versus the communication amounts over all clients. We observe that given the same MSE, the communication amounts in the other three methods are significantly higher than those in FUELS on both datasets. Specifically, when MSE reaches the convergence, FUELS will reduce about 97% communication overhead compared with FedProx. FUELS can yield superior prediction performance with the baselines and simultaneously reduce the communication cost to a great extent, which indicates the efficiency of FUELS.

6.2 Ablation Study

We compare FUELS with the following 4 variants. **(1) w/o inter**: Only the intra-client contrastive loss item is adopted; **(2) w/o intra** : Only the inter-client contrastive loss item is adopted; **(3) w/o p-aware** : Concatenation-based prototypes

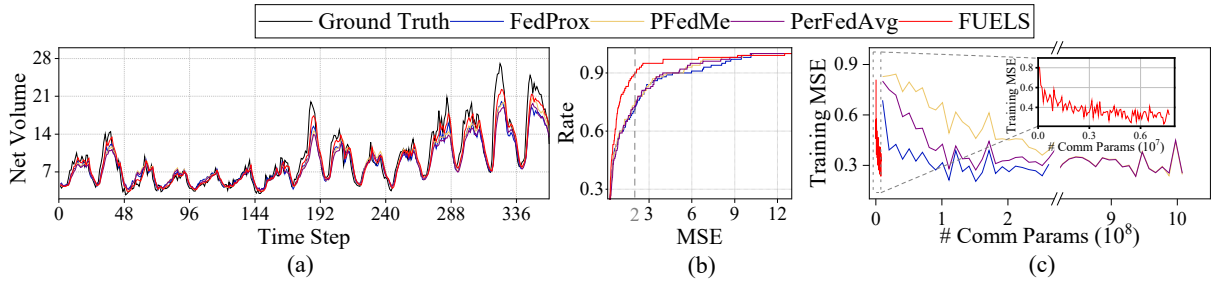

Figure 3: Effectiveness and efficiency comparison on Net dataset: (a) Visualization of forecasting results; (b) MSE CDFs over clients; (c) Training MSE versus communication amounts.

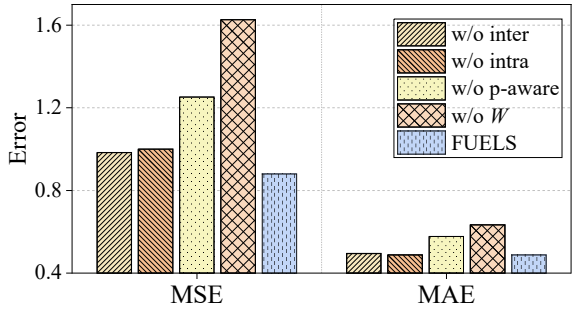

Figure 4: Performance comparison of FUELS and four variants on Net dataset.

are adopted instead of periodicity-aware prototypes; **(4) w/o *W*** : Dynamic hard negative filtering is omitted. The prediction performance of FUELS and the above variants on Net dataset is shown in Fig. 4. We have the following key observations. **(1)** The increased prediction errors of w/o inter and w/o intra compared with FUELS indicate that inter- and intra-client contrastive loss items can benefit the local training from different perspectives. **(2)** With concatenation-based prototypes, the prediction performance may be impacted by the noise in representations and "Comms" will significantly increase to 227928. **(3)** The higher error in w/o W indicates that the proposed dynamic filtering module can effectively filter out hard negatives and keep the semantic structure consistent.

6.3 Case Study

Relatedness Between Local Prototype and Traffic Data. We randomly select several clients to visualize the correlation between prototypes and local datasets. The results are shown in Fig. 2(a) and (b) of the Appendix. We calculate the similarity of local training datasets based on JSD and set the threshold as 0.00065, which is also the 50-th percentile of these JSD values. We observe that clients' correlation is consistent with the JSD-based results. Therefore, we claim that the designed local prototypes can effectively express client-specific knowledge.

Visualization of Dynamic Hard Negative Filtering. The illustration of parameters in W_n for a randomly selected client is presented in Fig. 3 of the Appendix. If a parameter in W_n is over 0, the corresponding value in Z will be over 0. Almost all the values on the diagonal are less than 0, which represents that positives are filtered out. The filtering approach can filter out hard negatives within or without closeness scope and simultaneously avoid the error filtering of true negatives within closeness range.

6.4 Combination with Privacy Protection Mechanisms

We incorporate FUELS with privacy-preserving mechanisms in case of the privacy leakage in the communicated prototypes. **We add different types of random noise with *Laplace*, *Guassian*, and *exponential* distribution to the local prototypes in FUELS.** The results indicate that there is no significant performance degradation after noise injection, which demonstrates that *FUELS is noise-robust and accommodates for differential privacy strategies*. Details are presented in Appendix 3.3.

6.5 Additional Experimental Results in Appendix

Due to space limitation, we provide auxiliary experimental results in Appendix, including **effect of prototype size** (in Appendix 3.2), **hyperparameter investigation** (in Appendix 3.1), and **error bars** (in Appendix 4).

7 Conclusion

We propose FUELS for tackling the spatio-temporal heterogeneity by adaptively aligning the temporal and spatial representations according to semantic similarity for the supplemented intra- and inter-client contrastive tasks to preserve the spatio-temporal heterogeneity in the latent representation space. Note-worthily, a lightweight but efficient prototype is designed as the client-level representation for carrying client-specific knowledge. Experimental results demonstrate the effectiveness and communication efficiency of FUELS.

Limitations & Future Works. The potential limitation of FUELS is the increased computation cost for augment data. In the future work, we will explore more computation-efficient contrastive approaches. Moreover, since the communication carrier is independent of network structures, *FUELS can be built over heterogeneous local prediction models, pre-trained models, or even large models*, which will be explored in future studies.

Acknowledgments

This work is mainly supported by the National Key Research and Development Program of China under Grant

2021YFB2900102 and also by the National Natural Science Foundation of China under Grant 62072436, Grant 62472410, and Grant 62402414. This paper is also funded by the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

References

Arivazhagan, M. G.; Aggarwal, V.; Singh, A. K.; and Choudhary, S. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.

Bui, D.; Malik, K.; Goetz, J.; Liu, H.; Moon, S.; Kumar, A.; and Shin, K. G. 2019. Federated user representation learning. *arXiv preprint arXiv:1909.12535*.

Chen, D.; Yao, L.; Gao, D.; Ding, B.; and Li, Y. 2023. Efficient Personalized Federated Learning via Sparse Model-Adaptation. *arXiv preprint arXiv:2305.02776*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, 2089–2099. PMLR.

Deng, S.; and Yang, L. 2023. Prototype Contrastive Learning for Personalized Federated Learning. In *International Conference on Artificial Neural Networks*, 529–540. Springer.

Fallah, A.; Mokhtari, A.; and Ozdaglar, A. 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33: 3557–3568.

Ghosh, A.; Chung, J.; Yin, D.; and Ramchandran, K. 2020. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33: 19586–19597.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Ji, J.; Wang, J.; Huang, C.; Wu, J.; Xu, B.; Wu, Z.; Zhang, J.; and Zheng, Y. 2023. Spatio-temporal self-supervised learning for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4356–4364.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 5132–5143. PMLR.

Li, D.; and Wang, J. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Li, L.; Duan, M.; Liu, D.; Zhang, Y.; Ren, A.; Chen, X.; Tan, Y.; and Wang, C. 2021. FedSAE: A novel self-adaptive federated learning framework in heterogeneous systems. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–10. IEEE.

Li, M.; and Zhu, Z. 2021. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4189–4196.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, R.; Zhong, T.; Jiang, X.; Trajcevski, G.; Wu, J.; and Zhou, F. 2022. Mining spatio-temporal relations via self-paced graph contrastive learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 936–944.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, W.; and Wang, S. 2022. Federated meta-learning for spatial-temporal prediction. *Neural Computing and Applications*, 34(13): 10355–10374.

Li, Y.; Li, J.; and Wang, Y. 2021. Privacy-preserving spatiotemporal scenario generation of renewable energies: A federated deep generative learning approach. *IEEE Transactions on Industrial Informatics*, 18(4): 2310–2320.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations (ICLR '18)*.

Liang, P. P.; Liu, T.; Ziyin, L.; Allen, N. B.; Auerbach, R. P.; Brent, D.; Salakhutdinov, R.; and Morency, L.-P. 2020. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*.

Lin, J.; Chen, Y.; Zheng, H.; Ding, M.; Cheng, P.; and Hanzo, L. 2021. A data-driven base station sleeping strategy based on traffic prediction. *IEEE Transactions on Network Science and Engineering*.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351–2363.

Liu, Q.; Sun, S.; Liu, M.; Wang, Y.; and Gao, B. 2023. Online Spatio-Temporal Correlation-Based Federated Learning for Traffic Flow Forecasting. *arXiv preprint arXiv:2302.08658*.

Liu, X.; Hu, J.; Li, Y.; Diao, S.; Liang, Y.; Hooi, B.; and Zimmermann, R. 2024. UniTime: A Language-Empowered Unified Model for Cross-Domain Time Series Forecasting. In *Proceedings of the ACM Web Conference 2024*.

Liu, X.; Liang, Y.; Huang, C.; Zheng, Y.; Hooi, B.; and Zimmermann, R. 2022. When do contrastive learning signals

help spatio-temporal graph forecasting? In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–12.

Liu, Y.; James, J.; Kang, J.; Niyato, D.; and Zhang, S. 2020. Privacy-preserving traffic flow prediction: A federated learning approach. *IEEE Internet of Things Journal*, 7(8): 7751–7763.

Mahdy, B.; Abbas, H.; Hassanein, H. S.; Noureldin, A.; and Abou-zeid, H. 2020. A clustering-driven approach to predict the traffic load of mobile networks for the analysis of base stations deployment. *Journal of Sensor and Actuator Networks*, 9(4): 53.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Meng, C.; Rambhatla, S.; and Liu, Y. 2021. Cross-node federated graph neural network for spatio-temporal data modeling. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1202–1211.

Mu, X.; Shen, Y.; Cheng, K.; Geng, X.; Fu, J.; Zhang, T.; and Zhang, Z. 2023. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143: 93–104.

Perifanis, V.; Pavlidis, N.; Koutsiamanis, R.-A.; and Efraimidis, P. S. 2023. Federated learning for 5G base station traffic forecasting. *Computer Networks*, 235: 109950.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. *Advances in neural information processing systems*, 30.

T Dinh, C.; Tran, N.; and Nguyen, J. 2020. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33: 21394–21405.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022a. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022b. FedProto: Federated Prototype Learning across Heterogeneous Clients. In *AAAI Conference on Artificial Intelligence*.

Tan, Y.; Long, G.; Ma, J.; Liu, L.; Zhou, T.; and Jiang, J. 2022c. Federated learning from pre-trained models: A contrastive learning approach. *Advances in Neural Information Processing Systems*, 35: 19332–19344.

Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; and Hoi, S. 2022. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*.

Xia, M.; Jin, D.; and Chen, J. 2022. Short-term traffic flow prediction based on graph convolutional networks and federated learning. *IEEE Transactions on Intelligent Transportation Systems*, 24(1): 1191–1203.

Xie, M.; Long, G.; Shen, T.; Zhou, T.; Wang, X.; Jiang, J.; and Zhang, C. 2021. Multi-center federated learning. *arXiv preprint arXiv:2108.08647*.

Yang, M.; Wang, X.; Zhu, H.; Wang, H.; and Qian, H. 2021. Federated learning with class imbalance reduction. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 2174–2178. IEEE.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823.

Yu, Q.; Liu, Y.; Wang, Y.; Xu, K.; and Liu, J. 2022. Multimodal Federated Learning via Contrastive Representation Ensemble. In *The Eleventh International Conference on Learning Representations*.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8980–8987.

Zhang, C.; Dang, S.; Shihada, B.; and Alouini, M.-S. 2021. Dual attention-based federated learning for wireless traffic prediction. In *IEEE INFOCOM 2021-IEEE conference on computer communications*, 1–10. IEEE.

Zhang, L.; Zhang, C.; and Shihada, B. 2022. Efficient wireless traffic prediction at the edge: A federated meta-learning approach. *IEEE Communications Letters*, 26(7): 1573–1577.

Zhao, S.; Jiang, X.; Jacobson, G.; Jana, R.; Hsu, W.-L.; Rustamov, R.; Talasila, M.; Aftab, S. A.; Chen, Y.; and Borcea, C. 2020. Cellular network traffic prediction incorporating handover: A graph convolutional approach. In *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 1–9. IEEE.

Zhu, Y.; and Wang, S. 2021. Joint traffic prediction and base station sleeping for energy saving in cellular networks. In *ICC 2021-IEEE International Conference on Communications*, 1–6. IEEE.

Zhu, Y.; Xu, Y.; Yu, F.; Liu, Q.; Wu, S.; and Wang, L. 2021. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, 2069–2080.

Zhu, Z.; Hong, J.; and Zhou, J. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, 12878–12889. PMLR.