FairSTG: Countering Performance Heterogeneity via Collaborative Sample-Level Optimization

Gengyu Lin, Zhengyang Zhou, *Member, IEEE*, Qihe Huang, Kuo Yang, Shifen Cheng, and Yang Wang, *Senior Member, IEEE*

***Abstract*—Spatiotemporal learning plays a crucial role in mobile computing techniques to empower smart cites. While existing research has made great efforts to achieve accurate predictions on the overall dataset, they still neglect the significant performance heterogeneity across samples. In this work, we designate the performance heterogeneity as the reason for unfair spatiotemporal learning, which not only degrades the practical functions of models, but also brings serious potential risks to real-world urban applications. To fix this gap, we propose a model-independent Fairness-aware framework for SpatioTemporal Graph learning (FairSTG), which inherits the idea of exploiting advantages of well-learned samples to challenging ones with collaborative mix-up. Specifically, FairSTG consists of a spatiotemporal feature extractor for model initialization, a collaborative representation enhancement for knowledge transfer between well-learned samples and challenging ones, and fairness objectives for immediately suppressing sample-level performance heterogeneity. Experiments on four spatiotemporal datasets demonstrate that our FairSTG significantly improves the fairness quality while maintaining comparable forecasting accuracy. Case studies show FairSTG can counter both spatial and temporal performance heterogeneity by our sample-level retrieval and compensation, and our work can potentially alleviate the risks on spatiotemporal resource allocation for underrepresented urban regions.**

***Index Terms*—Fairness learning, spatiotemporal forecasting, representation learning, self-supervised learning.**

I. Introduction

With the rapid urbanization and increasing number of urban devices, we are now embracing a new era with a vast amount of valuable spatiotemporal data. Actually, spatiotemporal data plays a crucial role in mobile computing services for smart cities, including traffic police assignment [1], [2], [3], urban safety management [4], [5], [6], and numerical weather forecasting [7]. However, data collected from real-world is inevitably trapped into bias due to imbalance sampling, inherent low quality or under-representation in gender, race or other sensitive attributes.

TABLE I
The Unfairness Issue in Spatiotemporal Learning

		DCRNN	MTGNN	D^2STGNN
METR-LA	overall MAE	3.57	3.49	3.35
	MAE variance	51.61	47.68	49.47
	overall MAPE	10.40%	9.87%	9.43%
	MAPE variance	0.18	0.15	0.14
PEMS-BAY	overall MAE	1.96	1.96	1.98
	MAE variance	17.23	16.50	17.02
	overall MAPE	4.64%	4.67%	4.84%
	MAPE variance	0.05	0.04	0.06

Recently, fairness issue, which calls for equal opportunities on allocation and assigned has received increasing attention in the field of mobile computing. Such fairness issue has been studied for edge computing systems [8], crowdsensing platforms [9] as well as intelligent traffics [10] to achieve fair task assignment and computational resource allocation. As an emerging technique for mobile computing, spatiotemporal learning, mostly for location-based services, is more prone to suffer unfairness. Without explicitly considering the fairness issue, machine learning models will erroneously learn such bias and even exacerbate the unfairness, leading to misleading decisions on downstream tasks [11], [12].

Current literature on spatiotemporal learning mostly concentrates on the overall performance [3], [13], [14], overlooking the performance heterogeneity across different samples and regions. As shown in Table I, a preliminary experiment conducted on two well-known datasets verifies such serious performance heterogeneity. Consider the expectation of Mean Absolute Error (MAE) and variance of MAE as the indicator for overall performance and sample-level performance heterogeneity, we find that advanced spatiotemporal learning methods achieve satisfactory overall performance, but the MAE variance is approximately 14 times of MAE on METR-LA and 9 times of MAE on PEMS-BAY. We designate such prediction disparity as performance heterogeneity, resulting in unfair performance over locations and prediction steps.

Unfortunately, an unfair spatiotemporal model can induce risks on two aspects. First, overlooking the performance

Received 29 January 2024; revised 11 December 2024; accepted 18 December 2024. Date of publication 25 December 2024; date of current version 4 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 12227901 and Grant 62072427, in part by the Natural Science Foundation of Jiangsu Province under Grant BK.20240460, in part by the Project of Stable Support for Youth Team in Basic Research Field, CAS under Grant YSBR-005, and in part by the State Key Laboratory of Resources and Environmental Information System. Recommended for acceptance by T. X. Han. *(Corresponding authors: Yang Wang; Zhengyang Zhou.)*

Gengyu Lin, Qihe Huang, and Kuo Yang are with the University of Science and Technology of China, Anhui 230026, China (e-mail: lingengyu@mail.ustc.edu.cn; hqh@mail.ustc.edu.cn; yangkuo@mail.ustc.edu.cn).

Zhengyang Zhou is with the Suzhou Institute for Advanced Research, University of Science and Technology of China, Anhui 230026, China, and also with the State Key Laboratory of Resources and Environmental Information System, Beijing 100101, China (e-mail: zzy0929@ustc.edu.cn).

Shifen Cheng is with the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: chengsf@lreis.ac.cn).

Yang Wang is with the Suzhou Institute for Advanced Research, University of Science and Technology of China, Anhui 230026, China (e-mail: angyan@ustc.edu.cn).

Digital Object Identifier 10.1109/TMC.2024.3522476

disparity among samples can result in misleading guidance and erroneous decisions for urban management. Second, in a technical perspective, pure error-based metrics which only emphasize the average results but neglect the inconsistency in performance can trap such evaluation into one-sided report. To empirically explain the significance of performance fairness, we provide the numerical results in Table I, which illustrates the performance disparity (MAE/MAPE variance) among testing samples. The larger disparity, the more unfair performances among samples, thus the more risks of exploiting the model for making urban decisions. To this end, a fair spatiotemporal learning framework is essential to empower non-biased urban applications. In this work, we dissect the inherent factors behind unfairness and counter the heterogeneity of prediction results across spatiotemporal domain to increase the quality of urban decisions, hence deducing the system risks, especially on underrepresented groups and individuals.

Concretely, the performance heterogeneity in spatiotemporal learning can be attributed to two aspects, i.e., the adequacy of data representation and the inherent regularity within datasets. First, urban sensors are more concentrated on city centers than suburban areas, leading to inadequate representation of marginal areas in the overall dataset and increasing their learning difficulty. Second, different samples and regions exhibit diverse patterns, inducing different degrees of learning difficulty. Meanwhile, abundant local contexts in spatiotemporal data, such as geographical locations, functional regions, and temporal contexts, influence the data regularity in a complex and intertwined manner. To this end, we posit that samples with lower regularity and high learning difficulty pose a greater challenge for the model, and exhibit poorer predictive performance. Thus, improving the forecasting performance on these samples is crucial for countering the unfairness issue in spatiotemporal data.

Existing literature concerning this work can be summarized as two lines. The line of spatiotemporal learning takes efforts to model spatial and temporal heterogenerous information on observations [15], achieving personalized node-level directional aggregation [16], [17], but has never modeled the heterogeneity of prediction results, i.e., the prediction disparities on regions and temporal steps, directly resulting in prediction unfairness. While the line of fair machine learning investigates how to separate the influence of sensitive factors during training, and design various fairness objectives including inter-group and intra-group equality [18], all these techniques still have not been advanced to urban spatiotemporal learning. To this end, we argue that there are two specific challenges in constructing a fairness-aware spatiotemporal learning framework.

- Spatiotemporal data is equipped with complex and heterogeneous dependencies but lacks explicit sensitive factors which are considered definitely leading to unfair predictions. Thus the first challenge is how to exploit the implicit factors which can potentially and jointly result in unfair predictions, to accurately identify the specific challenging samples suffering unfair performances.
- On model design aspect, how to devise fairness-aware learning strategies and maximally exploit the available contexts and high-quality spatiotemporal representations to collaboratively enhance learning of challenging samples, becomes the second challenge.

To address above challenges, in this work, we propose a model-independent Fairness-aware SpatioTemporal Graph learning (FairSTG) to counter the spatiotemporal performance heterogeneity for fair learning. Our FairSTG takes series of each region as the minimal sample unit and collaboratively optimizes the spatiotemporal graph representation with an integrated framework, which allows location-based fairness and joint enhancement. Specifically, we design an auxiliary self-supervised task by additionally taking various external factors as inputs, to actively identify which samples are prone to unfair treatment. This solution repairs the lacking of explicit sensitive factors. To realize fairness-aware learning, FairSTG performs collaborative optimization from both the representation space and the objective space. We especially construct the compensatory sample sets for adaptive representation mix-up, exploiting advantages of well-learned representations to improve challenging samples. Regarding learning objectives, a variance-based objective is devised to directly force the model to optimize towards consistent performance, and a variance-based fairness metrics is introduced to complement the pure error metrics, which fixes shortcomings of function degeneration in current metrics. The contributions can be summarized as below.

- This is the first effort that summarizes the serious outcomes of spatiotemporal learning without fairness, and subsequently attributes such performance heterogeneity into defective objectives and inherent heterogeneity of data learning difficulty.
- We propose a novel fair mobile computing technique FairSTG, from both inherent factor of low-quality representation and direct factor of the absence of fairness objectives, advancing modeling observation heterogeneity towards performance heterogeneity. We minimize the variances across samples and improve prediction performance of the challenging samples by drawing common patterns from the well-matched and high-quality representations.
- We design a fairness metric adapting to spatiotemporal forecasting, which jointly evaluate learning frameworks with error-based metrics. Experiments show that our solution significantly improves the equality across performances, where the fairness metric in FairSTG outperforms backbones with a large margin, ranging from 0.21% to 20.05% with 16 out of 24 achieving improvement above 4%, and achieves comparable or better accuracy against baselines. Case studies demonstrate that FairSTG can alleviate the risks on urban resource allocation for underrepresented urban regions, and FairSTG can potentially become a paradigm of fair urban computing for sustainable mobile computing and urban intelligence.

II. Related Work

A. Spatiotemporal Learning

Spatiotemporal learning is a crucial technique to empower urban applications. In an early stage, researchers take spatiotemporal forecasting tasks as time-series predictions, and introduce

statistical solutions such as ARIMA [19], VAR [20] to achieve forecasting. These statistical approaches are easy to implement with reasonable interpretability, but fail to simultaneously capture both spatial and temporal dependencies. With the prosperity of deep learning and Graph Neural Networks (GNNs), deep GNNs naturally have the edge on learning non-Euclidean spatial data and can potentially accommodate to various spatiotemporal learning tasks [21], [22], [23]. Moreover, recent literature also proposes to rank the related records in dynamic and heterogeneous graphs [15], [24], [25], and training in a cooperative manner [26]. In fact, spatiotemporal heterogeneity, which refers to the varying patterns across different temporal or spatial ranges, has been widely recognized in such data and raised more attention in academia [21]. Specifically, ST-SSL [1] emphasizes the heterogeneity across temporal steps and spatial regions with a pair-wise learning in a self-supervised manner, while HA-STGN [27] is proposed by designing a direction-aware road network and imposing a time-aware graph attention mechanism to capture such heterogeneity. Even though existing works have explicitly taken the heterogeneous observations into account with different aggregation and representation strategies, the performance heterogeneity along both spatial and temporal dimensions has never been considered. We designate such imbalanced prediction performance as the prediction unfairness, and overcome such issue in this paper.

B. Fairness-Aware Machine Learning

The unfairness in machine learning, which may do harm to interests of a specific group or individual, has received extensive attention. The unfairness can primarily stem from the inherent data imbalance, and the unawareness of fairness in machine learning algorithms can further aggravate such bias, leading to severe unfair resource allocation and discrimination.

Plenty of literature has attempted to counteract the bias in both data and algorithms to achieve fair machine learning. Based three stages in machine learning systems, fairness-aware learning can be generally classified into three categories, methods during pre-processing, in-processing, and post-processing. First, the pre-processing solutions remove the underlying bias by augmenting and adjusting the training data, and then train a model on debiased data, where it can be considered as the data-aspect solution. Second, in-processing methods try to incorporate fairness metrics or propose adversarial learning objectives to obtain fair representations, which can be viewed as the model aspect paradigm to balance the accuracy and fairness during the learning process. For example, to protect the interest of minority news providers, ProFairRec [28] exploits a sensitive attribute discriminator to identify the provider-bias information while generators make indistinguishable provider-fair representations against the discriminator. And VFAE [29] investigates a variational autoencoder with Maximum Mean Discrepancy to generate regularized fair representations. The post-processing methods allow transformations on model outputs, such as label re-assignment [30], [31], re-ranking of output lists [24], [32], [33] and projection of representations onto debiased subspace [34], [35], to mitigate unfairness.

Even fair learning systems have been widely investigated, the unfairness issue in spatiotemporal learning, which directly leads to unfair resource allocation and underestimated risks, has still been under-explored. Such under-exploration can be attributed to two challenges. First, spatiotemporal data lacks definite sensitive information for fairness constraints. Second, this kind of data exhibits complex spatial and temporal heterogeneity and dynamic variations [1], contributing to the difficulty in capturing real statuses of inactive spatial regions and interactions between active and inactive regions.

C. Fairness in Mobile Computing Community

The trade-off between system efficiency and fairness, is also an emerging topic in traditional mobile computing community, where recent researches include fairness-aware resource allocation [36], [37], fair edge computing [9], [38] and fair routing for intelligent traffics [10]. More specifically, regarding edge computing systems, [8] develops a proportionally fair resource scheme for wireless-powered mobile edge computing system with time division multiple access and partial offloading. For a crowdsensing system, [9] proposes a fair assignment strategy utilizing Lyapunov optimization to achieve a trade-off between minimizing the assignment cost and satisfying the two fairness constraints of users and tasks simultaneously. And for intelligent traffic systems, [10] studies the road pricing in connection with fair routing for general road networks and develops a new convex program which interpolates between a fairness-promoting and an efficiency-promoting traffic-assignment objective. Overall, these pioneering researches point out that fairness issue is essential for mobile computing and is promising to investigate the trade-off among performance, fairness and efficiency.

Actually, spatiotemporal learning is also a core technique benefiting mobile computing community. The general significance of fairness-aware machine learning is to provide consistent performance for different individuals within a system while fulfills the global performance requirements. Different from existing literature [1], [17], our proposed FairSTG is a model-independent general framework especially designed for regression tasks in spatiotemporal learning scenarios. It aims at mitigating the performance disparity of the backbone model among different samples while maintaining comparable forecasting performance, thus alleviating the negative edge of technique and facilitating the sustainable urban computing.

III. Motivation

Fairness issue, which calls for consistent experiences and system assignment for different individuals or groups, has received increasing attention in mobile computing community. Recent spatiotemporal forecasting solutions only emphasize the mean performance on the overall dataset while neglect the performance disparity among different samples. A preliminary experiment conducted on two well-known datasets verifies such serious performance heterogeneity from both temporal and spatial aspect. As illustrated in Fig. 1(a), the model exhibits inconsistent forecasting performance across different regions, and the sensors near transportation hubs suffer poorer forecasting performance possibly due to the more complex traffic

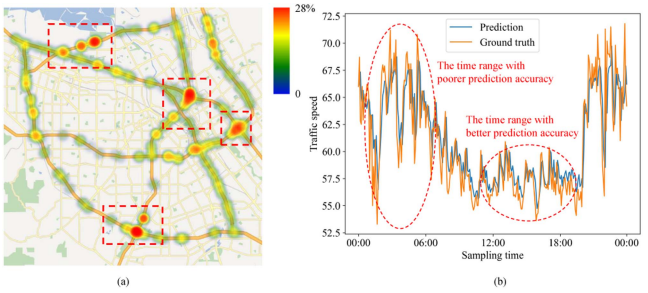


Fig. 1. The unfairness across spatial and temporal ranges. (a) The mean MAPE of each sensor in PEMS-BAY dataset generated by MTGNN, with red boxes indicating regions with significant errors. (b) The curve shows the traffic flow data and the corresponding predictions from MTGNN on sensor #64 in PEMS-BAY dataset. It can be observed that the prediction performance varies significantly at different timestamps even for the same sensor.

conditions. And in Fig. 1(b), the prediction performance varies significantly at different timestamps even for the same sensor, which demonstrates that there exists performance heterogeneity in temporal aspect.

More seriously, in a mobile computing system, the unfairness can potentially propagate along the processes of information collection, model learning and decision making [39]. The distribution of facilities and infrastructures in a city is usually associated with the socio-economic status. Then the data-driven models may overfit to strong biases in the source data related to corresponding socio-economic status and demographics [40]. For example, in a macro view, the deployed sensors for intelligent cities in aging and low-income communities may be fewer and with less quality than other well-developed communities. Then models without fairness constraints might ignore the patterns of these underrepresented communities and result in poorer forecasting performance and misleading guidance for resource allocation and welfare assignment, which exacerbates the unfairness for underrepresented regions and communities [39]. For a specific application of traffic management, when the task of traffic volume predictions suffers the performance heterogeneity, the underestimated regions tend to be allocated with less police force for manual management, thus increasing the risks of accidents and crowd incidents in corresponding regions. Consequently, fairness-aware spatiotemporal forecasting method is urgently needed to provide unbiased guidance and ensure equal opportunities for different individuals or groups.

IV. Problem Formulation

We focus on countering the unfairness issue in spatiotemporal learning while maintaining the performance of the backbone model.

Definition 1 (Spatiotemporal graph): A spatiotemporal graph (ST Graph) is formulated as $\mathbb{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_T\}$, to describe spatiotemporal data, where $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}, \mathbf{X})$. The node set $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ and edge set $\mathcal{E} = \{e_{ij} = (v_i, v_j)\}$ can be formulated respectively, where $N = |\mathcal{V}|$ is the number of nodes. It is worth noting that we do not force a predefined adjacency matrix for spatiotemporal graphs in our tasks, but it can be learned from input features. We then define $\mathbf{A} \in \mathbb{R}^{N \times N}$ as the virtual (learnable) adjacency matrix of our spatiotemporal graph when mentioned.

Besides, let $\mathbf{X}_{:,0:T-1} = \{\mathbf{X}_{:,0}, \mathbf{X}_{:,1}, \cdots, \mathbf{X}_{:,T-1}\} \in \mathbb{R}^{N \times T}$ denote a series of observed spatiotemporal graphs with N nodes and T time steps, where $\mathbf{X}_{:,t} = \{x_{0,t}, x_{1,t}, \cdots, x_{N-1,t}\} \in \mathbb{R}^N$ records the observations of these N nodes in $\mathcal{G}_t$ at time step t and $x_{i,t}$ represents the deterministic value of node i at time step t.

Problem 1 (Spatiotemporal forecasting): Given $\mathbf{X}_{:,t-w:t-1} \in \mathbb{R}^{N \times w}$, spatiotemporal forecasting aims to derive the following h steps of observations. Then, the spatiotemporal prediction problem can be formalized as follows,

$$\widehat{\mathbf{Y}}_{:,t:t+h-1} = f_\theta(\mathbf{X}_{:,t-w:t-1}; \mathcal{G}) \tag{1}$$

where f represents the model for spatiotemporal prediction, and $\theta \in \Theta$ denotes the learnable parameters.

We refer to the observations of the ith node with a temporal window of length w before time step t as a **spatiotemporal sample**, formalized as $\mathbf{X}_{i,t-w:t-1}$. And the corresponding ground truth is defined as $\mathbf{X}_{i,t:t+h-1}$. For ease of description, we denote the set of input spatiotemporal samples as $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M\}$, where $\boldsymbol{x}_i \in \mathbb{R}^w$ represents an individual sample and $M = |\mathcal{X}|$ denotes the number of samples. And the set of ground truth is formulated as $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M\}$, where $\boldsymbol{y}_i \in \mathbb{R}^h$. Then we can explicitly capture the spatial heterogeneity via our node-level defined spatiotemporal sample. Given above, the spatiotemporal forecasting problem can be reformulated as,

$$\hat{\mathcal{Y}} = f_\theta(\mathcal{X})$$

In this work, we take MAE as the error metric, and the objective of the forecasting task is to find an optimal set of parameters $\theta^* \in \Theta$ that minimizes the global MAE, i.e.,

$$\begin{aligned} \theta^* &= \arg_{\theta^* \in \Theta} \min \text{MAE}(\mathcal{Y}, \hat{\mathcal{Y}}) \\ &= \arg_{\theta^* \in \Theta} \min \text{MAE}(\mathcal{Y}, f_{\theta^*}(\mathcal{X})) \end{aligned} \tag{2}$$

Definition 2 (Fairness metrics in spatiotemporal forecasting): Existing fairness metrics in machine learning usually take an explicit sensitive attribute, such as gender and racial, as the grouping standard and evaluate the fairness by measuring the consistency of model performance across different groups. However, in spatiotemporal learning, there is lack of such explicit sensitive attributes and a strategy tailored for regression tasks is highly required. To fill the gap, we propose a variance-based fairness metric. We posit that a fair spatiotemporal prediction model should provide predictions with similar performance for different spatiotemporal samples, meaning that the prediction errors for various samples are close. We take the disparity of errors among different spatiotemporal samples to characterize the degree of the performance unfairness. And we employ the variance of errors to quantify such sample-level disparity, illustrated as $D(\mathcal{Y}, \hat{\mathcal{Y}}; \theta)$, where multiple metrics can be chosen to measure prediction errors. The larger the variance is, the more unfair the forecasts are.

Problem 2 (Fairness-aware spatiotemporal forecasting): In this work, we improve the spatiotemporal forecasting model

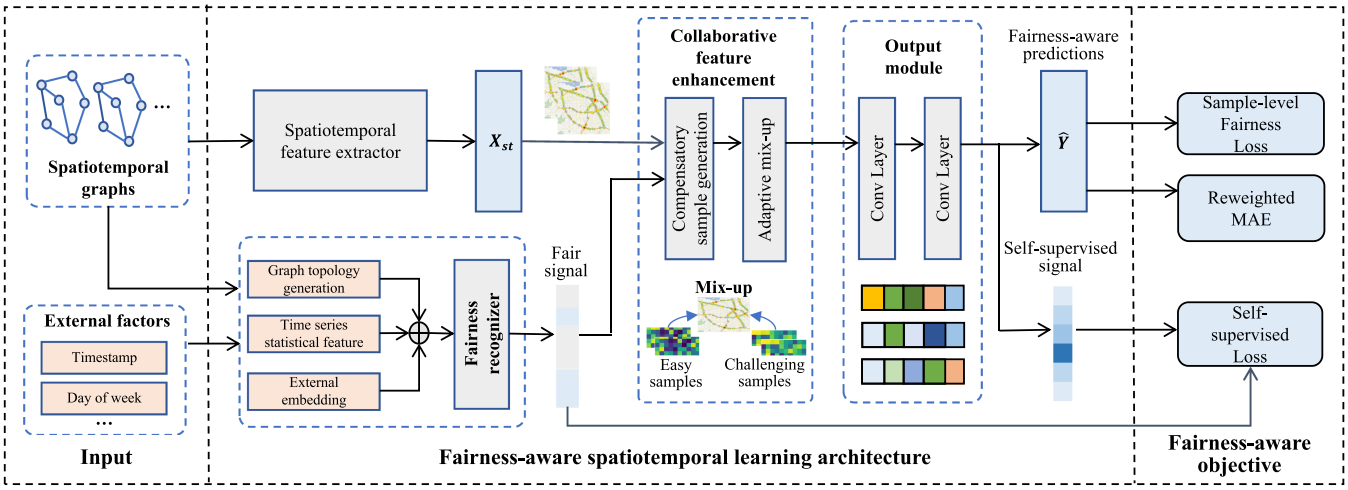


Fig. 2. Framework overview of FairSTG. The spatiotemporal feature extractor learns spatiotemporal representations from the original ST Graph. The fairness recognizer mines the learning difficulty and generates fairness signals in a self-supervised manner, and the collaborative feature enhancement adaptively transfers advantageous features from easy set to challenging set. Finally, the output module transforms the fused representations and produces the predictions.

f_{θ^*} to a fairness-aware version, which enforces the model to treat different spatiotemporal samples fairly and simultaneously maintains the original prediction performance. It can be formally formulated as,

$$\begin{aligned} &\tilde{\theta} = \arg_{\tilde{\theta}\in\Theta} \min D(\mathcal{Y}, \hat{\mathcal{Y}}; \tilde{\theta}) \\ &\text{s.t. MAE}(\mathcal{Y}, f_{\tilde{\theta}}(\mathcal{X})) \leq \delta \cdot \text{MAE}(\mathcal{Y}, f_{\theta^*}(\mathcal{X})) \end{aligned} \tag{3}$$

Here, $\delta > 0$ describes the trade-off between mitigating unfairness and the inevitable sacrifice in overall performance of the spatiotemporal prediction model. It also serves as an evaluation metric, measuring the overall performance sacrifice incurred by the alleviating the unfairness issue. If $\delta > 1$, it indicates that the model sacrifices some accuracy to ensure global fair predictions. If $0 < \delta \leq 1$, it suggests that the model simultaneously improves both fairness and forecasting performance.

V. Methodology

To counter the unfairness issue in spatiotemporal forecasting, we propose a novel Fairness-aware SpatioTemporal Graph learning (FairSTG), which ensures fair global predictions via collaborative feature transfer and fairness constraints. As illustrated in Fig. 2, FairSTG consists of four well-designed components, spatiotemporal feature extractor, fairness recognizer, collaborative feature enhancement and output module. Specifically, spatiotemporal feature extractor generates the spatiotemporal representations from the input ST Graphs. Fairness recognizer identifies the learning difficulty of samples and generates fairness signals in a self-supervised manner. Then, based on the fairness signals, we propose a collaborative feature enhancement to adaptively improve the informativeness of representations via transferring the advantageous representations from well-learned samples to those difficult to learn. Finally, we output fair forecasts.

A. Spatiotemporal Feature Extractor

The spatiotemporal feature extractor captures sequential patterns and spatial correlations from the input ST Graph. Existing literature designs various models for feature extraction based on specific application scenarios, such as RNN-based, CNN-based and GNN-based models. Our FairSTG is a model-independent framework, allowing different models to work as the backbone for the extractor, according to different application scenarios and data characteristics. In our work, we select two superior spatiotemporal GNN-based models MTGNN [13] and D^{2}STGNN [14] as the backbone model of the extractor. MTGNN designs a graph learning layer to extract uni-directional relations among nodes, and a novel mix-hop propagation layer and a dilated inception layer are further proposed to capture the spatial and temporal dependencies within the time series. D^{2}STGNN decouples and handles the diffusion and inherent traffic information separately. Note that in our work, the backbone can be easily replaced with any other prevalent models. For ease of description, we denote the spatiotemporal feature extractor as g_{st}, and the output spatiotemporal feature as $\mathbf{X}^{st}$, illustrated as,

$$\begin{aligned} \mathbf{X}^{st} &= g_{st}(\mathcal{X}) \\ &= T(G(\mathcal{X}; \mathbf{W}_G), \mathbf{W}_T) \end{aligned} \tag{4}$$

where we use T to represent the function capturing temporal correlations, G to represent the function capturing spatial correlations, and $\mathbf{W}_T$ and $\mathbf{W}_G$ denote learnable parameters. Equation (4) signifies that the feature extractor mines temporal and spatial correlations and generates spatiotemporal features.

B. Fairness-Aware Spatiotemporal Architecture

With the backbone of spatiotemporal learning well-trained, we introduce the our fairness-aware spatiotemporal learning architecture. Current machine learning solutions to fairness usually disentangle sensitive factors and derive a learning objective

to learn the representations independent of all sensitive factors [28], [41], [42]. However, given the tasks of spatiotemporal learning, they usually lack definite sensitive attributes, increasing the difficulty to identify which samples are prone to suffering unfair treatment. Therefore, from the representation perspective, we argue that fairness-aware spatiotemporal learning can be decomposed into two aspects. First, it is critical to explicitly identify which samples are easy-to-learn and otherwise, which are difficult-to-learn. Second, how to sufficiently exploit the advantages of well-learned representations to improve the quality of samples which are prone to unfair treatment. To this end, our fairness-aware spatiotemporal architecture is composed of two components, the fairness recognizer and the collaborative feature enhancement to systematically address above issues.

1) Fairness Recognizer: To remedy the lacking sensitive attributes in spatiotemporal datasets, we propose a fairness recognizer to identify whether the samples are difficult-to-learn. We first formally introduce the definition of easy samples and challenging samples.

Definition 3 (Easy samples and challenging samples): We characterize the learning difficulty of different spatiotemporal samples through the model's fitting degree of each sample. To determine the well-learned easy samples and challenging samples, given a sample $\boldsymbol{x}_i$, we utilize the error between the prediction and ground truth to quantify the model's fitting degree. Specifically, samples with K-smallest errors are categorized as the easy samples, denoted as $\mathcal{S}_e$, while other samples are categorized as the challenging samples, denoted as $\mathcal{S}_c$.[1]

To proactively identify the samples into easy ones and challenging ones, we propose a learnable **fairness recognizer**, which is motivated by the concept of computational identifiability [44], [45]. Given a family of binary functions $\mathcal{F}$, it is said that a subgroup $\mathcal{S}$ is computationally-identifiable if there exists a function $f : \mathbf{X} \to \{0, 1\}$ in $\mathcal{F}$ such that $f(\boldsymbol{x}) = 1$ if and only if $f(\boldsymbol{x}) \in \mathcal{S}$. Building on this definition, we propose a learnable fairness recognizer to identify the computationally-identifiable regions with relative high errors. Our learning-based recognizer establishes the connections between spatiotemporal samples and the identification of easy or challenging judgement. We therefore devise a self-supervised task to guide the learning difficulty of samples and generate the **fairness signals**, where the fairness signals indicate whether the given sample is challenging to learn, correspondingly to whether they are prone to be treated unfairly.

In the training phase, we obtain the prediction error of each sample where we can rank the samples by MAE errors and make a partition based on the rankings where easy samples with lower errors and challenging ones with higher errors. Then we can take these partition results as self-supervised signals and construct a self-supervised task via formulating a binary classification. We assign $z = 1$ as the self-supervised signal for easy samples, while assigning the labels for challenging samples as $z = 0$.

After then, we are going to determine the inputs of learnable fairness recognizer. First, recent literature has revealed that the spatiotemporal heterogeneity is mostly associated with the external factors (such as sampling time, sampling locations, weekdays, weather, etc.) [4], [46], we thus introduce these external factors into fairness recognizer to learn finer-grained features. Second, since spatiotemporal samples are essentially time series, adding statistical information of time series (such as mean, variance, etc.) will bring in informative knowledge including sequence trends and data distributions, empowering our framework to better characterize the patterns of time series. To this end, the auxiliary input of the fairness recognizer can be summarized as two aspects, the external spatiotemporal factors and statistical information of time series. We concatenate all these auxiliary factors into a fixed-length vector $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_M\}$ as the inputs of our fairness recognizer, where $\boldsymbol{c}_i = [\boldsymbol{x}_i^{st}; \boldsymbol{e}_i; \mu_i; \sigma_i^2] \in \mathbb{R}^{d_c}$, where $\boldsymbol{x}_i^{st}$, $\boldsymbol{e}_i$, μ_i and σ_i^2 respectively represent the spatiotemporal features, external factor embedding, sequence mean and sequence variance of a given spatiotemporal sample $\boldsymbol{x}_i$. We will elaborate on how to process external factors in the section of experiments.

Architecture design of the fairness recognizer: In fact, the learning difficulty is also concerned with spatiotemporal patterns, we thus exploit the Graph Convolutional Network (GCN) blocks as the basic architecture to instantiate our fairness recognizer. To be specific, we do not force our ST Graphs a predefined adjacency matrix, as the predefined topology may lack a direct relationship with the task, leading to significant bias. Inspired by previous study [13], we adopt an adaptive topology learning method to capture uni-directional relationships and generate the adjacency matrix as follows,

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} = \mathrm{ReLU}(\tanh(\mathbf{E}_1\mathbf{E}_2^\top - \mathbf{E}_2\mathbf{E}_1^\top)) \tag{5}$$

where $\mathbf{E}_1$, $\mathbf{E}_2$ represent randomly initialized node embedding. To reduce repeated and redundant calculations during the iterative training process, we directly produce $\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ rather than computing $\mathbf{A}$ and its Laplacian matrix. Thus, the lth GCN layer with adaptive topology learning can be formulated as,

$$\mathbf{H}^{(l)} = \mathrm{ReLU}(\mathbf{I} + \mathrm{ReLU}(\tanh(\mathbf{E}_1\mathbf{E}_2^\top - \mathbf{E}_2\mathbf{E}_1^\top)))\mathbf{H}^{(l-1)}\mathbf{W}^l \tag{6}$$

where $\mathbf{H}^{(l)}$ represents the hidden states in the lth GCN layer, $\mathbf{W}^l$ represents the parameter in the lth GCN layer, where in our work, the fairness recognizer consists of three stacked GCN layers. By denoting the fairness recognizer as g_{fa}, the output fairness signal of a given spatiotemporal sample $\boldsymbol{x}_i$ can be formulated as,

$$\hat{\boldsymbol{z}}_i = g_{fa}(\boldsymbol{c}_i|\mathcal{G}) \tag{7}$$

It is worth noting that the architecture design of the fairness recognizer plays an important role in controlling the granularity of computationally-identifiable regions. A more expressive recognizer architecture leads to a finer-grained identification but suffers more risks of overfitting to outliers. We will further analyze the impact of different network architecture designs of the fairness recognizer in the section of experiments.

2) Collaborative Feature Enhancement: After predicting the learning difficulty of spatiotemporal samples, the next challenge is how to proactively compensate for the perceived challenging

[1] Based on Pareto principle [43], we set K to 20%.

samples and improving the model's performance on this subgroup. We posit that spatiotemporal samples with similar patterns also exhibit similar representations. Based on this observation, for challenging samples, we obtain compensatory samples with similar patterns but better learning quality and design an attention-based strategy for advantageous feature transfer and fusion, which achieves adaptive transfer and enhancement of advantageous representations between easy set and challenging set.

Compensatory sample generation: Spatiotemporal samples with similar observations often exhibit similar patterns in representation spaces. In prevailing researches of spatiotemporal learning, surge of literature derives the node-level similarity via taking a series of node observations of nodes into similarity measurement functions or learning time-invariant node embeddings [2], [13], which ignores the temporal heterogeneity and results in indistinguishability between different samples. These solutions fail to satisfy the representation enhancement in sample levels. To capture finer-grained sample-level relationships, we establish a similarity matrix $\mathbf{S}$ between sample pairs within the current batch as follows,

$$\mathbf{S}_{ij} = \begin{cases} \text{SIM}(\boldsymbol{x}_i^{st}, \boldsymbol{x}_j^{st}) & \text{for } i \neq j \text{ and } \boldsymbol{x}_j \in \mathcal{S}_e \\ 0 & \text{others} \end{cases} \tag{8}$$

where $\boldsymbol{x}_i^{st}$ denotes the spatiotemporal feature of sample $\boldsymbol{x}_i$, and SIM represents the similarity measurement and we instantiate that as the cosine similarity in our work. We select the top-k_c samples both well-learned and most similar to it by learning in a collaborative manner, where the selected well-learned samples are designated as compensatory samples. For a given challenging sample $\boldsymbol{x}_i$, we select top-k_c most similar samples as its compensatory samples, denoted as $\{\boldsymbol{u}_{(i,1)}, \cdots, \boldsymbol{u}_{(i,k)}\}$. Then we aggregate the compensatory samples and denote the compensatory representation for $\boldsymbol{x}_i$ as $\boldsymbol{u}_{st}^i$ as,

$$\boldsymbol{u}_i^{st} = \text{AGGREGATE}(\{\boldsymbol{u}_{(i,1)}^{st}, \cdots, \boldsymbol{u}_{(i,k)}^{st}\}) \tag{9}$$

where $\boldsymbol{u}_{(i,j)}^{st}$ represents the spatiotemporal feature of compensatory sample $\boldsymbol{u}_{(i,j)}$. There are many options for AGGREGATE function, and we instantiate it as MEAN-POOLING.

Adaptive mix-up for fair representation: We then achieve the awareness of the learning difficulty of each sample and generate compensatory representation for each challenging sample. Due to the spatiotemporal heterogeneity among different samples, we must devise a personalized fusion method to allow the adaptive representation aggregation. Therefore, we propose an attention-based mix-up strategy. Formally, we can derive the mix-up strategy as follows,

$$\begin{aligned} \mathbf{Q} &= \boldsymbol{x}_i^{st}\mathbf{W}_q \\ \mathbf{K} &= \boldsymbol{u}_i^{st}\mathbf{W}_k \\ \alpha &= \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \\ \alpha' &= \text{MLP}(\alpha \boldsymbol{u}_i^{st}) \end{aligned} \tag{10}$$

where $\boldsymbol{x}_i^{st} \in \mathbb{R}^d$, $\boldsymbol{u}_i^{st} \in \mathbb{R}^d$, $\mathbf{W}_q \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}_k \in \mathbb{R}^{d \times d_k}$ are learnable parameters, and we constrain α' within the range (0,0.5) to preserve the intrinsic representation.

Finally, for any challenging sample $\boldsymbol{x}_i$, we obtain the mix-up representation $\boldsymbol{x}_i^{com}$ by fusing the intrinsic representation $\boldsymbol{x}_i^{st}$ and the compensatory representation $\boldsymbol{u}_i^{st}$ by,

$$\boldsymbol{x}_i^{com} = (1 - \alpha')\boldsymbol{x}_i^{st} + \alpha \boldsymbol{u}_i^{st} \tag{11}$$

$\boldsymbol{x}_i^{com}$ is the well-remedied representation generated with sample-level joint optimization and knowledge transfer.

In summary, at representation level, the fairness recognizer and the collaborative feature enhancement can work cooperatively, initially employing the concept of computationally-identifiable subgroup to design an auxiliary self-supervised task for perceiving the learning difficulty of diverse spatiotemporal samples. Thus our FairSTG can overcome lacking sensitive attributes. For challenging samples which are prone to suffering unfairness, samples with similar patterns but well-learned representation are selected for compensation, enhancing the overall performance.

C. Fairness-Aware Learning Objective

Recent literature has demonstrated that incorporating fairness metrics into learning objective can effectively remove discrimination during the training process [47]. To this end, in addition to collaboratively enhancing representations of challenging samples, our FairSTG imposes fairness constraints to improve the learning objective, which forces the model to provide fair treatment for diverse samples. After obtaining the compensated spatiotemporal representations, the output module produces the final predictions at one forward step, and the fairness-aware optimization objective constrains the model to optimize towards treating all samples fairly. Besides, considering the quality of fairness recognizer is highly relied on the feature extractor, we thus further devise a two-stage training strategy with a warm-up phase to pre-generate the extractor parameters. Therefore, in this subsection, we first introduce our output module and then formally present our optimization objective, and finally provide the two-stage training strategy, serving for fairness-aware learning.

1) Output Module: The output module is instantiated with two 1×1 standard convolutional layers, which transforms the input channel dimension to the output dimension in an one-step-forward manner rather than a step-by-step style. For a spatiotemporal sample $\boldsymbol{x}_i \in \mathbb{R}^w$, the resulting output is $\hat{\boldsymbol{y}}_i \in \mathbb{R}^h$, where h is the fixed dimension for outputs.

2) Optimization Objective: Given the fairness-aware learning architecture, we can systematically derive the objective of our FairSTG, which can be summarized as three aspects, i.e., reweighted loss for main spatiotemporal task by emphasizing the challenging samples, sample-level fairness-aware loss, and self-supervised objective for learning a fairness recognizer.

Reweighted loss for main task: Actually, from the perspective of learning difficulty, each sample does not share the same importance in the same training batch towards final optimizations, e.g., some samples are easy to learn while others suffer higher errors. We thus introduce a reweighted loss to regularize the

main task. We first present the initial main task of spatiotemporal learning, which is instantiated as a regression problem associated with MAE,

$$\text{MAE}(\mathcal{Y}, \hat{\mathcal{Y}}) = \frac{1}{M}\sum_{i=1}^{M} |\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i| \tag{12}$$

To increase the emphasis of difficult-to-learn samples, we assign higher weights to samples with higher errors, which guides the model to pay more attention on these subgroups. The weights are normalized within batches to prevent gradient explosions. Specifically, we perform a normalization step that rescales the error to [0, 1] and produce the cost-sensitive weight λ_i, and we center the weight and add 1 to ensure that all training samples to make sense to final loss. For any spatiotemporal sample $\boldsymbol{x}_i$, the corresponding weight λ_i can be formulated as follows,

$$\lambda_i = 1 + \frac{|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i|}{\sum_{j=1}^{M} |\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j|} \tag{13}$$

Furthermore, we construct the reweighted loss based on the cost-sensitive weights as,

$$\mathcal{L}_r = \frac{1}{M}\sum_{i=1}^{M} \lambda_i \cdot |\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i| \tag{14}$$

This reweighted loss can enable the learning objective more sensitive to challenging samples, thus increasing the attention for modeling challenging ones. With the process of model training, the λ will be dynamic with learning process and adjust the learning in fine-grained manner.

Fairness loss: Second, to explicitly ensure the fairness of prediction performance across different samples, we exploit the variance of MAE as the fairness constraint term to directly guide the model's learning process, i.e.,

$$\mathcal{L}_f = \frac{1}{M}\sum_{i=0}^{M-1} [|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i| - \frac{1}{M}\sum_{j=0}^{M-1} |\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j|]^2 \tag{15}$$

With this fairness loss, we actually introduce the awareness of fairness into our training process from the perspective of immediate strategy, which guarantees the equitable performance across samples.

Self-supervised loss: Third, the self-supervised learning objective of our fairness recognizer is instantiated with the Balanced Cross-Entropy loss (BCE Loss). To be specific, we take the ranking of prediction performance as the self-supervised signals by assigning $z = 1$ and $z = 0$ for easy and challenging samples respectively. Then the fairness recognizer output $\hat{z}$ can be compared with self-supervised signal z with the following contrastive loss, i.e.,

$$\mathcal{L}_s(\hat{\boldsymbol{z}}, \boldsymbol{z}) = -\omega * (\boldsymbol{z} * \ln \hat{\boldsymbol{z}}) + (1 - \boldsymbol{z}) * \ln(1 - \hat{\boldsymbol{z}}) \tag{16}$$

where ω represents the weight for positive samples, which is set based on the proportion of easy and challenging samples set. In our work, we set $\omega = 4$ via experimental trials.

Overall fairness-aware learning loss: By summarizing the above loss terms, the final loss function can be written as,

$$\mathcal{L} = \mu_r \mathcal{L}_r + \mu_f \mathcal{L}_f + \mu_s \mathcal{L}_s \tag{17}$$

TABLE II
THE STATISTICS OF DATASETS

Dataset	Number of samples	Number of nodes	Sampling interval
Metr-LA	34272	207	5 minutes
PEMS-BAY	52116	325	5 minutes
KnowAir	11688	184	3 hours
ETT	17420	7	1 hour

where μ_r, μ_f and μ_s are hyper-parameters.

It is worth highlighting that we introduce fairness constraints at the level of optimization objectives. On the one hand, we design a reweighted loss for the main task based on cost-sensitive weight, to emphasize more attention on challenging samples which are prone to unfair treatment and make timely adjustment. On the other hand, we devise a fairness loss term to directly minimize the performance difference across samples during the whole learning pipeline. These two objectives can work cooperatively to improve prediction fairness technically, not only making necessary adjustment to sample weights, but also minimizing the performance heterogeneity with an explicit objective constraint.

3) Training Strategy: Given that the accuracy of fairness recognizer is highly relied on the quality of the representations generated by the spatiotemporal feature extractor, we design a two-stage training strategy with a warm-up phase and a fairness-aware learning phase.

- Warm-up phase: We only train the spatiotemporal feature extractor and output module. In this phase, we set $\mu_r \neq 0$, $\mu_f = 0$, $\mu_s = 0$, and $\lambda_i = 1$ for all samples.
- Fairness-aware learning: All modules are trained and we set $\mu_r \neq 0$, $\mu_f \neq 0$ and $\mu_s \neq 0$. We will discuss the selection of hyper-parameters in detail in the experiment section.

With such two-stage training strategy, we can formulate a well-learned fairness recognizer to produce accurate fairness signals and then the representations can be collaboratively enhanced with our fairness-aware learning architecture.

VI. EXPERIMENTS

A. Dataset

We select four real-world mobility-related spatiotemporal datasets, from human mobility, air quality to smart grids, to evaluate the effectiveness and generality of our FairSTG. We summarize statistics of benchmark datasets in Table II, where the samples refer to the input-output pair for the model.

- METR-LA [13]: It indicates the traffics and inter-city human mobility, containing the average traffic speed measured by 207 loop sensors on the highways of Los Angeles County, from Mar 2012 to Jun 2012.
- PEMS-BAY [13]: It is the mobility dataset, composed of average traffic speed measured by 325 sensors in the Bay Area, ranging from Jan 2017 to May 2017.
- KnowAir [48]: This dataset records the PM2.5 concentrations, one of the important air quality index, in 184 main cities in China, where it is collected every 3 hours from Sep 2016 to Jan 2017.

- ETT [49]: The ETT (Electricity Transformer Temperature) is a crucial indicator of electric power in cities. The dataset is collected every 1 h from July 2016 to July 2018 where it consists of 7 features, and each consumer is considered as a distinct node.

B. Implementation Details and Evaluation Metrics

Implementation details: Following practices of classical time-series data split [3], [13], we split all the available data into training, validation, and testing parts with the ratio of 7:2:1 in chronological order. We normalize the data with standard normalization to ensure the stability of training process. We set the input window $w = 12$ and the output window $h = 12$. The model is trained by the Adam optimizer with gradient clip 5. The learning rate is set as 0.001. In our experiments, the dimension of the spatiotemporal representations and compensatory representations is fixed to 64. We examine the number of compensatory samples k_c in range of $\{5, 10, 20\}$ and choose 5 for all datasets. Regarding optimization objective, the parameter μ_r is fixed to 1 while μ_s is fixed to 0.1, and we examine the trade-off fairness parameter μ_f in range of $\{0.01, 0.1, 0.5, 1.0, 1.5\}$ for different datasets, we finally set 0.5 for METR-LA and PEMS-BAY, 0.1 for KnowAir and ETT.

Regarding some baselines requiring the predefined adjacent matrix, we extract the pairwise distances to measure the node-wise proximity as the adjacencies. For METR-LA and PEMS-BAY, following previous work [3], we compute distances between sensors within the road network and build the adjacency matrix using thresholded Gaussian kernel. For KnowAir, we compute the geographical distance between sampling points to create an adjacency matrix, and retain the top 20% of nodes with the highest weights as neighbors for each node to ensure the adjacency sparsity. For ETT, we take the most recent month to construct the adjacency with cosine similarity.

Concerning external factors, we separate them into two groups, i.e., continuous and categorical features. Continuous factors including temperature and timestamps are directly concatenated into a vector $\boldsymbol{e}^{con}$. Categorical factors including weekday, weather conditions are separately projected into a low-dimensional continuous vector space through embedding layers. These embeddings are then concatenated to form the vector $\boldsymbol{e}^{cat}$. Specifically, to enable the fairness recognizer to characterize the learning difficulty more accurately, we combine not only spatiotemporal external factors but also the sequence statistical information. Without loss of generality, for each sample $\boldsymbol{x}_i$ and the corresponding spatiotemporal representation $\boldsymbol{x}^i_{st}$, we compute the statistical information, mean μ_i and variance σ_i^2 of corresponding sample $\boldsymbol{x}_i$ in the element level. And we convert the timestamps of day into continuous values as $\boldsymbol{e}^{con}$. For indicator of day of week and node indexes, we organize them as categorical vector $\boldsymbol{e}^{cat}$. And continuous embedding $\boldsymbol{e}^{con}$ and categorical embedding $\boldsymbol{e}^{cat}$ are concatenated to form the external embedding $\boldsymbol{e_i}$ for the corresponding sample. In addition, those vectors are concatenated as the input of our fairness recognizer, denoted as $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_M\}$, where $\boldsymbol{c}_i = [\boldsymbol{x}^i_{st}; \boldsymbol{e}_i; \mu_i; \sigma_i^2] \in \mathbb{R}^{d_c}$.

Evaluation metrics: The goal of FairSTG is to improve the sample-level prediction fairness with simultaneously retaining the overall performances, and we also argue that the error-based metrics without considering prediction heterogeneity deduce the practical function of model evaluation. To this end, we incorporate five evaluation metrics that can be divided into two aspects, prediction accuracy evaluation and quality of fairness learning, where the fairness quality evaluation helps remedy the function degeneration of only error metrics.

We take Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) as evaluation metrics for accuracy, while introduce the variance of sample-level errors over the testing set, i.e., sample-level MAE-var and MAPE-var, as the fairness metrics.

C. Baselines

We compare our proposed framework with 6 state-of-the-art baselines as follows.

- DCRNN [3]: A graph-based recurrent neural network, which combines graph diffusion convolutions with recurrent neural network.
- STGCN [22]: A spatiotemporal graph convolutional network, incorporating graph convolutions with 1D convolutions.
- AGCRN [2]: A spaiotemporal graph convolutional network, which captures node-specific patterns and infers the inter-dependencies among time series automatically.
- MTGNN [13]: A spatiotemporal graph convolutional network, which integrates adaptive graph learning, graph convolutions and temporal convolutions.
- D^{2}STGNN [14]: A decoupled dynamic spatiotemporal neural networks, which decouples and handles the diffusion and inherent information separately.
- ST-SSL [1]: A spaiotemporal learning traffic prediction framework, which designs two self-supervised auxiliary tasks with a contrastive learning objective, to gain awareness of spatial-temporal heterogeneity and supplement the main forecasting task.

D. Experimental Results

Since our FairSTG is a general pluggable spatiotemporal learning framework alleviating the unfairness issue, we choose two advanced spatiotemporal learning models, MTGNN and D^{2}STGNN, as the backbones for the spatiotemporal feature extractor. We implement 3-step, 6-step and 12-step prediction, which is illustrated as horizons in our result tables, and all the solutions are evaluated based on five metrics regarding both fairness metrics and forecasting performance. Table III elaborates the forecasting performance and fairness metrics across different models. The best results are **bolded** and the runner-up are underlined. And to further illustrate the improvements of our FairSTG over the corresponding backbone model, we calculate the fairness and forecasting performance improvements of FairSTG to the corresponding backbone model in Table IV.

Performance on fairness metrics: Our FairSTG significantly surpasses all other baselines on fairness metrics across all

TABLE III
MAIN COMPARISON RESULTS

	horizon-3					horizon-6					horizon-12				
	MAE	MAPE	RMSE	MAE var	MAPE var	MAE	MAPE	RMSE	MAE var	MAPE var	MAE	MAPE	RMSE	MAE var	MAPE var
METR-LA															
DCRNN	2.7022	6.96%	5.2389	23.8967	0.0755	3.1126	8.51%	6.3414	36.0244	0.1201	3.5705	10.40%	7.5242	51.6100	0.1825
STGCN	3.4535	8.51%	7.7567	56.4673	0.0757	4.4417	11.45%	10.1028	96.2961	0.1430	5.9071	15.75%	12.9456	155.6183	0.2589
AGCRN	3.3413	8.33%	7.5762	54.1650	0.0789	4.0211	10.15%	9.3950	84.2929	0.1240	4.9653	12.53%	11.6030	128.5844	0.1820
MTGNN	2.6793	6.90%	5.1800	22.6463	0.0738	3.0409	8.19%	6.1705	33.3740	0.1121	3.4952	9.87%	7.2361	47.6894	0.1598
D^2STGNN	**2.5602**	**6.36%**	4.9837	23.1246	0.0524	**2.9194**	**7.72%**	6.0272	34.8865	0.0932	**3.3567**	9.43%	7.1357	49.4757	0.1431
ST-SSL	2.8243	7.39%	5.4742	26.1182	0.0816	3.2722	9.14%	6.6587	39.7393	0.1382	3.7280	10.67%	7.6999	53.5600	0.1781
FAirST+MTGNN	2.7316	7.02%	5.0509	**21.5644**	0.0684	3.1291	8.39%	6.0534	**31.9005**	0.1037	3.5521	10.17%	7.1207	45.5186	0.1525
FAirST+D^2STGNN	2.6020	6.47%	**4.8756**	21.6349	**0.0499**	2.9602	7.80%	**5.8432**	32.0360	**0.0859**	3.4316	**9.40%**	**6.8938**	**44.9188**	**0.1254**
PEMS-BAY															
DCRNN	1.3152	**2.75%**	2.7785	5.9908	0.0083	**1.6521**	3.71%	3.7743	11.5165	0.0229	1.9634	4.64%	4.5921	17.2332	0.0579
STGCN	1.3731	2.93%	2.8451	6.2096	0.0128	1.8013	4.16%	4.0325	13.0175	0.0409	2.3241	5.49%	5.1715	21.3443	0.0642
AGCRN	1.4135	3.09%	3.0233	7.1421	0.0132	1.7500	4.05%	3.9967	12.9109	0.0311	2.0721	4.91%	4.7852	18.6050	0.0521
MTGNN	1.3361	2.80%	2.8079	6.0993	0.0082	1.6687	3.80%	3.7648	11.3900	0.0249	1.9899	4.77%	4.5535	16.7757	0.0508
D^2STGNN	**1.3093**	2.83%	**2.7424**	5.8562	0.0089	1.6540	3.92%	3.7751	11.3713	0.0343	1.9818	4.84%	4.5864	17.0290	0.0624
ST-SSL	1.4030	2.91%	3.0104	7.0943	0.0081	1.8160	4.16%	4.0912	13.4398	0.0313	2.1250	5.17%	4.7808	18.3406	0.0640
FAirST+MTGNN	1.3312	2.81%	2.7445	**5.7608**	**0.0078**	1.6550	**3.70%**	**3.6613**	**10.6667**	**0.0228**	**1.9628**	**4.58%**	**4.3885**	**15.4070**	**0.0448**
FAirST+D^2STGNN	1.3678	2.96%	2.7773	5.8435	0.0086	1.6977	4.02%	3.7382	11.0251	0.0331	2.0149	4.97%	4.4479	15.6454	0.0589
KnowAir															
DCRNN	19.4362	39.03%	30.3832	545.3596	0.5182	24.6547	55.71%	37.3885	790.0374	1.2729	32.7574	86.90%	47.4751	1180.8402	3.3537
STGCN	13.7359	40.48%	20.6039	235.8448	0.4574	16.2138	51.05%	23.4688	287.8995	0.6644	27.7782	87.89%	35.0219	454.9049	1.8565
AGCRN	14.9210	49.35%	23.9165	349.5797	0.7642	17.2767	60.84%	26.7793	418.9212	1.2171	19.7528	73.67%	29.7286	493.9549	1.7813
MTGNN	11.9372	38.93%	19.3084	230.4142	0.4255	14.3814	47.84%	22.9801	321.3987	0.6921	17.0799	59.89%	26.5632	414.0658	1.1253
D^2STGNN	12.0410	37.85%	**19.0750**	228.8712	0.4370	14.7446	48.29%	23.1617	319.0642	0.7677	18.3538	63.46%	28.3847	468.8282	1.3700
ST-SSL	11.8982	38.78%	19.2251	228.0394	0.4399	14.2621	47.54%	**22.5824**	316.5560	0.6987	17.1275	62.69%	26.6144	414.9765	1.2057
FAirST+MTGNN	**11.8854**	**37.03%**	19.2816	**225.6162**	**0.3517**	**14.1500**	**46.20%**	23.1401	**305.4699**	**0.0600**	**16.7561**	**58.98%**	**26.0058**	**395.7106**	**1.0603**
FAirST+D^2STGNN	12.4411	39.86%	19.4670	226.3241	0.4174	14.8216	48.49%	23.1093	314.5615	0.7352	17.8283	60.69%	27.5562	441.7820	1.2329
ETT															
DCRNN	2.0154	21.03%	3.0606	6.8057	0.1713	2.5277	26.65%	4.8280	10.5346	0.3474	2.9636	30.76%	4.3930	13.5997	0.5401
STGCN	3.7973	47.16%	5.0410	10.9922	0.5943	3.4983	30.87%	11.0721	3.4983	0.1272	3.1639	25.57%	4.5034	10.2705	0.0789
AGCRN	1.4100	9.14%	2.0400	2.2083	0.0275	1.6100	10.32%	2.3900	3.1247	0.0249	1.7500	13.15%	2.6177	3.9166	0.0250
MTGNN	1.5023	12.81%	2.1853	2.5811	0.0267	1.8136	14.45%	2.7287	4.2237	0.0304	1.9305	15.44%	2.9455	4.8330	0.0392
D^2STGNN	**1.2199**	**10.35%**	**1.8181**	1.9608	0.0181	**1.5175**	**12.27%**	**2.1841**	2.8447	0.0222	1.7474	13.38%	2.6303	3.9519	0.0231
ST-SSL	1.4541	12.83%	2.0891	2.3073	0.0265	1.9052	14.19%	2.8121	4.3810	0.0189	1.9843	14.97%	2.8741	4.4265	0.0247
FAirST+MTGNN	1.3625	10.95%	2.0019	2.2038	**0.0146**	1.6359	12.49%	2.4444	3.3768	0.0187	1.7832	12.89%	2.7225	4.3247	0.0223
FAirST+D^2STGNN	1.3081	10.92%	1.8956	**1.9292**	0.0155	1.5397	12.33%	2.2631	**2.8174**	**0.0185**	**1.6725**	**12.77%**	**2.5542**	**3.8085**	**0.0201**

TABLE IV
THE PERFORMANCE AND FAIRNESS COMPARISON BETWEEN FAIRSTG AND ITS CORRESPONDING BACKBONE

	horizon-3				horizon-6				horizon-12			
	MAE	forecasting improvement	MAE var	fairness improvement	MAE	forecasting improvement	MAE var	fairness improvement	MAE	forecasting improvement	MAE var	fairness improvement
METR-LA												
MTGNN	2.6793		22.6463		3.0409		33.3740		3.4952		47.6894	
FairSTG+MTGNN	2.7316	-1.9520%	21.5644	4.7774%	3.1291	-2.9005%	31.9005	4.4151%	3.5521	-1.6279%	45.5186	4.5520%
D^2STGNN	2.5602		23.1246		2.9194		34.8865		3.3567		49.4757	
FairSTG+D^2STGNN	2.6020	-1.6327%	21.6349	6.4421%	2.9602	-1.3975%	32.0360	8.1708%	3.4316	-2.2314%	44.9188	9.2104%
PEMS												
MTGNN	1.3361		6.0993		1.6687		11.3900		1.9899		16.7757	
FairSTG+MTGNN	1.3312	0.3667%	5.7608	5.5498%	1.6550	0.8210%	10.6667	6.3503%	1.9628	1.3619%	15.4070	8.1588%
D^2STGNN	1.3093		5.8562		1.6540		11.3713		1.9818		17.0290	
FairSTG+D^2STGNN	1.3678	-4.4680%	5.8435	0.2169%	1.6977	-2.6421%	11.0251	3.0445%	2.0149	-1.6702%	15.6454	8.1250%
KnowAir												
MTGNN	11.9372		230.4142		14.3814		321.3987		17.0799		414.0658	
FairSTG+MTGNN	11.8854	0.4339%	225.6162	2.0823%	14.1500	1.6090%	305.4699	4.9561%	16.7561	1.8958%	395.7106	4.4329%
D^2STGNN	12.0410		228.8712		14.7446		319.0642		18.3538		468.8282	
FairSTG+D^2STGNN	12.4411	-3.3228%	224.3241	1.1129%	14.8216	-0.5222%	314.5615	1.4112%	17.8283	2.8632%	441.7820	5.7689%
ETT												
MTGNN	1.5023		2.5811		1.8136		4.2237		1.9305		4.8330	
FairSTG+MTGNN	1.3625	9.3057%	2.2038	14.6178%	1.6359	9.7982%	3.3768	20.0511%	1.7832	7.6301%	4.3247	10.5173%
D^2STGNN	1.2199		1.9608		1.5175		2.8447		1.7474		3.9519	
FairSTG+D^2STGNN	1.3081	-7.2301%	1.9292	1.6116%	1.5397	-1.4629%	2.8174	0.9597%	1.6725	4.2864%	3.8085	3.6286%

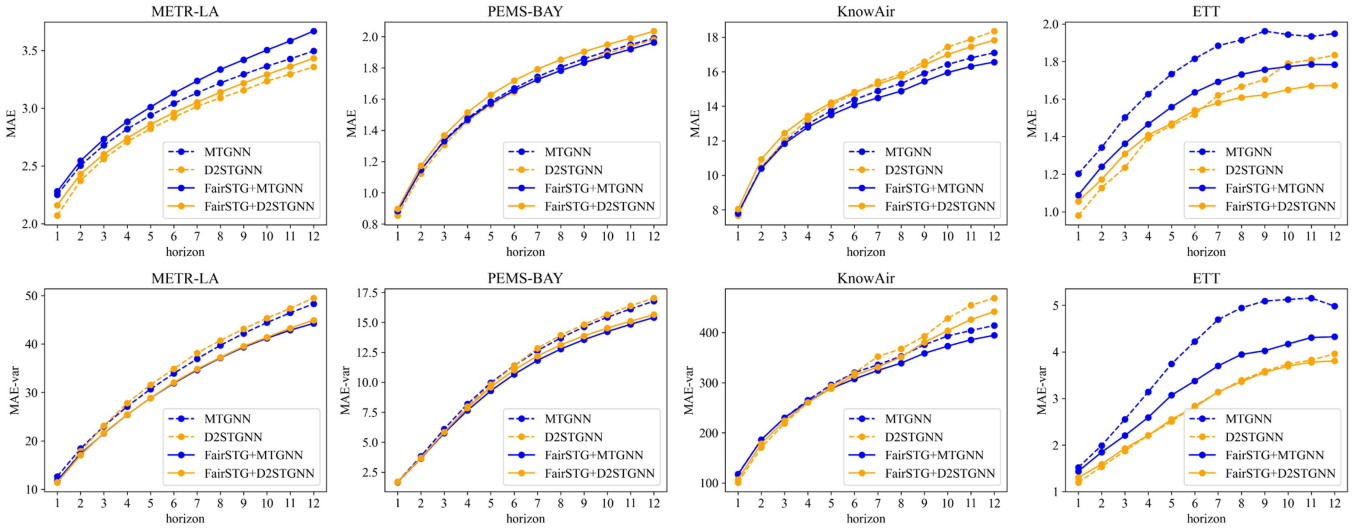


Fig. 3. Forecasting and fairness performance comparison at each horizon. The top and bottom lines respectively indicate the prediction accuracy and fairness performances.

tasks. Overall, as illustrated in Table IV, FairSTG improves fairness metric ranging from 0.21% to 20.05% with 16 out of 24 achieving improvement above 4%. Specifically, for the farthest step prediction, the MAE-var improvements at horizon-12 over the best baseline are 5.59%, 8.15%, 4.43% and 2.76% on METR-LA, PEMS-BAY, KnowAir and ETT, respectively, and the improvements on MAPE-var are 12.36%, 12.01%, 5.77% and 12.98% on METR-LA, PEMS-BAY, KnowAir and ETT. We attribute such superiority to two aspects, i.e., 1) Baseline models primarily focus on improving overall performance, inducing the overfitting of samples with lower difficulty and neglecting the challenging subgroup. 2) Our FairSTG is equipped with advantageous compensation mechanism and introduces fairness constraints on both representation space and optimization objective, thereby enhancing the model's fairness performance.

Performance on forecasting metrics: We observe that our FairSTG framework achieves comparable forecasting accuracy in all datasets, noting that there usually exists a trade-off between fairness and accuracy [50], [51], [52]. Particularly, our FairSTG achieves optimal forecasting performance at horizon-12 on all datasets. As illustrated in Table IV, the forecasting achieves better performances in several horizons, with improvements from 0.43% to 9.79%, while slightly declines ranging from 0.52% to 2.90% with only one 7.23% sacrifice on ETT-horizon-3. Given the satisfactory improvement on fairness, we believe the performance loss is exactly within an acceptable range. The superiority of forecasting metrics can be two-folded. First, as the forecasting horizon increases, the regularity between predictions and inputs weakens and the learning difficulty increases. Thus, learning without fairness can be trapped into a lazy learning mode, directly neglecting the samples with higher learning difficulty. Second, in addition to fairness-aware learning objective, our FairSTG is capable of suppressing the prediction accuracy diversity via the joint optimization of sample-level representation. i.e., collaborative feature enhancement and adaptive mix-up for fair representation. Therefore, our FairSTG reasonably achieves superior performance on both evaluations of accuracy and fairness.

Generality and plug-ability of FairSTG: It is worth noting that our FairSTG is a model-independent fairness guarantee framework where its performance is constrained by the backbone model, thus we should pay more attention on the improvement against the corresponding backbone. We further compare the empirical prediction performances of our FairSTG with two backbone models. As shown in Fig. 3, the fairness metrics are consistently superior to backbones along with the increases of forecasting horizons. Besides, we observe that our FairSTGs carrying with different backbones can achieve similar and comparable performance with corresponding backbones, illustrating the stability of the whole architecture of of FairSTG. It is worth noting that our FairSTG achieves optimal performance in both KnowAir and ETT datasets. The reason behind this phenomenon lies in that these two datasets have relatively small scales, where the collaborative feature enhancement of FairSTG can exactly alleviate data sparsity and improve the overall performance. Such satisfactory joint enhancement further validates the intuition that mitigating forecasting heterogeneity contributes to the promotion of overall performance. In summary, our integrated FairSTG can enable the fairness for different backbone models and maintain comparable forecasting performance, which verifies the generality of our FairSTG framework.

E. The Improvement in Challenging Samples

One of the goals of our FairSTG is to alleviate the unfairness on challenging samples which are difficult to learn, thus it is necessary to verify whether FairSTG improves the forecasting performance on such subgroup. For an explicit comparison, we select samples with performance at the top 30% as easy ones, and samples with performance at the bottom 30% as

TABLE V
THE COMPARISON ON EASY SET AND CHALLENGING SET

	easy set MAE	easy set MAE-var	challenging set MAE	challenging set MAE-var
METR-LA				
MTGNN	0.1687	0.0355	10.0551	105.2392
D^2STGNN	**0.1026**	**0.0201**	9.5583	112.0150
FairSTG+MTGNN	0.1834	0.0419	9.9270	**95.2383**
FairSTG+D^2STGNN	0.1102	0.0231	**9.4708**	96.2490
PEMS-BAY				
MTGNN	0.2111	0.0160	5.2211	40.5645
D^2STGNN	**0.2037**	**0.0153**	5.3478	41.3632
FairSTG+MTGNN	0.2176	0.0169	**5.1770**	37.1559
FairSTG+D^2STGNN	0.2273	0.0185	5.2904	**36.5698**
KnowAir				
MTGNN	3.1115	3.3080	38.2072	679.0028
D^2STGNN	3.2667	3.6180	41.5705	724.8887
FairSTG+MTGNN	**3.0611**	**3.2038**	**36.7819**	**669.3046**
FairSTG+D^2STGNN	3.1327	3.3571	40.4428	677.5437
ETT				
MTGNN	0.2987	0.0369	4.4653	6.5106
D^2STGNN	0.3179	0.0412	4.0504	5.0291
FairSTG+MTGNN	**0.2648**	**0.0268**	4.0310	**5.4197**
FairSTG+D^2STGNN	0.2657	0.0292	**3.8328**	5.4260

challenging ones. We then compare the integrated FairSTGs with corresponding vanilla backbones. As shown in Table V, our FairSTG outperforms all baseline models on challenging set at both accuracy metric MAE and fairness metric MAE-var, while maintaining comparable performance on easy set with satisfactory MAE and MAE-var. Additionally, for each group learned by FairSTG, performances on the challenging set overwhelmingly surpass the corresponding backbone models on accuracy metric, i.e., MAE, which refers to that our FairSTG indeed improves model's expressive capacity on challenging samples which are prone to suffer unfair treatment, and thus mitigating the forecasting heterogeneity and unfairness issue in spatiotemporal learning. Moreover, on datasets of KnowAir and ETT, our FairSTG simultaneously achieves optimal results on forecasting accuracy and fairness metrics, on both easy set and challenging set. As analyzed earlier, since KnowAir and ETT are with a relatively small scale, the collaborative feature enhancement in our FairSTG can alleviate the data sparsity and thus improving the overall forecasting and fairness performance. In brief, through the analysis on inter-group and intra-group prediction performance and fairness learning quality, we can further verify that FairSTG is equipped with the capacity of alleviating unfairness on challenging samples. Therefore, our FairSTG can exactly compensate the samples with lower regularity and under-representation, potentially suppressing the discrimination in urban intelligent system.

F. The Analysis of Fairness Recognizer

The effectiveness of the fairness recognizer: We first verify the correctness of our intuition, i.e., the learning difficulty of a spatiotemporal sample should be associated with its sequential features and external factors. Even though spatiotemporal datasets lack explicit supervision of annotations on challenging samples, we design a fairness recognizer, to learn the difficulty of different spatiotemporal samples during training and make inference during testing. Due to the self-supervised design without explicit labels in both training and testing phases, in Fig. 4, we report the accuracy of the self-supervised binary classification task on both training set and testing set in FairSTG with the backbone D^2STGNN, as the accuracy on both sets not only provides empirical guarantee, but also confirms the consistent results of our proposed fairness recognizer between training and testing sets.

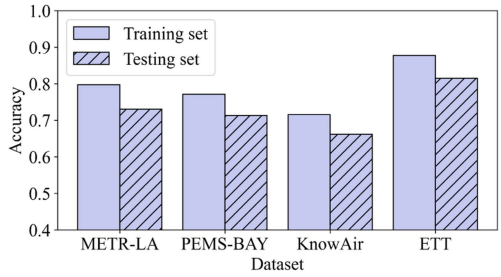


Fig. 4. The accuracy of fairness recognizer in self-supervised classification.

TABLE VI
THE RESULTS OF DIFFERENT ARCHITECTURES FOR THE FAIRNESS RECOGNIZER

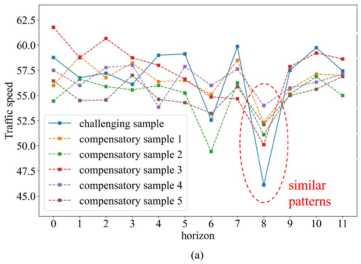
	MAE	MAE-var	Training accuracy	Testing accuracy
METR-LA				
FairSTG-GCN(3)	**3.4316**	**44.9188**	**0.7972**	0.7309
FairSTG-Linear(3)	3.4509	46.5286	0.7860	**0.7471**
PEMS-BAY				
FairSTG-GCN(3)	**2.0349**	**15.6454**	**0.7714**	0.7131
FairSTG-Linear(3)	2.0352	15.9652	0.7171	**0.7296**
KnowAir				
FairSTG-GCN(3)	**17.8283**	**441.7820**	**0.7159**	**0.6619**
FairSTG-Linear(3)	18.356	460.5098	0.7054	0.625
ETT				
FairSTG-GCN(3)	**1.6725**	**3.8085**	0.8774	**0.8149**
FairSTG-Linear(3)	1.7056	4.0518	**0.8779**	0.7852

The accuracy of fairness recognizer ranges from 72.20% to 87.40% on the training set, while accounts for the ranges from 66.19% to 81.49% on the test set. The accurate and high-quality predictions on all four datasets indicate that challenging samples can be well-identified in the absence of explicitly labeled sensitive attributes and can exactly satisfy further fair predictions.

The architecture of fairness recognizer: We investigate how the neural network architecture influences the performance of fairness recognizer. In our main experiments, we utilize a three-layer GCN as the architecture to capture the spatiotemporal correlations. We then take a linear three-layer MLP as an alternative for the impact analysis, where the initialized and modified one are designated as FairSTG-GCN(3) and FairSTG-Linear(3). Our implementation is based on D^2STGNN at horizon $= 12$ and the results are reported in Table VI. In terms of the accuracy of self-supervised classification, FairSTG-GCN(3) surpasses FairSTG-Linear(3) in almost all scenarios. This is probably

TABLE VII
THE DETAILED INFORMATION ABOUT THE TOP FIVE NODES WITH THE LARGEST IMPROVEMENT

Node ID	Longitude	Latitude	Δe_i
10	-121.96	37.26	1.77
46	-121.95	37.27	1.13
150	-121.85	37.34	1.01
45	-121.99	37.41	0.95
29	-122.03	37.30	0.87



(a)

	Node	Sample time
challenging sample	21	Wed. 13:05
compensatory sample 1	39	Wed. 13:05
compensatory sample 2	179	Wed. 13:05
compensatory sample 3	55	Fri. 19:10
compensatory sample 4	179	Tue. 09:00
compensatory sample 5	84	Wed. 13:05

(b)

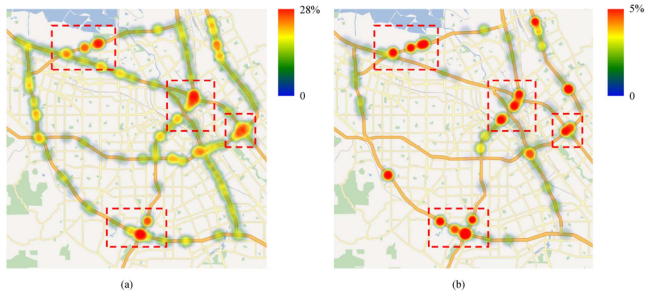
Fig. 5. Case study in compensatory samples. (a) The raw time series of a challenging sample and its compensatory samples. (b) The list of sampled timestamps and nodes.

because the GCN structure excels in capturing correlations in graph structures, thus perceiving the learning difficulty of different samples more accurately. Moreover, for forecasting performance and fairness metrics, FairSTG-GCN(3) slightly outperforms FairSTG-Linear(3), which further delivers that the GCN-based fairness recognizer should be more suitable for fair spatiotemporal forecasting as the inherent regularity of observations are associated with underlying spatial dependencies.

G. Case Study

In this subsection, we investigate how compensatory samples generate and collaboratively enhance the representation of challenging-to-learn samples through a toy case.

The generation of compensatory samples: Given a challenging sample, the compensatory samples are derived by computing the similarity between challenging ones and well-learned ones, where the compensatory set is extracted from different spatiotemporal graphs. We first plot the original time series of a challenging sample and the generated compensatory samples derived by our FairSTG. As shown in Fig. 5(a), all compensatory samples exhibit similar patterns with the challenging one, indicating that our FairSTG can capture effective samples with in a high-quality manner. Besides, we further investigate the sampling timestamps and nodes of these samples. As shown in Fig. 5(b), for compensatory samples 1, 2, and 5, where they come from the same spatiotemporal graph to the challenging sample, while compensatory samples 3 and 4 are sampled at other timestamps. It delivers that our compensatory sample generation can span the search space of advantageous representations from a single ST Graph to the samples within whole batch and significantly extends the scope of advantageous representation transfer. Consequently, the model can delve deeper into exploring and leveraging the cross-step spatiotemporal correlations within observations, alleviating the unfairness across temporal



(a) (b)

Fig. 6. Case study on challenging samples. (a) The forecasting performance (MAPE) of each sensor in PEMS-BAY dataset. (b) The improvement value of each sensor, and only sensors with improvement are plotted.

perspective. Actually, generating more informative representations from compensatory set for challenging samples can compensate for the insufficiency of model's expressive power, and further gain both fairness quality and prediction accuracy within spatiotemporal learning.

Visualization of performance improvement on challenging samples: We visualize the spatial distribution of prediction results based on our FairSTG, to evaluate its capability in improving performance of the challenging subgroup. We first plot the error (here we exploit MAPE) of each sensor at horizon 12 based on MTGNN in PEMS-BAY, where the red boxes highlight regions with greater errors. In Fig. 6(a), it can be observed that sensors located at transportation hubs often suffer more serious prediction errors, possibly due to their complex traffic patterns and higher learning difficulty, inducing susceptibility to unfair treatment. Additionally, we further visualize the improvement of an error metric in Fig. 6(b). To be specific, the improvement of MAPE on ith sensor can be illustrated as $\Delta e_i = e_i - e'_i$, where e_i and e'_i respectively indicate the MAPE of the backbone and our FairSTG. Noting that $\Delta e_i > 0$ means FairSTG outperforms the backbone model in making more accurate predictions. We also provide detailed information about the top five nodes with the largest improvement in (Table VII). As shown, sensors accounting for significant improvement are also concentrated at transportation hubs, consistent with the spatial distribution of sensors with larger errors. If the prediction performance is not fair, e.g., the predicted results are with much higher volumes than real-world ground-truth, the assigned officers on duty will be a waste of human resources. With our fair prediction, such waste will be significantly avoided on these regions and more informed decisions can be made. This case delivers us that our FairSTG can distinguish the challenging samples from the spatial perspective by constraining the samples into node-levels, while simultaneously enhance the performance of the backbone model on challenging subgroup. Therefore, in a real-world urban applications, such as traffic status or road risk prediction systems, with our FairSTG, the prediction system can emphasize more on underrepresented urban regions, which reflects more consistency with facts, and allocate the resources in a positively fairness-aware manner.

TABLE VIII
ABLATION EXPERIMENTS

	MAE	MAE-var	MAPE	MAPE-var
METR-LA				
FairSTG	3.6221	**45.5186**	10.17%	0.1525
FairSTG-w/o-FE	3.6103	45.8562	**10.15%**	**0.1519**
FairSTG-w/o-FO	**3.5379**	48.9587	10.17%	0.1666
PEMS-BAY				
FairSTG	**1.9628**	**15.4070**	**4.58%**	**0.0448**
FairSTG-w/o-FE	1.9925	15.9100	4.64%	0.0457
FairSTG-w/o-FO	2.0100	16.9199	4.70%	0.0476
KnowAir				
FairSTG	16.7561	**395.7106**	**58.98%**	**1.0603**
FairSTG-w/o-FE	16.9433	421.5523	59.94%	1.1146
FairSTG-w/o-FO	**16.7029**	402.2793	60.09%	1.1032
ETT				
FairSTG	**1.7832**	**4.3247**	**12.89%**	**0.0201**
FairSTG-w/o-FE	1.8216	4.5512	13.76%	0.0288
FairSTG-w/o-FO	1.7860	4.3745	13.55%	0.0267

H. Ablation Experiments

We conduct an ablation study to validate the effectiveness of key components that contribute to the improved outcomes of our proposed framework. We name FairSTG without different components as follows. 1) FairSTG-w/o-FE (FairSTG without feature enhancement): We remove our fairness recognizer and collaborative feature enhancement from FairSTG, and only remain fairness constraints in the optimization objective. 2) FairSTG-w/o-FO (FairSTG without fairness objective): We remove the fairness constraints from the optimization objective in our FairSTG and only take the original MAE loss and self-supervised BCE loss as the overall objective.

We choose MTGNN as the backbone and report the accuracy performance and fairness quality in Table VIII. We can observe that, 1) The fairness metrics of FairSTG-w/o-FE and FairSTG-w/o-FO both exhibit a decrease, indicating that collaborative feature enhancement at the representation level and fairness constraints at objective level can work collaboratively to mitigate the unfairness issue in backbone models. 2) For relatively large-scale datasets such as METR-LA and PEMS-BAY, FairSTG-w/o-FO exhibits a more prominent decrease in fairness quality. This suggests that fairness-aware learning objectives play a more crucial role by directly guiding the model's learning process for these large-scale datasets with complex topology and spatiotemporal contexts. 3) In contrast, for relatively small-scale datasets such as KnowAir and ETT, FairSTG-w/o-FE exhibits a noticeable decrease in fairness metrics, which suggests that extracting compensatory representations at the representation level plays a more pivotal role. Such contrastive phenomenon can be attributed to that small-scale datasets are with limited spatiotemporal information where collaborative enhancement becomes more significant, while large-scale datasets are more informative and fairness constraints is more dominated during learning process. Even so, our solution is a general and plug-and-play solution to spatiotemporal learning, where the collaborative enhancement and fairness-aware objective can respectively make sense under different scenarios.

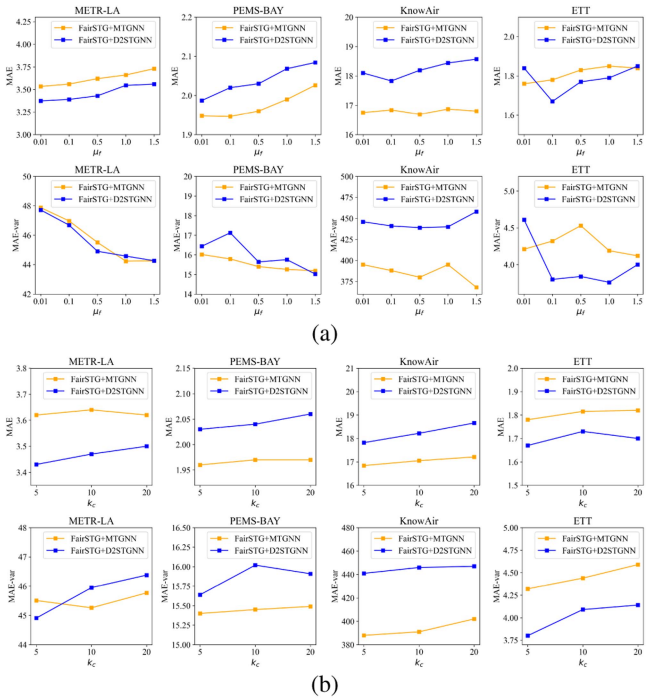


Fig. 7. Hyper-parameter study. (a) Performance variation on different μ_f. (b) Performance variation on different k_c.

I. Hyper-Parameters Analysis

We conduct the parameter study on two core hyper-parameters in our proposed FairSTG, μ_f within a range of $\{0.01, 0.1, 0.5, 1.0, 1.5\}$, which controls the proportion of fairness constraint in the learning objective, and k_c within a range of $\{5, 10, 20\}$ that defines the number of compensatory samples in collaborative representation enhancement.

We adjust the specific parameter and fix the others in each experiment. Fig. 7 reports the results of our hyper-parameter study. As shown in Fig. 7(a), the forecasting accuracy decreases while fairness quality increases with the increases of emphasis of fairness μ_f. This implies that the model tends to generate fair predictions for different samples, but at the cost of decline in overall performance. And it is further verified that there exists an inevitable trade-off between fairness and accuracy in fairness-aware learning. Besides, Fig. 7(b) demonstrates that a small number of compensatory samples will result in better forecasting performance and fairness degree. It is rational because increasing the number of compensatory samples will unavoidably introduce noises to the mix-up representations and leads to a performance degradation. Overall, we are required to adjust the parameters to achieve a tradeoff between the fairness and accuracy.

VII. Conclusion and Discussion

In this work, we uncover a new heterogeneity phenomenon designated prediction unfairness in spatiotemporal forecasting, and attribute such performance heterogeneity to data sufficiency and inherent regularity within observations. We then propose

a model-independent framework named FairSTG to systematically mitigate sample-level unfairness in spatiotemporal graph learning tasks, where the core idea is to conduct a joint optimization from both the representation space and optimization objectives. The substantial experimental results demonstrate that our FairSTG improves the fairness metrics significantly and ensures the comparable forecasting performance simultaneously. We believe our FairSTG, which provides more reliable guidance for decision-making process thus ensures the consistent treatments for different individuals and groups, can be a new paradigm of data-driven sustainable mobile computing.

For future work, we will design the approximate estimation algorithm of sample-pair similarity for a more computationally efficient paradigm and explore adaptive forecasting strategies based on different horizons to better balance performance and fairness. And it would be worthwhile to further explore the root causes of prediction unfairness from both model and data aspects, and develop data-adaptive fairness learning to accommodate different datasets.

REFERENCES

[1] J. Ji et al., "Spatio-temporal self-supervised learning for traffic flow prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 4356–4364.

[2] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17804–17815.

[3] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2017, *arXiv: 1707.01926*.

[4] Z. Zhou, Y. Wang, X. Xie, L. Chen, and C. Zhu, "Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3786–3799, Aug. 2022.

[5] B. An, A. Vahedian, X. Zhou, W. N. Street, and Y. Li, "HintNet: Hierarchical knowledge transfer networks for traffic accident forecasting on heterogeneous spatio-temporal data," in *Proc. 2022 SIAM Int. Conf. Data Mining*, 2022, pp. 334–342.

[6] J. Zheng, J. Wang, Z. Lai, C. Wang, and H. Zhang, "A deep spatiotemporal network for forecasting the risk of traffic accidents in low-risk regions," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 5207–5220, 2023.

[7] C. Shang, J. Chen, and J. Bi, "Discrete graph structure learning for forecasting multiple time series," 2021, *arXiv:2101.06861*.

[8] M. Poposka, Z. Hadzi-Velkov, and T. Shuminoski, "Proportional fairness in wireless powered mobile edge computing networks," in *Proc. IEEE 30th Int. Conf. Syst., Signals Image Process.*, 2023, pp. 1–5.

[9] Y. Liu et al., "A fair task assignment strategy for minimizing cost in mobile crowdsensing," in in *Proc. IEEE 26th Int. Conf. Parallel Distrib. Syst.*, 2020, pp. 44–53.

[10] D. Jalota, K. Solovey, M. Tsao, S. Zoepf, and M. Pavone, "Balancing fairness and efficiency in traffic routing via interpolated traffic assignment," *Auton. Agents Multi-Agent Syst.*, vol. 37, no. 2, 2023, Art. no. 32.

[11] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 25–34, Aug. 2020.

[12] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.

[13] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 753–763.

[14] Z. Shao et al., "Decoupled dynamic spatial-temporal graph neural network for traffic forecasting," 2022, *arXiv:2206.09112*.

[15] X. Li, B. Kao, Z. Ren, and D. Yin, "Spectral clustering in heterogeneous information networks," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 4221–4228.

[16] Z. Zhou, K. Yang, Y. Liang, B. Wang, H. Chen, and Y. Wang, "Predicting collective human mobility via countering spatiotemporal heterogeneity," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 4723–4738, May 2024.

[17] Z. Zhou et al., "Greto: Remedying dynamic graph topology-task discordance via target homophily," in *Proc. 11th Int. Conf. Learn. Representations*, 2022.

[18] H. He, Q. Zhang, S. Wang, K. Yi, Z. Niu, and L. Cao, "Learning informative representation for fairness-aware multivariate time-series forecasting: A group-based perspective," in *Proc. Int. Conf. Learn. Representation*, 2023.

[19] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[20] J. Geweke, "The dynamic factor analysis of economic time series," *Econometrica: J. Econometric Soc.*, vol. 45, no. 1, pp. 97–112, 1977.

[21] G. Jin, Y. Liang, Y. Fang, J. Huang, J. Zhang, and Y. Zheng, "Spatio-temporal graph neural networks for predictive learning in urban computing: A survey," 2023, *arXiv:2303.14483*.

[22] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3634–3640.

[23] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," 2019, *arXiv: 1906.00121*.

[24] M. Morik, A. Singh, J. Hong, and T. Joachims, "Controlling fairness and bias in dynamic learning-to-rank," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 429–438.

[25] Y. Li et al., "S2phere: Semi-supervised pre-training for web search over heterogeneous learning to rank data," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2023, pp. 4437–4448.

[26] Y. Li et al., "COLTR: Semi-supervised learning to rank with co-training and over-parameterization for web search," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12542–12555, Dec. 2023.

[27] M. Xu et al., "Learning to effectively model spatial-temporal heterogeneity for traffic flow forecasting," *World Wide Web*, vol. 26, no. 3, pp. 849–865, 2023.

[28] T. Qi et al., "ProFairRec: Provider fairness-aware news recommendation," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 1164–1173.

[29] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," 2015, *arXiv:1511.00830*.

[30] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3323–3331.

[31] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5684–5693.

[32] W. Liu, J. Guo, N. Sonboli, R. Burke, and S. Zhang, "Personalized fairness-aware re-ranking for microlending," in *Proc. 13th ACM Conf. Recommender Syst.*, 2019, pp. 467–471.

[33] N. Sonboli, F. Eskandanian, R. Burke, W. Liu, and B. Mobasher, "Opportunistic multi-aspect fairness through personalized re-ranking," in *Proc. 28th ACM Conf. User Model., Adapt. Personalization*, 2020, pp. 239–247.

[34] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, "Null it out: Guarding protected attributes by iterative nullspace projection," 2020, *arXiv: 2004.07667*.

[35] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4356–4364.

[36] R. Dipinkrishnan and V. B. Kumaravelu, "Enhancing sum spectral efficiency and fairness in NOMA systems: A comparative study of metaheuristic algorithms for power allocation," *IEEE Access*, vol. 12, pp. 85165–85177, 2024.

[37] M. A. Rahman, S. W. Tan, A. T. Asyhari, I. F. Kurniawan, M. J. Alenazi, and M. Uddin, "IoT-enabled intelligent garbage management system for smart city: A fairness perspective," *IEEE Access*, vol. 12, pp. 82693–82705, 2024.

[38] A. Rashid, L. U. Khan, N. Khan, H. Min, A. Ahmad, and S. Ahmad, "Resource-optimized vehicular edge networks with fairness constraints," *IEEE Access*, vol. 12, pp. 67924–67934, 2024.

[39] G. Wang, S. Pan, and S. Xu, "Decoupling the unfairness propagation chain in crowd sensing and learning systems for spatio-temporal urban monitoring," in *Proc. 8th ACM Int. Conf. Syst. Energy-Efficient Buildings, Cities, Transp.*, 2021, pp. 200–203.

[40] A. Yan and B. Howe, "Fairness-aware demand prediction for new mobility," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1079–1087.

[41] Y. Wang, Y. Zhao, Y. Dong, H. Chen, J. Li, and T. Derr, "Improving fairness in graph neural networks via mitigating sensitive attribute leakage," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 1938–1948.

[42] C. Wu, F. Wu, X. Wang, Y. Huang, and X. Xie, "Fairness-aware news recommendation with decomposed adversarial learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 4462–4469.
[43] M. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Phys.*, vol. 46, pp. 323–351, 2004.
[44] Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. Rothblum, "Multicalibration: Calibration for the (computationally-identifiable) masses," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1944–1953.
[45] P. Lahoti et al., "Fairness without demographics through adversarially reweighted learning," 2020, *arXiv: 2006.13114*.
[46] Y. Sang, Y. Xu, B. Ning, and Z. Xie, "GCMT: A graph-contextualized multitask spatio-temporal joint prediction model for cellular trajectories," *World Wide Web*, vol. 26, pp. 1649–1665, 2022.
[47] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious classification: A data scientist's guide to discrimination-aware classification," *Big Data*, vol. 5, no. 2, pp. 120–134, 2017.
[48] S. Wang, Y. Li, J. Zhang, Q. Meng, L. Meng, and F. Gao, "PM2.5-GNN: A domain knowledge enhanced graph neural network for PM2. 5 forecasting," in *Proc. 28th Int. Conf. Adv. Geographic Inf. Syst.*, 2020, pp. 163–166.
[49] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11106–11115.
[50] A. Bose and W. Hamilton, "Compositional fairness constraints for graph embeddings," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 715–724.
[51] E. Dai and S. Wang, "Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, 2021, pp. 680–688.
[52] Z. Liu, T. K. Nguyen, and Y. Fang, "On generalized degree fairness in graph neural networks," in *Proc. 37th AAAI Conf. Artif. Intell.*, Washington, USA, 2023, pp. 7–14.

Gengyu Lin received the BS degree from Beijing Normal University, in 2022. She is currently working toward the MS degree with University of Science and Technology of China. Her mainly research interests include spatiotemporal data mining and fairness algorithms.

Zhengyang Zhou (Member, IEEE) received the PhD degree from the University of Science and Technology of China, in 2023. He is now an associate researcher with the Suzhou Institute for Advanced Research, University of Science and Technology of China (USTC). He has published over 20 papers on top conferences and journals, such as *Data Mining and Knowledge Discovery*, ICLR, *IEEE Transactions on Knowledge and Data Engineering*, WWW, AAAI, *SDMIMD Journal of Management*, and ICDE. His mainly research interests include human-centered urban computing, and mobile data mining.

Qihe Huang received the BS degree from the Nanjing University of Information Science and Technology, in 2022. He is currently working toward the MS degree with the University of Science and Technology of China. His research centers on spatiotemporal data mining and mobile edge computing.

Kuo Yang received the BE degree from Northeastern University at Qinhuangdao, China, in 2021. He is currently working toward the PhD degree with the School of Data Science, USTC. His mainly research interests are data-driven urban management, spatiotemporal data mining, and mobile computing.

Shifen Cheng received the PhD degree from the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. He is an associate professor of State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. His research interests include spatiotemporal data mining, urban computing, and intelligent transportation.

Yang Wang (Senior Member, IEEE) received the PhD degree from the University of Science and Technology of China, in 2007. He is now an associate professor with the School of Computer Science and Technology, School of Software Engineering, and School of Data Science in USTC. Since then, he keeps working at USTC till now as a postdoc and an associate professor successively. Meanwhile, he also serves as the vice dean of school of software engineering of USTC. His research interest mainly includes wireless (sensor) networks, distribute systems, data mining, and machine learning, and he is also interested in all kinds of applications of AI and data mining technologies especially in urban computing and AI4Science.