Balancing Imbalance: Data-Scarce Urban Flow Prediction via Spatio-Temporal Balanced Transfer Learning

Xinyan Hao[1], **Huaiyu Wan**[1,2], **Shengnan Guo**[1,2] and **Youfang Lin**[1,2*]

[1]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing, China
{xinyanhao, hywan, guoshn, yflin}@bjtu.edu.cn

Abstract

Advanced deep spatio-temporal networks have become the mainstream for traffic prediction, but the widespread adoption of these models is impeded by the prevalent scarcity of available data. Despite cross-city transfer learning emerging as a common strategy to address this issue, it overlooks the inherent distribution imbalances within each city, which could potentially hinder the generalization capabilities of pre-trained models. To overcome this limitation, we propose a Spatio-Temporal Balanced Transfer Learning (STBaT) framework to enhance existing spatio-temporal prediction networks, ensuring both universality and precision in predictions for new urban environments. A Regional Imbalance Acquisition Module is designed to model the regional imbalances of source cities. Besides, to promote generalizable knowledge acquisition, a Spatio-Temporal Balanced Learning Module is devised to balance the predictive learning process. Extensive experiments on real-world datasets validate the efficacy of our proposed approach compared with several state-of-the-art methods.

1 Introduction

Accurate prediction of urban spatio-temporal data is crucial for smart city development, serving as an important topic in the field of urban computing that supports various tasks [Lv *et al.*, 2014]. Although deep learning has significantly advanced spatio-temporal prediction tasks like predicting traffic flow [Qu *et al.*, 2023; Wang *et al.*, 2024a] and taxi demand [Geng *et al.*, 2019], it faces challenges in scenarios where training data is scarce due to privacy concerns and varying urban developments [Wang *et al.*, 2018]. These challenges exist in many cities such as Hong Kong [HKGov, 2025] and Liverpool [Kono, 2022], and they could potentially hindering the developments of smart city.

Inspired by transfer learning, researchers have proposed *spatio-temporal transfer learning* that aims to extract transferable knowledge from data-rich (source) domains to address urban computing tasks in data-scarce (target) domains,

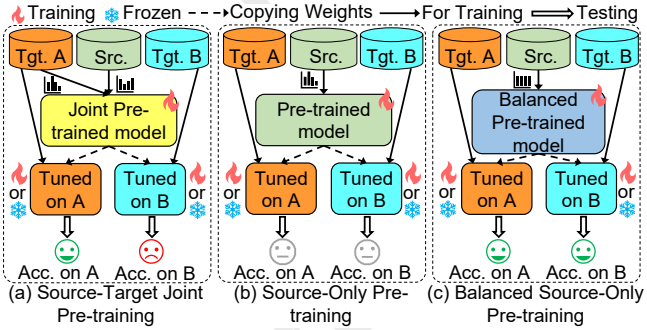


Figure 1: Comparison of different cross-city prediction paradigms in transferable knowledge learning

such as traffic prediction, air quality prediction [Wei *et al.*, 2016], point of interest (POI) recommendation [Ding *et al.*, 2020], and trajectory simulation [Wang *et al.*, 2024b]. The strategies of existing work on learning transferable knowledge can be divided into two categories. (1) Source-target joint pre-training: Utilizing data from both source and target cities in pre-training phase, managing to learn knowledge beneficial to specific target by using some matching strategies [Wang *et al.*, 2021; Tang *et al.*, 2022; Jin *et al.*, 2022]. (2) Source-only pre-training: Employing large-scale source data to train a prediction model by decoupling target from pre-training procedure, aiming to be used to any other city [Wang *et al.*, 2018; Yao *et al.*, 2019; Lu *et al.*, 2022; Liu *et al.*, 2023; Yuan *et al.*, 2024].

An ideal cross-city prediction model should be widely applicable to numerous unseen cities (i.e., *universality*) while also demonstrating high prediction accuracy in the specific cities where it is applied (i.e., *precision*). Existing research in this field often prioritizes either universality or precision in model design, posing challenges when attempting to achieve both simultaneously. Source-target joint pre-training methods tend to predict precisely on the single selected city due to the introduction of *specific distribution biases from that target* during pre-training, but they struggle to generalize extensively to other unselected cities and remain narrowly adaptability, as shown in Figure 1(a). Source-only pre-training methods often require multiple datasets and higher training costs to exhibit universality of unseen domains, yet they absorb *distribution biases of the source cities* inevitably, which

*Corresponding author.

may lead to sub-optimal target performances even with fine-tuning [Jin *et al.*, 2022], as shown in Figure 1(b). **Consequently, how to reduce the impact of harmful distribution biases from specific domains during pre-training, thereby achieving both universality and precision in generalizing to unseen cities, remains a challenging problem.**

Harmful distribution biases in specific cities arise from inherent disparities between them and unseen cities. The disparities between cities are rooted in the unique characteristics of each city [Yuan *et al.*, 2024]. Accordingly, the adverse effects of distribution biases stem from, and can be alleviated by removing this uniqueness of the data source cities. Despite the intricate factors contributing to city uniqueness, the uneven distribution of regional traffic patterns, i.e., *regional imbalance*, serves as its fundamental cause. For example, Washington has more cultural centers than transportation hubs, whereas Chicago is the opposite; New York exhibits a higher proportion of areas with significantly high or low traffic flow, in contrast to Chicago. **Therefore, the key to building a predictive model that encapsulates universality and precision lies in identifying and calibrating regional imbalances for balanced source-only pre-training**, capable of achieving the effect of "pre-train once, adapt everywhere", as illustrated in Figure 1(c).

To this end, this paper introduces ***STBaT***, **s**patio-**t**emporal **ba**lanced **t**ransfer learning, for data-scarce urban flow prediction. To quantify the regional imbalance of different cities, we propose a novel Regional Imbalance Acquisition Module (RIAM). It effectively obtains regional pattern density and inter-regional pattern similarity, serving as indicative information for balanced pre-training. After that, we devise a Spatio-Temporal Balanced Learning Module (STBLM) to train the prediction network. It facilitates knowledge transfer between similar regions and re-weights the intake of predictive supervision information to enhance the model's ability of generalization. To achieve information flow between RIAM and STBLM, we design a bidirectional complementary iterative learning algorithm to pre-train our spatio-temporal prediction module, promoting the reliable regional imbalance awareness and balanced predictive learning. In summary, we make the following contributions in this paper:

- *Novel Study Perspective*: To our knowledge, we are the first to study balanced cross-city transfer learning, offering a new perspective to achieve universality and precision in data-scarce urban flow prediction.
- *Advanced Methodology*: We propose STBaT, a novel spatial-temporal balanced transfer learning framework for urban flow prediction. By linking a RIAM and a STBLM through a bidirectional complementary iterative learning algorithm, it identifies and corrects regional imbalances, enabling the model's generalization capabilities to unseen cities.
- *Strong empirical evidence*: We conduct cross-city experiments across a multitude of tasks on three real world datasets. The results show the superiority of STBaT in addressing the data scarcity of urban flow prediction. The code is made publicly available at https://github.com/ShaunHao/stbat.

2 Related Work

Deep learning-based traffic prediction techniques can extract complex spatio-temporal relationships and accurately predict various traffic data, such as urban flow [Zhang *et al.*, 2017; Gong *et al.*, 2022; Chen *et al.*, 2022; Liang *et al.*, 2021; Ji *et al.*, 2023; Xia *et al.*, 2023; Wen *et al.*, 2023] and traffic speed [Yu *et al.*, 2017; Li *et al.*, 2018; Guo *et al.*, 2019; Zheng *et al.*, 2020; Song *et al.*, 2020; Cirstea *et al.*, 2022; Jia *et al.*, 2023], playing a crucial role in intelligent transportation systems. However, a common issue is data scarcity, leading to decreased accuracy in deep models. To address this, cross-city transfer learning methods have been proposed, encompassing two main categories: source-target joint pre-training method and source-only pre-training method. The former leverages sufficient source data along with a small amount of target data for pre-training, using techniques such as domain adaptation [Wang *et al.*, 2021; Fang *et al.*, 2022], domain adversarial networks [Tang *et al.*, 2022; Ouyang *et al.*, 2024], selective meta-learning [Jin *et al.*, 2022] and knowledge distillation [Jin *et al.*, 2023] to alleviate data scarcity of specific target city. In contrast, the latter pre-trains models on rich data from source cities, employing strategies such as region matching [Wang *et al.*, 2018], meta-learning [Yao *et al.*, 2018; Lu *et al.*, 2022], pattern bank construction [Liu *et al.*, 2023], parameter generation [Yuan *et al.*, 2023] and prompt learning [Yuan *et al.*, 2024] to acquire knowledge that enables the model to adapt to more cities.

However, these two types of work focus solely on performance in specific city or generalization across a wide range of cities, lacking a combination of both. This paper differs from existing works by aiming to learn transferable knowledge that is independent of the target cities while removing distribution biases from the source cities, thereby achieving universality and precision in predictions for unseen cities.

3 Preliminaries

Definition 1 (Region)**.** *We divide city c into a grid map of size $W_c \times H_c$ based on latitude and longitude, which contains W_c rows and H_c columns. Each grid is defined as a cell region r_c, and all the grids form a set of cell regions R_c.*

Definition 2 (Spatio-temporal Urban Flow Series)**.** *In city c, we represent the time range as a set $T_c = \{t_c - |T_c| + 1, \ldots, t_c\}$, consisting of $|T_c|$ evenly split non-overlapping time intervals, where t_c represents the latest timestamp of city c. Then, the spatio-temporal urban flow series in city c is represented as $X_c = \left\{x_{r_c}^{(t)} | r_c \in R_c, t \in T_c\right\}$, where $x_{r_c}^{(t)}$ denotes the flow data of region r_c at time t, e.g., the number of taxi pickups or drop-offs.*

Definition 3 (Few-Shot Urban Flow Prediction)**.** *Given the urban flow series data $X_{S_1}, X_{S_2}, \ldots, X_{S_N}$ from source cities $S_1, S_2, \ldots, S_N$ and X_T from a data-scarce target city T, our goal is to train a prediction model, such that it can minimize the prediction error on the test data X'_T of T.*

Definition 4 (Zero-Shot Urban Flow Prediction)**.** *Given $X_{S_1}, X_{S_2}, \ldots, X_{S_N}$ from $S_1, S_2, \ldots, S_N$, our goal is to train a prediction model that minimizes the prediction error on the unseen target city test data, even with no training data of it.*

4 Methodology

In this section, we provide a detailed explanation of our proposed method, STBaT, and how it addresses the problem of spatio-temporal balanced transfer learning. The overall method, as shown in Figure 2, consists of two major components: the Regional Imbalance Acquisition Module (RIAM), and the Spatio-Temporal Balanced Learning Module (STBLM). First, we introduce our novel RIAM and explain how it comprehensively captures the regional imbalance characteristics of source cities from both functionality and flow perspectives. Next, we present our newly designed STBLM and describe how it assists the learning process of the spatio-temporal prediction network to acquire balanced and generalizable prediction knowledge. Finally, we outline the training algorithm for the above two components, which ensures the reliability of the acquired regional imbalance and the generalization of the learned balanced knowledge through a bidirectional complementary iterative learning algorithm.

4.1 Regional Imbalance Acquisition Module

Capturing the imbalance inherent in regional traffic patterns is a prerequisite for eliminating distribution biases in cities. In diverse urban environments, traffic patterns are influenced by regional functionality and traffic flow factors, thus displaying distinctive imbalances. To concretize these imbalances, an intuitive approach involves modeling and accessing the distribution information of the regional traffic patterns. To implement the above intuitions, a Regional Imbalance Acquisition Module (RIAM) is proposed. RIAM utilizes a regional pattern autoencoder to extract region embeddings of source cities, representing their regional traffic patterns. A density estimator and a similarity measure are used to obtain regional pattern density and inter-regional pattern similarity, which then serve as imbalance indicators.

Regional Pattern Autoencoder

As aforementioned, the traffic pattern is influenced by both functional and flow-related factors. To simultaneously capture these two factors, a regional pattern autoencoder consisting of a regional pattern encoder and a decoder has been developed. Given normalized POI distributional tensor $P_{poi} = \{p_{r_s} \mid r_s \in R_s\} \in \mathbb{R}^{|R_s| \times K}$ with K categories of a city s, the encoder takes it into a feed-forward network to obtain the region embedding tensor $\Phi_s = \{\phi_{r_s} \mid r_s \in R_s\} \in \mathbb{R}^{|R_s| \times D}$ that reflect traffic patterns, in which embedding of each region as

$$\phi_{r_s} = f_{\text{enc}}(p_{r_s}; \theta_{\text{enc}}) = FC(ReLU(FC(p_{r_s}))), \quad (1)$$

where θ_{enc} denotes the learnable parameters of encoder, $FC(\cdot)$ denotes a fully connected layer, and $ReLU(\cdot)$ denotes a ReLU activation layer. Then, ϕ_{r_s} is parallelly fed into a POI classifier and a spatio-temporal feature distribution predictor, which jointly form the decoder. The POI classifier outputs the prediction of p_{r_s} as $\hat{p}_{r_s} \in \mathbb{R}^K$, which is optimized through cross-entropy loss, aiming at encoding regional functionality into ϕ_{r_s}. The spatio-temporal feature distribution predictor predicts the mean $\hat{\mu}_{r_s} \in \mathbb{R}^J$ and standard deviation $\hat{\sigma}_{r_s} \in \mathbb{R}^J$ of the J-dimensional regional flow features, which are refined through Kullback-Leibler Divergence loss $\mathcal{L}_{\text{KLD}}$ to align with the actual feature statistics μ_{r_s} and σ_{r_s}, thereby integrating flow pattern into ϕ_{r_s}. The details of μ_{r_s} and σ_{r_s} are introduced in Section 4.2. Formally, the decoder can be defined as

$$\hat{p}_{r_s}, \hat{\mu}_{r_s}, \hat{\sigma}_{r_s} = f_{\text{dec}}(\phi_{r_s}; \theta_{\text{dec}}), \quad (2)$$

$$\hat{p}_{r_s} = softmax(FC(ReLU(FC(\phi_{r_s})))), \quad (3)$$

$$\hat{\mu}_{r_s} = FC(ReLU(FC(\phi_{r_s}))), \hat{\sigma}_{r_s} = FC(ReLU(FC(\phi_{r_s}))), \quad (4)$$

where θ_{dec} denotes the learnable parameters. The total loss for the regional pattern autoencoder can be expressed as

$$\mathcal{L}_{\text{emb}} = -\sum_{k}^{K} \left(p_{r_s}^{(k)} \log \hat{p}_{r_s}^{(k)} \right) + \beta \mathcal{L}_{\text{KLD}}, \quad (5)$$

where $\mathcal{L}_{\text{KLD}} = \sum_j^J \left(\log \frac{\sigma_{r_s}^{(j)}}{\hat{\sigma}_{r_s}^{(j)}} + \frac{(\hat{\sigma}_{r_s}^{(j)})^2 + (\hat{\mu}_{r_s}^{(j)} - \mu_{r_s}^{(j)})^2}{2(\sigma_{r_s}^{(j)})^2} - \frac{1}{2} \right)$, and β is a balancing hyper-parameter.

Density Estimator

As a fundamental imbalance indicator, it is essential to obtain the density of regional traffic patterns of the source cities. The density estimator derives the density of the traffic pattern distribution of a city s based on its region embeddings ϕ_{r_s} extracted from the regional pattern autoencoder.

Since the region embedding space is a continuous space, each region's pattern is not independent of others. The more neighboring embeddings a region has in the embedding space, the higher the density of regional traffic patterns corresponding to it. Based on this intuition, the density estimator employs kernel density estimation [Rosenblatt, 1956; Parzen, 1962] to estimate the regional pattern density smoothly. It takes into account each ϕ_{r_s} and uses a specific kernel as a weight to infer the probability density function of the traffic pattern. For any given region $r_s \in R_s$, its regional pattern density can be inferred as

$$\rho_{r_s} = \frac{1}{h|R_s|} \sum_{i \in R_s} K\left(\frac{\phi_{r_s} - \phi_i}{h} \right), \quad (6)$$

where $K(\cdot)$ is the normal kernel function, and h is the bandwidth parameter. The ρ_{r_s} will further support balanced learning of the STBLM. For specific implementation details, please refer to Section 4.2.

Similarity Measure

Considering the continuous nature of traffic patterns ϕ_{r_s}, proximity and distance concepts are essential characteristics. Explicitly measuring inter-regional pattern similarity enhances the understanding of shared knowledge among regions, empowering balanced learning within the subsequent STBLM. The similarity measure calculates how close any two region embeddings are to obtain the inter-regional pattern similarity. Given any two regions $i, j \in R_s$, the similarity measure obtains the similarity of their traffic patterns as

$$\delta(\phi_i, \phi_j) = sim(\phi_i, \phi_j), \quad (7)$$

where $sim(\cdot)$ denotes the cosine similarity metric. For further details on how $\delta(\phi_i, \phi_j)$ is used in STBLM, please refer to Section 4.2.

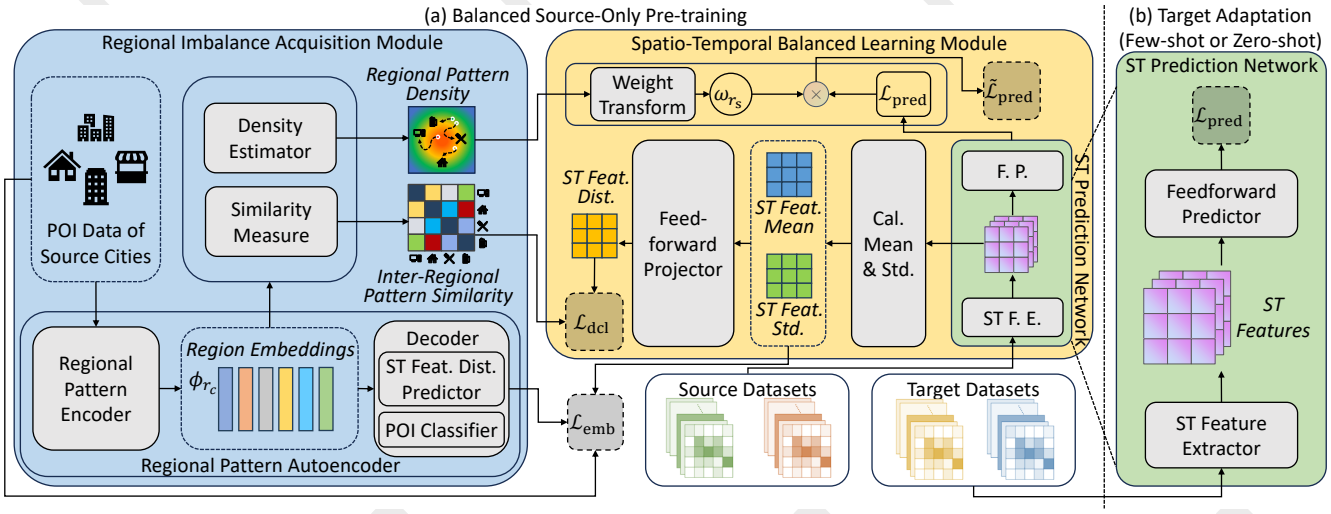


Figure 2: The overview of our Spatio-Temporal Balanced Transfer Learning framework

4.2 Spatio-Temporal Balanced Learning Module

Under the influence of distribution biases, training on unbalanced source data absorbs detrimental knowledge for generalizing to unseen cities. For instance, imbalanced learning may lead to a skewed distribution of extracted spatio-temporal features in regions with lower traffic pattern density, impacting the efficacy of predictive knowledge acquisition in these areas. To address these challenges, we introduce a Spatio-Temporal Balanced Learning Module (STBLM) with two functions: facilitating knowledge transfer among similar regions to rectify imbalances in feature extraction and re-weighting label information to promote balanced predictive learning.

Spatio-Temporal Prediction Network

The spatio-temporal prediction network, comprising a spatio-temporal feature extractor and a feedforward predictor, can be formally expressed as

$$z_{r_s} = f_{\text{feat}}([x_{r_s}^{(t-\tau)}, \ldots, x_{r_s}^{(t-1)}]; \theta_{\text{feat}}), \tag{8}$$

$$\hat{x}_{r_s}^{(t)} = f_{\text{pred}}(z_{r_s}; \theta_{\text{pred}}) = FC(ReLU(FC(z_{r_s}))), \tag{9}$$

where z_{r_s} is the extracted spatio-temporal feature, $\hat{x}_{r_s}^{(t)}$ is the future traffic prediction, θ_{feat}, θ_{pred} denotes the learnable parameters of spatio-temporal feature extractor and feedforward predictor. Notice that STBaT has no specific structural assumptions on the spatio-temporal feature extractor, as long as it does not alter the spatial dimensions of the input data. Thus, various advanced spatio-temporal prediction models can be compatible with STBaT.

Balanced Learning for Spatio-Temporal Feature Extraction

Regions with similar traffic patterns share common knowledge, which should be learned by the spatio-temporal feature extractor. By adjusting regional feature distributions based on inter-regional pattern similarity, knowledge can be transferred between similar regions, thus calibrating feature learning in low-density regions.

First, STBLM computes the mean μ_{r_s} and standard deviation σ_{r_s} of the spatio-temporal features z_{r_s} of region r_s, serving as regional spatio-temporal feature statistics

$$\mu_{r_s} = \frac{1}{B}\sum_{i=1}^{B} z_{r_s}, \sigma_{r_s} = \sqrt{\frac{\sum_{i=1}^{B}(z_{r_s} - \mu_{r_s})^2}{B-1}} \tag{10}$$

where B is number of samples in a training batch.

Then, a feedforward projector is utilized to fuse the μ_{r_s} and σ_{r_s} and extract the representation of the regional feature distribution d_{r_s}, formalized as

$$d_{r_s} = f_{\text{proj}}(\mu_{r_s}||\sigma_{r_s}; \theta_{\text{proj}}) \tag{11}$$

$$= FC(ReLU(FC(\mu_{r_s}||\sigma_{r_s}))), \tag{12}$$

where θ_{proj} denotes the learnable parameters of feedforward predictor, and $||$ is the concatenation operation.

Next, a newly developed *Similarity-aware Distribution Contrastive Learning* (SimDCL) is employed to adjust regional feature distributions based on inter-regional pattern similarity, thereby optimizing the spatio-temporal feature extractor. The core idea is, the feature distributions d_i and d_j of two regions i and j should be closer when there is a higher similarity $\delta(i, j)$ between them. For a city s with region set R_s, the SimDCL loss can be defined as

$$\mathcal{L}_{\text{dcl}} = \frac{1}{|R_s|(|R_s|-1)} \sum_{i}^{|R_s|}\sum_{k}^{|R_s|} \mathbb{1}_{k\neq i} \mathcal{L}_{\text{dcl}}^{(i,k)}, \tag{13}$$

where $\mathcal{L}_{\text{dcl}}^{(i,k)} = -\log \frac{\exp(sim(d_i, d_k)/\tau_{\text{dcl}})}{\sum_{j}^{|R_s|} \mathbb{1}_{j\neq i, \delta(\phi_i,\phi_k)\geq\delta(\phi_i,\phi_j)} \exp(sim(d_i, d_j)/\tau_{\text{dcl}})}$, τ_{dcl} is the temperature hyper-parameter, and $\mathbb{1}_{\cdot}$ is the indicator function.

Recalling from Section 4.1, the feature statistics are also utilized to optimize the decoder, in order to encode the regional flow patterns into ϕ_{r_s}. Thus, we feed the μ_{r_s} and σ_{r_s} obtained from Eq.10 into the RIAM to optimize Eq.5, promoting RIAM's comprehensive modeling of regional imbalances.

Balanced Learning for Spatio-Temporal Traffic Prediction

In existing work, the equal weighting of prediction losses across regions results in an imbalance in the model's attention to various traffic patterns, leading to a greater focus on accurately predicting high-density traffic patterns rather than low-density ones. By assigning equal weights to each traffic pattern within the loss function, rather than to each region, a more balanced focus can be achieved across diverse traffic patterns.

To balance the influence of regions with varying densities on the overall prediction loss while ensuring numerical stability, STBLM calculate the multiplicative inverse of the regional pattern density ρ_{r_s} and scale it to transform it into the regional weight

$$\omega_{r_s} = \frac{|R_s|}{\sum_{r_s}^{|R_s|} \frac{1}{\rho_{r_s}}} \frac{1}{\rho_{r_s}}. \tag{14}$$

Next, the ω_{r_s} is utilized to reweight the prediction loss, resulting in a density-aware weighted prediction loss, defined as

$$\widetilde{\mathcal{L}}_{\text{pred}} = \sum_{r_s \in R_s} \sum_{t \in T_s} \omega_{r_s} \mathcal{L}_{\text{pred}}(\hat{x}_{r_s}^{(t)}, x_{r_s}^{(t)}), \tag{15}$$

where $\hat{x}_{r_s}^{(t)}$ is the prediction for the future traffic state, $\mathcal{L}_{\text{pred}}(\cdot)$ is the typical prediction loss, e.g., mean squared error loss, etc.

4.3 Training Process

To streamline training for RIAM and STBLM without additional pre-training, synchronizing imbalance information acquisition with balanced prediction knowledge learning, we devised a bidirectional complementary iterative learning algorithm with the following steps:

1. **Optimize STBLM and Obtain the Regional Spatio-Temporal Feature Statistics**. Using the current STBLM with $\theta_{\text{STBLM}} = \{\theta_{\text{feat}}, \theta_{\text{pred}}, \theta_{\text{proj}}\}$ and $\varrho = \{\rho_{r_s} \mid r_s \in R_s\}, \Delta = \{\delta(\phi_i, \phi_j) \mid i, j \in R_s\}$ derived from the current RIAM, the losses $\mathcal{L}_{\text{dcl}}$ and $\widetilde{\mathcal{L}}_{\text{pred}}$ are computed to optimize θ_{STBLM}. Then, the $\mathcal{N} = \{\mu_{r_s} \mid r_s \in R_s\}, \Sigma = \{\sigma_{r_s} \mid r_s \in R_s\}$ are obtained to provide guidance for the subsequent optimization of RIAM.
2. **Optimize RIAM and Derive the Regional Imbalance Information**. Using the current RIAM with $\theta_{\text{RIAM}} = \{\theta_{\text{enc}}, \theta_{\text{dec}}\}$ and the obtained $\mathcal{N}, \Sigma$ from the current STBLM, the loss $\mathcal{L}_{\text{emb}}$ is calculated to optimize θ_{RIAM}. After this, ϱ and Δ are derived from the current RIAM to aid in the next optimization of STBLM.

By iteratively executing the above two steps, a bidirectional connection is established between RIAM and STBLM, contributing to a balanced pre-training process of spatio-temporal prediction network, i.e., θ_{feat} and θ_{pred}. As a result, the pre-trained θ_{feat} and θ_{pred} can be adapted to urban flow prediction task in new cities with limited or no training data.

City	# Regions	Time span (m/d/y)	# Taxis	# Bikes
New York (NY, N)	460	1/31/2016-12/31/2016	133M	13.8M
Chicago (CHI, C)	476	1/31/2016-12/31/2016	24.5M	3.5M
Washington (DC, D)	420	1/31/2016-12/31/2016	10M	2.7M

Table 1: Detailed statistics of the datasets

5 Experiments

5.1 Experiment Settings

Datasets

Following previous work, we take urban flow prediction as an example task, and evaluate our proposed framework on real-world public datasets of three cities: New York (NY), Chicago (CHI), and Washington (DC), which contain vehicle pickup and drop-off records of bike and taxi. Each dataset covers a time range of one year with time intervals of one hour. The detailed statistics of datasets are shown in Table 1. Additionally, we also use public POI data of each city. All data are collected and opened by [Jin *et al.*, 2022].

Task Settings

We use a similar few-shot traffic prediction setting to [Wang *et al.*, 2021; Jin *et al.*, 2022]. We choose each of three cities as the target city, while using one of (New York or Chicago) or all of the remaining cities as the source cities. This creates a total of eight tasks of cross-city transfer. For each source city, we divided the training and validation sets in a 2:1 ratio. For each target city, we allocate the last 2 months for testing, the preceding 2 months for validation, and the last 3 days or 0 days before the validation data for training, forming two data-scarce scenarios: few-shot and zero-shot. We first pre-train our framework on source training data, then fine-tune it on target training data. After that, we evaluate it on target test data. Min-max normalization is applied for data preprocessing.

Baselines

We compare the performance of STBaT and a number of baselines for urban flow prediction. Based on whether requires source data, the baselines can be classified into *non-transfer* baselines and *transfer* baselines.

- *Non-transfer Methods*: The non-transfer baselines consist of three methods, including ARIMA [Box and Jenkins, 1968], a statistical method, and two deep learning methods, i.e., ST-net [Yao *et al.*, 2019] and PDFormer [Jiang *et al.*, 2023].
- *Transfer Methods*: For *transfer* baselines, we select two types of learning-based approaches: four source-only pre-training methods, including vanilla finetuning (Fine-tuned), MAML [Finn *et al.*, 2017] and MetaST [Yao *et al.*, 2019]; and two source-target joint pre-training methods, consisting of ST-DAAN [Wang *et al.*, 2021] and CrossTReS [Jin *et al.*, 2022].

For ARIMA models, we employ 6 autoregressive steps, 1 moving average step, and 1 integration step. Following previous work, we use the official code and hyper-parameters reported by the original papers of PDFormer, MAML, MetaST, ST-DAAN, RegionTrans and CrossTReS.

Task	Baselines	Few-Shot RMSE				Few-Shot MAE				Zero-Shot RMSE				Zero-Shot MAE			
Bike	Target	DC		CHI	NY	DC		CHI	NY	DC		CHI	NY	DC		CHI	NY
	ARIMA	3.384		2.551	9.900	1.481		0.968	5.167	-		-	-	-		-	-
	ST-net	2.945		2.187	10.126	1.394		0.832	5.186	-		-	-	-		-	-
	PDFormer	2.919		2.183	9.774	1.385		0.819	5.097	-		-	-	-		-	-
	Source	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI
	Finetuned	2.584	2.530	2.153	9.109	1.304	1.291	0.852	4.810	3.454	3.605	2.495	15.968	1.270	1.764	0.917	6.813
	MAML	2.860	2.697	2.175	8.981	1.352	1.408	1.065	4.960	3.596	3.645	2.470	16.349	1.282	1.659	1.222	6.954
	MetaST	2.565	2.462	1.990	8.984	1.225	1.258	0.830	4.641	3.429	3.593	2.433	15.672	1.264	1.304	0.782	6.837
	RegionTrans	2.557	2.496	2.031	8.846	1.119	1.273	0.826	4.704	3.447	3.604	2.489	15.880	1.272	1.773	0.913	6.815
	ST-DAAN	2.538	2.421	1.938	8.810	1.106	1.177	0.784	4.513	-	-	-	-	-	-	-	-
	CrossTReS	2.482	2.405	1.895	8.779	1.058	1.070	0.731	4.445	-	-	-	-	-	-	-	-
	STBaT (ours)	**2.358**	**2.363**	**1.857**	**8.570**	**0.952**	**0.992**	**0.682**	**4.154**	**3.343**	**3.521**	**2.306**	**15.182**	**1.145**	**1.205**	**0.710**	**6.741**
	Improvements	+5.0%	+1.8%	+2.0%	+2.4%	+10.0%	+7.3%	+6.7%	+6.6%	+2.5%	+2.0%	+5.2%	+3.1%	+9.4%	+7.6%	+9.2%	+1.1%
Taxi	Target	DC		CHI	NY	DC		CHI	NY	DC		CHI	NY	DC		CHI	NY
	ARIMA	5.194		8.435	23.307	1.839		2.610	7.127	-		-	-	-		-	-
	ST-net	5.412		9.922	23.988	1.851		3.349	7.923	-		-	-	-		-	-
	PDFormer	5.271		9.483	23.390	1.846		2.966	7.302	-		-	-	-		-	-
	Source	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI	NY	CHI
	Finetuned	5.015	4.890	7.871	21.809	1.794	1.675	2.642	6.836	7.558	6.049	17.553	37.513	2.177	1.901	4.769	11.212
	MAML	5.333	4.867	7.913	21.531	1.762	1.710	2.636	7.001	7.580	6.108	17.283	39.194	2.212	1.906	4.743	11.839
	MetaST	4.767	4.863	7.875	21.272	1.780	1.814	2.609	6.532	7.525	6.018	17.023	37.277	2.197	1.885	4.645	11.211
	RegionTrans	4.810	4.678	7.744	21.367	1.775	1.691	2.732	6.741	7.550	6.039	17.536	37.507	2.181	1.901	4.764	11.203
	ST-DAAN	4.675	4.694	7.687	21.209	1.636	1.683	2.583	6.469	-	-	-	-	-	-	-	-
	CrossTReS	4.598	4.585	7.673	20.899	1.599	1.561	2.421	6.285	-	-	-	-	-	-	-	-
	STBaT (ours)	**4.391**	**4.415**	**7.625**	**20.545**	**1.423**	**1.413**	**2.196**	**5.558**	**7.374**	**5.912**	**16.030**	**36.560**	**2.107**	**1.780**	**4.229**	**10.715**
	Improvements	+4.5%	+3.7%	+6.3%	+1.7%	+11.0%	+9.5%	+9.3%	+11.6%	+2.0%	+1.8%	+5.8%	+1.9%	+3.4%	+5.6%	+9.0%	+4.4%

Table 2: Performance of all methods on cross-city setting, with the best results highlighted in bold and the suboptimal results underlined

Implementation Details

We implement STBaT using PyTorch. For the dimensions of FCs in the regional pattern autoencoder, we set them to (16, 32) for the encoder, (32, 14) for the POI classifier and (32, 256) for the spatio-temporal feature distribution predictor. We choose the ST-net model as the spatio-temporal feature extractor, stacking three residual blocks with 64 channels and a single-layer LSTM with 128 hidden units. For the STBLM, the dimensions of the FC layers of the feedforward predictor are set to 256 and 1, and the dimensions of the feedforward projector are set to 256 and 256. To capture temporal dependencies, we set the horizon τ to 6, meaning that the input data consisted of observations from the previous 6 intervals, to predict the urban flow in next time step. We train the model until the validation error does not decrease for 5 consecutive epochs on the source data, and then select the model with the lowest source validation error as the initialization for target adaptation. We evaluate the target performance using root mean squared error (RMSE) and mean absolute error (MAE). The mean error for both the pickup and drop-off predictions is reported.

5.2 Evaluations

Performance Comparison of Few-Shot and Zero-Shot Prediction

We evaluate each method in both few-shot and zero-shot scenarios, and report the average results of 5 independent runs in cross-city tasks in Table 2.

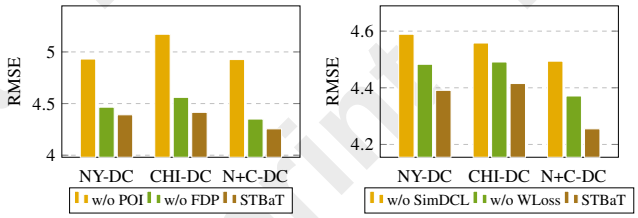
In few-shot prediction scenario, deep learning methods based solely on target data i.e., ST-net and PDFormer, exhibit unsatisfactory performance. Among the transfer baselines, source-target joint pre-training methods (e.g., ST-DAAN and CrossTReS) generally outperform source-only pre-training methods (e.g., MetaST and RegionTrans). This advantage stems from the utilization of few-shot target data during pre-training to help adapting to specific target city. Notably, STBaT does not utilize any target data during pre-training, whereas it reduces RMSE by up to 6.3% and MAE by up to 11.6% compared to the best source-target joint pre-training method. This improvement can be attributed to the balanced learning strategy, which acquires a favorable initialization that internalizes knowledge beneficial for generalization, thus facilitating precise transfer to the target domain.

In zero-shot prediction scenario, the non-transfer baselines and some transfer baselines that utilize target for pre-training are not applicable. Among the source-only pre-training methods, MetaST yields lower error compared to other baselines, resulting in suboptimal performance. This may be attributed to its model design, which includes memory units that can capture transferable patterns from the source cities. Thanks to the capability to identify and calibrate distribution biases arising from imbalance in the source cities, STBaT once again achieves optimal performance in zero-shot scenarios. It enhances RMSE by 5.8% and MAE by 9.4% compared to the best source-only pre-training method, showcasing the superiority of our proposed framework in situations of severe data scarcity.

Across all baselines, although source-only pre-training methods can fit to more data-scarce scenarios, they fail to outperform source-target joint pre-training methods across all tasks. Consequently, no single baseline achieves both universality and precision, except for STBaT, which consistently outperforms other methods in all tasks and scenarios.

Impact of Components

To assess the efficacy of each design within STBaT, we create two variants of our framework.



(a) Impact of model structure (b) Impact of training loss

Figure 3: Performance comparison of different variants of STBaT

Method	Finetuned	MetaST	CrossTReS	STBaT (ours)
Peak Memory (GB)	1.57	3.39	7.65	1.95
Training Time (h)	0.25	2.34	1.28	0.39

Table 3: Training memory and time costs of different approaches

- **Variants of model structure**: Including removing the POI classifier (*w/o POI*) and removing the spatio-temporal feature distribution predictor (*w/o FDP*).
- **Variants of training loss**: Including removing the similarity-aware distribution contrastive learning loss (*w/o SimDCL*) and removing the density-aware weighted prediction loss (*w/o WLoss*).

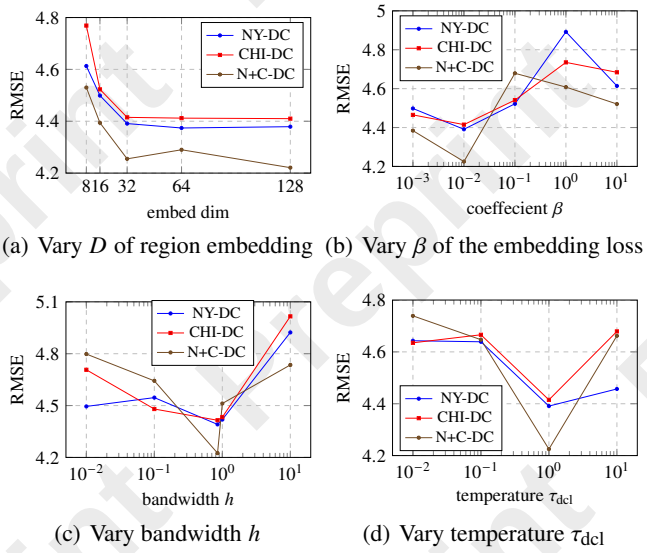
Figure 3 depicts the RMSE of STBaT variants in few-shot taxi flow prediction task of DC. For the model structure, STBaT w/o POI and w/o FDP both lead to significant performance drop, highlighting the importance of capturing regional functionalities and flow characteristics for modeling urban regional imbalances. For the training loss, both STBat w/o SimDCL and w/o WLoss result in accuracy degradation, attributed to unbalanced learning in feature extraction and traffic prediction, respectively.

Impact of Hyper-parameters

To assess the impact of the hyper-parameters on STBaT's performance, we conduct analyses in few-shot taxi flow prediction tasks of DC. The results are shown in Figure 4. (1) We vary the dimension D of the region embedding, and set it to 32. This allows us to effectively capture the traffic patterns. (2) We tune the parameter β of embedding loss and find that a value of 0.01 consistently yields the best performance in different prediction tasks. (3) We adjust the bandwidth h of the RIAM's density estimator, and set it to 0.85 to achieve the best performance. (4) We set the temperature τ_{dcl} to 1 to better learn balanced spatio-temporal feature extraction, adapting effectively to different target city prediction tasks.

Impact of Data Scarcity

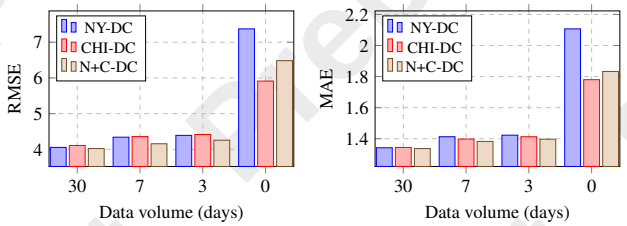
We conducted an experiment based on taxi flow prediction tasks to investigate the performance of our method under different scarcities of target data. Figure 5 shows that as the amount of available data in the target city increases, the performance on different cross-city tasks gradually improves and eventually approaches saturation. The significant improvement from 0 days to 3 days of data indicates that model can adapt well to the target city, even with minimal data.



(a) Vary D of region embedding (b) Vary β of the embedding loss

(c) Vary bandwidth h (d) Vary temperature τ_{dcl}

Figure 4: Performance of STBaT with different hyper-parameters



(a) RMSE of different target data volumes in different tasks (b) MAE of different target data volumes in different tasks

Figure 5: Performance with different scarcities of data

Efficiency Study

We studied the training efficiency of STBaT against several representative baselines, with detailed comparisons shown in Table 3. The results demonstrate that while vanilla fine-tuning achieves the lowest training costs, its performance lags significantly behind competitors; STBaT substantially outperforms MetaST and CrossTReS in both memory and time consumptions through its effective learning algorithms.

6 Conclusion

This paper introduces a spatio-temporal balanced transfer learning framework (STBaT), a novel paradigm to enhance the universality and precision of pre-training a model for data-scarce urban flow prediction. By leveraging a Regional Imbalance Acquisition Module and a Spatio-Temporal Balanced Learning Module, STBaT demonstrates its superiority in few-shot and zero-shot scenarios of diverse prediction tasks. For future work, integrating the idea of STBaT into various urban computing tasks, such as graph-based predictions, shows promise. Our study paves an encouraging path for future spatio-temporal generalization learning.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62272033).

References

[Box and Jenkins, 1968] George EP Box and Gwilym M Jenkins. Some recent advances in forecasting and control. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 17(2):91–109, 1968.

[Chen *et al.*, 2022] Wei Chen, Huaiyu Wan, Shengnan Guo, Haoyu Huang, Shaojie Zheng, Jiamu Li, Shuohao Lin, and Youfang Lin. Building and exploiting spatial–temporal knowledge graph for next poi recommendation. *Knowledge-Based Systems*, 258:109951, 2022.

[Cirstea *et al.*, 2022] Razvan-Gabriel Cirstea, Bin Yang, Chenjuan Guo, Tung Kieu, and Shirui Pan. Towards spatio-temporal aware traffic time series forecasting. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2900–2913. IEEE, 2022.

[Ding *et al.*, 2020] Jingtao Ding, Guanghui Yu, Yong Li, Depeng Jin, and Hui Gao. Learning from hometown and current city: Cross-city poi recommendation via interest drift and transfer learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(4), September 2020.

[Fang *et al.*, 2022] Ziquan Fang, Dongen Wu, Lu Pan, Lu Chen, and Yunjun Gao. When transfer learning meets cross-city urban flow prediction: Spatio-temporal adaptation matters. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2030–2036. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[Geng *et al.*, 2019] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3656–3663, 2019.

[Gong *et al.*, 2022] Yongshun Gong, Zhibin Li, Jian Zhang, Wei Liu, and Yu Zheng. Online spatio-temporal crowd flow distribution prediction for complex metro system. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):865–880, 2022.

[Guo *et al.*, 2019] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019.

[HKGov, 2025] HKGov. Hong kong smart city blueprint. https://www.smartcity.gov.hk/, January 2025.

[Ji *et al.*, 2023] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Junbo Zhang, and Yu Zheng. Spatio-temporal self-supervised learning for traffic flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 4356–4364, 2023.

[Jia *et al.*, 2023] Yuxin Jia, Youfang Lin, Xinyan Hao, Yan Lin, Shengnan Guo, and Huaiyu Wan. Witran: Water-wave information transmission and recurrent acceleration network for long-range time series forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Jiang *et al.*, 2023] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4365–4373, Jun. 2023.

[Jin *et al.*, 2022] Yilun Jin, Kai Chen, and Qiang Yang. Selective cross-city transfer learning for traffic prediction via source city region re-weighting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 731–741, 2022.

[Jin *et al.*, 2023] Yilun Jin, Kai Chen, and Qiang Yang. Transferable Graph Structure Learning for Graph-based Traffic Forecasting Across Cities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1032–1043, August 2023.

[Kono, 2022] Vitor Kono. Sustainable Travel Innovations by Liverpool John Moores University. https://vivacitylabs.com/sustainable-travel-innovation-liverpool/, January 2022.

[Li *et al.*, 2018] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

[Liang *et al.*, 2021] Yuxuan Liang, Kun Ouyang, Junkai Sun, Yiwei Wang, Junbo Zhang, Yu Zheng, David Rosenblum, and Roger Zimmermann. Fine-Grained Urban Flow Prediction. In *Proceedings of the Web Conference 2021*, pages 1833–1845, April 2021.

[Liu *et al.*, 2023] Zhanyu Liu, Guanjie Zheng, and Yanwei Yu. Cross-city few-shot traffic forecasting via traffic pattern bank. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1451–1460, 2023.

[Lu *et al.*, 2022] Bin Lu, Xiaoying Gan, Weinan Zhang, Huaxiu Yao, Luoyi Fu, and Xinbing Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1162–1172, 2022.

[Lv *et al.*, 2014] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction

with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014.

[Ouyang *et al.*, 2024] Xiaocao Ouyang, Yan Yang, Wei Zhou, Yiling Zhang, Hao Wang, and Wei Huang. Citytrans: Domain-adversarial training with knowledge transfer for spatio-temporal prediction across cities. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):62–76, 2024.

[Parzen, 1962] Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076, 1962.

[Qu *et al.*, 2023] Hao Qu, Yongshun Gong, Meng Chen, Junbo Zhang, Yu Zheng, and Yilong Yin. Forecasting fine-grained urban flows via spatio-temporal contrastive self-supervision. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8008–8023, 2023.

[Rosenblatt, 1956] Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837, 1956.

[Song *et al.*, 2020] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-Temporal Synchronous Graph Convolutional Networks: A New Framework for Spatial-Temporal Network Data Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):914–921, April 2020.

[Tang *et al.*, 2022] Yihong Tang, Ao Qu, Andy HF Chow, William HK Lam, SC Wong, and Wei Ma. Domain adversarial spatial-temporal network: a transferable framework for short-term traffic forecasting across cities. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1905–1915, 2022.

[Wang *et al.*, 2018] Leye Wang, Xu Geng, Xiaojuan Ma, Feng Liu, and Qiang Yang. Cross-city transfer learning for deep spatio-temporal prediction. *arXiv preprint arXiv:1802.00386*, 2018.

[Wang *et al.*, 2021] Senzhang Wang, Hao Miao, Jiyue Li, and Jiannong Cao. Spatio-temporal knowledge transfer for urban crowd flow prediction via deep attentive adaptation networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(5):4695–4705, 2021.

[Wang *et al.*, 2024a] Binwu Wang, Jiaming Ma, Pengkun Wang, Xu Wang, Yudong Zhang, Zhengyang Zhou, and Yang Wang. Stone: A spatio-temporal ood learning framework kills both spatial and temporal shifts. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2948–2959, 2024.

[Wang *et al.*, 2024b] Yu Wang, Tongya Zheng, Yuxuan Liang, Shunyu Liu, and Mingli Song. COLA: Cross-city Mobility Transformer for Human Trajectory Simulation. In *Proceedings of the ACM Web Conference 2024*, pages 3509–3520, May 2024.

[Wei *et al.*, 2016] Ying Wei, Yu Zheng, and Qiang Yang. Transfer Knowledge between Cities. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1905–1914, August 2016.

[Wen *et al.*, 2023] Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23, 2023.

[Xia *et al.*, 2023] Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 37068–37088. Curran Associates, Inc., 2023.

[Yao *et al.*, 2018] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Yao *et al.*, 2019] Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *The world wide web conference*, pages 2181–2191, 2019.

[Yu *et al.*, 2017] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[Yuan *et al.*, 2023] Yuan Yuan, Chenyang Shao, Jingtao Ding, Depeng Jin, and Yong Li. Spatio-Temporal Few-Shot Learning via Diffusive Neural Network Generation. In *The Twelfth International Conference on Learning Representations*, October 2023.

[Yuan *et al.*, 2024] Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4095–4106, August 2024.

[Zhang *et al.*, 2017] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Zheng *et al.*, 2020] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1234–1241, Apr. 2020.