# Multi-Source Transfer Learning With Spatial-Temporal Graph Neural Network for Short-Term Bicycle Traffic Prediction

Xiamei Wen, Megha Khosla, and Serge Hoogendoorn

***Abstract*—Bicycle transportation, a low-carbon option, is essential for promoting sustainable urban mobility. However, predicting bicycle traffic is challenging due to limited investments in data collection, especially in smaller cities. This paper proposes a multi-source transfer learning spatial-temporal graph neural network (Multi-TLSTGCN) for accurate bicycle traffic prediction in target cities with limited available data. This study first examines how to transfer knowledge from single source domain to the target domain while mitigating the risk of negative transfer. Following this, a multi-source adaptive transfer learning approach is developed to optimize traffic prediction in the target domain by adaptively integrating knowledge from multiple sources. Finally, the performance of the Multi-TLSTGCN model is evaluated under various levels of target data scarcity and compared with models that do not incorporate source domain knowledge. The experimental results demonstrate several key insights: 1) Models fine-tuned with a single-cluster pre-trained source model where the clusters are formed based on similar traffic patterns are more effective at minimizing negative knowledge transfer than those fine-tuned with single-city pre-trained source models. 2) The proposed Multi-TLSTGCN outperforms baseline models in bicycle traffic prediction, showing promise for accurate predictions in data-scarce environments; and 3) The Multi-TLSTGCN model remains robust across varying levels of data scarcity, exhibiting only a slight decrease in accuracy as the availability of target data decreases, in contrast to models relying solely on target domain data. These findings highlight the Multi-TLSTGCN model as an effective and promising solution for bicycle traffic prediction with limited data availability.**

***Index Terms*—Bicycle traffic, transfer learning, traffic prediction.**

## I. Introduction

ACCURATE short-term traffic predictions are crucial for efficient transportation operations in urban areas. These predictions enable proactive decision-making by traffic management authorities, helping to mitigate traffic delay, reduce travel time, and enhance overall road network efficiency [1], [2]. Traffic prediction has been approached through simulation, parametric, and machine learning methods. Simulation techniques like discrete event simulation (DES) [3] and agent-based modeling (ABM) [4] replicate traffic dynamics but require extensive system knowledge and may lack generalizability. Parametric methods, including regression models [5], autoregressive integrated moving average (ARIMA) [6], and Kalman filters [7], offer quantitative insights but struggle with complex, non-linear traffic patterns.

Machine learning has shifted traffic prediction towards data-driven models that capture intricate patterns and adapt to dynamic environments. Methods such as k-nearest neighbors (KNN) [8], support vector machines (SVM) [9], and artificial neural networks (ANN) [10] improve prediction but often fail to capture deep traffic dependencies. Deep learning further enhances accuracy by modeling spatial-temporal dependencies. Recurrent neural networks (RNNs) [11], long short-term memory networks (LSTMs) [12], and attention mechanisms [13] extract temporal trends, while graph neural networks (GNNs) [14] and convolutional neural networks (CNNs) [15] model spatial interactions. By integrating both, deep learning provides more precise and comprehensive traffic forecasts.

The effectiveness of machine learning models depends on data quality and quantity. Despite abundant traffic data from IoT [16], real-world datasets often suffer from missing values, sensor errors, and inconsistencies. Resource constraints further limit traffic monitoring, especially in smaller cities. Privacy concerns may also restrict data access. Bicycle traffic faces even greater challenges due to inadequate monitoring infrastructure, as urban planning has historically prioritized motorized transport. Data scarcity and quality issues undermine prediction accuracy, making reliable insights into bicycle traffic increasingly critical as cities shift toward reduced car dependency.

Transfer learning addresses data scarcity by leveraging knowledge from data-rich environments (source domain) to improve predictions in data-limited areas (target domain) [17]. In transportation, it enhances traffic prediction by transferring patterns from well-monitored cities to smaller ones. Some studies achieve this by freezing lower model layers while fine-tuning upper layers with target data [18], while others mitigate domain shifts using adversarial training [19], maximum mean discrepancies (MMD) [20], or Dynamic Time Warping (DTW) [21]. Despite progress, transfer learning for traffic prediction, especially bicycle traffic, remains an ongoing challenge.

### A. Current Research Gaps and Our Contributions

One major limitation of existing transfer learning models is their narrow focus on single-source and single-target domains,

Received 5 March 2025; revised 8 July 2025 and 7 October 2025; accepted 12 November 2025. Date of publication 26 November 2025; date of current version 18 February 2026. This work was supported by China Scholarship Council. The Associate Editor for this article was X. Ma. *(Corresponding author: Xiamei Wen.)*

Xiamei Wen and Serge Hoogendoorn are with the Department of Transport and Planning, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: X.Wen@tudelft.nl; s.p.hoogendoorn@tudelft.nl).

Megha Khosla is with the Department of Intelligent Systems, Delft University of Technology, 2628 XE Delft, The Netherlands (e-mail: m.khosla@tudelft.nl).

Digital Object Identifier 10.1109/TITS.2025.3633930

with limited exploration of ensemble transfer learning. This is particularly problematic for bicycle traffic, where sensor distribution is generally sparse, and data from a single city may not provide sufficient coverage for accurate predictions. The lack of research into multi-source transfer learning, and specifically what information should be transferred, leaves a significant gap, especially for bicycle traffic. In motorized traffic, road structures are relatively consistent across cities, leading to similar distribution patterns that can be easily transferred between urban environments. In contrast, bicycle traffic is far more complex, with no clear road hierarchies, and traffic patterns vary widely from city to city. This variability makes it difficult to identify consistent trends that can be transferred effectively across different urban environments.

To overcome the above challenges, we develop a robust multi-source transfer learning framework which we refer to as Multi-TLSTGCN tailored to the requirements of data-scarce spatial-temporal bicycle traffic prediction. Such a framework would be instrumental in overcoming the limitations of current models and providing accurate predictions. Our approach enhances traffic prediction accuracy in data-scarce target cities by leveraging information learned from multiple source domains with more abundant data. Our approach begins with a traffic pattern clustering technique, which groups traffic patterns from various source cities, enabling more effective single-source knowledge transfer to the target cities. Subsequently, to maximize the utility of knowledge from all sources, we employ an adaptive transfer learning approach that integrates the most relevant aspects from the multi-source models, tailoring them for bicycle traffic prediction in the target city. We conducted extensive experiments using real-world bicycle traffic datasets from six cities to validate the proposed framework. Results indicate that our proposed Multi-TLSTGCN model outperforms both baseline models and single-source transfer approaches. Notably, as target city data availability decreases, our model's robustness surpasses that of models without transfer support, demonstrating its effectiveness for accurate bicycle traffic prediction in data-scarce urban settings.

The paper is organized as follows. Section II provides an overview of existing research on spatial-temporal traffic flow prediction, followed by an exploration of research on transfer learning in transportation. Section III offers detailed insights into the proposed Multi-TLSTGCN model. Section IV introduces the methodology of this study, including the spatial-temporal prediction model, clustering approach and the propose adaptive transfer learning approach. Section V presents the experimental setup, datasets used, experiment results, discussions, and an ablation study. Finally, Section VI concludes this work.

## II. Related Work

### A. Spatial-Temporal Traffic Flow Prediction

In recent years, the burgeoning availability of traffic data has sparked a growing interest in deep learning traffic prediction research. Deep learning models for traffic prediction, For example, Recurrent Neural Networks (RNNs) [11] can leverage vast amounts of historical traffic data to learn complex nonlinear temporal patterns and make accurate predictions. Graph Neural Networks (GNNs) [25] are tailored to handle non-Euclidean spatial structural data, making them well-suited for modeling the intricate structure of traffic networks [26]. By integrating GNNs with temporal dependencies, it becomes promising to capture the complexity and non-linear spatial-temporal traffic patterns effectively. Bao et al. [27] investigated spatial–temporal accurate traffic flow prediction based on Spatial–Temporal Complex Graph Convolution Network (ST-CGCN). It combines spatial and temporal features while considering node correlations and external interferences. By leveraging dynamic weights and complex correlation matrices, ST-CGCN outperforms existing models in real-world datasets. Shi et al. [28] designed a novel Attention-based Periodic-Temporal neural Network (APTN) to address the challenges of accurate traffic forecasting in smart cities. By leveraging attention mechanisms, APTN effectively captures spatial, short-term, and long-term periodical dependencies in traffic data. The experiments demonstrated significant improvements of APTN in traffic forecasting applications over existing state-of-the-art models based on real-world datasets PeMSD4 and PeMSD8 from California. However, the effectiveness of these advanced methods can decline when trained on limited data, potentially resulting in overfitting when the dataset is not sufficiently large.

In recent years, significant amounts of motorized traffic data have been gathered, particularly in large metropolitan areas with abundant resources. This is largely due to the variety of data collection sources available, such as traffic cameras, GPS devices integrated into vehicles, and advanced smart infrastructure sensors [29]. These data sources have enabled detailed analysis and prediction of traffic patterns for cars and other motorized vehicles in large metropolitan areas [27], [30]. However, when it comes to bicycle traffic, the situation is different. In countries like the Netherlands, where bicycle is a primary mode of urban transport, particularly for short-distance travel and eco-friendly commuting [31], the collection of bicycle traffic data has been problematic. The lack of dedicated sensors and tracking systems, along with frequent malfunctions of the existing ones, significantly hampers the ability to gather comprehensive bicycle traffic data [32]. The problem is even more pronounced in smaller municipalities, where limited resources make sustained data collection efforts difficult, further complicating the analysis and prediction of bicycle traffic patterns. While general bicycle traffic data is difficult to obtain, bike-sharing systems offer a notable exception. In large metropolitan areas, data from these systems is abundant, thanks to the robust digital infrastructure that supports them [33]. However, this ease of data collection is not applicable to general bicycle traffic, which remains largely unmonitored. Consequently, most existing studies on bicycle traffic flow prediction focus on bike-sharing systems [34], leaving general bicycle traffic, despite its significance as a major urban transportation mode, largely neglected.

### B. Transfer Learning in Transportation

Transfer learning is a technology focuses on utilizing knowledge gained from one domain or task and applying it to a different but related domain or task [17]. In the context of traffic prediction, transfer learning has shown significant promise in addressing challenges related to traffic data scarcity, variability in traffic patterns, and the need for robust predictive models. Li et al. [18] applied transfer learning techniques with multiple transfer strategies to machine learning methods for short-term traffic prediction using data from highways England road networks in the UK. To address the challenge of data scarcity in small cities, Huang et al. [21] proposed a cross-city traffic prediction approach called TrafficTL. This method leverages big data from other cities to assist data-scarce cities by identifying similarities between data sets and mitigating negative transfer effects caused by differences in data distributions from distant locations. Yao et al. [35] introduces an Adversarial Domain Adaptation with Spatial-Temporal Graph Convolutional Network (Ada-STGCN) model designed to predict traffic indicators for data-scarce target road networks by transferring knowledge from data-rich source road networks. Experimental results on real-world traffic datasets demonstrate that Ada-STGCN outperforms state-of-the-art baseline methods in traffic flow prediction tasks, delivering superior prediction accuracy. These studies have explored various methods for transferring knowledge from data-rich cities to data-scarce ones. However, these methods are typically designed for motorized traffic, which benefits from abundant traffic data both temporally and spatially, with extensive sensor coverage across the road network. In contrast, the sensor distribution for private bicycle traffic is often sparse, resulting in limited source data that may cover only a small portion of the road network. This lack of diverse traffic patterns in the source data makes it challenging to transfer knowledge effectively. Consequently, existing methods may not be suitable for private bicycle traffic prediction.

Single-source transfer may not perfectly align with the features of the target domain. In contrast, multi-source transfer can encompass a broader range of scenarios and subdomains within the target domain, leveraging multiple related domains to improve the effectiveness of transfer learning. Li et al. [37] proposed a physics-guided multi-source transfer learning method for multi-region traffic flow, utilizing adversarial training and MFD-based weighting. This approach achieves domain adaptation and assigns weights based on traffic network properties. To mitigate the impact of noise or potentially negative knowledge from source cities, Jin et al. [38] proposed a selective transfer learning framework called CrossTReS for traffic prediction. This framework adaptively re-weights source regions to enhance fine-tuning in the target domain. Mo and Gong [39] proposed a transfer learning method for traffic prediction called Cross-city Multi-Granular Adaptive Transfer Learning (MGAT). The model is trained on multiple source cities using meta-learning algorithm to achieve a strong initialization. It then captures multi-granular regional characteristics and employs an Adaptive Transfer module with Spatial-Attention and Multi-head Attention mechanisms to automatically select the most relevant features, ensuring optimal knowledge transfer. The significant success of these studies highlights the potential of multi-source transfer learning. However, existing approaches primarily focus on multi-region transfer, with a balanced distribution of source data across different regions. For bicycle traffic, where source data may be sparse within each city, this assumption may not hold, making traditional multi-region transfer less effective.

## III. Preliminaries and Problem Definition

This section provides an overview of the preliminaries relevant to transfer learning traffic prediction. First, we define the structure of the traffic network, followed by an explanation of traffic networks in both the source and target domains. Finally, we outline the specific research problem addressed in this work.

### A. Physical Traffic Network Description

To simplify bicycle spatial correlations and account for the absence of bidirectional traffic flow data, we define the bicycle traffic network as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ is the set of nodes representing sensors that record the traffic flow of bicycle on the road network; $\mathcal{E}$ is a set of edges; and $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the adjacency matrix of $\mathcal{G}$ with $N$ nodes recorded as a weighted adjacency matrix. Note that the traffic flow feature collected by each sensor changes over time. Specifically, $\mathbf{X}_t = (\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \ldots, \mathbf{x}_{t,i}, \ldots, \mathbf{x}_{t,N}) \in \mathbb{R}^{F\times N}$ denotes the traffic feature of the graph $\mathcal{G}$ at time $t$, where $\mathbf{x}_{t,i}$ indicates the feature value of the $i-th$ node, and $F$ represents the total number of traffic features for each node.

In this paper, sensors are depicted as nodes within a graph, and the Dijkstra algorithm [40] is used to determine the shortest path between nodes within the road network. The connectivity of the topological structure is defined by these shortest paths between nodes. Considering that the traffic patterns observed at a particular sensor can be influenced by the behavior of neighboring sensors, we establish weighted connections between sensors using a distance-based adjacency matrix $\mathbf{A}$. This entails applying a Gaussian kernel weighting function [41], with weights determined by sensor distances, as elaborated in the following.

$$\mathbf{A}(i,j) = \begin{cases} \exp\left(-\frac{|\mathrm{dist}(i,j)|^2}{2\hat{\theta}^2}\right) & \text{if } dist(i,j) < H \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $|\mathrm{dist}(i,j)|$ indicates the shortest path between node $i$ and node $j$. In this context, $\hat{\theta}$ represents the standard deviation of distances, and $H$ is utilized as the threshold parameter.

*1) Traffic Networks of Source Domain:* We define the multi-source domains graph as $\mathcal{G}^S = \mathcal{G}^{S_1}, \ldots, \mathcal{G}^{S_i}, \ldots, \mathcal{G}^{S_M}$ to describe the source road network of the cities, where $M$ represent the number of source cities. $\mathcal{G}^{S_i} = (\mathcal{V}^{S_i}, \mathcal{E}^{S_i}, \mathbf{A}^{\mathbf{S_i}})$ denotes the graph of $i-th$ source city, where $\mathcal{V}^{S_i}$ is the set of nodes representing sensors that record the traffic flow on the road network of $i-th$ source city; $\mathcal{E}^{S_i}$ is a set of edges; and $\mathbf{A}^{\mathbf{S_i}} \in \mathbb{R}^{N^{S_i}\times N^{S_i}}$ is the adjacency matrix of $\mathcal{G}^{S_i}$ in city $i$ with $N^{S_i}$ nodes recorded as a weighted adjacency matrix. $\mathbf{X}_{\mathbf{t}}^{\mathbf{S_i}} \in \mathbb{R}^{N^{S_i}}$ is the feature matrix of $i-th$ source domain graph at time $t$.

*2) Traffic Networks of Target Domain:* We define the target domain graph as $\mathcal{G}^T = (\mathcal{V}^T, \mathcal{E}^T, \mathbf{A}^T)$, where $\mathcal{V}^T$ is the set of nodes representing sensors that record the traffic flow of the road network; $\mathcal{E}^T$ is a set of edges; and $\mathbf{A}^T \in \mathbb{R}^{N^T\times N^T}$ is the

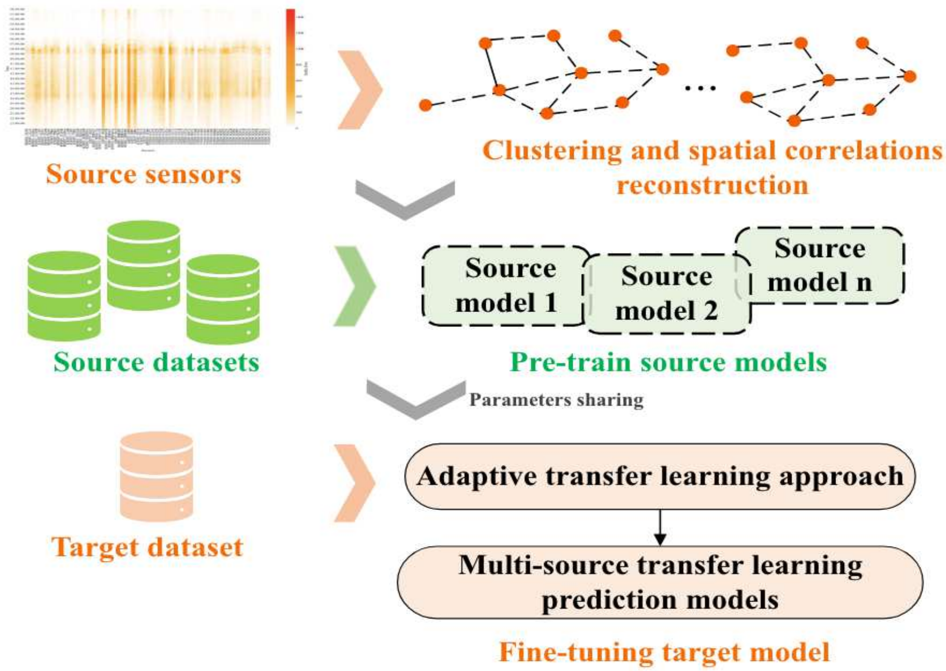



Fig. 1. Overall framework of Multi-TLSTGCN model. Cluster-based multi-source transfer with reconstructive spatial network.

adjacency matrix of $\mathcal{G}^T$ with $N^T$ nodes recorded as a weighted adjacency matrix of target city. The traffic flow feature on the target domain graph changes over time. $\mathbf{X}_t^T \in \mathbb{R}^{N^T}$ is the feature matrix of target domain graph at time $t$.

*B. Problem Definition*

Our objective is to develop a precise bicycle traffic prediction model for a data-scarce target road network within a city by leveraging insights gained from other cities. In this paper, both the source and target domains share a common task, which involves predicting future traffic state based on historical observations.

For both the source and target domains, the historical bicycle traffic flow feature for each input of each node includes data from one hour earlier, which can be represented as $\mathbf{X} = (\mathbf{X}_{t-k+1}, \mathbf{X}_{t-k+2}, \ldots, \mathbf{X}_t) \in \mathbb{R}^{F \times T_k}$, where $T_k$ are the time steps of each slice source, $\mathbf{X}_t = (\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \ldots, \mathbf{x}_{t,N}) \in \mathbb{R}^{\mathbb{F}}$. $F$ indicates the vector features of each node at time $t$. Our goal is to find a function $f(\cdot)$ to predict the following $T$ time steps data $\hat{\mathbf{X}} = (\hat{\mathbf{X}}_{t+1}, \hat{\mathbf{X}}_{t+2}, \ldots, \hat{\mathbf{X}}_{t+T})$, that is:

$$(\hat{\mathbf{X}}_{t+1}, \hat{\mathbf{X}}_{t+2}, \ldots, \hat{\mathbf{X}}_{t+T}) = f_\theta((\mathbf{X}_{t-k+1}, \ldots, \mathbf{X}_t), \mathbf{A}_t) \quad (2)$$

To align with the need for immediate traffic management and planning, such as adjusting traffic signals or deploying resources to manage traffic, our approach involves utilizing the most recent one-hour historical data as input for our model to predict the traffic for the subsequent hour, where $\theta$ represents the learnable parameters of the function.

## IV. Methodology

To address the challenge of data scarcity in traffic flow prediction, we introduce the Multi-TLSTGCN model, which leverages abundant data from other cities to inform learning in the target city. The core idea is to first build a pre-trained model using sensor data from multiple cities, capturing a wide variety of traffic patterns and dynamics. By leveraging this diverse datasets, the model gains a richer understanding of traffic behaviors, leading to better generalization. Once pre-trained, the model is fine-tuned on the limited data from the target city, resulting in significantly improved predictive performance, even with scarce target data. The overall framework is shown in **Fig. 1**.

The *first* challenge in realizing our objective is that, unlike motorized traffic, which follows organized road hierarchies with dense sensor coverage, bicycle traffic lacks such structure, and the irregular distribution of sensors often fails to cover all regions. Additionally, road usage patterns vary significantly between cities, making it difficult to effectively transfer regional models to different bicycle networks. To overcome this, we first cluster sensors from multiple cities into groups based on similar bicycle traffic patterns. Clustering sensor data ensures that those with comparable behaviors are grouped together, enabling us to train a separate source model for each cluster. This process (detailed in Section IV-A) improves the likelihood that models will identify familiar patterns when transferred to the target city, thereby reducing the risk of negative knowledge transfer, where mismatched knowledge could degrade performance. Furthermore, clustering helps manage the complexity and variability of bicycle traffic, making models more adaptable and focused within each cluster.

However, even within a cluster of sensors with similar patterns, local differences in traffic behavior still exist at the individual sensor level, resulting in the *second* challenge. To address these variations, we incorporate data-driven and adaptive spatial correlations within each cluster (detailed in

Section IV-B), allowing the model to adjust for local discrepancies. This ensures that predictions remain accurate and context-sensitive, as the model considers how traffic at one location influences other locations with similar traffic behaviors. (refer to Section IV-C)

Relying solely on single-source knowledge of bicycle traffic may not provide sufficient predictive performance in the target domain. This is primarily because bicycle traffic flow can vary significantly across different regions due to a multitude of factors, including population density, urban layout, and the availability of transportation infrastructure for bicycle. Moreover, the distribution of sensors used to collect bicycle traffic data is often sparse, particularly in lower-priority or less densely populated areas. To overcome this *third* challenge, we introduce an adaptive multi-source transfer learning approach that aggregates contributions from multiple source domains to improve prediction accuracy in the target domain (details in Section IV-D).

### A. Bicycle Traffic Pattern Clustering

To capture the overall traffic patterns for each sensor, we calculated the average traffic flow at 5-minute intervals throughout the day, creating a general representation of daily traffic trends for each sensor. To categorize sensors with similar daily traffic patterns into distinct groups, we implemented a two-step approach that combined deep learning with clustering techniques. This approach enabled us to effectively group sensors based on their unique traffic flow characteristics.

First, we used a temporal convolutional model to extract key temporal features from the general daily traffic patterns of each sensor. Specifically, a 1D convolutional neural network (1D-CNN) was designed, where the input channel size corresponded to each sensor's half-day time-series traffic patterns. The model processed the average traffic flow daily patterns of each sensor through two 1D-CNN layers with ReLU activation in between, capturing essential temporal characteristics. This process yields a reduced-dimensional matrix encapsulating the critical temporal dynamics of each sensor's traffic flow. Second, we applied Z-score normalization [42] to the transformed data, ensuring that all features contributed equally to the clustering process. Last, we employed the K-means algorithm [43] to cluster sensors with similar traffic patterns.

To determine the optimal number of clusters, we used the Elbow Method and Silhouette Score analysis [44], as detailed in Equations 3 and Equations 4. The Elbow Method helped identify the point where adding more clusters no longer significantly reduced the sum of squared distances, while the Silhouette Score assessed the cohesion and separation of the clusters. We determine the optimal number of clusters by identifying the point where the sum of squared errors begins to decrease more gradually, while also considering the Silhouette Score, aiming for a higher score to indicate better clustering quality.

$$SSE = \sum_{i=1}^{k_c} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2, \tag{3}$$

where $SSE$ is the within-cluster sum of squared errors; $k_c$ is the number of clusters; $C_i$ represents the $i$-th cluster; $x_j$ is a data point within cluster $C_i$; $\mu_i$ is the centroid of cluster $C_i$.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{4}$$

where $s(i)$ is Silhouette Score of data point $i$; $a(i)$ is the average distance from data point $i$ to all other points in the same cluster; $b(i)$ is the minimum average distance from the data point $i$ to all the points in the nearest different cluster.

This combined approach allowed us to effectively cluster sensors based on their daily traffic flow patterns, ensuring that the clusters were both meaningful and representative of the underlying traffic behaviors.

### B. Virtual Spatial Correlations Reconstruction Within Each Cluster

Since road network structures vary significantly across cities, directly transferring spatial correlations based on geographical proximity from source to target domains can lead to ineffective or even negative knowledge transfer. To address this challenge, we introduce virtual spatial correlations, which are defined by traffic pattern similarity rather than physical proximity in source domains. By focusing on similarity, we more effectively identify relevant spatial correlations in the target domain, improving the overall performance of transfer learning. Specifically, we *first* construct a data-driven graph that captures spatial correlations within each cluster based on the traffic attributes of sensors. This graph reflects the underlying similarities in traffic patterns, allowing for more meaningful connections between sensors. *Second*, we implement a parameter-based adaptive graph mechanism that dynamically adjusts spatial correlations over time to obtain overall graph spatial correlations. This adaptive graph continuously refines the model's understanding of spatial dynamics within each cluster, enabling it to account for temporal changes and local variations. This approach allows for more accurate and context-aware predictions. Below we explain the key two steps in detail.

*1) Data-Based Graph Spatial Correlations Learning:* To capture the data-based spatial correlations in bicycle, we begin by employing a 2D convolutional neural network (2D-CNN) to analyze the traffic flow information from each sensor based on the input data. Once the feature representations for each node are obtained, we calculate spatial correlations between nodes using the dot product, as shown in [45]. Instead of using a dynamic adjacency matrix for each input, which can complicate computations and introduce noise, we simplify the process by averaging the values across each batch to generate a stable data-based graph spatial correlations. We compute the the data-based adjacency matrix corresponding to the batch, $\mathbf{A}_{attri}$ as follows:

$$\mathbf{A}_{attri} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{ReLU}(M_i M_i^{\top}), \tag{5}$$

where $M_i$ is the traffic flow representations of all nodes in $i$-th batch; $M_i^{\top}$ is the transpose of $M_i$; $n$ denotes the batch size.

*2) Parameter-Based Adaptive Graph Spatial Correlations Learning:* To refine the dynamics of data-based graph spatial correlations over time and uncover hidden spatial relationships, we employ a parameter-based adjacency matrix derived from two learnable embedding dictionaries $E_1, E_2 \in \mathbb{R}^{N\times P}$. These dictionaries are used to capture and model the evolving spatial dependencies among nodes, enabling a more adaptive and nuanced representation of spatial correlations.

$$\mathbf{A}_a = E_1 E_2^\top \tag{6}$$

where $P$ is the hidden dimensions of each node. Both $E_1$, $E_2$ are the source node embedding dictionaries.

*3) Overall Graph Spatial Correlations Learning:* The overall graph spatial correlations, as defined in Equations 7, are obtained by multiplying the data-based graph spatial correlations with the parameter-based adaptive graph spatial correlations. Rectified Linear Unit activation function (ReLU) is then applied to ensure that the resulting spatial correlation values remain non-negative.

$$\mathbf{A}_{all} = ReLU(\mathbf{A}_{attri}\mathbf{A}_a) \tag{7}$$

### *C. Spatial-Temporal Dependency Learning*

Considering the fact that bicycle traffic network can encompass temporal attributes and diverse spatial features to portray different facets of spatial and temporal dynamics within the transportation system. In this study, we utilize an integrated attention-based spatial-temporal graph convolutional neural network known as ASTGCN, combining both the multi-head self-attention mechanism and graph convolutional neural network, to effectively capture the spatial-temporal dependencies inherent in traffic data of source and target domains.

*1) Temporal Dependency Capture Module:* Traffic flow patterns exhibit intricate temporal dependencies, where the flow at any given time is intricately linked to preceding time intervals. The multi-head attention mechanism with position encoding as shown in Equations 11 and 12 adeptly captures these traffic flow dependencies, this mechanism permits the model to discern various facets of the input sequence through multiple attention heads, facilitating the comprehensive capture of both short-term fluctuations and long-range trends within traffic flow data. In this scenario, we employ the multi-head self-attention mechanism [46], as depicted in Equations 8 and 9, to capture the temporal feature representation of the traffic flow data.

$$Multi-head(\mathbf{Q},\mathbf{K},\mathbf{V}) = Concat(head_1,\ldots,head_h) \tag{8}$$

where $head_i$ represents a self-attention mechanism, particularly a Scaled Dot-Product Attention [46], as outlined in Equation 10. The traffic flow input sequence $\mathbf{X}$ intended for prediction undergoes projection onto distinct learned linear subspaces, yielding the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ for the attention mechanism, respectively.

$$head_i = attention(\mathbf{X}_i\mathbf{W}_i^Q, \mathbf{X}_i\mathbf{W}_i^k, \mathbf{X}_i\mathbf{W}_i^V) \tag{9}$$

let $\mathbf{Q}_i = \mathbf{X}_i\mathbf{W}_i^Q$, $\mathbf{K}_i = \mathbf{X}_i\mathbf{W}_i^K$, and $\mathbf{V}_i = \mathbf{X}_i\mathbf{W}_i^V$, where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$ represent the learnable weights of the linear projection layers.

$$attention(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) = softmax\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_m}}\right)\mathbf{V}_i \tag{10}$$

where $d_m$ represents the scaling factor utilized to balance the complexity and capacity of the model.

$$P_{(pt,2dim)} = sin(pt/10000^{2dim/d_m}) \tag{11}$$

$$P_{(pt,2dim+1)} = cos(pt/10000^{2dim/d_m}) \tag{12}$$

The traffic flow input sequence $\mathbf{X}$ intended for prediction undergoes projection onto distinct learned linear subspaces, where $pt$ represents the position index of the traffic flow input sequence, while $dim$ corresponds to the $dim_{th}$ dimension of the positional encoding vector. Additionally, $d_m$ denotes the dimension of the positional encoding, aligning with the dimension of input embeddings to enable element-wise addition.

*2) Spatial Dependency Capture Module:* Bicycle traffic network frequently display intricate spatial connections and irregular geometries, rendering them non-Euclidean in nature. This complexity poses challenges for conventional Euclidean-based methods when applied to graph-based traffic prediction tasks. Spatial-based GCN [47] is a neural network architecture tailored to address this issue. GCN employ convolutional operations within graph structures to aggregate information from neighboring nodes, enabling them to learn representations that encapsulate both local and global spatial characteristics.

In this study, we model the traffic network as a graph, where nodes represent individual sensors and edges represent the connections or correlations between them. Each node is associated with traffic features extracted from its historical traffic flow data. To capture the spatial dependencies, a GCN is employed, which applies a series of graph convolutions to the traffic features of each node. These convolutions use the adjacency matrix $\hat{\mathbf{A}}_t$ to update a node's feature representation by aggregating information from its neighboring nodes. As a result, the GCN produces updated feature representations for each node, reflecting both local and global traffic patterns. The operation of the spatial-based GCN is described by Equation 13.

$$GCN(\mathbf{X}_t) = ReLU\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}_t\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}_t\mathbf{W}\right) \tag{13}$$

where $\hat{\mathbf{A}}_t$ represents the graph adjacency matrix. $\hat{\mathbf{D}}$ is a diagonal matrix where the $i_{th}$ element on the diagonal corresponds to the degree of the $i_{th}$ node, defined as $\hat{\mathbf{D}}_{ii} = \sum_j \mathbf{A}_{ij}$, here $j$ loops through all nodes in the graph. $\mathbf{W}$ represents the trainable weight matrix, while $ReLU(\cdot)$ denotes the activation function.

*3) Residual Connection and Feedforward Networks:* During the training of deep learning neural networks, particularly those with numerous layers, the gradients of the loss function concerning the parameters may diminish significantly as they propagate backward through the layers. Consequently, the earlier layers' weights may cease to learn effectively due to their gradients approaching zero. To mitigate the vanishing gradient problem and enhance model performance, we introduce residual connections between each layer. This mechanism enables the input of a specific layer to bypass one or more layers, directly adding it to the subsequent layer's output, as demonstrated in Equation 14. Additionally, we integrate a fully connected feedforward network [46] into each encoder layer. This incorporation empowers the model to capture intricate input-output relationships and introduce nonlinearity, as depicted in Equation 15.

$$\mathbf{X}^{l+1} = layer(\mathbf{X}^{l}) + \mathbf{X}^{l} \tag{14}$$

$$FeedForward(\mathbf{X}) = ReLU(\mathbf{X}\mathbf{W}_0 + b_0)\mathbf{W}_1 + b_1 \tag{15}$$

where $layer(\mathbf{X}^{l})$ represents the traffic flow embedding output of layer $l$, $\mathbf{X}^{l}$ denotes the traffic flow embedding input of layer $l$, and $\mathbf{X}^{l+1}$ signifies the traffic flow output after the residual connection. The function $layer(\cdot)$ denotes the operation responsible for capturing temporal or spatial information.

*4) Multi-Step Traffic Prediction:* In this study, we adopt a multi-step traffic prediction approach, which offers a holistic perspective on future traffic conditions. Additionally, within our model architecture, we include a fully connected layer in the reference section. This layer is designed to forecast traffic flow across multiple future time steps simultaneously, preventing the accumulation of prediction errors from prior time steps.

### *D. Bicycle Multi-Source Adaptive Transfer Learning Approach*

Multi-source data can help overcome the scarcity of bicycle data by providing a broader and more comprehensive dataset. Carefully selecting data from multiple sources can help mitigate the risk of negative transfer, ensuring that the model benefits from relevant and diverse information.

*1) Adaptive Transfer Learning Approach:* In this context, we propose an adaptive transfer learning approach. First, each cluster source model is trained separately on the target domain dataset. In the loss function as shown in Equation 16, we adaptively aggregate the contributions from all cluster-based transfer models using learnable weighted parameters to fine-tune the models. This method allows the model to adapt to the unique characteristics of the target domain while effectively utilizing source domain knowledge, minimizing the risk of negative transfer. The formula for combining the prediction results is provided in Equation 17.

$$\mathcal{L}_{\text{loss}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{16}$$

where $N$ is the number of samples; $y_i$ represents the traffic flow true value of the $i$-th sample; $\hat{y}_i$ represents the traffic flow predicted value of the $i$-th sample.

$$\hat{y}_i = \sum_{i=j}^{n} \alpha_j \cdot \hat{y}_{i,j} \tag{17}$$

where $\hat{y}_{i,j}$ represents the fine-tuned traffic flow prediction result of the $j$-th source model for the $i$-th sample on the target dataset; $\alpha_j$ denotes the learnable weight assigned to the $j$-th source model, which is initialized randomly using a standard normal distribution; $n$ is the total number of cluster-based transfer models.

**Algorithm 1** The Multi-TLSTGCN Model

**Input:** Bicycle traffic features $\mathbf{X}$ over $T_k$ historical timesteps; Number of training epochs $E_r$; learning rate $\eta$; batch size $B_z$; Number of prediction steps $T$; prediction output $\hat{\mathbf{X}}$; Learnable weights $\alpha_j$; Optimizer: Adam.
**Output:** Trained Multi-TLSTGCN model.
**Source Domain Training:**;
Cluster the source bicycle traffic patterns into $k_c$ clusters;
Reconstruct virtual spatial correlations within each cluster;
**foreach** *cluster in source domain* **do**
    **foreach** *epoch* $r \in \{1, 2, \ldots, E_r\}$ **do**
        Shuffle the training data;
        **foreach** *batch in training data* **do**
            $\hat{\mathbf{X}}_s \leftarrow f_\theta(\mathbf{X}_s)$;
        **end**
    **end**
**end**
**Target Domain Adaptation:**;
Load pretrained source models;
**foreach** *epoch* $r \in \{1, 2, \ldots, E_r\}$ **do**
    Shuffle the training data;
    **foreach** *batch in training data* **do**
        **foreach** *pretrained source model* **do**
            $\hat{\mathbf{X}}_r \leftarrow f_\theta(\mathbf{X}_r)$;
        **end**
        Compute $\hat{\mathbf{X}}$ using Equation (17);
        Update learnable parameters by minimizing the loss in Equation (16) using Adam;
    **end**
**end**

*2) Transfer Learning Process:* The model training process comprises two primary phases. Initially, the first phase occurs predominantly within the source domains, focusing on acquiring knowledge from the source domains. Subsequently, the second phase involves transferring the acquired knowledge from the source domains to the target domains for accurate traffic prediction. The overall workflow of Multi-TLSTGCN is outlined in Algorithm 1.

*a) Source domain training:* In the pre-training phase of multi-source transfer learning, sensors from various source cities are grouped into distinct clusters, each representing a different source domain, denoted as $\mathcal{S}_i$. Separate models are trained on the data within each of these source domains. Once trained, these source models are transferred to the target domain for subsequent fine-tuning.

*b) Target domain training:* After acquiring pre-trained models from multiple source domains, we fine-tune them using the target dataset. An adaptive transfer learning approach is then applied to combine the outputs from these fine-tuned models, producing the final prediction results, as described in Equation 17.

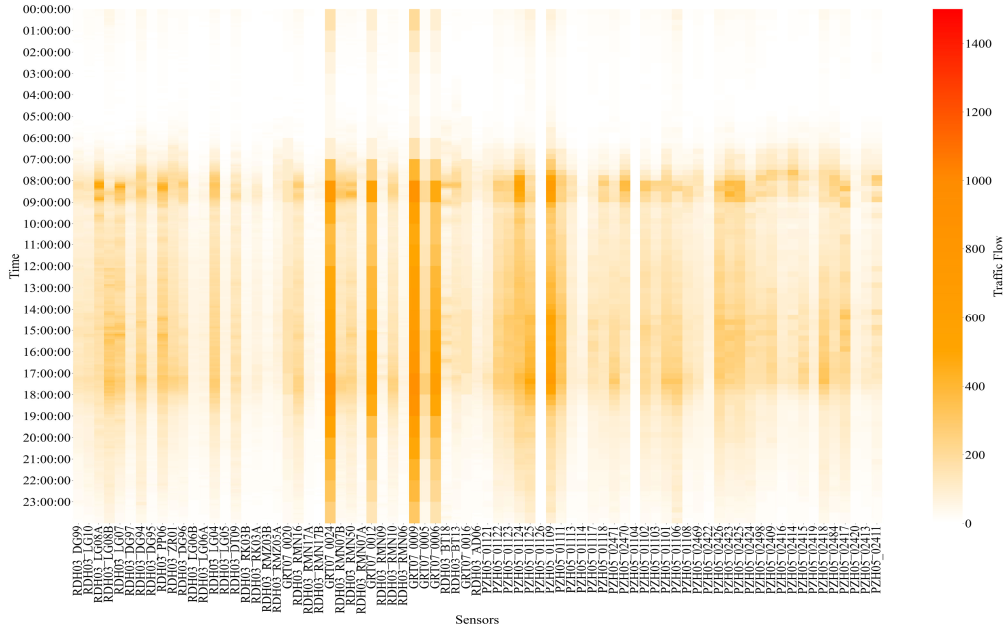



Fig. 2. Aggregated daily average bicycle traffic patterns in five-minute intervals.

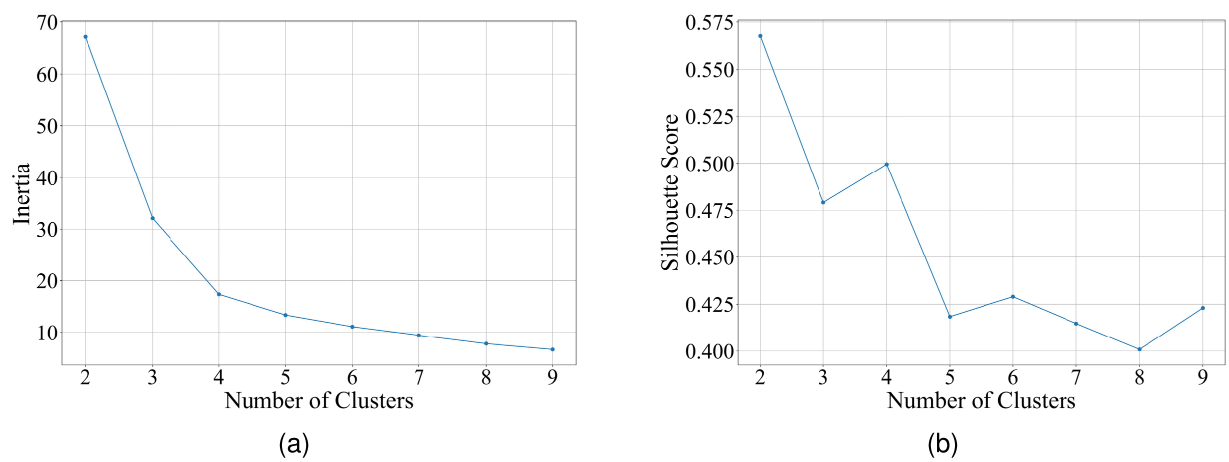



Fig. 3. The optimal number of clusters. (a) The sum of squared errors (SSE), the rate of SSE decline significantly slows at 4 clusters. (b) Silhouette score, the 4-cluster configuration achieved the second-highest score.

## V. Experiments

### A. Dataset Description

In this section, we first introduce the datasets used in both the source and target domains for our analysis. We then detail the cluster patterns identified within the source datasets using the K-means algorithm, which formed the basis for our subsequent modeling.

*1) Overall Dataset Description:* To evaluate the effectiveness of our proposed model in predicting bicycle traffic with scarce datasets, we conducted experiments using bicycle datasets from six cities in the Netherlands: Delft, The Hague, Rotterdam, Leiden, Dordrecht, and Gouda. Notably, the Delft dataset was collected by smart cameras located on the TU Delft campus, while the datasets for the other five cities were sourced from publicly accessible bicycle traffic flow data provided by the National Road Traffic Data Portal of the Netherlands, which were collected using loop detector sensors. We utilized the datasets from The Hague, Rotterdam, Leiden, and Dordrecht as source datasets. These datasets include data from 18, 21, 21, and 17 sensors respectively, recorded between January 1, 2022, and December 31, 2022. To fine-tuning the target model, we conducted experiments separately using two datasets from Delft and Gouda. Both target datasets include data from 10 sensors, recorded between April 1, 2022, and May 31, 2022. Additionally, all datasets contain geographic information regarding the sensor locations.

The bicycle traffic flow datasets are aggregated into 5-minute intervals, providing a balance that smooths out fluctuations while still capturing significant traffic patterns. This results in 12 data points per hour, allowing for detailed analysis. To improve model convergence and training stability, we apply Z-score standardization to normalize the input datasets. Our goal is to perform multi-step bicycle traffic prediction by using one hour of historical data to forecast traffic flow for the next 60, 30, and 15 minutes. All datasets are sequentially

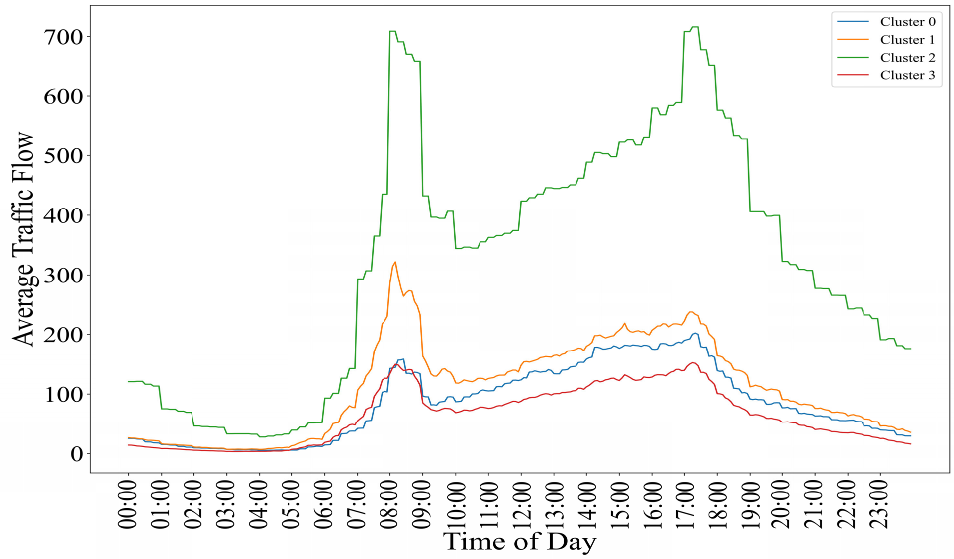



Fig. 4. Cluster result. Clustering based on aggregated daily average bicycle traffic patterns.

divided into training, validation, and test sets, following a 6:2:2 split ratio, respectively.

*2) Clustering Source Datasets:* To divide the source bicycle traffic flow data into distinct clusters, we begin by analyzing the overall average daily traffic patterns for each source sensor. As shown in **Fig.** 2, the morning peak traffic patterns are clearly pronounced across these sensors, while the evening peak appears more dispersed. Additionally, distinct traffic patterns emerge throughout the day. For example, some sensors consistently record high traffic volumes throughout the day without displaying significant peak hours, whereas other sensors show little to no traffic throughout the entire day. Next, we design a 1D convolutional neural network (CNN) to capture the unique traffic patterns of each sensor, which are then used as the initial features for K-means clustering. To determine the optimal number of clusters, we apply K-means clustering across a range of 2 to 9 clusters, calculating the sum of squared errors (SSE) for each scenario and performing a silhouette score analysis. As depicted in **Fig.** 3, the rate of SSE decline significantly slows at 4 clusters, and the silhouette score is also high at this point. This suggests that clustering the traffic patterns into 4 distinct groups is the most effective choice.

Based on the above analysis, we grouped sensors from all source cities into four distinct clusters using the K-means algorithm, as illustrated in **Fig.** 4. Specifically, Cluster0 and Cluster3 exhibit similar peak traffic in the morning but differ in their afternoon peak patterns. In contrast, Cluster1 shows more pronounced morning peak hours compared to Cluster0 and Cluster3. Cluster2, however, stands out with significantly higher peak traffic, and unlike the other clusters, its morning peak begins slightly later.

## *B. Evaluation Metrics*

In this paper, the prediction results of bicycle traffic flow are evaluated by Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Weighted Absolute Percentage Error (WAPE). We use the relative performance degradation of WAPE between source-target city and target-only city to evaluate the negative transfer learning. The formulations to calculate these metrics are shown below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{18}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{19}$$

$$WAPE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n}|y_i|}, \tag{20}$$

$$RP_{WAPE} = \frac{WAPE_{source-target} - WAPE_{target-only}}{WAPE_{target-only}}, \tag{21}$$

where $\hat{y}_i$ and $y_i$ represent the predicted value and ground truth data of the sample $i$, respectively. $n$ is the number of sample values.

## *C. Baseline Models*

To assess the performance of our proposed Multi-TLSTGCN model in predicting bicycle traffic flow, we compare it with several widely recognized traffic prediction models, from simple historical and statistical models (HA, ARIMA) to advanced machine learning and deep learning techniques commonly applied in traffic prediction (SVR, LSTM, STGCN and ASTGCN). Through this comparison, we aim to highlight Multi-TLSTGCN's capability in forecasting bicycle traffic, particularly in situations where the data is scarce or noisy. The baseline models used for comparison are as follows.

- HA: The Historical Average (HA) model predicts traffic flow by taking the average value of historical data.
- SVR: Support Vector Regression (SVR) [48] is a regression algorithm that aims to minimize the discrepancy

between the predicted value and a predefined margin, which is an extension of Support Vector Machines (SVM) for regression tasks.

- LSTM: Long Short-Term Memory (LSTM) [49] is a specialized variant of Recurrent Neural Networks (RNNs) specifically designed to effectively capture and model long-term dependencies in sequential data.
- STGCN: Spatial-Temporal Graph Convolutional Network (STGCN) [50] is proposed to tackle the time series traffic prediction problem by harnessing comprehensive spatial-temporal correlations.
- ASTGNN: The attention-based spatial-temporal graph neural network (ASTGNN) [51] is designed to capture the dynamics of traffic data across both temporal and spatial dimensions.
- TransGTR: TransGTR [52] is a transferable traffic forecasting framework that jointly learns and transfers graph structures and forecasting models across cities. It comprises a node feature network trained with knowledge distillation for city-agnostic feature extraction, a structure generator with temporal decoupled regularization to ensure spatial features share similar distributions across cities, enabling effective knowledge transfer, and a forecasting model.

### D. Experimental Settings

The experiments were conducted using Google Colab, a cloud-based Python environment. The computing environment included a Tesla L4 GPU with a CUDA version of 12.0. The CPU used was an Intel(R) Core(TM) i9-9900KS clocked at 4 GHz. We implemented all the deep learning models using the PyTorch framework in Python. The models were optimized using the Adam optimizer. The hyperparameters for all deep learning models were carefully tuned through a validation set. Specifically, for the Multi-TLSTGCN model, we set the model dimension $d_{model}$ to 64, the number of attention heads $h$ to 8, the convolution kernel size to 3, The learning rate is reduced by a factor of 0.1 from 0.0001 to 0.00001, and batch size is 64. To prevent overfitting and ensure optimal model performance, we implemented an early stopping mechanism during the training process. Specifically, we used an early stopping criterion with a patience value of 20. This means that if the performance of the model on the validation set does not improve by at least 0.000001 for consecutive 20 epochs, the training process is halted.

### E. Experimental Results and Discussion

In this section, we first evaluate the performance of single-city and cluster-based source transfer learning for predicting bicycle traffic flow in a target city. To assess the effectiveness of multi-source transfer learning in enhancing prediction accuracy, we conduct an experiment using the proposed Multi-TLSTGCN model. The performance of this model is compared against widely-used traffic prediction models to demonstrate that multi-source transfer learning is a promising approach for bicycle traffic prediction, particularly when dealing with scarce bicycle traffic data.

TABLE I
HYPERPARAMETERS OF THE MULTI-TLSTGCN

| Hyperparameters | Values |
|---|---|
| batch size | 64 |
| $d_{model}$ | 64 |
| kernel size | 3 |
| initial learning rate | 0.0001 |
| patience value | 20 |
| attention heads | 8 |

*1) Single-Source Transfer Results:* As all datasets are sourced from cities across the Netherlands, traffic patterns such as peak and off-peak hours are generally consistent across these locations, with the exception of Delft. This consistency supports a balanced assessment of the model's effectiveness. Delft, however, presents a unique case due to its educational land use and campus-centered sensor distribution, resulting in multiple traffic peaks that correspond to the academic calendar. These characteristics make Delft's traffic patterns distinct from those of other cities in the dataset. To thoroughly evaluate the robustness of our approach, we include Delft as a target city in our analysis.

*a) Single-city transfer:* To evaluate the effectiveness of knowledge transfer from individual source cities to the target city and to investigate the influence of source-specific features on transfer performance, we first implement single-source transfer learning for each target city independently. This approach involves training models on datasets from individual source cities to capture their respective traffic patterns and subsequently fine-tuning these models using the target city's dataset. The goal is to improve the accuracy of bicycle traffic flow predictions for the target city. As shown in **Table** II, using The Hague, Leiden, and Dordrecht as source cities improves the prediction accuracy for Delft, with Leiden achieving the highest performance 5.88% in WAPE compared to the baseline. In contrast, employing Rotterdam as a source city adversely impacts the accuracy of predictions over a 60-minute horizon with 0.86% degradation on WAPE, indicating a negative transfer effect. This suggests that Rotterdam's traffic patterns diverge significantly from those in Delft, particularly over longer time frames. This discrepancy may be due to Rotterdam's larger and more complex bicycle traffic network, which complicates long-term forecasting.

For the target city Gouda, as indicated in **Table** II, The Hague, Leiden, and Dordrecht contribute to improved prediction accuracy for bicycle traffic flow. Notably, Leiden achieves the highest performance among the source cities, with a 8.26% reduction in WAPE compared to using Gouda's own dataset alone. This improvement may be attributed to the similarity in sensor distribution between Leiden and Gouda, where sensors are predominantly located on ring roads around the city, resulting in comparable traffic patterns. In contrast, using Rotterdam as a source leads to a 2.60% increase in WAPE, which is worse than using Gouda's own dataset. This negative transfer effect suggests that while general traffic patterns may appear similar,

TABLE II
PERFORMANCE OF TRAFFIC PREDICTION USING SINGLE-CITY SOURCE TRANSFER

| Cities | 15 min | | | 30 min | | | 60 min | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | WAPE | MAE | RMSE | WAPE | MAE | RMSE | WAPE |
| **Delft (Target)** | **0.90** | **2.10** | **34.68%** | **0.98** | **2.28** | **37.82%** | **1.09** | **2.56** | **42.03%** |
| The Hague–Delft | 0.87 | 2.04 | 33.56% | 0.95 | 2.23 | 36.35% | 1.04 | 2.44 | 39.91% |
| Rotterdam–Delft | 0.87 | 2.00 | 33.56% | 0.97 | 2.24 | 37.18% | 1.10 | 2.58 | 42.39% |
| Leiden–Delft | 0.87 | 2.10 | 33.36% | 0.94 | 2.27 | 36.15% | 1.03 | 2.48 | 39.56% |
| Dordrecht–Delft | 0.86 | 1.93 | 33.04% | 0.94 | 2.13 | 36.01% | 1.04 | 2.40 | 40.14% |
| **Gouda(Target)** | **17.37** | **39.23** | **21.87%** | **21.16** | **44.99** | **26.65%** | **24.41** | **50.19** | **30.76%** |
| The Hague–Gouda | 17.12 | 35.46 | 21.54% | 20.22 | 40.33 | 25.45% | 23.08 | 45.69 | 29.09% |
| Rotterdam–Gouda | 17.76 | 38.59 | 22.36% | 21.60 | 44.80 | 27.20% | 25.05 | 50.51 | 31.56% |
| Leiden–Gouda | 15.41 | 33.17 | 19.39% | 19.19 | 38.60 | 24.16% | 22.39 | 44.73 | 28.22% |
| Dordrecht–Gouda | 16.53 | 36.49 | 20.81% | 20.63 | 43.06 | 25.97% | 24.16 | 49.15 | 30.45% |

TABLE III
PERFORMANCE OF TRAFFIC PREDICTION USING SINGLE-CLUSTER SOURCE TRANSFER

| Cities | 15 min | | | 30 min | | | 60 min | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | WAPE | MAE | RMSE | WAPE | MAE | RMSE | WAPE |
| **Delft (Target)** | **0.90** | **2.10** | **34.68%** | **0.98** | **2.28** | **37.82%** | **1.09** | **2.56** | **42.03%** |
| Cluster1–Delft | 0.85 | 1.96 | 32.57% | 0.92 | 2.12 | 35.18% | 1.01 | 2.33 | 38.76% |
| Cluster2–Delft | 0.85 | 1.99 | 32.71% | 0.92 | 2.15 | 35.34% | 1.02 | 2.37 | 39.11% |
| Cluster3–Delft | 0.85 | 1.93 | 32.61% | 0.92 | 2.12 | 35.45% | 1.01 | 2.34 | 38.99% |
| Cluster4–Delft | 0.86 | 1.92 | 32.84% | 0.93 | 2.11 | 35.76% | 1.03 | 2.36 | 39.83% |
| **Gouda(Target)** | **17.37** | **39.23** | **21.87%** | **21.16** | **44.99** | **26.65%** | **24.41** | **50.19** | **30.76%** |
| Cluster1–Gouda | 15.83 | 34.30 | 19.92% | 19.54 | 39.81 | 24.60% | 22.70 | 45.62 | 28.61% |
| Cluster2–Gouda | 15.91 | 33.81 | 20.02% | 19.71 | 39.37 | 24.81% | 22.93 | 45.44 | 28.89% |
| Cluster3–Gouda | 15.64 | 34.37 | 19.69% | 19.52 | 39.69 | 24.58% | 22.79 | 45.71 | 28.71% |
| Cluster4–Gouda | 16.17 | 34.77 | 20.35% | 20.23 | 41.20 | 25.47% | 23.61 | 47.31 | 29.75% |

variations in traffic distribution and sensor-specific patterns across cities can hinder effective knowledge transfer. These findings highlight the importance of selecting source cities with compatible sensor layouts and traffic characteristics to ensure successful transfer learning.

*b) Single-cluster transfer:* To mitigate the risk of negative transfer in bicycle traffic prediction, it is crucial to account for the sparse distribution of sensors and the complexity of bicycle road networks across different cities. Unlike motorized traffic, which follows a more structured and hierarchical road system, bicycle networks tend to be more intricate and irregular. Directly transferring spatial-temporal traffic data to the target city may lead to negative transfer due to these structural differences. In this study, we cluster sensors from all source cities to identify traffic patterns in the source domains that closely resemble those in the target city. This approach makes knowledge transfer more relevant and effective, reducing the likelihood of negative transfer and ensuring that the transferred information is beneficial. We then transfer information from each single-source cluster individually to improve traffic predictions in the target city. The results in **Table** III show that each single-source cluster positively contributes to bicycle traffic predictions in the target cities.

*c) Comparison of single-city and single-cluster source transfer:* **Fig.** 5 demonstrates that single-city source transfer learning generally performs worse compared to single-cluster source transfer learning. For example, in the Delft prediction task, the WAPE difference between the best single-city transfer (Dordrecht→Delft, 39.91%) and the best cluster-based transfer (Cluster-based, 38.76%) is 1.15 percentage points, this corresponds to a 2.88% relative improvement, which is meaningful given the inherent noise and variability in traffic data. While certain individual source-target city pairs (e.g., Leiden→Gouda) may exceed cluster-based performance in specific cases, cluster-based transfer demonstrates more consistent average performance and offers reduced risk of

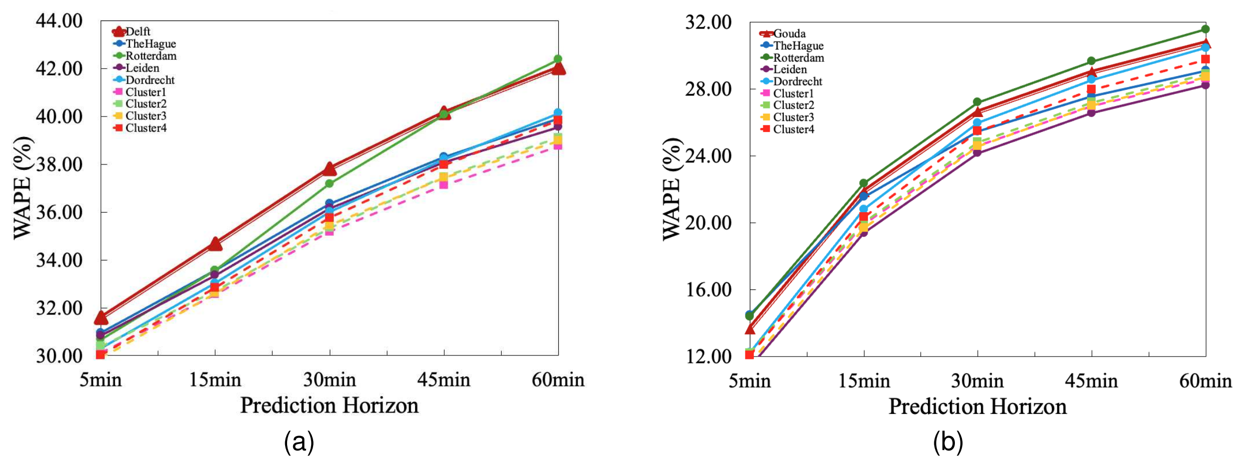



Fig. 5. Performance comparison between single-city source and single-cluster source. (a) WAPE(Delft). (b) WAPE(Gouda).

TABLE IV
PERFORMANCE COMPARISON OF MULTI-STEP BICYCLE TRAFFIC PREDICTION

| Target | Baselines | 15 min | | | 30 min | | | 60 min | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE | RMSE | WAPE | MAE | RMSE | WAPE | MAE | RMSE | WAPE |
| **Delft** | HA | 3.29 | 5.22 | 104.96% | 3.29 | 5.22 | 104.96% | 3.29 | 5.22 | 104.96% |
| | SVR | 1.07 | 2.41 | 41.03% | 1.22 | 2.65 | 47.03% | 1.56 | 3.19 | 60.22% |
| | LSTM | 1.01 | 2.31 | 38.92% | 1.10 | 2.54 | 42.34% | 1.23 | 2.86 | 47.31% |
| | STGCN | 0.93 | 2.08 | 35.66% | 1.00 | 2.31 | 38.51% | 1.10 | 2.61 | 42.37% |
| | ASTGNN | 0.86 | 2.04 | 32.94% | 0.93 | 2.25 | 35.86% | 1.03 | 2.49 | 39.55% |
| | TransGTR | 1.28 | 1.96 | 49.08% | 1.33 | 2.08 | 51.00% | 1.39 | 2.26 | 53.51% |
| | **Multi-TLSTGCN** | **0.82** | **1.85** | **31.56%** | **0.89** | **2.03** | **34.20%** | **0.97** | **2.24** | **37.52%** |
| **Gouda** | HA | 69.83 | 104.51 | 76.91% | 69.82 | 104.50 | 76.94% | 69.81 | 104.49 | 76.99% |
| | SVR | 24.25 | 51.73 | 30.52% | 26.48 | 54.51 | 33.34% | 29.44 | 59.20 | 37.10% |
| | LSTM | 18.15 | 39.75 | 22.84% | 22.77 | 47.35 | 28.67% | 27.29 | 55.43 | 34.39% |
| | STGCN | 18.01 | 37.27 | 22.66% | 20.98 | 42.56 | 26.42% | 23.49 | 47.38 | 29.60% |
| | ASTGNN | 16.38 | 35.53 | 20.61% | 19.87 | 40.18 | 25.01% | 22.99 | 45.73 | 28.97% |
| | TransGTR | 17.28 | 37.11 | 21.74% | 20.56 | 42.39 | 25.89% | 22.83 | 46.32 | 28.77% |
| | **Multi-TLSTGCN** | **15.21** | **32.88** | **19.15%** | **19.17** | **38.60** | **24.14%** | **22.50** | **44.82** | **28.35%** |

negative transfer. This makes it especially advantageous in scenarios where prior knowledge of structural or behavioral similarities between cities is unavailable.

Single-city transfer relies heavily on the alignment between the source and target cities' road networks and sensor distributions. When well-matched, it can yield high accuracy. However, mismatches, such as complex or dissimilar traffic patterns, can introduce noise, leading to negative transfer and reduced predictive performance.

In contrast, single-cluster transfer draws from more focused and homogeneous subsets of traffic patterns. This helps maintain alignment with the target city's dominant characteristics, mitigates the introduction of irrelevant or conflicting information, and supports more stable performance overall. Given the sparse and complex nature of bicycle traffic data, cluster-based transfer learning emerges as a more robust and scalable approach for active mode traffic prediction.

*2) Multi-Source Transfer Learning Results:* In this study, we clustered bicycle data from The Hague, Rotterdam, Leiden, and Dordrecht into different groups to serve as sources supporting predictions in Delft and Gouda. The results, presented in **Table** IV, demonstrate that the proposed Multi-TLSTGCN significantly outperforms other baselines and the single-cluster source model. This superior performance highlights the effectiveness of incorporating knowledge from multiple cities, making it a promising approach for improving bicycle traffic prediction. In comparison, the baseline HA model exhibits poor performance compared to other models. This is because HA relies on predicting traffic states by averaging past observations, which fails to capture the inherent fluctuations and variability of bicycle traffic. SVR and LSTM can effectively

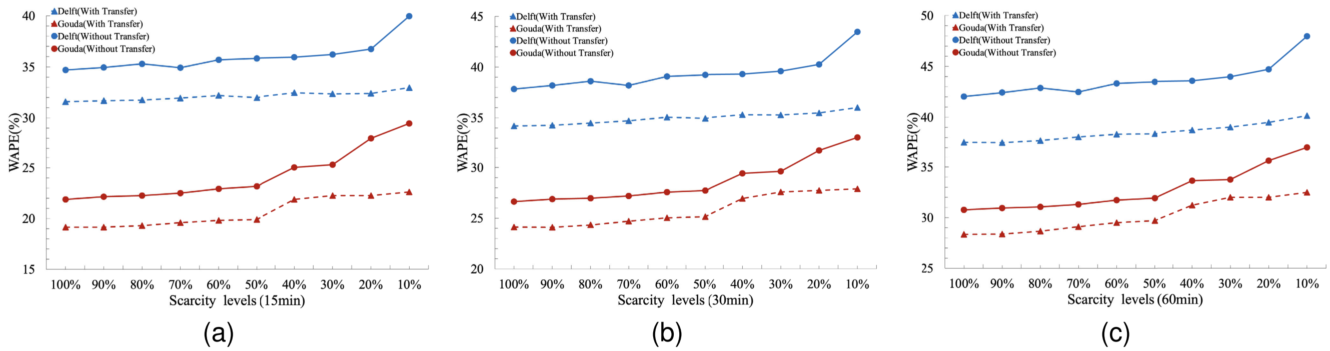



Fig. 6. Performance across varying levels of target data sparsity.

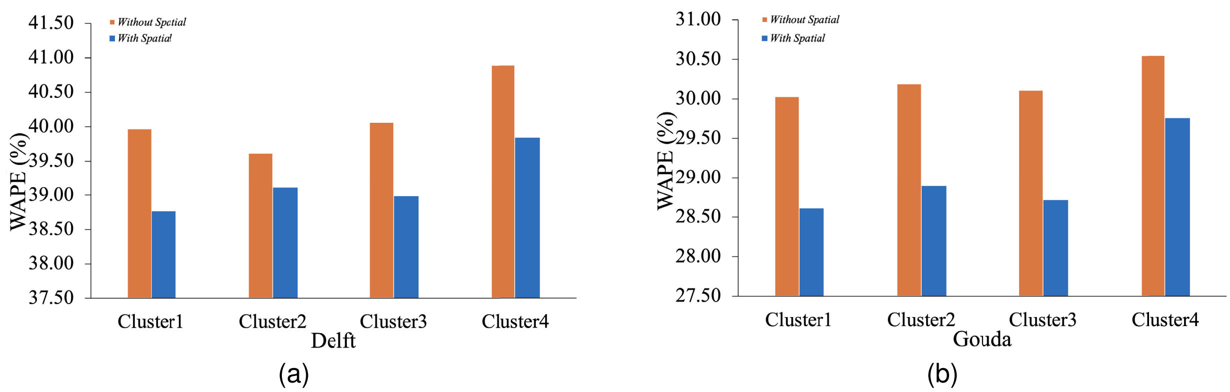



Fig. 7. Virtual spatial correlations ablation study. (a) WAPE(Delft). (b) WAPE(Gouda).

capture both linear and nonlinear patterns, making them well-suited for the complexities of bicycle traffic flow data. However, their focus on temporal features alone limits their performance, as spatial correlations among sensors in the bicycle road network are also crucial. STGCN and ASTGNN address this by using graph convolutional neural networks to capture spatial correlations of bicycle traffic, thereby improving prediction accuracy. Nevertheless, these deep learning models require large datasets to capture features accurately, posing a significant challenge for bicycle traffic prediction due to the inherent data scarcity. We also compared the performance of Multi-TLSTGCN with the state-of-the-art transfer learning method TransGTR. Although TransGTR incorporates a node feature extractor and a structure generator to capture city-agnostic features, it still struggles with bicycle traffic prediction—particularly when Delft is the target city, where bicycle flow is strongly influenced by academic schedules, which are difficult to generalize across cities.

*3) Multi-Source Transfer Learning With Different Levels of Target Data Scarcity:* To assess the performance of the proposed Multi-TLSTGCN model in real-world situations with varying levels of bicycle flow data availability, we designed a series of diverse scenarios for traffic prediction of target cities, each with different levels of bicycle target dataset scarcity. By reducing the available target training dataset from 90% down to 0%, we aim to evaluate the model's robustness and generalization across different level of data-scarce environments.

For each target city, we conduct two scenarios: the first involves training the model under varying levels of data scarcity in the target city without leveraging knowledge from source models, while the second trains the model using Multi-TLSTGCN, benefiting from the transfer knowledge acquired from source models.

As shown in **Fig.** 6, the scenarios without transfer knowledge exhibit increasing error rates as the amount of training data decreases. In Gouda, the error begins to rise dramatically when only 50% of the training dataset is included. For Delft, the error steadily increases from the full 100% training dataset down to 20%; however, once the training dataset falls below 10%, the error escalates sharply. In contrast, the scenarios utilizing Multi-TLSTGCN, which incorporates transfer knowledge, show a more gradual increase in error. For Gouda, a noticeable increase occurs when only 50% of the training dataset is utilized, but the error stabilizes when the dataset is reduced to 30%, in comparison to scenarios without transfer knowledge. For Delft, throughout the various levels of data scarcity, the error only slightly increases as the training dataset decreases. The WAPE values differ across cities Gouda and Delft, primarily due to variations in traffic flow stability, sensor coverage, and urban layout.

These findings suggest that Multi-TLSTGCN effectively transfers valuable information to the target domain, enabling accurate traffic predictions even when target data is limited. This underscores the model's robustness in supporting bicycle traffic prediction under data-scarce conditions.

Our assessment further indicates that the proposed model can be applied across a variety of real-world scenarios. As shown in Figure X, the model maintains strong performance,

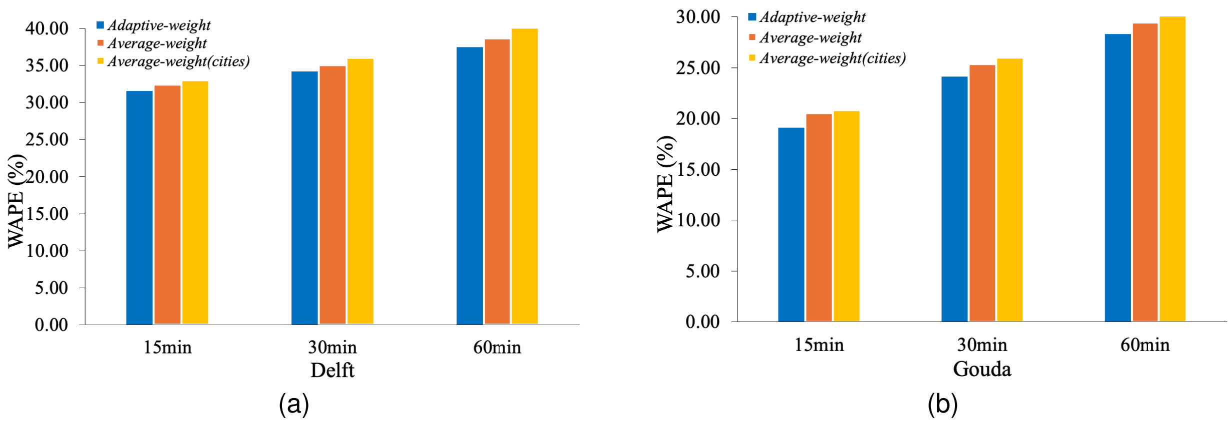



Fig. 8. Multi-source transfer learning approach ablation study. (a) WAPE(Delft). (b) WAPE(Gouda).

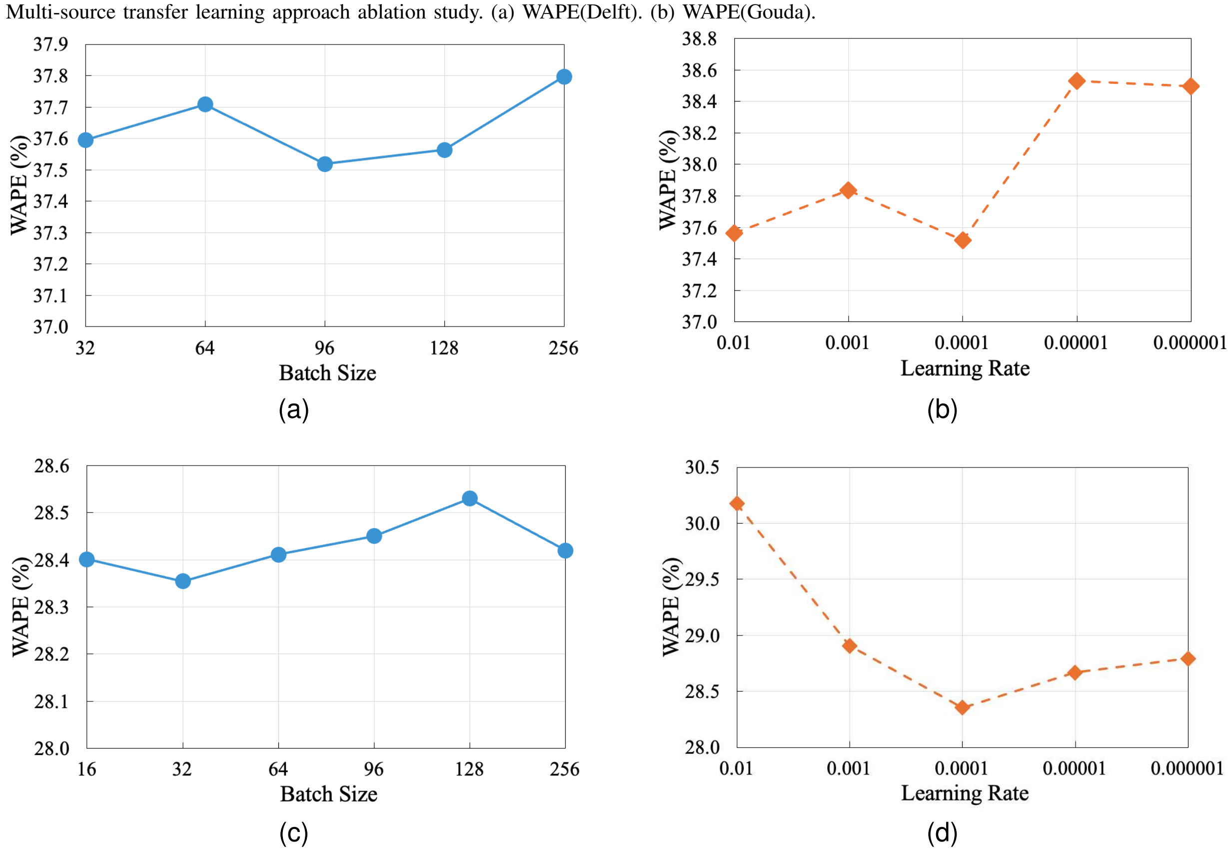



Fig. 9. Hyperparameter analysis. (a) Batch size (Delft). (b) Learning rate (Delft). (c) Batch size (Gouda). (d) Learning rate (Gouda).

with only a slight degradation as the size of the training dataset decreases. This suggests that the approach is robust even under limited data conditions.

For practical deployment, the amount of training data required will depend on two key factors: (i) the availability of local data and (ii) the level of predictive accuracy desired. These two aspects must be balanced in real-world applications. In data-rich environments, larger datasets will naturally support higher accuracy, while in data-scarce contexts, the model can still be operationalized with smaller datasets, but with somewhat reduced precision.

### *F. Ablation Study*

To evaluate the impact of virtual spatial correlations from each cluster-based transfer model on traffic prediction in the target domain and to assess the effectiveness of the adaptive transfer learning approach in multi-source transfer learning, we conducted ablation experiments. These experiments were carried out using the same bicycle datasets previously described. Finally, we perform a comprehensive hyperparameter analysis to demonstrate the rationale behind our hyperparameter selection process.

*1) The Impact of Virtual Spatial Correlations:* Given the complexity of bicycle road networks and their variability across different cities, we grouped sensors from all cities into distinct clusters based on traffic patterns, disregarding geographical spatial correlations. However, even within the same cluster, traffic patterns among sensors can vary. To account for these differences and capture potential spatial correlations within each cluster, we established a virtual spatial correlation among sensors. To evaluate whether this virtual correlation effectively captures useful spatial information from source data to enhance traffic prediction in target cities, we conducted ablation experiments.

As shown in **Fig.** 7, we evaluated the impact on target city. Generally, source models that incorporated virtual spatial correlations outperformed those without them in enhancing traffic predictions for the target city. However, in Delft, the inclusion of virtual spatial dependencies did not significantly improve performance. This can be attributed to Delft's limited sensor network coverage, where traffic patterns across different locations are relatively similar, leading the model to rely more on temporal correlations rather than spatial ones. In contrast, Gouda's sensors are spread across a larger, more complex network, and the results indicate that leveraging virtual spatial information significantly improves traffic prediction accuracy in Gouda.

*2) The Effectiveness of the Multi-Source Adaptive Transfer Learning Approach:* To effectively transfer knowledge from multiple data sources, we propose a multi-source adaptive transfer learning approach that combines prediction outputs from various source models using target domain data. As shown in **Fig.** 8, our method significantly outperforms a basic aggregation approach for all cluster-based models and city-based models, which merely averages the parameters of all source models. These results underscore the strength of our adaptive transfer learning strategy in harnessing diverse knowledge sources to enhance prediction accuracy for the target city.

*3) Hyperparameter Analysis:* Hyperparameters play a crucial role in controlling the training dynamics of deep learning models, including the proposed Multi-TLSTGCN. As detailed in the details in Section V-D, all hyperparameters were carefully tuned using a validation set to ensure optimal model performance and generalization. We analyze the model's sensitivity using two key hyperparameters as examples: batch size and learning rate.

The impact of batch size and learning rate on the WAPE is visualized in **Fig.** 9. The results demonstrate that the optimal configuration for the source city of Delft is achieved with a batch size of 96 and a learning rate of 0.0001. In contrast, for the source city of Gouda, the model achieves its best performance with a batch size of 32 and the same learning rate of 0.0001.

## VI. Conclusion

Given the scarcity of bicycle data and the significant requirements for accurate traffic prediction, we propose the Multi-TLSTGCN model, which leverages datasets from other cities through transfer learning to enhance prediction accuracy in the target city. The proposed process begins by clustering sensors from different source cities into distinct groups based on their traffic patterns. Next, we reconstruct the spatial correlations within each group based on the similarity of traffic patterns among sensors. These pre-trained source group models are then fine-tuned using the target city's dataset. Finally, adaptive transfer learning approach is used to aggregate the predictions from these multiple source models, optimizing the traffic predictions for the target city.

In this study, we conducted three key experiments to evaluate the feasibility and robustness of our model in predicting bicycle traffic under data scarcity conditions. First, we perform single-city and single-cluster experiments to explore how traffic knowledge transfer can avoid negative transfer. Second, we implement a multi-source transfer learning experiment using Multi-TLSTGCN for traffic prediction, assessing the model's performance under data scarcity. The results demonstrate that our approach outperforms baselines trained solely on the target dataset. We evaluated the robustness of Multi-TLSTGCN by training the model under varying levels of data scarcity. Comparing this to the same model without transfer learning, we observe that Multi-TLSTGCN consistently outperforms its counterpart. As the amount of target training data decreases, the accuracy of the non-transfer learning model declines significantly, whereas Multi-TLSTGCN maintains its predictive performance, demonstrating its effectiveness in data-scarce environments.

Overall, this study demonstrates the potential for achieving accurate bicycle traffic prediction in cities with limited data collection devices or insufficient traffic data. While the Multi-TLSTGCN framework is designed to be data-driven and flexible, its generalizability to cities with less developed cycling infrastructure may be influenced by variations in sensor availability, traffic dynamics, and urban topology. Nevertheless, the clustering-based transfer learning strategy and the virtual spatial correlation module are model-agnostic, offering adaptability to datasets from cities outside the Netherlands. Moreover, the proposed approach is designed for typical operating scenarios and is not specifically tailored to conditions such as extreme weather, irregular events, unique urban topologies, or distinct policy interventions. Exploring these challenging scenarios could be an important work for future research. In addition, this paper does not incorporate time-related factors such as day of the week, seasonality, or weather conditions. These factors could be integrated in real-world deployments and have the potential to further improve prediction accuracy. For deployment purposes, the real-time computational cost is primarily associated with the target model, which is initialized using the pretrained source model. Since the pretrained model can be generated offline and reused across tasks, the runtime cost during deployment remains modest and practical for real-world applications. In future work, we plan to evaluate the model on data from cities with diverse cycling cultures and infrastructure characteristics to further validate its robustness and applicability. Additionally, we aim to enhance the framework by developing synthetic data generation techniques tailored for regions with extremely sparse historical records. These enhancements will expand the model's utility, offering effective solutions for traffic prediction in a wide range of urban settings challenged by data scarcity and variability.

## Acknowledgment

The Delft dataset used in this study is sourced from the Outdoor Mobility Digital Twin (OMDt) project at TU

Delft (https://micd.tudelftcampus.nl/projects/outdoor-mobility-digital-twin/). The authors gratefully acknowledge their support in providing this data for this research.

## REFERENCES

[1] Y. Jia, J. Wu, and M. Xu, "Traffic flow prediction with rainfall impact using a deep learning method," *J. Adv. Transp.*, vol. 2017, pp. 1–10, Aug. 2017.

[2] M. C. Poelman, "Predictive traffic signal control under uncertainty: Analyzing and reducing the impact of prediction errors," Ph.D. thesis, Delft Univ. Technol., Delft, The Netherlands, Oct. 2024, pp. 1–14.

[3] S. Yu, Y. Li, G. Sheng, and J. Lv, "Research on short-term traffic flow forecasting based on KNN and discrete event simulation," in *Proc. 15th Int. Conf. Adv. Data Mining Appl.*, Dalian, China, Nov. 2019, pp. 853–862.

[4] B. Zhao, K. Kumar, G. Casey, and K. Soga, "Agent-based model (ABM) for city-scale traffic simulation: A case study on San Francisco," in *Proc. Int. Conf. Smart Infrastruct. Construct. (ICSIC)*, Jan. 2019, pp. 203–212.

[5] I. Alam, D. M. Farid, and R. J. Rossetti, "The prediction of traffic flow with regression analysis," in *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018*, vol. 2. Singapore: Springer, 2019, pp. 661–671.

[6] C. Xu, Z. Li, and W. Wang, "Short-term traffic flow prediction using a methodology based on autoregressive integrated moving average and genetic programming," *Transport*, vol. 31, no. 3, pp. 343–358, Sep. 2016.

[7] S. V. Kumar, "Traffic flow prediction using Kalman filtering technique," *Proc. Eng.*, vol. 187, pp. 582–587, May 2017.

[8] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul. 1985.

[9] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Apr. 1998.

[10] J. Zou, Y. Han, and S. S. So, "Overview of artificial neural networks," in *Artificial Neural Networks: Methods and Applications*. Totowa, NJ, USA: Humana Press, 2009, pp. 14–22.

[11] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design Appl.*, vol. 5, nos. 64–67, p. 2, 2001.

[12] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *Proc. ESANN*, 2015, p. 89.

[13] S. Hao, D.-H. Lee, and D. Zhao, "Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system," *Transp. Res. C, Emerg. Technol.*, vol. 107, pp. 287–300, Oct. 2019.

[14] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2008.

[15] J. Gu et al., "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[16] M. Ge, H. Bangui, and B. Buhnova, "Big data for Internet of Things: A survey," *Future Gener. Comput. Syst.*, vol. 87, pp. 601–614, Oct. 2018.

[17] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big data*, vol. 3, pp. 1–40, May 2016.

[18] J. Li, F. Guo, A. Sivakumar, Y. Dong, and R. Krishnan, "Transferability improvement in short-term traffic prediction using stacked LSTM network," *Transp. Res. C, Emerg. Technol.*, vol. 124, Mar. 2021, Art. no. 102977.

[19] Z. Fang, D. Wu, L. Pan, L. Chen, and Y. Gao, "When transfer learning meets cross-city urban flow prediction: Spatio-temporal adaptation matters," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 2030–2036.

[20] S. Wang, H. Miao, J. Li, and J. Cao, "Spatio-temporal knowledge transfer for urban crowd flow prediction via deep attentive adaptation networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4695–4705, May 2022.

[21] Y. Huang, X. Song, Y. Zhu, S. Zhang, and J. J. Q. Yu, "Traffic prediction with transfer learning: A mutual information-based approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8236–8252, Aug. 2023.

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

[23] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal ARIMA model with limited input data," *Eur. Transp. Res. Rev.*, vol. 7, no. 3, pp. 1–9, Sep. 2015.

[24] D.-W. Xu, Y.-D. Wang, L.-M. Jia, Y. Qin, and H.-H. Dong, "Real-time road traffic state prediction based on ARIMA and Kalman filter," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 2, pp. 287–302, Feb. 2017.

[25] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[26] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi, and B. Yin, "Deep learning on traffic prediction: Methods, analysis, and future directions," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4927–4943, Jun. 2022.

[27] Y. Bao et al., "Spatial–temporal complex graph convolution network for traffic flow prediction," *Eng. Appl. Artif. Intell.*, vol. 121, May 2023, Art. no. 106044.

[28] X. Shi, H. Qi, Y. Shen, G. Wu, and B. Yin, "A spatial–temporal attention approach for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4909–4918, Aug. 2021.

[29] A. Naghib, N. Jafari Navimipour, M. Hosseinzadeh, and A. Sharifi, "A comprehensive and systematic literature review on the big data management techniques in the Internet of Things," *Wireless Netw.*, vol. 29, no. 3, pp. 1085–1144, Apr. 2023.

[30] Y. Xu, X. Cai, E. Wang, W. Liu, Y. Yang, and F. Yang, "Dynamic traffic correlations based spatio-temporal graph convolutional network for urban traffic prediction," *Inf. Sci.*, vol. 621, pp. 580–595, Apr. 2023.

[31] E. Heinen, K. Maat, and B. van Wee, "The effect of work-related factors on the bicycle commute mode choice in The Netherlands," *Transportation*, vol. 40, no. 1, pp. 23–43, Jan. 2013.

[32] M. El Esawey, A. I. Mosa, and K. Nasr, "Estimation of daily bicycle traffic volumes using sparse data," *Comput., Environ. Urban Syst.*, vol. 54, pp. 195–203, Nov. 2015.

[33] K. Kim, "Investigation of modal integration of bike-sharing and public transit in Seoul for the holders of 365-day passes," *J. Transp. Geography*, vol. 106, Jan. 2023, Art. no. 103518.

[34] D. Pan, Y. Zhao, J. Liu, X. Jia, and M. Wang, "Spatio-temporal neural structural causal models for bike flow prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 4242–4249.

[35] Z. Yao, S. Xia, Y. Li, G. Wu, and L. Zuo, "Transfer learning with spatial–temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8592–8605, Aug. 2023.

[36] X. Ouyang, Y. Yang, W. Zhou, Y. Zhang, H. Wang, and W. Huang, "CityTrans: Domain-adversarial training with knowledge transfer for spatio-temporal prediction across cities," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 62–76, Jan. 2024.

[37] J. Li, C. Liao, S. Hu, X. Chen, and D.-H. Lee, "Physics-guided multi-source transfer learning for network-scale traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 17533–17546, Nov. 2024.

[38] Y. Jin, K. Chen, and Q. Yang, "Selective cross-city transfer learning for traffic prediction via source city region re-weighting," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 731–741.

[39] J. Mo and Z. Gong, "Cross-city multi-granular adaptive transfer learning for traffic flow prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 11, pp. 11246–11258, Nov. 2022.

[40] N. A. Ojekudo and N. P. Akpan, "Anapplication of Dijkstra's algorithm to shortest route problem," *IOSR J. Math.*, vol. 13, no. 3, p. 14, 2017.

[41] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[42] S. Patro, "Normalization: A preprocessing stage," 2015, *arXiv:1503.06462*.

[43] K. Krishna and M. N. Murty, "Genetic K-means algorithm," *IEEE Trans. Syst., Man, Cybern., B Cybern.*, vol. 29, no. 3, pp. 433–439, Mar. 1999.

[44] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and K-means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 17–24, 2014.

[45] X. Ta, Z. Liu, X. Hu, L. Yu, L. Sun, and B. Du, "Adaptive spatio-temporal graph neural network for traffic forecasting," *Knowl.-Based Syst.*, vol. 242, Apr. 2022, Art. no. 108199.

[46] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[47] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

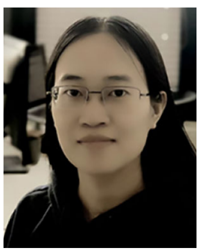
[48] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, pp. 155–161.

[49] A. Graves and A. Graves, "Long short-term memory," in *Supervised Sequence Labelling With Recurrent Neural Networks*. Berlin, Germany: Springer, 2012, pp. 37–45.

[50] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.

[51] S. Guo, Y. Lin, H. Wan, X. Li, and G. Cong, "Learning dynamics and heterogeneity of spatial–temporal graph data for traffic forecasting," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 11, pp. 5415–5428, Nov. 2022.

[52] Y. Jin, K. Chen, and Q. Yang, "Transferable graph structure learning for graph-based traffic forecasting across cities," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2023, pp. 1032–1043.



**Xiamei Wen** received the B.Sc. degree in transportation engineering from Chang'an University, Xi'an, China, in 2017, and the M.Sc. degree in transportation engineering from Wuhan University of Technology, Wuhan, China, in 2021. She is currently pursuing the Ph.D. degree with the Department of Transport and Planning, Delft University of Technology. During the M.Sc. degree, she has conducted research on traffic behavior at stop-controlled intersections using video-based trajectory data. At Delft University of Technology, her research is focused on developing a spatial-temporal graph neural network traffic prediction model based on federated learning. The research aims to obtain high prediction accuracy while ensuring data security.

**Megha Khosla** received the Ph.D. degree from the Algorithms and the Complexity Group, Max Planck Institute for Informatics (MPII), Saarbrücken, Germany. Prior to her being a Faculty Member with TUD, she was a Senior Researcher with the L3S Research Centre and Leibniz University Hannover. She has also managed several collaborative projects both in academia and in the industry, where she spent a three-year stint after the Ph.D. She is currently an Assistant Professor with the Multimedia Computing Group, Intelligent Systems Department. Her research interests include machine learning on graph structured data. Specifically, she develops algorithms to enable effective, interpretable, and privacy-preserving learning on graphs.

**Serge Hoogendoorn** received the Ph.D. degree from Delft University of Technology in 1999. He is currently (one of the four) Distinguished Professor of smart urban mobility with Delft University of Technology. He has a part-time appointment at Monash University and is an Honorary Professor with the School of Transportation, Southeast University, China. His current research evolves around smart urban mobility, with focal areas such as theory, modeling, and simulation of traffic and transportation networks, including cars, pedestrian, cyclists and novel public transport services; development of methods for integrated management of these networks (regional network management, crowd and bicycle management; public transport operations); impact of uncertainty of travel behavior and network operations; impact of ICT on network flow operations, robustness and resilience, and urban data and its applications. In all these topics, his work considers both recurrent and non-recurrent (emergency) situations.