i-CLTP: Integrated contrastive learning with transformer framework for traffic state prediction and network-wide analysis

Citation for the original published paper (version of record):
Jia, R., Gao, K., Liu, Y. et al (2025). i-CLTP: Integrated contrastive learning with transformer framework for traffic state prediction
and network-wide analysis. Transportation Research, Part C: Emerging Technologies, 171.
http://dx.doi.org/10.1016/j.trc.2024.104979

Transportation Research Part C 171 (2025) 104979

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

i-CLTP: Integrated contrastive learning with transformer framework for traffic state prediction and network-wide analysis☆

Ruo Jia [a], Kun Gao [a,*], Yang Liu [a], Bo Yu [b,*], Xiaolei Ma [c], Zhenliang Ma [d]

[a] Department of Architecture and Civil Engineering, Chalmers University of Technology, Goteburg SE-412 96, Sweden
[b] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, College of Transportation Engineering, Tongji University, Shanghai, 201804, China
[c] School of Transportation Science and Engineering, Beihang University, Beijing 100191, China
[d] Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm, 10044, Sweden

ARTICLE INFO

Keywords:
Traffic state prediction
Contrastive learning
Transformer
Soft clustering
Fundamental diagram

ABSTRACT

Traffic state predictions are critical for the traffic management and control of transport systems. This study introduces an innovative contrastive learning framework coupled with a transformer architecture for spatiotemporal traffic state prediction, designed to capture the spatio-temporal heterogeneity inherent in traffic. The transformer structure functions as the upper level of the prediction framework to minimize the prediction errors between the input and predicted output. Based on the self-supervised contrastive learning, the lower level in the framework is proposed to discern the spatio-temporal heterogeneity and embed the latent characteristic of traffic flow by regenerating the augmentation features. Then, a soft clustering problem is applied between the upper level and lower level to category the types of traffic flow characteristics by minimizing the joint loss across each cluster. Subsequently, the proposed model is evaluated through a real-world highway traffic flow dataset for bench marking against several latest existing models. The experimental results affirm that the proposed model considerably enhances traffic state prediction accuracy. In terms of precision metrics, the model records a Mean Absolute Error of 13.31 and a Mean Absolute Percentage Error of 7.85%, reflecting marked improvements of 2.0% and 14.5% respectively over the latest and most competitive baseline model. Furthermore, the analysis reveals that capacity of the proposed method to learn the cluster patterns of spatio-temporal traffic dynamics reflected by calibrated fundamental diagrams.

1. Introduction

Traffic state prediction has consistently been a pivotal aspect of planning, operations, management, and control in the context of Intelligent Transportation Systems (ITS). Traditionally, the acquisition of traffic state information relies on fixed and infrastructure-based sensors, which are then utilized to facilitate direct applications of traffic state prediction. Recently, the advent of various data collection methodologies coupled with advancements in artificial intelligence has catalyzed a proliferation of applications that innovatively address a multitude of challenges associated with traffic state prediction. These applications, exemplifying the integration of Advanced Traveler Information Systems and Advanced Traffic Management Systems, predominantly provide travelers with real-time updates on traffic conditions, thereby enhancing user-oriented and intelligent services within the domain of ITS.

☆ This article belongs to the Virtual Special Issue on "VSI: Data Analytics and ML".

* Corresponding author.
E-mail addresses: gkun@chalmers.se (K. Gao), boyu@tongji.edu.cn (B. Yu).

https://doi.org/10.1016/j.trc.2024.104979
Received 15 July 2024; Received in revised form 30 October 2024; Accepted 14 December 2024
Available online 23 December 2024

Traffic state refers to the various conditions or states of traffic flow at the road segment level or road network level, quantified by metrics such as the levels of congestion, speed, density, and traffic flow volume. Since last century, many studies in the field of traffic state estimation and prediction have primarily focused on the state at road segment level (Zhan et al., 2020), where model-based techniques were widely used, such as auto-regressive integrated moving average (Williams and Hoel, 2003), Kalman Filter models (Guo et al., 2014), and Hidden Markov Models (Qi and Ishak, 2014). The increase of multi-data sources and traffic big data have led to the emergence of data-driven approaches, and deep learning models have emerged as the popularity in traffic state prediction (Lv et al., 2014; Yin et al., 2021; Liu et al., 2022; Zhong et al., 2023). Kumar and Raubal (2021) claimed that machine learning models could achieve high levels of predictive accuracy in forecasting traffic states. Despite the rapid advancements in deep learning that have notably enhanced predictive capabilities, most current approaches still struggle to capture the nonlinear spatio-temporal characteristics inherent in traffic state in a network-wide level analysis. Several studies have attempted to tackle the complexities of network-wide traffic states by transforming them into graph representations that analyze adjacency and incorporating traffic flow propagation into spatio-temporal features using deep learning architectures (Yu et al., 2017b; Cui et al., 2020; Shi et al., 2020). However, these efforts primarily utilize data-driven models as the core predictive mechanism (Li et al., 2020; Yu et al., 2017a; Lan et al., 2022). While traffic flow states are effectively represented through model-based approaches, few studies have explored the potential of integrating these model-based strategies into data-driven deep learning frameworks. The integration of data-driven and model-based approaches in traffic state prediction is complicated by several inherent challenges. Data-driven models often struggle with issues such as data sparsity and imbalance, while model-based approaches grapple with inherent uncertainties. Additionally, simply using graph techniques in deep learning, for instance, graph neural network or graph convolutional network for spatial representation can further amplify the uncertainties tied to an already skewed traffic state distribution across various levels of spatiotemporal detail (Kim et al., 2024). These complexities underscore the need for more holistic and integrative approaches that can effectively tackle these multidimensional challenges in the field.

Data augmentation approaches offer the possibility of integrating data-driven and model-based methods within the same framework, where data is augmented by the calibrated model improving the representation of the physical world and spatiotemporal data. Zhou et al. (2020) introduced a data augmentation method that leverages prior knowledge to address short-term traffic forecasting. Similarly, Wang et al. (2021, 2024) and Liu et al. (2022b) have designed loss functions that address problems inherent to physical models to capture geographical spatio-temporal correlations and tackle data imbalance. However, the challenge of obtaining a robust spatio-temporal representation of sparse network-wide traffic state at the model design stage remains unresolved and warrants further exploration.

Besides, in traffic state prediction, the spatial and temporal traffic states are very complicated with high dimension features. The tensor decomposition based model (Xie et al., 2023; Zhu et al., 2023) is a representative one that used neighborhood information as graph regularizer to perform traffic state estimation. The tensor decomposition based models always organize the traffic state data as a third-order or higher order tensor (e.g., road segment × day × time interval) which lacks of uncertainty representation, and the calculation of these high-dimensional tensors requires high memory and computing power. Inspired by the idea from urban computing (Zhang et al., 2017), many studies combine the spatial and temporal factors into the same feature by using a spatio-temporal encoder which provides a way for spatial–temporal traffic states analysis (Yu et al., 2017a; Cui et al., 2020; Chen et al., 2022). These traffic prediction methods model the temporal dynamics with a shared parameter space for all time periods, which can hardly precisely preserve the temporal heterogeneity in the latent embedding space. However, this strategy does not consider the consistency in spatio-temporal dimensions. In fact, the difference in temporal and spatial dimensions may be discussed separately. For example, the propagation of congestion in spatiotemporal graphs shows distinct patterns of spreading across different road segments and times, yet this aspect is seldom discussed.

Motivated by the above gaps, this study aims to explore a novel traffic state prediction approach, addressing the heterogeneity of data and leveraging the coherence of traffic flow model, to enhance the efficacy of spatiotemporal traffic state prediction. To achieve this goal, there are three principle tasks: (1) devising an appropriate model for traffic state prediction that accounts for the non-linear dynamics and intrinsic interrelations within the traffic data; (2) incorporating spatial information and network connectivity into the model with different categories to augment system performance; (3) representing the spatio-temporal feature of traffic state efficiently, and jointly considering the spatio-temporal heterogeneity to improve prediction precision. To fulfill these tasks, we introduce a new framework termed the contrastive learning (CL) with transformer model with threefold principal contributions:

(1) We have developed a data augmentation approach that facilitates the integration of node level and graph level heterogeneity patterns of traffic states. It integrates prior physical information at both node and graph levels through data augmentation, utilizing historical traffic volumes and time occupancy data from network-wide data. This method enhances the granularity and precision of traffic prediction by incorporating real-world traffic dynamics into the modeling process, and the data augmentation techniques is leveraged to solve the complexity of high dimension feature in traffic state data.

(2) A contrastive learning approach has been designed to optimize spatial and temporal losses holistically, thereby enhancing prediction precision. This method separately analyzes spatial and temporal features to identify similarities and heterogeneity, and then contrasts them together to minimize the combined impact on the predictive performance of model. This approach classifies traffic flow states based on the principles outlined in the traffic flow fundamental diagram. This dual focus on spatial and temporal dimensions enables a more comprehensive understanding of traffic state patterns, facilitating more accurate traffic state predictions.

(3) An comprehensive analysis is conducted to validate the effectiveness of our proposed methodology in distinguishing the spatiotemporal differences in traffic state dynamics through soft clustering based on contrastive learning. Additionally, we propose a comparison with the best state-of-the-art methods. The results demonstrate that the integration of data augmentation and contrastive

learning significantly enhances the interpretability of our model, besides contributing to improvements in prediction accuracy. This highlights the robust capability of our approach to capture and analyze complex spatio-temporal traffic patterns effectively.

The structure of following sections is organized as follows. Section 2 reviews related work in the field. Section 3 describes the methodology employed in this study. Section 4 discusses data description. Section 5 delves into deeper explorations of the performances of the proposed method. Finally, Section 6 concludes the study and outlines directions for future research.

2. Literature review

2.1. Model-based methods

As a fundamental task in transport management and control, considerable research has been dedicated to proposing various models for traffic state prediction. Hamed et al. (1995) developed a time-series model to predict future traffic flow on urban network, utilizing the Box-Jenkins method to determine whether the time series was stationary/seasonality or not. Lee and Fambro (1999) implemented the subset Auto-Regressive Integrated Moving Average (ARIMA) model for short-term freeway traffic volume forecasting. Similarly, Williams and Hoel (2003) employed seasonal ARIMA processes to model uni-variate traffic condition data streams. The findings suggested that while heuristic forecast generation methods significantly enhanced the performance of non-parametric regression, they did not surpass the efficacy of seasonal ARIMA models. Additionally, it was observed that traffic condition data tends to exhibit stochastic rather than chaotic characteristics. More recently, Shahriari et al. (2020) introduced an ensemble ARIMA model that integrated bootstrap techniques with the traditional parametric ARIMA framework to enhance prediction accuracy while adhering to theoretical principles. The proposed methodology involved generating a collection of ARIMA models based on random subsamples of the data, and their findings suggested that this ensemble strategy yielded significant improvements in the precision of predictions.

The Kalman filter method was first developed for road traffic volume estimation by Okutani and Stephanedes (1984) due to its efficiency and robustness in recursive scenarios with noisy data. Xie et al. (2007) explored the integration of Wavelet decomposition with the Kalman filter for traffic speed prediction. The result demonstrated that the consistently surpassed the basic Kalman filter model in accuracy and stability for traffic speed prediction. Kwon and Murphy (2000) employed coupled hidden Markov models to model and predict traffic speeds on freeways, distinguishing between two traffic states—congestion and free flow—based on average speed. Inspired by coupled hidden Markov model, Qi and Ishak (2014) introduced a method that characterizes traffic states within a two-dimensional framework, utilizing both first-order (mean) and second-order (contrast) statistical analyses of speed data. Their approach facilitates the modeling of freeway traffic dynamics through state transition probabilities, enabling model to deduce the most probable sequence of traffic states from a series of speed observations. These model-based methods are generally designed to integrate the periodic characteristics of traffic states with recursive representation and learning techniques, thereby enhancing both accuracy and robustness. However, although the extended methods have been proven to be quite promising in inferring periodicity, unfortunately, they continue to exhibit limitations in addressing the nonlinear dynamics inherent in time series data.

2.2. Sequence and graph machine learning

The increase of multi-data sources and traffic big data have led to the emergence of data-driven approaches, and deep learning models have emerged as the popularity in traffic state prediction (Yin et al., 2021). Kumar and Raubal (2021) claimed that machine learning models could achieve high levels of predictive accuracy in forecasting traffic states. Lv et al. (2014) utilized a Deep Belief Network for traffic prediction, which was one of the pioneering application of Deep Neural Networks (DNNs) for traffic state prediction. Recurrent neural networks (RNNs) and long short-term memory (LSTM) models have been utilized to address the temporal dependencies of traffic prediction (Yu et al., 2017b; Cui et al., 2020). Nevertheless, the majority of existing frameworks predominantly rely on stacking naive Long Short-Term Memory (LSTM) units in a many-to-many structure for sequential modeling. While LSTMs are adept at capturing temporal dependencies, this approach of structuring the model by merely layering multiple LSTM units in a many-to-many configuration presents several significant limitations (Cui et al., 2020; Li et al., 2020). For instance, within such a structure, the length of the target sequence is constrained to be equal to or shorter than the input sequence. This limitation severely restricts the flexibility and generalization capabilities of the model, particularly when the target and input sequences are of different lengths. Furthermore, the conventional many-to-many structure does not process the entire input sequence when generating intermediate outputs (outputs before the final step). This results in limitations and lacks rationality in several multiple-step-ahead prediction tasks, especially when employing the typically default unidirectional LSTM configuration. Afterwards, Li et al. (2020) introduced a novel architecture combining Graph Convolutional Neural Network (GCN), Gated Recurrent Unit (GRU), and Fully Connected Neural Network for traffic state prediction. This approach leverages the integration of diverse traffic data features, achieving enhanced performance through the synergistic fusion of these models. Furthermore, among LSTM, GRU or other optimized variants of time series forecasting techniques, their performance may still degrade to some extent as the input sequence lengthens. However, the development of the Sequence-to-Sequence encoder–decoder architecture has had profound implications for sequential modeling tasks in recent years. By offering a more flexible and scalable framework and the attention mechanism addressing the bottleneck faced by simple Seq2seq models in capturing long-range dependencies, this architecture has not only found extensive application in traditional deep learning tasks but has also garnered increasing attention (Liu et al., 2019). Furthermore, the incorporation of a multihead attention mechanism and stacked layers enables the Transformer to learn dynamic and hierarchical

features in sequential data. This method offers a promising solution to overcome the limitations posed by the predefined adjacency matrix, as highlighted in Yan et al. (2021).

Leveraging road network structures, various studies have evolved the conventional CNN and RNN frameworks into graph-based counterparts for more accurate traffic state prediction, including graph convolution GRU (Yu et al., 2017a; Li et al., 2017) and graph attention (Zhang et al., 2018). These methodologies broaden the scope of traffic prediction from the straightforward Euclidean spaces to the more complex and non-Euclidean configurations of road networks. Yet, these advancements often rest on the assumption of static similarities among roads based on distance, structure, or semantics, which can lead to inaccuracies. For instance, two stadiums might share semantic similarities, but their real-time activities could differ significantly, potentially skewing predictive accuracy. Furthermore, relying solely on physical proximity or semantic resemblance overlooks the comprehensive spatial dynamics at play, such as connectivity, which could have a substantial impact on traffic patterns. Qin et al. (2017) employed input attention to discern correlations across various time series data. Similarly, Liang et al. (2018) have developed a model that utilizes global spatial attention to understand how the time series data of one sensor correlates with that of others, complemented by local spatial attention that delves into the correlations within the time series of a single sensor. These attention frameworks are designed to accurately determine the influence of different sensors on the target sensor for predictions. Shi et al. (2020) introduced a groundbreaking model that adeptly managed dynamic spatial relationships and navigated both short-term and long-term temporal dependencies, offering a more nuanced approach to forecasting traffic states.

2.3. Self-supervised learning and contrastive learning

Self-supervised learning (SSL) represents a machine learning approach where models derive their own supervisory signals by predicting parts of their input, diverging from traditional methods that rely on externally provided labels (Zhu et al., 2020; Peng et al., 2020). This technique is particularly valuable in domains such as traffic state prediction, where inherent uncertainties make it difficult to obtain reliable labels. For instance, Wang et al. (2024) noted that traffic accident prediction presented unique challenges due to the variable distribution of traffic accidents across different spatial regions and time periods. SSL proves excellent to enable models to learn useful representations directly from the unlabeled data itself, especially when labeled data are scarce or entirely absent (Li et al., 2022; Ji et al., 2023). There are three main types of self-supervised method: auto-associative, contrastive, and non-contrastive. Auto-associative models learn to reconstruct their inputs. Contrastive models focus on telling the difference between similar and different data samples. Non-contrastive models try to learn representations without explicitly contrasting data points (Xie et al., 2022). Each of these types has its own benefits when it comes to improving the power of unlabeled data for building complex models that can understand patterns (Xie et al., 2022). This is especially important in fields like traffic state prediction, especially for the data augmentation and contrastive learning, where the changing nature of the data creates extra prediction issues. Zhou et al. (2020) introduced an auto-associative data augmentation method that leveraged prior knowledge to address the zero-inflation issue in short-term traffic forecasting. Similarly, Wang et al. (2021) and Liu et al. (2022b) modified the loss function to tackle data imbalance. These approaches represent significant advancements in data augmentation and loss function redesign. However, the challenge of obtaining a robust spatio-temporal representation of sparse traffic data during the model design stage at both node level and graph level, remains unresolved and warrants further exploration.

Contrastive learning (CL) has become a promising approach for learning graph representations without supervision, particularly effective in prediction tasks (Zhu et al., 2021). As a prominent self-supervised learning technique, contrastive learning is increasingly relevant in spatio-temporal prediction research. It is often paired with graph neural networks to enhance the feature representation learning of spatio-temporal data. Applications of this technique are found in areas such as traffic forecasting and weather prediction. Liu et al. (2022a) investigated the integration of contrastive learning into spatio-temporal graph prediction tasks, demonstrating improved prediction accuracy. Additionally, Li et al. (2022) proposed an innovative adaptive graph construction strategy known as Self-Paced Graph Contrast Learning. This strategy distinguishes between positive and negative neighbors by maximizing the margin, leading to the generation of an optimal graph through a self-paced approach. Wang et al. (2024) proposed a traffic accident prediction method with contrastive learning, within the proposed method, the CL can adaptive construct graph structures to learn global spatial correlations among urban regions. However, most of the CL approach only focus on one dimension features which cannot capture the variation of traffic states (Wei et al., 2024; Qu et al., 2023). Besides, for their research, the graph representation is defined by grids which lacks the adjacent matrix analysis which lacks the representation of road network structure.

Despite progress in self-supervised learning techniques, the development of specific graph modification strategies, has been somewhat overlooked in traffic state prediction scenarios. Many existing methods follow the common workflow for data augmentation, like randomly removing connections or mixing up attributes, and apply contrastive learning to extract the spatio-temporal information to develop robust representations against minor changes to less important parts of the graph (Qu et al., 2023; Wei et al., 2024; Zhu et al., 2021). However, in the context of transportation, it is often overlooked that this reconstruction of network-wide relationships might neglect the propagation of traffic flow (the natural feature of traffic flow) when nodes and links are removed from the network (Wang et al., 2024). It is crucial to emphasize that data augmentation in transportation should prioritize the addition of virtual links between nodes, rather than their removal, to ensure that both the physical and dynamic characteristics of traffic flow are maintained in the augmented data and modeling. For instance, the propagation of a traffic shockwave may experience delays across adjacent nodes and links due to spatial and temporal differences. However, removing nodes from the model would prevent the learning of this propagation effect, potentially leading to inaccuracies in predictions. Moreover, in terms of spatial and temporal representation, the learning tasks can be categorized into two distinct groups to separately enhance prediction heterogeneity (Ji

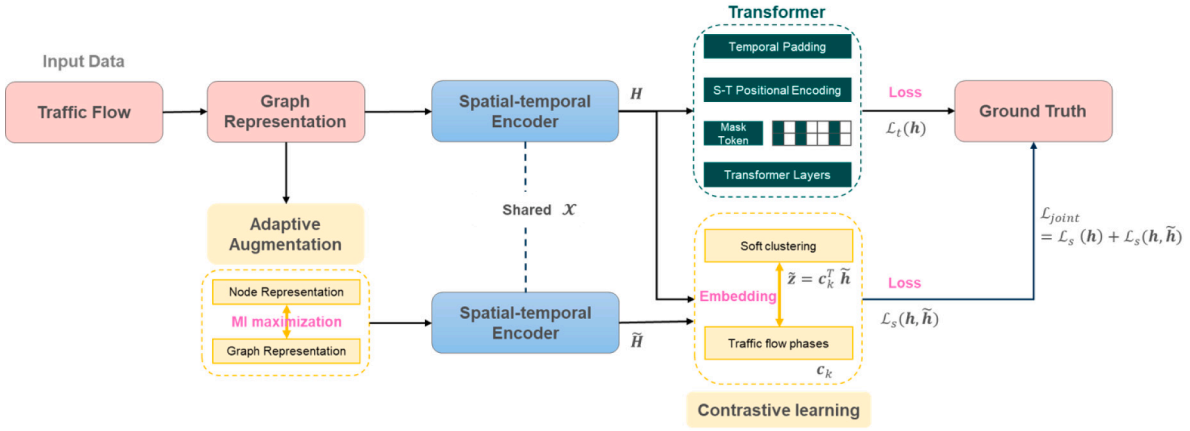


Fig. 1. Architecture of contrastive learning transformer model.

et al., 2023). This involves the development of a contrastive learning technique that effectively minimizes the overall loss between spatial and temporal dimensions.

To address these challenges, this study employs data augmentation to train deep learning models using unlabeled data, thereby circumventing the need for extensively annotated datasets. This approach not only enhances model robustness but also enables the exploration of complex patterns in traffic data without the constraints of labeled training examples. This study proposes an integrated contrastive learning with transformer framework for traffic state prediction, fully considering spatio-temporal heterogeneity and hidden traffic flow dynamics. This framework leverages data augmentation techniques to capture both spatio-temporal patterns and hidden propagation relationships. Additionally, it incorporates a soft clustering component to categorize traffic flow into fundamental diagram clusters, reflecting theoretical traffic flow perspectives in different spatial locations. These components are unified through a joint loss function, which is optimized within the deep learning architecture.

3. Methodology

This study proposes an integrated contrastive learning with transformer framework for traffic state prediction, fully considering spatio-temporal heterogeneity and hidden traffic flow dynamics. This framework leverages data augmentation techniques to capture both spatio-temporal patterns and hidden propagation relationships. Additionally, it incorporates a soft clustering component to categorize traffic flow into fundamental diagram clusters, reflecting theoretical traffic flow perspectives in different spatial locations. These components are unified through a joint loss function, which is optimized within the deep learning architecture. The schematic structure of the proposed methodology is depicted in Fig. 1. The framework comprises four primary components:

- Graph Representation Layer: The initial step in network-wide traffic state prediction involves transforming the traffic state data into a structured format to better serve subsequent method applications. This layer focuses on converting nodes into low-dimensional and dense embeddings that capture both the attributes and the structural characteristics of the graph.
- Adaptive Augmentation Layer: Designed to extract distinctive traffic state characteristics from network-wide features, enhancing the model to represent complex traffic dynamics.
- Spatio-temporal Encoder Layer: This layer is equipped with a spatial representation mechanism to effectively capture dynamic spatio-temporal relationships within the traffic network.
- Transformer Layer: It employs a multi-head attention mechanism along with positional encoding to accurately capture global time dependencies, crucial for predicting traffic states over time.
- Contrastive Learning Layer: Focuses on training the encoders to generate contrastive representations, utilizing the combined data from spatial augmentations and temporal transformations.

The methodology framework is structured into three distinct parts for the sake of elaboration. In the first part, we utilize graph representation to depict traffic states. In the second part, we develop data augmentation techniques aimed at uncovering latent information within the traffic state data and identifying potential new nodes and links to address spatio-temporal data imbalances. Following this, based on the data augmentation, we establish a self-supervised learning framework that focuses on discerning latent traffic flow patterns. This framework incorporates contrastive learning to minimize the loss between spatial and temporal predictions on traffic states.

3.1. Graph representation of road network

The initial step in network-wide traffic state prediction involves transforming the traffic state data into a structured format. Graph representation learning, particularly through graph neural network (GNN), has become a significant approach for analyzing

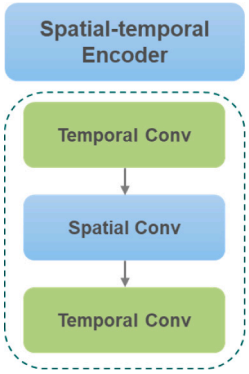


Fig. 2. Architecture of spatio-temporal encoder.

graph-structured data. It focuses on converting nodes into low-dimensional and dense embeddings that capture both the attributes and structure of the graph. However, for traffic state on networks, which primarily focus on traffic conditions of road segments, the definition of nodes and links differs from traditional traffic assignment models. In traffic state estimation, nodes are assumed to correspond to individual road segments, while links represent the connections between these segments. Furthermore, the traffic state on a road segment is not only related to its own temporal conditions but is also influenced by the states of adjacent segments. To capture this, we convert the network-wide information into a graph representation. We model the network-wide traffic state as a directed graph $G = (V, E, \mathbf{A})$, where $V = \{v_1, \ldots, v_{|V|}\}$ represents traffic state on road segment with the size of $|V| = N$, and $E = \{e_1, \ldots, e_{|E|}\}$ denotes the set of its edges and we denote the adjacency matrix of G by $\mathbf{A} \in \mathbb{R}^{N \times N}$, where $A_{ij} = 1$ if $(v_i, v_j) \in E$ for $1 \leq i, j \leq N$. The traffic state to be estimated can be denoted the feature matrix of G by $\mathbf{X} \in \mathbb{R}^{N \times d}$, and $d = 3$ denotes the dimension of traffic flow data including three features: traffic volume, traffic occupy and traffic speed. At time slice t, observations on the road network G can be described as $\mathbf{X_t} = (x_t^1, x_t^2, \ldots, x_t^n) \in \mathbb{R}^{N \times d}$. Further, the traffic flow prediction problem is formulated as given the historical traffic flow $(\mathbf{X}_{t-T+1}, \mathbf{X}_{t-T+2}, \ldots, \mathbf{X}_t) \in \mathbb{R}^T$ and for the traffic network on graph $G = (V, E, \mathbf{A})$ to predict the traffic information for graph at the t-th time slot. We aim to learn a predictive function which accurately estimates the traffic flow features $\mathbf{X}_{t+1} \in \mathbb{R}^{N \times d}$ at the future time step $t + 1$.

As for this network G, particularly noting that geographically proximate nodes often exhibit interconnected traffic dynamics. To fully leverage the topological characteristics of the traffic network for state prediction, we employ the approach of using graph convolution based on Chebyshev polynomial approximation. This method effectively transforms and propagates geographical information across the network, capturing the spatial relationships among nodes and links (Lan et al., 2022), which is highly suitable for network-wide study, enabling a more precise analysis of traffic state. The structural information of a road network is captured using a graph representation method that facilitates the analysis of traffic state information across various nodes, while accounting for the complexity of each node. For instance, the scaled Laplacian matrix is applied particularly in cases where a node has multiple adjacent connections, thereby enhancing its spatio-temporal relevance. This scaled Laplacian matrix, when used in conjunction with Chebyshev polynomials, becomes a widely utilized technique in deep learning. This technique can be formulated as

$$\tilde{L} = \frac{2}{\lambda_{\max}}(\mathbb{D} - \mathbb{A}) - I_N, \tag{1}$$

where $\mathbb{A}$ is the adjacency matrix representing the link structure of the traffic network, I_N is the $N \times N$ identity matrix, and $\mathbb{D} \in \mathbb{R}^{N \times N}$ is the degree matrix, which reflects the complexity of different nodes in the traffic network. Each diagonal element D_{ii} is computed as $D_{ii} = \sum_j A_{ij}^*$. $\lambda_{\max}$ is the maximum eigenvalue of the Laplacian matrix L, defined as $L = \mathbb{D} - \mathbb{A}$. This structure is used to normalize the graph information. By using the Laplacian matrix $\tilde{L}$, we can capture the structural properties of the traffic network graph G, which are crucial for predicting the network-wide traffic state, facilitating more accurate and robust traffic state predictions.

3.2. Spatio-temporal encoder

Using graph representation techniques, we model the spatial information of traffic states at the road network level. However, effectively updating temporal information within the spatial framework remains challenging. Besides, temporal information is also crucial for understanding traffic states, as it reflects varying traffic state over time because of different phases in traffic flow theory. To tackle this challenge, we have developed a spatio-temporal encoder that effectively integrates spatial and temporal information, providing a more comprehensive understanding of traffic dynamics.

Based on data augmentation layer, traffic state patterns for nodes and edges were captured through hidden pattern extraction. For encoding the temporal traffic patterns, we adopt the 2-D causal convolution along the time dimension with a gated mechanism GRU. Specifically, our temporal convolution takes the traffic flow tensor as the input and outputs a time-aware embedding for each region.

As depicted in Fig. 1, the spatio-temporal pattern is primarily extracted via a spatio-temporal encoder block, as showcased in Fig. 2. Previous research (Yu et al., 2017b; Zhang et al., 2017) indicates that integrating spatial and temporal patterns via a convolutional neural network is an efficient approach. Furthermore, to unify the spatial and temporal features of traffic states at one level, we adopt a "sandwich" structure in the encoder block to simultaneously capture the traffic state features from nodes and graphs. For graph convolution, node information is derived from adjacent matrix as well. To integrate the dynamic attributes of the nodes, we aggregate the input from the graph signal $x = x_t \in \mathbb{R}^N$ at each time step using the Kth order Chebyshev polynomial T_k, as follows:

$$g_\theta * Gx = g_\theta(L)x = \sum_{k=0}^{K-1} \theta_k \left(T_k(\tilde{L}) \circ P^{(k)}\right) x, \tag{2}$$

where g_θ signifies the approximate convolution kernel, which is typically a function designed to capture the local graph structure around each node based on the properties encoded in the Laplacian matrix. $*$ denotes the graph convolution operation, and $\theta \in \mathbb{R}^K$ is the learnable coefficient vector of the polynomial, which is iteratively updated during training to optimize the model performance. $P^{(k)} \in \mathbb{R}^{N \times N}$ is the spatio-temporal attention matrix corresponding to the kth head, which dynamically weights the significance of different nodes or time steps in the spatio-temporal data. For the multi-channel input $x^{(l)} \in \mathbb{R}^{N \times c^{(l-1)} \times M}$, the feature dimension of each node is $c^{(l-1)}$, and $g_\theta \in \mathbb{R}^{K \times c^{(l)} \times c^{(l)}}$ represents the convolutional kernel parameters. These parameters are crucial for transforming node features across successive layers (Lan et al., 2022). Consequently, each node aggregates information from the 0 to $(K-1)$th order adjacent nodes. It effectively models the sequential patterns of traffic data across various time steps, as well as the geographical correlations among different spatial regions.

3.3. Transformer layer

Another important component in our study is the transformer layer, as illustrated in Fig. 1 integrated before the merge layer. This integration enhances our ability to derive insights and improve predictions by leveraging the combination of time series data with spatial augmentation. The cornerstone of the transformer is the multi-head attention mechanism, which captures temporal dependencies (Yan et al., 2021). This mechanism can generally be described as mapping a query Q, and a set of key K - value V pairs to an output:

$$Q = W_Q \mathcal{X}, \quad K = W_K \mathcal{X}, \quad V = W_V \mathcal{X} \tag{3}$$

Here, each of W_Q, W_K, and W_V transforms the encoder output $\mathcal{X}$ into the appropriate dimensions for queries, keys, and values, respectively. This facilitates subsequent operations within the attention mechanism.

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4}$$

This formulation allows the model to dynamically weigh the importance of different features based on the interactions between Q and K. In the multi-head attention setup, different attention "heads" can focus on different features:

$$\mathcal{H} = \left[\text{Att}(W_Q^{(1)}\mathcal{X}, W_K^{(1)}\mathcal{X}, W_V^{(1)}\mathcal{X}), \ldots, \text{Att}(W_Q^{(n_h)}\mathcal{X}, W_K^{(n_h)}\mathcal{X}, W_V^{(n_h)}\mathcal{X})\right] W_M \tag{5}$$

Here, each $\text{Att}(W_Q^{(i)}, W_K^{(i)}, W_V^{(i)})$ corresponds to a different 'head' in the multi-head setup, allowing the model to capture various aspects of the data simultaneously, thereby enhancing the depth and breadth of the analysis. The output of the transformer layer, denoted as $\mathcal{H}$, is obtained by applying $\text{MultiAtt}(Q, K, V)$ to the encoder output $\mathcal{X}$. This mechanism effectively synthesizes the spatial and temporal aspects of the data into a comprehensive analysis tool, suited for complex prediction tasks. Then, the model is optimized by minimizing the loss function below:

$$\mathcal{L}_t = \sum_{n=1}^{N} \left| X_{t+1,n} - \hat{X}_{t+1,n} \right| \tag{6}$$

where $\mathbf{X}_{t+1,n}$ is the ground truth of traffic state at $t+1$ and $\hat{\mathbf{X}}_{t+1,n}$ is the predicted state.

3.4. Adaptive augmentation

Building on the graph representation and spatio-temporal encoder steps, we have transformed the spatio-temporal traffic state into low-dimensional dense embedding that preserves both the attributive and structural features of the graph. However, Unlike conventional traffic networks, where links typically represent weighted connections based on factors like distance or traffic flow, the links E in our model may only provide the connectivity without accounting for weights or the directional sequence of traffic flow in network level. To address this issue, our study tackles data sparsity and heterogeneity through adaptive augmentation in node level and graph level. Specifically, we dynamically reconstruct the graph information of the traffic network to better capture its inherent complexity, as detailed in the following section.

Zhu et al. (2021) introduces an adaptive data augmentation strategy that operates at both the structural and attribute levels of the graph. Drawing inspiration from this work, our study in traffic state prediction aims to keep important patterns of traffic states even when random changes happen. To achieve this, we initially generate two distinct graph views by performing stochastic graph

augmentation at the node and graph levels. Subsequently, we implement adaptive augmentation, which dynamically modifies the graph structure during training to enhance the learning process. This process includes adding, removing, or re-weighting nodes and edges based on criteria that the model learns over time, under the assumption of random sampling from a Bernoulli distribution to find the optimal augmentation. The objective is to ensure that the encoded embeddings of each node in the two views remains consistent while being distinguishable from the embedding of other nodes.

3.4.1. Node level augmentation

Node level augmentation addresses data sparsity, particularly for nodes with limited traffic state data might be under-represented in the embedding layer. Following the concept of adaptive augmentation, we introduce noise into node attributes to identify potentially significant but sparse nodes. The adaptive augmentation is achieved by randomly masking a fraction of the dimensions in node features with zeros and then ranking and selecting the probability of augmentation importance. The process at the node level involves applying an augmentation operator to select a subset of the traffic tensor $\mathcal{X}_{t-T:t}$. We specifically pick parts of a mask m from a Bernoulli distribution with a success probability $1-\hat{p}_i^f$ for each feature dimension, where m represents randomness into the feature selection during training. By node augmentation, this method results in a modified feature set $\hat{\mathbf{X}}$, enhancing the model to generalize with limited or imperfect data.

$$m \sim \text{Bernoulli}(1-\hat{p}_i^f) \tag{7}$$

The probability $\hat{p}_i^f$ reflects the importance of the ith feature dimension across nodes. To estimate the importance, we aggregate the weights across all nodes for each feature dimension as follows:

$$w_i^f = \sum_{v\in V} x_{vi}\cdot\varphi(v), \tag{8}$$

where $\phi(v)$ is a centrality measure that quantifies the importance of node v, and $x_{vi}\in\{0,1\}$ indicates the presence of the ith feature dimension in node v. According to Zhou et al. (2020), they turn to measure the magnitude of feature value at dimension i of node v, and calculate its absolute value x_{vi}:

$$w_i^f = \sum_{v\in V} |v_{ui}|\cdot\phi(v). \tag{9}$$

Following this, we normalize the importance weights to obtain a probability that signifies feature importance:

$$\hat{p}_i^f = \min\left(\frac{s_{max}^f - s_i^f}{s_{max}^f - s_{\min}^f}, p_f\cdot p_t\right), \tag{10}$$

where p_f is a hyper-parameter that controls the overall magnitude of feature augmentation and p_t is a threshold probability as illustrated in the Refs. Zhou et al. (2020) and Zhu et al. (2021). $s_i^f = \log w_i^f$, $s_{\max}^f$, μ_s^f is the maximum and the average value of s_i^f respectively. In this way, a node level augmented graph is defined as G_1, which provide a augmentation improvement considering the importance of feature dimension and randomly replaced with zeros. By training the whole prediction method, the value of p_f is iterated in order to find the optimal feature dimension and node.

3.4.2. Graph level augmentation

Beyond node level augmentation, the framework also includes graph level augmentation to capture the physical characteristics of the traffic network. Given the dynamics of traffic states, traffic flow can propagate across multiple links, potentially integrating joint state information from various nodes. By incorporating prior physical knowledge at the graph level, we assess whether adjacency relations effectively retain useful information or whether new links may need to be identified in the dataset.

Graph-level augmentation differs from node-level in that, rather than adding noise, we analyze the structural impact of removing links within the graph. By examining the effects of link removal, we re-weight and adjust the graph connections to improve representation. To refine graph topology, we randomly remove edges based on their importance, ensuring critical connections are retained. Edge importance is determined by edge centrality, derived from the centrality of the two nodes it connects. Edge centrality is calculated as:

To quantify the significance of an edge within the network, we utilize edge centrality, which is derived from the centrality of the two nodes it connects. Specifically, edge centrality is calculated as:

$$w_{uv} = \frac{\phi(u)+\phi(v)}{2} \tag{11}$$

where w_{uv} represents the edge centrality of edge (u,v), and $\phi(u)$ and $\phi(v)$ are the centralities of node u and v, respectively. The probability of retaining edge (u,v) in the augmented edge set $\tilde{E}$ is given by:

$$P((u,v)\in\tilde{E}) = 1-p_{uv}^e, \tag{12}$$

where p_{uv}^e is the calculated probability of removal, determined as

$$p_{uv}^e = \min\left(\frac{s_{max}^e - s_{uv}^e}{s_{max}^e - s_{\min}^e}, p_t\right) \tag{13}$$

where s^e_{uv} represents the standardized edge centrality score, $s^e_{\max}$ and $s^e_{\min}$ are the maximum and minimum centrality scores in the network, and p_t serves as a threshold to maintain the graph structure. By carefully adjusting the composition of edges based on their calculated importance, a graph level augmented graph, denoted as G_2, is defined. This process enhances augmentation by incorporating the physical characteristics of the traffic network into the graph level structure, thereby optimizing it for improved predictive accuracy.

3.4.3. Loss of adaptive augmentation

To align the encoded embeddings of each node across different augmented views while distinguishing them from other embeddings, a contrastive objective is used. For each node v_i, we consider two augmented views: the node-level augmented graph $\tilde{G}_1$ and the graph-level augmented graph $\tilde{G}_2$. Let u_i represent the embedding of v_i in $\tilde{G}_1$, and let v_i represent the embedding of the same node in $\tilde{G}_2$. The adaptive augmentation loss aims to maximize the mutual information between these two views, thereby capturing the shared underlying structure of the data. The loss is defined as:

$$l(u_i, v_i) = \log \frac{e^{\theta(u_i,v_i)/\tau}}{e^{\theta(u_i,v_i)/\tau} + e^{\theta(u_i,v_k)/\tau} + e^{\theta(u_k,v_i)/\tau}} \tag{14}$$

where each similarity $e^{\theta(u_i,v_i)}$ is based on cosine similarity, which measures the alignment between two embedding vectors. This function considers the anchor-positive pair (u_i, v_i) and negative pairs $e^{\theta(u_i,v_k)/\tau}$ and $e^{\theta(u_k,v_i)/\tau}$, where u_k and v_k are embeddings from other nodes. By maximizing the similarity between the anchor-positive pairs and minimizing the similarity with negative pairs, we effectively maximize the mutual information between the two augmented views. Since the views are symmetric, the overall objective is:

$$\mathcal{L}_g = \frac{1}{2N} \sum_{i=1}^{N} \ell \left(l(u_i, v_i) + l(v_i, u_i) \right) \tag{15}$$

This objective encourages alignment between views for each node while preserving distinctions from other nodes, thus enhancing representation quality across the augmented views $\tilde{G}_1$ and $\tilde{G}_2$. By maximizing mutual information between the two views, this method captures the shared information while preserving node-level uniqueness.

3.5. Contrastive learning

Drawing on the spatial information acquired from the encoder $\mathcal{X}$ and the temporal predictions representation by Transformer $\mathcal{H}$ from the preceding layer of our framework, it is crucial to effectively synthesize and capture these dimensions in our traffic state predictions. This integration acknowledges the complexity of traffic dynamics, which are depicted by various phases in the fundamental traffic flow diagram. As for traffic state estimation, influenced by spatial–temporal heterogeneity and external factors, often deviates from typical daily traffic patterns (Hou et al., 2013). To address these variations and enhance the accuracy of our models, clustering is employed for better calibration of traffic states (Gu et al., 2018). To navigate these complexities, we employ self contrastive learning approach designed to identify potential clusters patterns or ranking within the traffic state dataset. By comparing both spatial and temporal differences, contrastive learning is then used in the framework, enabling a more nuanced understanding and prediction of traffic state for prediction. Specifically, contrastive learning is employed to facilitate comparison and training, enhancing prediction precision between $\mathcal{H}$ and $\mathcal{X}$. In general, the key components of a contrastive learning framework include transformations that generate multiple views from a given graph, encoders that compute the representation for each view, and the learning objective to optimize parameters in encoders.

Given the heterogeneity-aware augmented encoder, we aim to enable the traffic state embeddings to effectively preserve the spatial heterogeneity with contrastive learning. To achieve this goal, we design a clustering-based contrastive learning task over traffic patterns to map them into multiple latent representation spaces corresponding to diverse traffic flow phases. It is noted that this clustering is only related to the spatial pattern of traffic states and irrelevant to the temporal dimension. Specifically, we generate K cluster embeddings $\{c_1, \ldots, c_K\}$ (indexed by k) as latent factors for traffic pattern clustering. Formally, the clustering process is performed with $\tilde{z}_{n,k} = c_k^\top \hat{h}_n$. Here, $\hat{h}_n \in \mathbb{R}^D$ is the traffic pattern embedding n encoded from previous adaptive augmentation $\tilde{G}$. $\tilde{z}_{n,k}$ represents the estimated relevance score between phase n embedding and the embedding c_k of the kth cluster as a representation of different phases in traffic states. Afterwards, the cluster assignment of phase n is generated with $\tilde{z}_n = (\tilde{z}_{n,1}, \ldots, \tilde{z}_{n,K})^\top$.

$$\ell(h_n, \hat{z}_n) = -\sum_k \hat{z}_{n,k} \log \frac{\exp(\hat{z}_{n,k})}{\sum_j \exp(\hat{z}_{n,j})} \tag{16}$$

The overall self-supervised objective over all regions is defined as follows by combined with Eq. (15):

$$\mathcal{L}_s = \sum_{n=1}^{N} \ell(h_n, \hat{z}_n) + \mathcal{L}_g \tag{17}$$

By incorporating the supervision on h_n with the heterogeneity-aware cluster assignment $\hat{z}_n$, we make the region embedding h_n to be reflective of spatial heterogeneity within the traffic state in the network. To enhance the efficacy of contrastive learning through a heterogeneity-aware soft clustering paradigm, we have designed an auxiliary learning task focused on prediction. Subsequently, we propose a soft clustering approach tailored to provide differentiated learning signals for augmentation. This method aims to harness the inherent diversity within the data, allowing for more nuanced feature extraction and improved model performance.

Two augmented graph views, G_1 and G_2, are generated by applying both topology- and node-attribute-level augmentations. These views facilitate contrastive learning by providing varied contexts, each characterized by its own unique set of probabilities for the different augmentation techniques applied.

3.6. Merge

Incorporating the approaches from the layers discussed, we integrate contrastive learning with the transformer architecture by defining a loss function that captures the effects of both components within our framework. We then compute the overall loss by combining the losses from self-supervised spatial and temporal heterogeneity modeling, as specified in Eq. (18). This comprehensive approach allows us to formulate a joint learning objective that leverages the strengths of both modeling strategies:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_t(h) + \mathcal{L}_s(h, \hat{h}) \tag{18}$$

Accordingly, the model then is subsequently trained using the back-propagation algorithm, to minimize the joint loss $\mathcal{L}_{\text{joint}}$ ensuring that both spatial and temporal dimensions are effectively optimized to improve the accuracy and robustness of traffic state predictions.

4. Data for model validation and comparison

4.1. Data description

We conduct validation and analysis of our model using the PeMS08 dataset, obtained from the Caltrans Performance Measurement System (PeMS). PeMS captures traffic flow data every 30 s along major highways throughout California's metropolitan areas, which is then aggregated into 5-min intervals for analytical convenience. Specifically, the PeMS08 dataset derives from Region 8, encompassing traffic flow, speed, and occupancy data across 170 nodes (i.e. road segments) in Los Angeles County. To refine our analysis, we also extracted number of lanes information from the raw station data within PeMS, allowing us to better account for traffic flow and density per lane, which is crucial in determining fundamental diagram in different road segments. The dataset covers the period from 1 July 2016 to 31 August 2016, spanning 62 days. For our experiments, we divided the data into a training set consisting of 38 days (1 July to 7 August 2016), a validation set of 12 days (8 August to 19 August 2016), and a test set of 12 days (20 August to 31 August 2016).

4.2. Spatial heterogeneity in traffic flow states

Due to varying capacities and conditions among different road segments, traffic flow characteristics in different road segments exhibit both similarities and differences, namely spatial heterogeneity. Fig. 3 presents an illustration of this phenomenon, based on traffic flow data from four distinct sensors in different road segments within the PeMS dataset, highlighting the spatial variability in traffic patterns. As shown, Nodes 1 and 127 demonstrate very similar traffic flow profiles, with notable peaks during the late night (1 AM to 5 AM) and midday (10 AM to 2 PM) hours. Despite minor differences in scale from 0 AM to 4 AM, the overall trend remains consistent between the two nodes, exemplifying similarities among some nodes in our analysis. Conversely, Nodes 7 and 169 show significant differences in their traffic flow trends. These nodes experience main peak flows during daylight hours from 4 AM to 5 PM and the intensity of traffic flow volumes differs. The clear distinctions in traffic flow scale, peak hours, and trends among these four nodes underscore the heterogeneity of the traffic flow in different locations. In our study, to solve this problem, we employed a self-supervised learning framework to classify such heterogeneity. This method not only groups nodes with similar patterns into the same cluster but also maximizes the differences between clusters to enhance classification accuracy and improve predictive precision account for the spatial heterogeneity.

4.3. Temporal heterogeneity in traffic flow states

In addition to spatial heterogeneity, our analysis reveals distinct temporal patterns within the traffic flow. Similar to other time-series forecasting challenges, the traffic flow in our study exhibits hourly, daily, and weekly trends. Specifically focusing on Node 127, we illustrate temporal heterogeneity in Fig. 4. Analysis indicates that traffic flows on Tuesdays and Wednesdays are remarkably similar, mirroring the patterns observed between Saturdays and Sundays. However, a pronounced divergence is evident between in weekday and weekend,particularly during peak commuting time, as depicted in Fig. 4. Weekdays are characterized by a significant morning rush hour spike, starkly contrasting with the more evenly distributed traffic observed on weekends, where such pronounced fluctuations are notably absent. This divergence highlights the temporal heterogeneity in traffic flow state dynamics across different time periods, underscoring the complexity of predicting traffic flow and the importance of considering temporal heterogeneity in prediction models.

4.4. Evaluation metrics

The prediction performance of different models in predicting traffic flow states is quantified by the mean absolute error (MAE) and mean absolute percentage error (MAPE). MAE quantifies the absolute prediction error, while the MAPE captures the relative prediction error. Given the variability in traffic flow across different nodes and times, both metrics are utilized to assess the comprehensive performance of the proposed model in this study compared to other existing models.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|v_i - \hat{v}_i\right| \tag{19}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{v_i - \hat{v}_i}{X_i}\right| \tag{20}$$

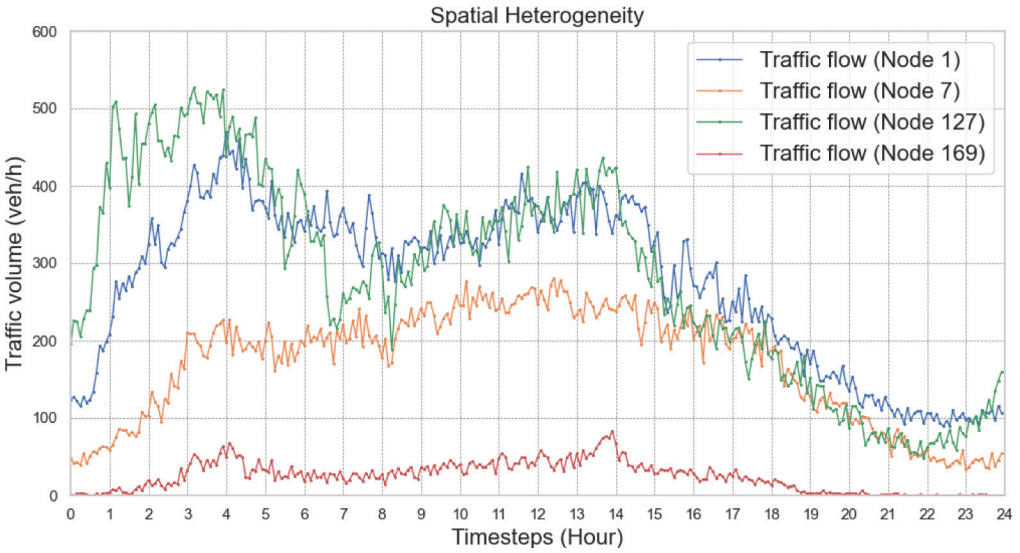


Fig. 3. Spatial heterogeneity.

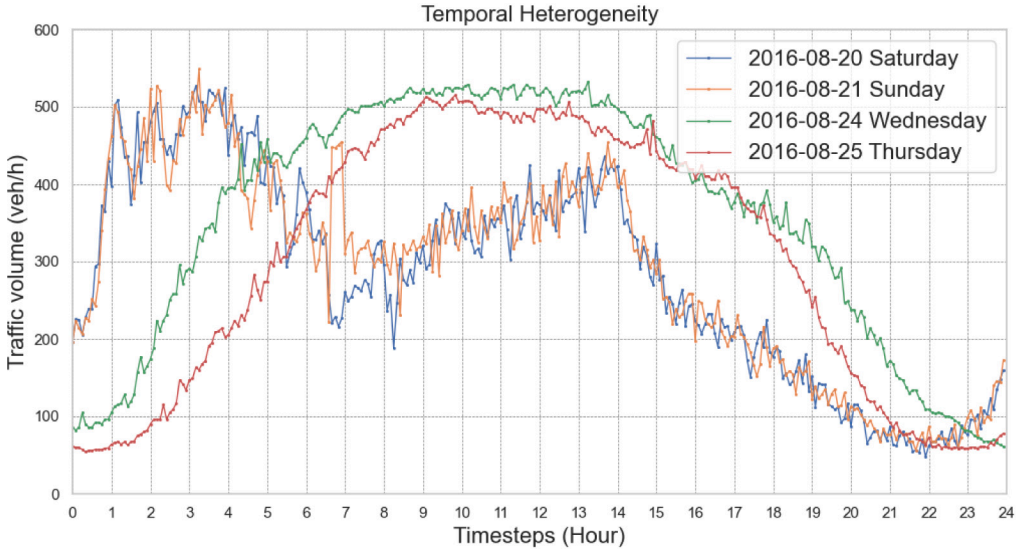


Fig. 4. Temporal heterogeneity.

4.5. Validating spatio-temporal patterns learned by the contrastive learning

In Section 3.5, we introduce clustering-based contrastive learning task over traffic pattern to discern the similarities and differences in traffic flow state dynamics. To validate these approaches, we implement a verification step to analyze the effectiveness of our designed model. Initially, we utilize t-distributed Stochastic Neighbor Embedding (t-SNE) technique as a tool to visualize and explain the clustering results obtained from the contrastive learning component. The input for clustering consists of the predicted traffic flow characteristics at each node, including traffic volume, speed, and density, within our predicted time intervals. For instance, if we are predicting the next 5 min, the input for clustering will be the traffic flow characteristics predicted for these 5 min. Similarly, if the prediction spans the next 10 min, the input will be the traffic flow characteristics for that period. The same principle applies to other predicted time intervals. For the t-SNE techniques, given a set of N high-dimensional objects $\{x_1, \ldots, x_N\}$, t-SNE computes the probabilities p_{ij} that are proportional to the similarity of traffic state objects x_i and x_j including.

For $i \neq j$, define

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{21}$$

and set $p_{i|i} = 0$. The normalization ensures that $\sum_j p_{j|i} = 1$ for all i. The symmetrized joint probabilities p_{ij} are defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \tag{22}$$

The similarity from data point x_j to point x_i is expressed as the conditional probability $p_{j|i}$, such that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i. The bandwidth of the Gaussian kernel σ_i is chosen to match a predefined entropy of the conditional distribution, which is related to the perplexity of the data and can be adjusted using the bisection method. The goal of t-SNE is to learn a map $\{y_1, \ldots, y_N\}$ in a lower-dimensional space that reflects the pairwise similarities p_{ij}, using a similar approach for the low-dimensional counterparts q_{ij}, defined by a Student-t distribution as follows:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|y_k - y_l\|^2)^{-1}}, \tag{23}$$

where y is the predicted cluster index. To better distinguish the difference between clusters, we further verified the clusters of traffic state in original dimension, which conclude the traffic volume, speed and density. We use a dual-regime modified Greenshields fundamental diagram expressed proposed by Hou et al. (2013), which consist of six parameters: the density break point k_{bp}, free-flow speed u_f, speed intercept v_f, minimum traffic speed v_0, and the jam density k_{jam}. k_{jam} and v_0 were assumed 225 vehicle per mile per lane for all the network. The equation is shown as follow:

$$v_i = \begin{cases} u_f, & 0 < k_i < k_{op}, \\ v_0 + (v_f - v_0)\left[1 - \left(\frac{k_i}{k_{jam}}\right)^{\alpha}\right], & k_{op} \leq k_i < k_{jam}. \end{cases} \tag{24}$$

where v_i is the speed, and k_i is density. To calibrate FDs against field data, curve fitting was performed on a least square method to minimize the root mean squared error expressed as:

$$\min RMSE = \frac{1}{N} \sum_{i=1}^{n} (v_i - \hat{v}_i)^2 \tag{25}$$

where $\hat{v}_i$ is the predicted speed, and N is the number of observations. Based on Eq. (25), the dual-regime modified Green-shields fundamental diagram is calibrated, which gives a smooth joint point at the break point density. The least square calibration is implemented to find the optimal solution.

5. Experiments and results

We compare our proposed methodology with several existing traffic state prediction models, which are regarded as our comparison baseline. Further, the effects of spatial augmentation percentage and hyper-parameter τ on prediction errors are examined. To further demonstrate the advantages of our approach, particularly in uncovering latent patterns amid data sparsity and imbalance, the fundamental diagram classification and calibration are also discussed in this section. It helps verify the similarity and heterogeneity of traffic states, providing a deeper understanding effectiveness of our method.

As discussed in Section 3, our objective is to learn a predictive model that accurately estimates future traffic state $\mathbf{X}_{t+1} \in \mathbb{R}^{N \times d}$ at the future time step $t + 1$. Initially focusing on a one-step prediction (5 min), we extend our model to generate predictions for longer periods, ranging from 2 steps to 6 steps (i.e. 10 min to 30 min). We trained and tested our models on Pytorch. We conduct computational experiments to assess the performance of our proposed deep learning framework on an computer, with Intel(R) Core(TM) i9-10980XE CPU and NVIDIA RTX A4000 GPU. The training time for i-CLTP is less than 40 min for 100 epochs.

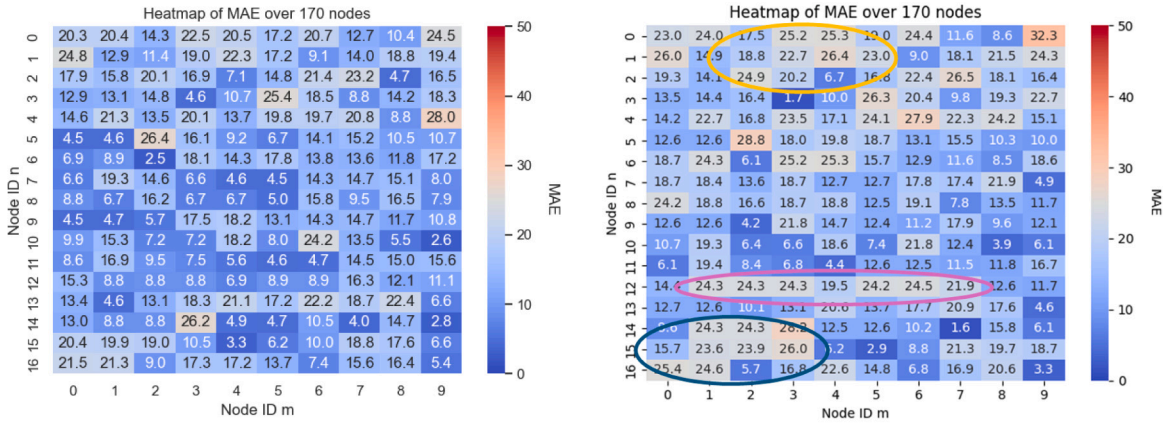
5.1. Model comparison

To verify the accuracy of our predictions, we compare the performance metrics with those of other baseline models derived from the latest research in traffic flow prediction. Specifically, our comparisons are based on methods designed for the PeMS dataset, as detailed in Thunder (2023). This allows us to evaluate precision more accurately against established benchmarks in the field. We consider several advanced baseline methods from a range of modern deep learning approaches, each representing the current best-performing techniques in the field, including

- DCRNN (Li et al., 2017): Diffusion Convolutional Recurrent Neural Network, which utilizes a diffusion sequence-to-sequence architecture to model spatial dependencies.
- STGCN (Yu et al., 2017a): Spatio-Temporal Graph Convolutional Networks which integrates a novel gated CNN module and captures hidden spatial dependencies through a data-driven graph, further fused with given spatial graphs.
- STFGNN (Li and Zhu, 2021): Spatio-Temporal Fusion Graph Neural that combines a data-driven graph with a sophisticated neural network architecture for enhanced predictive accuracy.
- DSTAGNN (Lan et al., 2022): Dynamic Spatio-Temporal Aware Graph Neural Network by combining a data-driven graph and a sophisticated neural network architecture
- STDMAE (Gao et al., 2023): Spatial-Temporal-Decoupled Masked Pre-training, which uses two decoupled masked autoencoders to reconstruct spatio-temporal series along spatial and temporal dimensions.

Table 1
Model comparison on PeMS dataset in terms of MAE and MAPE (%).

Model	PEMS08 MAE	PEMS08 MAPE
DCRNN (Li et al., 2017)	17.86	11.45
STGCN (Yu et al., 2017a)	18.02	11.40
STFGNN (Li and Zhu, 2021)	16.64	10.60
DSTAGNN (Lan et al., 2022)	15.67	9.94
STDMAE (Gao et al., 2023)	13.44	8.76
i-CLTP (Ours)	13.27	7.63



(a) MAE result of i-CLTP (b) MAE result of STDMAE

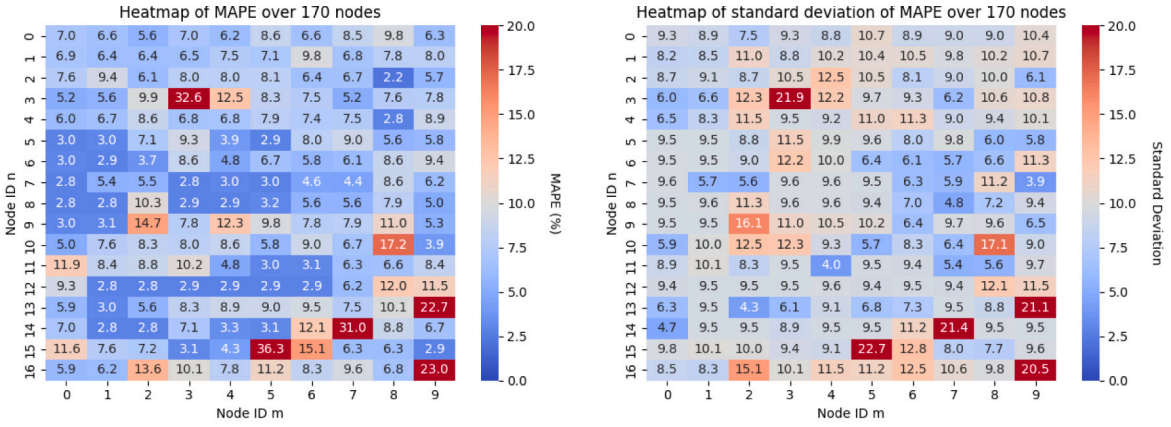
Fig. 5. Node level prediction MAE of i-CLTP and STDMAE.

The prediction performance of models based on identical input features and training data are summarized in Table 1. Our method reaches the best MAE (13.27) and MAPE (7.63%) which surpasses all the prediction models. The proposed model outperforms the best of the baseline comparison model (i.e. STDMAE) (Gao et al., 2023; Lan et al., 2022) in MAE and MAPE by 1.3% and 12.9%, respectively. This result indicates a notable improvement in prediction performance achieved by the proposed model, highlighting its efficacy in traffic state prediction across the network. Specifically, the improvement in MAPE by 12.9% is particularly significant for predicting traffic states at individual nodes. This aspect is crucial for achieving a more uniform prediction quality across various nodes, including those with sparse data. The calculation of MAPE is profoundly affected by the absolute values of traffic flow data, making it especially sensitive to areas with few traffic data recorded by sensors.

Our methodology leverages data augmentation and self-supervised learning to address challenges posed by data sparsity and heterogeneity, aiming to improve prediction metrics across all nodes. We have specifically evaluated the prediction accuracy for various nodes within the network, considering different degrees of spatio-temporal heterogeneity. Fig. 5 illustrates the node level prediction errors using our proposed method and baseline model. Figs. 5(a) and 5(b) display the node level MAE of the i-CLTP and STDMAE methods, respectively. The 170 nodes are arranged in a 17×10 matrix. The Node ID is calculated as $10 \times n + m$ where n and m represent the Y-axis and X-axis indices. Following the same organization, Fig. 6 displays the node level MAPE for the i-CLTP and STDMAE.

For the proposed i-CLTP method, the MAE values for all 170 nodes are depicted in Fig. 5(a), ranging from 2.6 to 28. In contrast, the prediction MAE for nodes using the baseline model ranges from 1.6 to 32.3, as shown in Fig. 5(b). Examining the standard deviation of the MAE results across all 170 nodes reveals that i-CLTP achieves a standard deviation of 5.95, marking an improvement over the 6.61 observed in STDMAE. This indicates that i-CLTP provides a more effective capture of spatial heterogeneity, emphasizing variations at network levels. The broader range of MAE in the baseline model suggests a greater variation in prediction accuracy across different nodes. Furthermore, it is evident that the overall MAE values and the colors in the heatmap of i-CLTP are slightly lower than those for STDMAE. Besides, Fig. 5(b) illustrates spatial heterogeneity in prediction clearly, with three primary clusters circled where the MAE values are notably similar and generally higher than those of surrounding nodes. This pattern indicates that the prediction errors of the STDMAE model are particularly pronounced in these areas, likely due to the limited capacity to capture specific spatio-temporal heterogeneity characteristics. Conversely, the MAE values of i-CLTP are more balanced, as indicated by similar colors representing similar values, highlighting the capacity of proposed i-CLTP to balance the prediction in different nodes especially in nodes with sparse data.

The MAPE results are depicted in Fig. 6. The MAPE values of i-CLTP vary from 2.2 to 36.3 across different nodes as shown in Fig. 6(a). For the STDMAE, MAPE ranges from 4.0 to 22.5 as illustrated in Fig. 6(b). The standard deviation of MAPE across all nodes displays a trend similar to that seen with MAE. For i-CLTP, the standard deviation is 2.93, which is significantly better than



(a) MAPE result of i-CLTP (b) MAPE result of STDMAE

Fig. 6. Node level prediction MAPE for i-CLTP and STDMAE.

the 6.97 observed for STDMAE. The overall MAPE of i-CLTP is significantly better than that of STDMAE, with the improvement in MAPE reaching 12.9%. This improvement can be attributed to the better performance at other nodes. As shown in Fig. 6, there are five nodes (Node 33, Node 139, Node 147, Node 155, and Node 169) with very high MAPE ranging from 22.7 to 36.3. A similar issue occurs in Fig. 6(b) but with a range from 21.1 to 22.7, indicating that predictions for these five nodes are unlikely to achieve significant accuracy. Given that the overall MAE values in STDMAE is relatively high, the smaller percentage error represented by the MAPE is deemed acceptable. Excluding these nodes, it is evident that the prediction precision of i-CLTP markedly surpasses that of STDMAE, particularly in terms of spatial variation and relative values.

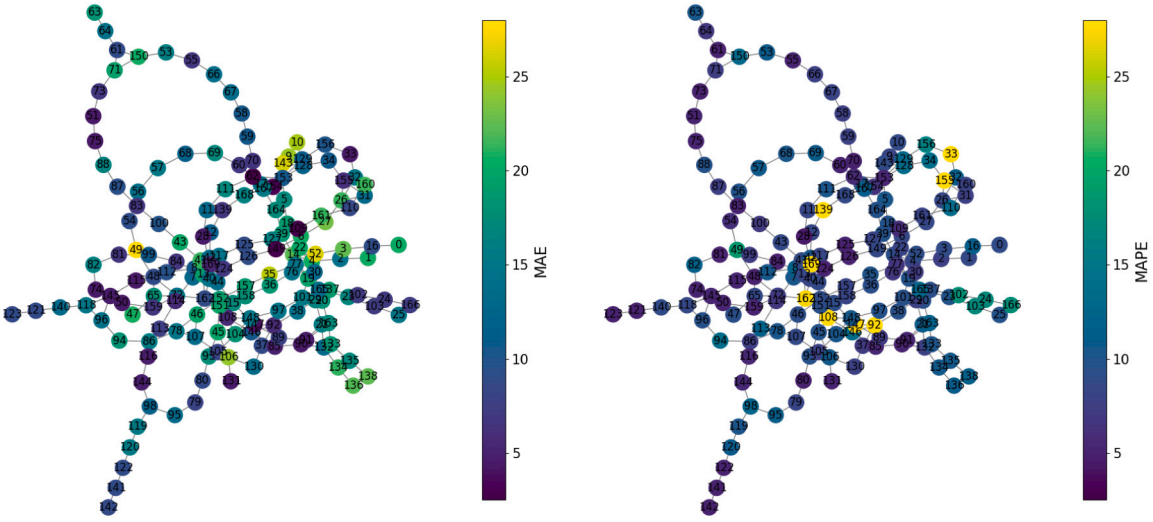
Fig. 7 illustrates the spatial distribution of MAE and MAPE across the network, utilizing the link adjacency matrix and relationships. Notably, nodes characterized by high connection complexity exhibit elevated prediction errors. For instance, Nodes 35, 49, 52, and 143 display the highest MAE values, indicating lower prediction accuracy. This diminished accuracy likely comes from multiple adjacencies of these nodes and significant centrality and connectivity within the road networks. Such characteristics can induce substantial variations and uncertainties in traffic states, which are influenced by adjacent node activities, thereby complicating the prediction of future traffic states. Besides, as previously discussed, the five nodes (Node 33, Node 139, Node 147, Node 155, and Node 169) with very high MAPE, ranging from 22.7 to 36.3, show a similar spatial distribution in Fig. 7(b). Notably, Nodes 139, 147, and 169 are located on adjacent links and are connected to many surrounding nodes result in prediction accuracy challenging to achieve. A similar pattern occurs with Nodes 33 and 155, which also exhibit a high degree of connectivity in the network. Conversely, nodes with simpler adjacency relationships show significantly better prediction performance in terms of MAE. For example, Nodes 63, 64, 134, and 136 are located at the edge of the network, meaning they have relatively simpler adjacency relationships, which facilitates achieving high prediction accuracy. Similarly, nodes that exist on only one link, such as Nodes 66 and 67, consistently demonstrate MAE values below 15. This high level of prediction accuracy is likely due to the similarity in their traffic state features.

5.2. Prediction performance of proposed method

In Fig. 8, we compare the 5-min predicted traffic flow on Node 7 and Node 127 using i-CLTP and the true value from sensors. It indicates that the models demonstrate superior predictive performance and are able to capture the variation in weekday and weekend. Despite the apparent differences in traffic flow patterns between the two nodes, the i-CLTP model accurately mirrors the temporal variations, highlighting its effectiveness in accommodating diverse traffic state dynamics.

Although our proposed model is primarily designed for short-term traffic state prediction (e.g., the next 5 min) and does not encompass an external framework for long-term forecasting (e.g., the next 10 or 15 min), we nevertheless evaluate its performance over extended intervals in comparison to existing methods that specialize in long-term predictions. The comparative results are illustrated in Fig. 9. Our model demonstrates superior predictive accuracy for intervals of 5, 10 and 15 min, achieving MAPE of 7.63, 7.86, and 8.56 respectively. In contrast, the STDMAE model (the best model compared) shows MAPE values of 8.76 for 5 min, 8.88 for 10 min, and 8.98 for 15 min, respectively. Relative to STDMAE, i-CLTP achieves improvements of 12.89%, 11.49%, and 4.68% for the respective prediction time intervals. These findings affirm that, despite its focus on short-term forecasting, our model maintains predictive capabilities for intervals of up to 15 min.

We further evaluate the performance of our model over longer prediction intervals (i.e. over 15 min), comparing it to existing methods that specialize in long-term predictions. The comparative results are depicted in Fig. 9. The accuracy of our model matches that of the best existing model (i.e., STDMAE) for predicting traffic states at the 20-min mark. However, the proposed model falls short compared to STDMAE (with mechanisms for long-term forecasting) in prediction scenarios characterized by relatively



(a) MAE visualization in network (b) MAPE visualization in network

Fig. 7. Node network and its prediction metrics.

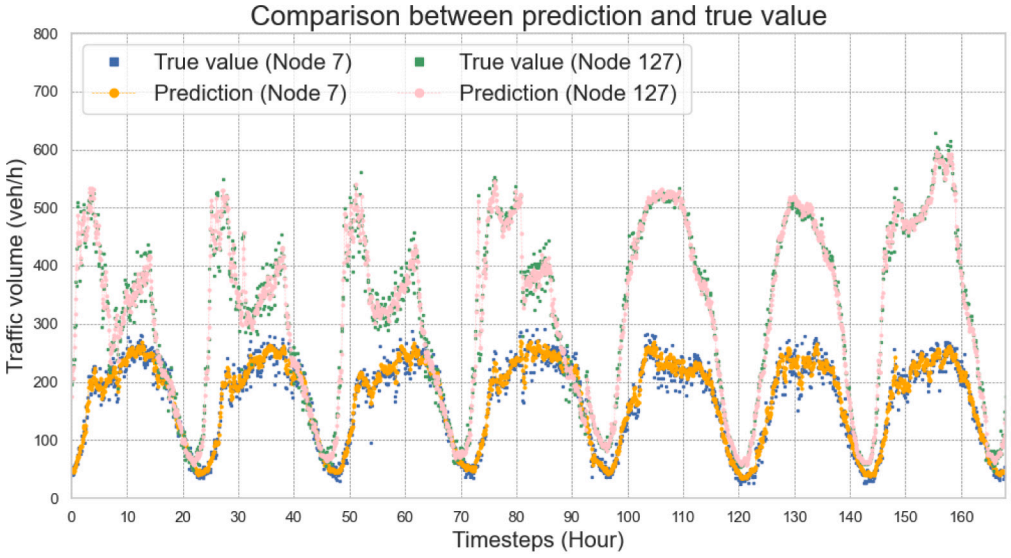


Fig. 8. Time-series of traffic flow prediction (Node 7 and Node 127).

prolonged periods over 20 min. For the prediction accuracy for 25 min and 30 min time interval, the performance of our model is 1.9%, and 14.75% lower than the best baseline model STDMAE, respectively. However, our proposed model still has better predictive accuracy compared other existing popular methods until the prediction interval is up to 30 min (see Fig. 9).

It is worth noting that STDMAE has better predictive performance over 20 min interval as it focus on accommodating frequent temporal fluctuations of traffic states in its model design, but it is significantly impacted by the magnitude of changes in traffic states (Gao et al., 2023; Chang et al., 2023). This may impair the capacity of STDMAE in predicting uncertainties in notable change in traffic states such as traffic congestion due to an accident or cut-in driving behavior. In such as scenario, our proposed model focusing on short-term prediction shows more powerful capacity compared to STDMAE, evidenced by Fig. 9. Meanwhile, it is pertinent to acknowledge that performance of our proposed models may not remain optimal for predictions for long-term prediction as the foundational aims of our proposed model priorities short-term predictive accuracy over long-term forecasts. The rationale behind this design choice stems from a deliberate emphasis on enhancing short-term prediction capabilities, possibly at the expense of integrating features and architectural considerations that benefit long-term forecasting in extensive research.

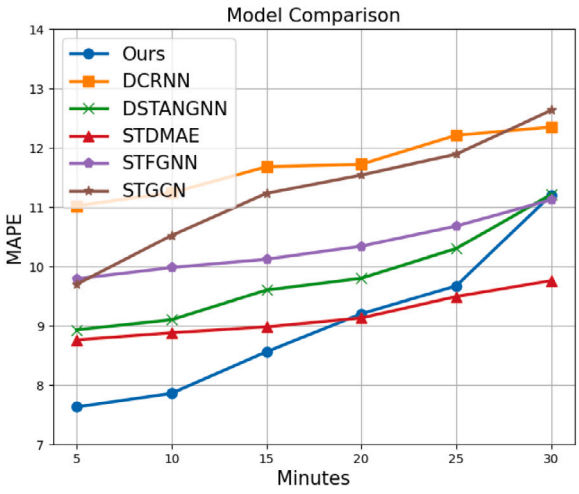
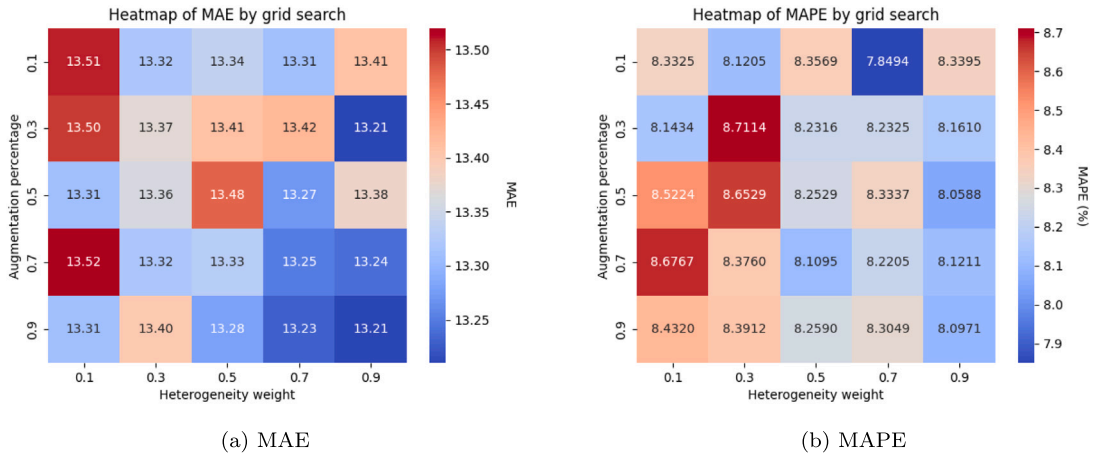


Fig. 9. MAPE comparison.



(a) MAE (b) MAPE

Fig. 10. Sensitivity analysis on spatial heterogeneity.

5.3. Ablation analysis

In our proposed model, there are several critical hyperparameters to be determined, which may affect the model performance significantly. Therefore, this section elaborates the potential impacts of these hyperparameters. Firstly, a sensitivity analysis is conducted to examine the effect of adjusting the augmentation percentage, which determines the proportion of nodes and traffic flow features omitted in our analysis. An augmentation percentage of 0.1 implies that the data augmentation process retains 90% of the original features, modifying only 10%. Another critical parameter in our model is the weight of spatial heterogeneity, denoted as τ, which plays a significant role in calculating the overall contrastive learning objective across all regions, as described in Eqs. (7) and (16). To systematically explore the combined effects of these parameters, we employ a grid-search strategy and organize our analysis within a two-dimensional space in Fig. 10(a). The x-axis represents the augmentation percentage which range from 0.1 to 0.9, while the y-axis corresponds to the weight of spatial heterogeneity, which also varied from 10% to 90%.

Fig. 10(a) shows that a higher spatial heterogeneity ratio τ correlates with an improvement in MAE prediction precision. A similar pattern is observed with the augmentation percentage, where increased values correspond to enhanced predictive precision. This phenomenon underscores the efficacy of our augmentation approach and the consideration of spatial heterogeneity in boosting the precision of contrastive learning predictions. Notably, higher values indicate that the spatial pattern and label regeneration layer yield superior performance compared to the transformer component. However, our ablation analysis reveals that entirely omitting the Transformer architecture does not yield optimal results, which highlights its contributory role in the overall model effectiveness.

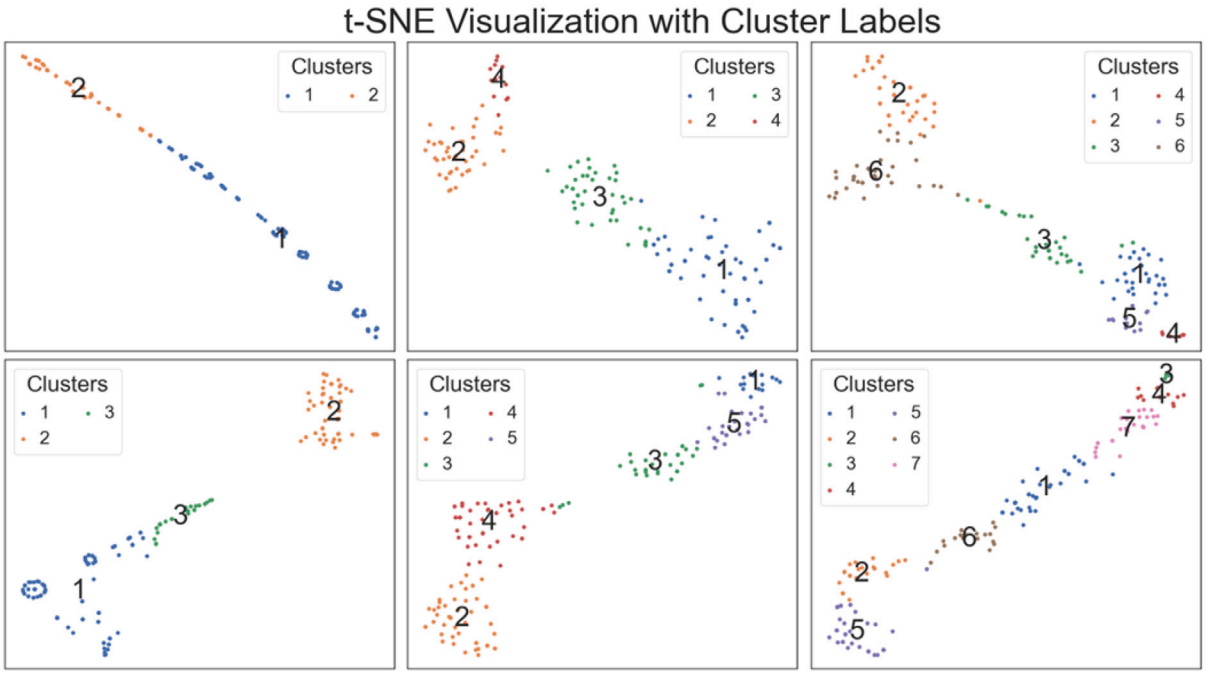


Fig. 11. tSNE with different clusters.

5.4. Validation of spatio-temporal pattern learned by i-CLTP

As illustrated in Section 3.5, in the contrastive learning step, we developed a soft clustering approach to identify the inner clusters of traffic states in the high-dimensional tensors. An unsupervised learning classification algorithm was utilized to deduce the properties of each node. In this section, we demonstrate the clusters of patterns in contrastive learning by t-SNE as illustrated in Section 4.5. These properties may influenced by various traffic-related attributes such as the number of lanes, speed regulations which are not shown directly from traffic state, and more importantly, different phases of traffic flow theory. We empirically determined the optimal number of clusters, exploring $k = 2, 3, 4, 5, 6, 7$ to evaluate the predictive performance of the model. Apparently, when $k = 5$, the t-SNE reach the highest performance as shown in Fig. 11 where the distinction between clusters is particularly found. This clarity in separation suggests a robust grouping that significantly aligns with the underlying traffic dynamics. This optimal clustering not only enhances our understanding of traffic patterns but also improves the model's ability to predict various traffic conditions effectively. The clusters classified within the traffic state network are shown in Fig. 12, where the differences between nodes extend beyond simple adjacency relationships. The i-CLTP method effectively captures and computes latent patterns, facilitating clustering through spatiotemporal heterogeneity. This approach allows for a more nuanced differentiation of nodes, accounting for both spatial and temporal variations in traffic states.

5.5. Difference in traffic flow characteristics in different nodes

Based on the clusters determined through CL, we further focus on their traffic state in traffic flow, and verify the performance in the original sample space, such as traffic volume, speed and occupy. Fig. 13a demonstrates the relations of traffic density and traffic flow volumes in all 170 nodes, showing the overall fundamental diagram patterns. The overall fundamental diagram displays an triangular structure in general. However, as discussed in Section 4.3, the scale and the trend of traffic flow characteristics in different nodes may be quite different due to spatio-temporal heterogeneity. To better understand the difference among different nodes in traffic flow characteristics, which are fully considered in our model, we further analyze the fundamental diagram based on the clusters that contrastive learning classify on the augmented feature. Based on the outcomes of self-supervised learning, five distinct categories of nodes are clustered. The relationship between traffic flow and density exhibits distinct trends across different clusters shown in Fig. 13.

In Fig. 13(b), the gray scatter plots represent the traffic flow–density relations for all nodes in Cluster 1, where the peak values of traffic flow volume indicate the capacity of the link. Traffic flow–density relations of three sample nodes (Nodes 5, 8, and 12) are differentiated and displayed in different colors in Fig. 13(b). The traffic flow characteristics fall in different regions of the flow–density diagram. Specifically, traffic flow characteristics of Nodes 8 and 12 mainly fall in the free flow regime, indicating lower density and free flow traffic conditions. Conversely, traffic flow characteristics of Node 12 are located in the congested flow regime with higher traffic density. This underscores the variability within nodes in a cluster, reflecting the complex traffic flow dynamics.

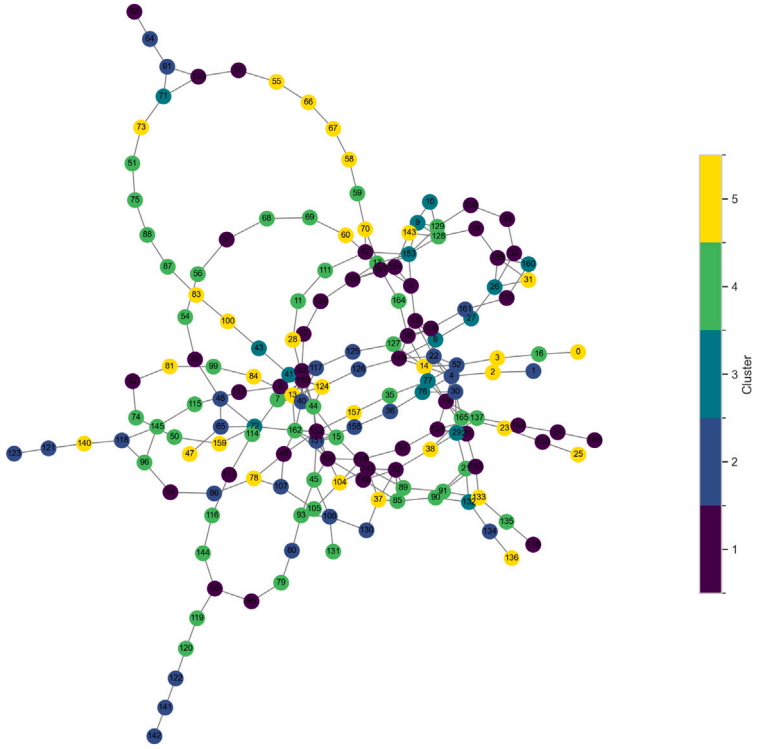
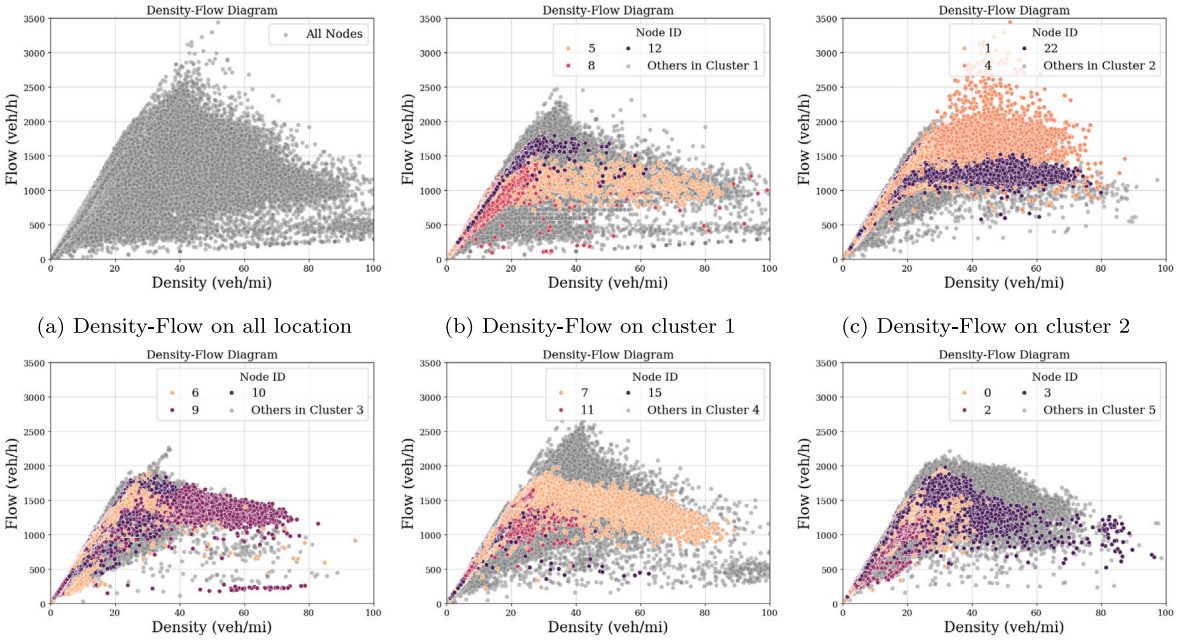


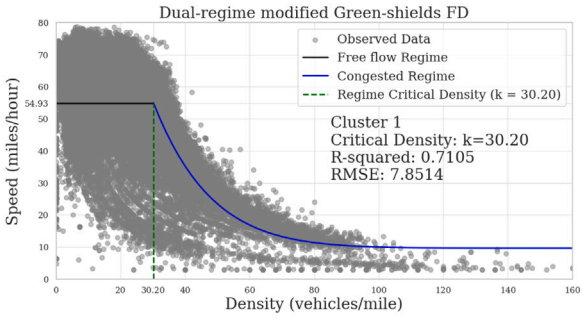
Fig. 12. Traffic state network with different clusters learned from i-CLTP.



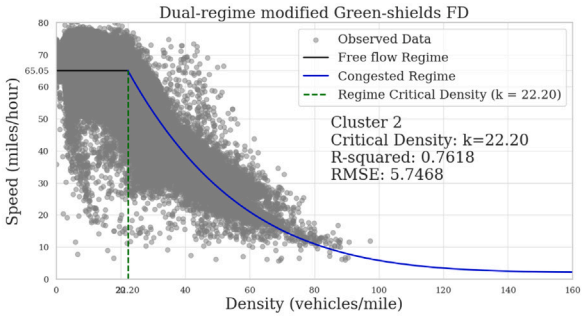
(a) Density-Flow on all location (b) Density-Flow on cluster 1 (c) Density-Flow on cluster 2

(d) Density-Flow on cluster 3 (e) Density-Flow on cluster 4 (f) Density-Flow on cluster 5

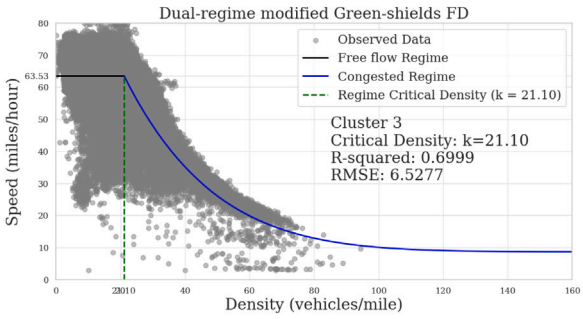
Fig. 13. Density–Flow diagram for each clusters.



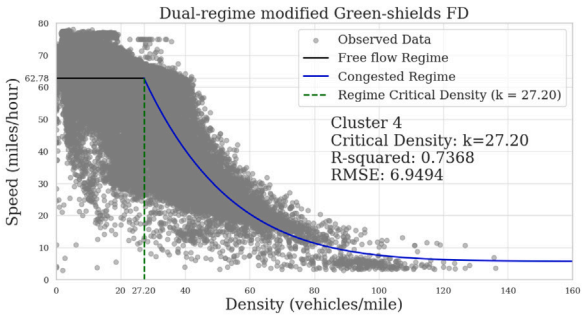
(a) FD difference for three nodes on cluster 1



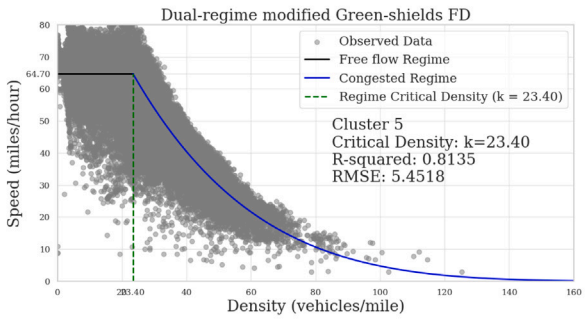
(b) FD difference for three nodes on cluster 2



(c) FD difference for three nodes on cluster 3



(d) FD difference for three nodes on cluster 4



(e) FD difference for three nodes on cluster 5

Fig. 14. Density–Speed difference on clusters.

More importantly, the flow–density relationship of Cluster 2 in Fig. 13(c) markedly contrasts with that observed in Cluster 1, featuring generally lower density and speed. Within cluster 2, Nodes 1, 4, and 22 exhibit similar trends. However, a notable distinction is that Node 4 experiences higher variability in the congested regime. Further analysis of Clusters 3, 4, and 5, depicted in Figs. 13(d), 13(e), and 13(f) respectively, indicates each cluster exhibits distinct fundamental diagram trends (i.e. different traffic flow characteristics). Crucially, traditional classification based solely on observed traffic flow and density would not group some nodes in a cluster due to outliers and variability of some nodes in a cluster. Our method involves an advanced self-supervised classification process in stead of using basic traffic parameters without spatio-temporal heterogeneity. It incorporates a computed hidden layer, enhanced by contrastive learning and refined through supervised learning techniques, facilitating more accurate cluster categorization. Consequently, despite their apparent differences, some nodes are classified into the same cluster, affirming a consistent underlying trend in their fundamental flow–density relationships.

Further, we calibrate the speed–density relations of different clusters using a dual-regime modified Green-shields model. In Cluster 1 in Fig. 14(a), the apex of the fundamental diagram curve signifies the uppermost traffic flow volume (i.e., capacity) where the speed and density are 54.93 miles per hour and 30.2 veh per mile, respectively. This break point shows a capacity of 1658.89 veh per hour (i.e. 54.93 54.92×30.2) marked by a vertical delineation, distinguishes between the free flow and congested flow phases of fundamental diagram. This break point density in Cluster 1 is the lowest recorded among the five clusters, which means for Cluster 1 the roads have the less capacity compared with others. For Cluster 2 in Fig. 13(c), Node 4 exhibits significant variability in flow–density relationship. However, the speed–density calibration of Cluster 2 in Fig. 14(b) yields the best robust R-square and

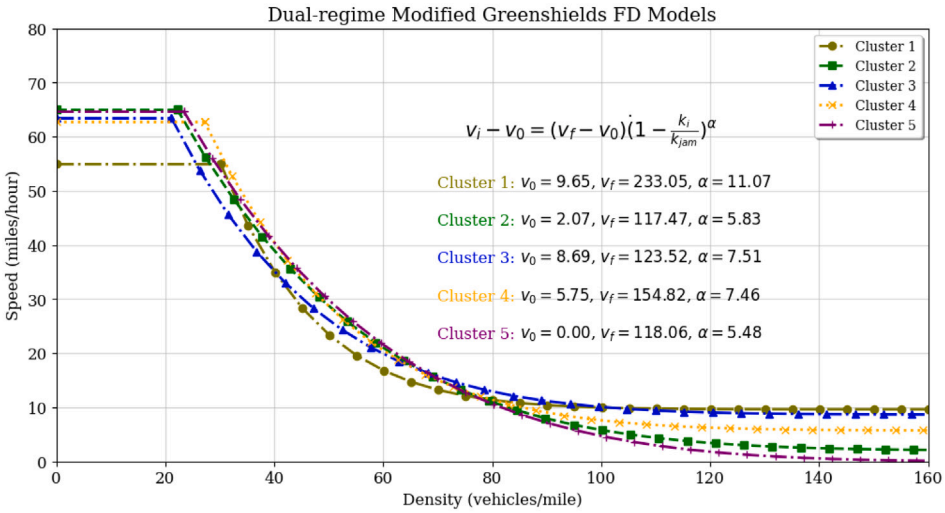


Fig. 15. Density–Speed calibration.

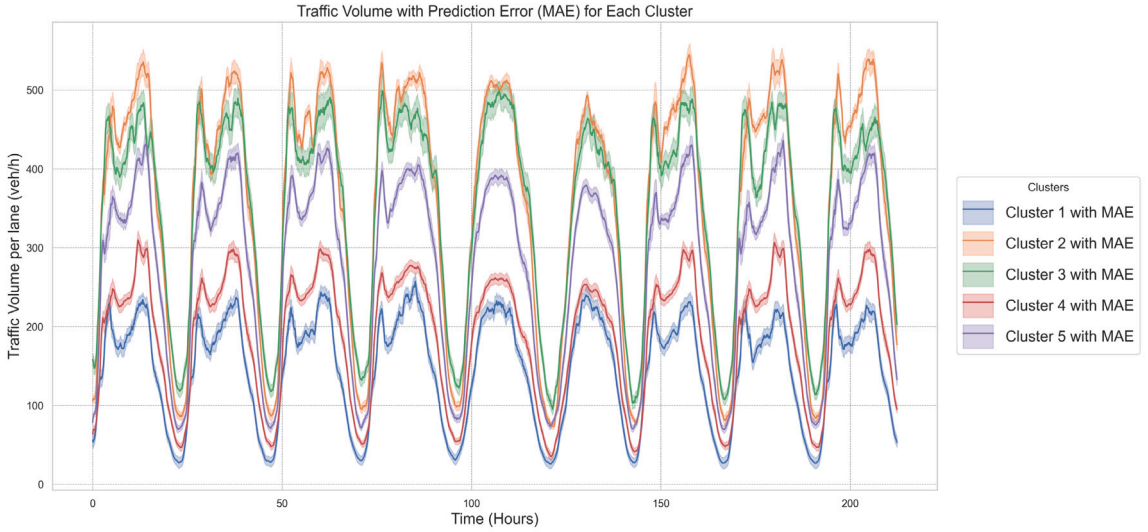


Fig. 16. Temporal heterogeneity for each clusters.

the minimal RMSE, indicating an exceptional model fit. It suggest that despite some points have high variation in Cluster 2, the average fitting precision is high. Further analysis of Clusters 3, 4, and 5 is illustrated in Figs. 14(c), 14(d), and 14(e), respectively. The k_{bp} equal to 21.20 (Cluster 3), 27.20 (Cluster 4), 23.40 (Cluster 5) and v_u equals to 63.53 (Cluster 3), 62.78 (Cluster 4) and 64.70 (Cluster 5), respectively. The calibrated speed–density relations of five clusters are summarized in Fig. 15. With the exception of Cluster 1, which exhibits a relatively low free-flow speed, the remaining maintains a free-flow speed in the vicinity of 65 miles per hour. Moreover, the expectations parameter α, indicative of the steepness of the speed–density relationship in the congested regime, ranges from 5.48 to 7.51 for most clusters. Distinctly, the α calibrated in Cluster 1 is 11.07, signifying a more pronounced decrease in speed with increasing density compared to the other clusters. This steeper gradient in Cluster 1 underscores a more sensitive congestion response within its density–speed dynamic. These result demonstrate disparities in the speed–density relations of different clusters, which is modeled in our proposed models. Meanwhile, the analysis of the density–speed relationship in clusters identified inside our model validates the capacity of our methodology to appropriately categorize the fundamental diagram (or traffic flow characteristics) within the data and to construct a high-dimensional feature relationship for each fundamental diagram pattern.

To further illustrate temporal heterogeneity of traffic states in different clusters obtained from the algorithm, we conducted an experiment to show the temporal dynamics of traffic states and prediction errors across different clusters, as depicted in Fig. 16. The average predicted traffic volume per lane for each cluster emphasizes the latent clustering pattern, while the shaded areas indicate

MAE over time, which helps to highlight the impact of temporal variation on different clusters as shown in Fig. 16. The temporal patterns of traffic states in terms of traffic volumes differ substantially among identified clusters, indicating distinct patterns and scales of temporal dynamics in traffic states in road segments of different clusters. For example, Clusters 2 and 3 demonstrate a single peak on Sundays, suggesting specific temporal traffic dynamics that could correspond to unique regional or activity-based characteristics. In contrast, Clusters 1, 4, and 5 exhibit two distinct peaks, indicative of a different pattern of temporal traffic patterns, which possibly reflects commuter versus non-commuter traffic or varying levels of road usage intensity across these regions. Additionally, the temporal heterogeneity between clusters is also clear from the MAE patterns. Notably, the MAE remains relatively consistent across different time periods and volumes, suggesting that our model maintains a steady level of predictive accuracy in temporal dimensions. This stability is particularly good given the variations in traffic volume in time series, as prediction accuracy often fluctuates with larger or smaller traffic volumes. Here, the proposed model achieves similar accuracy across high and low traffic volumes in time series, which speaks to its robustness in handling diverse temporal dynamics. The proposed approach not only accurately represents the spatio-temporal heterogeneity inherent in traffic states but also enhances the data features through augmentation. Such enhancements make even sparsely distributed data points correctly and logically assigned to their respective categories or clusters for more accurate prediction.

6. Conclusion

This study proposes a novel self-supervised learning approach with a transformer model to predict spatio-temporal traffic flow states. The transformer structure functions as the upper level of the prediction framework to minimize the prediction errors between the ground-truth input and predicted output. Based on the self-supervised contrastive learning, the lower level in this framework is proposed to discern the spatio-temporal heterogeneity and embed the latent characteristic of traffic flow by regenerating the augmentation features. Then, a soft clustering problem is applied between the upper level and lower level to category the types of traffic flow by minimizing the joint loss across each cluster. Experiment results indicate that the proposed model significantly improves traffic flow prediction performance compared to the latest models for the same tasks. Specifically, the proposed model reaches the best MAE (13.27) and MAPE (7.63%) in short term prediction, which surpasses all the baseline prediction models. The proposed model outperforms the best of the baseline comparison model (i.e. STDMAE) in MAE and MAPE by 1.3% and 12.9%, respectively. Meanwhile, empirical validation of the proposed model to precisely predict spatio-temporal traffic flow dynamics with consideration of divergent patterns of traffic in different locations, which is fully considered in the modeling structure. The FD calibration shows that contrastive learning effectively captures latent patterns, successfully classifying them into distinct clusters. This latent clustering enables each group to be predicted separately, improving overall prediction accuracy by addressing spatiotemporal heterogeneity. The outcomes of this study provide a promising alternative for traffic flow state prediction, supporting more efficient and sustainable traffic management and congestion mitigation at the network level.

The proposed model exhibits certain limitations that warrant future improvement as well. Firstly, due to the computational constraints, we deploy the spatio-temporal encoder as the prior layer to the transformer layer, which condenses the input into a lower-rank tensor that effectively reduces the computational complexity. However, there are some other techniques remains longer tensor directly input to transformer which will cost more computation complexity but keep more information from the source data and get higher precision. Despite the discussion on transformer layer is not the key contribution of our study, however, in the further research, it is interesting to explore the transformer architecture in greater depth to enhance precision. Secondly, our model evaluation on the PeMS data focus on short term prediction considering the data augmentation and spatio-temporal pattern. However, other studies may concentrate on predictions over extended time intervals and consider more features. Moreover, it is worth to extend our model for multi-step prediction and consider more external features, such as day type and weather conditions. Last but not the least, the efficacy of the proposed model, alongside the data augmentation and contrastive learning techniques, can be further examined and validated in other datasets to prove generality with more evidence. Lastly, adapting the prediction algorithm for uncommon yet critical scenarios, such as accidents and road closures, is a valuable direction for enhancing intelligent traffic management. These scenarios hold particular importance in optimizing real-time responses and ensuring robust traffic flow. However, due to the lack of traffic state data during accidents, we were unable to evaluate the performance of our proposed algorithm under such critical conditions. A promising avenue for future research involves leveraging datasets that include accident scenarios to test and refine the algorithm, thereby improving its applicability in managing complex traffic situations.

CRediT authorship contribution statement

Ruo Jia: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kun Gao:** Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Yang Liu:** Writing – original draft, Validation, Methodology, Investigation, Data curation. **Bo Yu:** Visualization, Validation, Software, Resources, Methodology, Investigation, Data curation. **Xiaolei Ma:** Writing – original draft, Supervision, Software, Resources, Methodology, Investigation, Data curation. **Zhenliang Ma:** Validation, Resources, Methodology, Investigation, Data curation.

Acknowledgments

This study was funded and carried out as part of the VINNOVA project (2023-01042) supported by Sweden's Innovation Agency, the Area of Advance Transport at Chalmers University of Technology, Sweden, and the e-MATS project (P2023-00029) funded by the Joint Programming Initiative Urban Europe. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the sponsors' views.

Data availability

Data will be made available on request.

References

Chang, Z., Koulieris, G.A., Shum, H.P., 2023. On the design fundamentals of diffusion models: A survey. arXiv preprint arXiv:2306.04542.

Chen, Z., Liu, K., Wang, J., Yamamoto, T., 2022. H-ConvLSTM-based bagging learning approach for ride-hailing demand prediction considering imbalance problems and sparse uncertainty. Transp. Res. C 140, 103709.

Cui, Z., Ke, R., Pu, Z., Wang, Y., 2020. Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values. Transp. Res. C 118, 102674.

Gao, H., Jiang, R., Dong, Z., Deng, J., Song, X., 2023. Spatio-temporal-decoupled masked pre-training for traffic forecasting. arXiv preprint arXiv:2312.00516.

Gu, Z., Saberi, M., Sarvi, M., Liu, Z., 2018. A big data approach for clustering and calibration of link fundamental diagrams for large-scale network simulation applications. Transp. Res. C 94, 151–171.

Guo, J., Huang, W., Williams, B.M., 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. Transp. Res. C 43, 50–64.

Hamed, M.M., Al-Masaeid, H.R., Said, Z.M.B., 1995. Short-term prediction of traffic volume in urban arterials. J. Transp. Eng. 121 (3), 249–254.

Hou, T., Mahmassani, H.S., Alfelor, R.M., Kim, J., Saberi, M., 2013. Calibration of traffic flow models under adverse weather and application in mesoscopic network simulation. Transp. Res. Rec. 2391 (1), 92–104.

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

Kim, Y., young Tak, H., Kim, S., Yeo, H., 2024. A hybrid approach of traffic simulation and machine learning techniques for enhancing real-time traffic prediction. Transp. Res. C 160, 104490. http://dx.doi.org/10.1016/j.trc.2024.104490, URL: https://www.sciencedirect.com/science/article/pii/S0968090X24000111.

Kumar, N., Raubal, M., 2021. Applications of deep learning in congestion detection, prediction and alleviation: A survey. Transp. Res. C 133, 103432.

Kwon, J., Murphy, K., 2000. Modeling Freeway Traffic with Coupled HMMs. Technical Report. Technical report, Univ. California, Berkeley.

Lan, S., Ma, Y., Huang, W., Wang, W., Yang, H., Li, P., 2022. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In: International Conference on Machine Learning. PMLR, pp. 11906–11917.

Lee, S., Fambro, D.B., 1999. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transp. Res. Rec. 1678 (1), 179–188.

Li, Z., Huang, C., Xia, L., Xu, Y., Pei, J., 2022. Spatial-temporal hypergraph self-supervised learning for crime prediction. In: 2022 IEEE 38th International Conference on Data Engineering. ICDE, IEEE, pp. 2984–2996.

Li, Z., Xiong, G., Tian, Y., Lv, Y., Chen, Y., Hui, P., Su, X., 2020. A multi-stream feature fusion approach for traffic prediction. IEEE Trans. Intell. Transp. Syst. 23 (2), 1456–1466.

Li, Y., Yu, R., Shahabi, C., Liu, Y., 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926.

Li, M., Zhu, Z., 2021. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4189–4196.

Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y., 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction. In: IJCAI. pp. 3428–3434.

Liu, Y., Jia, R., Ye, J., Qu, X., 2022. How machine learning informs ride-hailing services: a survey. Communications in Transportation Research 2, 100075.

Liu, X., Liang, Y., Huang, C., Zheng, Y., Hooi, B., Zimmermann, R., 2022a. When do contrastive learning signals help spatio-temporal graph forecasting? In: Proceedings of the 30th International Conference on Advances in Geographic Information Systems. pp. 1–12.

Liu, Y., Liu, Z., Jia, R., 2019. DeepPF: A deep learning based architecture for metro passenger flow prediction. Transp. Res. C 101, 18–34.

Liu, X., Zhang, Z., Lyu, L., Zhang, Z., Xiao, S., Shen, C., Philip, S.Y., 2022b. Traffic anomaly prediction based on joint static-dynamic spatio-temporal evolutionary learning. IEEE Trans. Knowl. Data Eng. 35 (5), 5356–5370.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.-Y., 2014. Traffic flow prediction with big data: A deep learning approach. IEEE Trans. Intell. Transp. Syst. 16 (2), 865–873.

Okutani, I., Stephanedes, Y.J., 1984. Dynamic prediction of traffic volume through Kalman filtering theory. Transp. Res. B 18 (1), 1–11.

Peng, Z., Huang, W., Luo, M., Zheng, Q., Rong, Y., Xu, T., Huang, J., 2020. Graph representation learning via graphical mutual information maximization. In: Proceedings of the Web Conference 2020. pp. 259–270.

Qi, Y., Ishak, S., 2014. A Hidden Markov Model for short term prediction of traffic conditions on freeways. Transp. Res. C 43, 95–111.

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., Cottrell, G., 2017. A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971.

Qu, Y., Rong, J., Li, Z., Chen, K., 2023. ST-A-PGCL: Spatiotemporal adaptive periodical graph contrastive learning for traffic prediction under real scenarios. Knowl.-Based Syst. 272, 110591.

Shahriari, S., Ghasri, M., Sisson, S., Rashidi, T., 2020. Ensemble of ARIMA: combining parametric and bootstrapping technique for traffic flow prediction. Transp. A: Transp. Sci. 16 (3), 1552–1573.

Shi, X., Qi, H., Shen, Y., Wu, G., Yin, B., 2020. A spatial–temporal attention approach for traffic prediction. IEEE Trans. Intell. Transp. Syst. 22 (8), 4909–4918.

Thunder, J., 2023. PEMS08 traffic flow prediction. URL: https://www.kaggle.com/code/jvthunder/pems08-traffic-flow-prediction. (Accessed 03 July 2023).

Wang, B., Lin, Y., Guo, S., Wan, H., 2021. GSNet: Learning spatial-temporal correlations from geographical and semantic aspects for traffic accident risk forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4402–4409.

Wang, S., Zhang, Y., Piao, X., Lin, X., Hu, Y., Yin, B., 2024. Data-unbalanced traffic accident prediction via adaptive graph and self-supervised learning. Appl. Soft Comput. 157, 111512.

Wei, T., Lin, Y., Guo, S., Lin, Y., Zhao, Y., Jin, X., Wu, Z., Wan, H., 2024. Inductive and adaptive graph convolution networks equipped with constraint task for spatial–temporal traffic data kriging. Knowl.-Based Syst. 284, 111325.

Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. J. Transp. Eng. 129 (6), 664–672.

Xie, D., Chen, S., Duan, H., Li, X., Luo, C., Ji, Y., Duan, H., 2023. A novel grey prediction model based on tensor higher-order singular value decomposition and its application in short-term traffic flow. Eng. Appl. Artif. Intell. 126, 107068.

Xie, Y., Xu, Z., Zhang, J., Wang, Z., Ji, S., 2022. Self-supervised learning of graph neural networks: A unified review. IEEE Trans. Pattern Anal. Mach. Intell. 45 (2), 2412–2429.

Xie, Y., Zhang, Y., Ye, Z., 2007. Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition. Comput.-Aided Civ. Infrastruct. Eng. 22 (5), 326–334.

Yan, H., Ma, X., Pu, Z., 2021. Learning dynamic and hierarchical traffic spatiotemporal features with transformer. IEEE Trans. Intell. Transp. Syst. 23 (11), 22386–22399.

Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., Yin, B., 2021. Deep learning on traffic prediction: Methods, analysis, and future directions. IEEE Trans. Intell. Transp. Syst. 23 (6), 4927–4943.

Yu, R., Li, Y., Shahabi, C., Demiryurek, U., Liu, Y., 2017b. Deep learning: A generic approach for extreme condition traffic forecasting. In: Proceedings of the 2017 SIAM International Conference on Data Mining. SIAM, pp. 777–785.

Yu, B., Yin, H., Zhu, Z., 2017a. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875.

Zhan, X., Li, R., Ukkusuri, S.V., 2020. Link-based traffic state estimation and prediction for arterial networks using license-plate recognition data. Transp. Res. C 117, 102660.

Zhang, J., Shi, X., Xie, J., Ma, H., King, I., Yeung, D.-Y., 2018. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. arXiv preprint arXiv:1803.07294.

Zhang, J., Zheng, Y., Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Zhong, C., Wu, P., Zhang, Q., Ma, Z., 2023. Online prediction of network-level public transport demand based on principle component analysis. Communications in Transportation Research 3, 100093.

Zhou, Z., Wang, Y., Xie, X., Chen, L., Liu, H., 2020. RiskOracle: A minute-level citywide traffic accident forecasting framework. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1258–1265.

Zhu, Z., Xu, M., Wang, K., Lei, C., Xia, Y., Chen, X.M., 2023. A non-local grouping tensor train decomposition model for travel demand analysis concerning categorical independent variables. Transp. Res. C 157, 104396.

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L., 2020. Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131.

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