

Relation-diffusion augmented network-wide traffic state estimation under sparse sensor deployment

Tao Guo^{1,2}, Ningkang Yang¹✉, Kevin Yu², Panagiotis Angeloudis², Constantinos Antoniou¹

 **Cite this article:** Guo T, Yang N, Yu K, et al. *Commun Transp Res* 2026, 6(3): 9640046. <https://doi.org/10.26599/COMMTR.2026.9640046>

ABSTRACT: Network-wide traffic state estimation (TSE) under sparse sensor deployment aimed to infer traffic conditions at unobserved road segments from limited sensor observations. Existing diffusion graph convolutional methods mainly propagate node features over physical topology or adaptive graphs, but the adaptive relationships involving unobserved nodes are often weakly supported because their traffic features are missing, zero-padded, or represented only by initialized embeddings. To address this limitation, this study proposed the relation-augmented diffusion graph convolution network (RADGCN), a

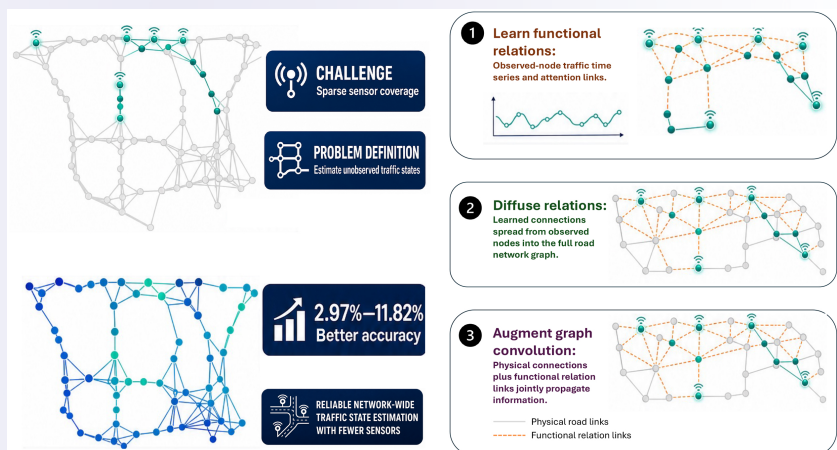
relation-aware adaptive diffusion framework for TSE under sparse observations. RADGCN first learned functional relationships among observed nodes using reliable temporal traffic features and graph-aware attention and then propagated these relationships to the full network through learnable bilateral diffusion kernels. The resulting relation matrix was incorporated into diffusion graph convolution as an additional transition operator, allowing traffic states to be estimated through both physical topology and learned functional dependencies. An auxiliary edge prediction task was further introduced to regularize node embeddings under inductive K -order neighborhood subgraph training. Experiments across three benchmark datasets showed that RADGCN consistently outperformed state-of-the-art baselines across different sparse-sensor settings, achieving average improvements of 2.97%–11.82% in mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). Additional ablation results verified the effectiveness of relation diffusion, relation-augmented feature propagation, and K -order subgraph sampling. Further experiments also validated RADGCN's generalizability across datasets and robustness to sensor sparsity.

KEYWORDS: traffic state estimation (TSE); traffic kriging; graph neural network (GNN); diffusion graph convolution network (DGCN); inductive learning

1 Introduction

Accurate and timely traffic information is a fundamental prerequisite for modern intelligent transportation systems (ITSs). Traffic data support a wide range of applications, including congestion monitoring, traffic control, route guidance, infrastructure planning, and environmental impact assessment. In particular, network-wide traffic information enables transport agencies to better understand spatiotemporal traffic dynamics and make informed operational decisions. To obtain such information, various sensing technologies have been deployed in

urban traffic networks, diversifying from traditional fixed detectors to a rich ecosystem of video, probe-vehicle, and communication-based data sources (Wright and Horowitz, 2016), but their roles are not yet equivalent. Fixed-location sensors form the backbone of most freeway monitoring systems because they provide long-term, calibrated, and regulator-accepted measurements of flow, speed, and occupancy that underpin historical databases, control algorithms, and model validation frameworks (Seo and Bayen, 2017). Newer vision and probe-based methods mainly act as supplements. They expand coverage, enable richer trajectory-level analysis, and help compensate for



¹ Chair of Transportation Systems Engineering, Technical University of Munich, Munich 80333, Germany. ² Department of Civil & Environmental Engineering, Imperial College London, London SW7 2AZ, UK.

✉ Corresponding author. E-mail: ningkang.yang@tum.de

Received: April 2, 2026; Revised: June 13, 2026; Accepted: August 6, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

spatial gaps, yet they often suffer from issues such as heterogeneous penetration rates (Lovisari et al., 2016), privacy constraints (Yang et al., 2025), or infrastructure compatibility, which limit their use as sole authoritative sources. As a result, even in recent traffic state estimation (TSE) and ITS studies that leverage multisource and big-data fusion, fixed infrastructure detectors still serve as the anchor dataset against which other modalities are integrated or evaluated, so infrastructure-based fixed sensors remain the mainstream technology for operational traffic sensing and control in practice.

However, deploying and maintaining traffic detectors is expensive and logistically demanding. The installation of fixed detectors, such as inductive loops, typically requires lane closures and intrusive construction, disrupting traffic operations and increasing deployment costs. In addition to the initial installation expenses, these sensors also require continuous maintenance, calibration, and periodic replacement due to wear, environmental conditions, or equipment failures. In a large-scale road network, only a small fraction of links or nodes are directly equipped with sensors, leaving large portions of road segments unmonitored (Nie et al., 2023a; Wu et al., 2021a; Zhou et al., 2025).

This limitation motivates the problem of network-wide TSE, which aims to infer traffic conditions at unobserved locations based on limited observations and spatial correlations within the network. This problem is sometimes referred to as the traffic kriging problem by analogy with spatial interpolation methods in geostatistics. In the remainder of this article, the terms TSE and traffic kriging problem will be used interchangeably. Reliable solutions to this problem are essential for enabling network-level traffic monitoring when dense sensing infrastructure is unavailable.

Earlier efforts on the TSE problem have predominantly focused on model-driven approaches, largely rooted in classical traffic flow theory, such as the Lighthill–Whitham–Richards (LWR) model (Lighthill and Whitham, 1955) and the cell transmission model (CTM) (Daganzo, 1994). Despite their substantial theoretical value and demonstrated success in general traffic estimation and interpretation tasks, these methods typically require sophisticated, case-specific hyperparameter calibration and rely on strong assumptions embedded in the mathematical formulation of traffic dynamics (Zhou et al., 2025). Such limitations hinder their generalizability to large-scale traffic networks and to scenarios characterized by severe data sparsity. Data-driven methods mitigate the limitations arising from strong mathematical assumptions about idealized scenarios by directly learning traffic spatiotemporal patterns from observed data. In particular, the emergence of graph neural networks (GNNs), whose graph topology and message-passing mechanisms naturally align with the structure of transportation networks and traffic flow dynamics, has significantly advanced this paradigm. By this means, road segments are abstracted as nodes in a graph, and their connectivity is represented as edges between nodes. Within this context, the diffusion graph convolution network (DGCN) (Li et al., 2017; Wu et al., 2021a), which models message passing as a diffusion process over nodes based on a Markov transition matrix, provides a natural analogy to traffic flow propagation across road segments. Subsequent studies have identified and addressed several limitations of classical DGCN models, including the lack of adaptive adjacency matrices (Nie et al., 2023a; Wei et al., 2024) and the reliance on zero-padding for unobserved nodes (road segments) (Zhou et al., 2025). However, important gaps remain in current DGCN-based TSE methods.

The first limitation arises from the construction of adaptive

node relationship matrices. Recent studies have focused on moving beyond purely topology-based diffusion by introducing either adaptive dependency learning or enhanced feature propagation on traffic graphs. Road segments with available measurements are termed observed nodes, whereas those without measurements are referred to as unobserved nodes. In this line of research, adaptive relations are typically constructed in two ways. One is to learn a dynamic affinity matrix from fully available auxiliary features at all nodes, such as the speed-aware adaptive matrix in the study by Nie et al. (2023b), which is computed by attention over complete speed records assumed to be available for all road segments. The other is to derive adaptive adjacency from learnable node embeddings, as in the work of Wei et al. (2024), where embeddings of unknown nodes are first initialized as zeros and then updated through diffusion before adaptive relations are formed. In the abovementioned two routes, relationships associated with unobserved nodes are not inferred directly from observed traffic dynamics but are instead introduced indirectly through fully sampled auxiliary signals, feature diffusion, or parameterized embeddings. This indirect construction weakens the observational support for learned relations, particularly for nodes far from observed regions, thereby limiting the reliability of the inferred relational structure and adversely affecting subsequent state estimation. Another major research gap arises from the contradiction between inductive training and the completeness of graph structure for unbiased message-passing mechanism learning. Despite its benefits in an inductive training setting, random subgraph sampling inherently results in structural incompleteness of the training graph, which not only truncates the local neighborhood context but also distorts multihop dependency propagation and creates a mismatch between the training-time subgraph and inference-time full-graph structure. As a result, the learned node embeddings may fail to preserve the structural semantics that are critical for reliable graph-based estimation. Balancing the practicality of inductive training against the structural information loss caused by subgraph sampling remains a key challenge. These limitations motivate a shift from feature-centric diffusion to relation-aware diffusion. A feasible way to address this issue is to first infer relationships from observed nodes, where traffic patterns are reliable, and then propagate such relational information to the rest of the network, rather than directly propagating node features over a graph whose adaptive structure is only weakly supported in unobserved regions. Based on this idea, this research proposes the relation-augmented diffusion graph convolution network (RADGCN), a relation-aware diffusion framework that explicitly separates relationship inference from traffic state propagation. This design moves beyond conventional feature completion on graphs toward relation completion over graphs, yielding a more reliable inductive bias for TSE under sparse observations. In addition, to support inductive training while alleviating the structural incompleteness caused by subgraph sampling, a K -order neighbor subgraph sampling strategy and an auxiliary edge prediction task with negative-edge sampling are further developed.

The main contributions of this research are summarized as follows:

Observed-node relation learning and diffusion. We propose a relation-diffusion mechanism that first learns adaptive functional relationships only from observed nodes, where traffic signals are reliable, and then propagates these relationships to the full traffic network through learnable bilateral diffusion kernels. This design avoids directly constructing adaptive relations from zero-padded or weakly supported representations of unobserved nodes.

Relation-augmented traffic state propagation. We integrate the learned and propagated relation matrix into diffusion graph convolution as an additional transition operator. The resulting RADGCN model jointly leverages physical road connectivity and learned functional dependencies, enabling more reliable TSE in sparsely observed, spatially fragmented networks.

Inductive training with structural regularization. We incorporate K -order neighborhood subgraph sampling and an auxiliary edge-prediction task to reduce structural representation disruption caused by random subgraph training. The ablation experiments further validate the benefits of the auxiliary edge-prediction task and K -order sampling over random node sampling.

The remainder of the paper is organized as follows: Section 2 summarizes the related works in the TSE problem, particularly the studies utilizing the GNN paradigm and research gaps. Section 3 illustrates the problem formulation and provides the preliminary knowledge of DGCN, Graphormer, and K -order subgraph sampling, which are not claimed as the core innovation in this paper but are intensively used in the proposed framework. Section 4 presents the methodology of the paper, followed by the experimental design in Section 5 and the results analysis in Section 6. The paper concludes in Sections 7 and 8 by summarizing the key findings and identifying the limitations.

2 Related work

TSE has been studied under diverse problem settings, depending on the target variables, sensing technologies, and spatial coverage of available observations. Prior studies have estimated speed (Nie et al., 2023b; Wei et al., 2024; Wu et al., 2021a; Zhou et al., 2025), flow (Nie et al., 2023b; Xing et al., 2024), density (Zhang et al., 2024), or multiple traffic variables (Liu et al., 2024; Saffari et al., 2022; Shi et al., 2022) using fixed detectors (Wei et al., 2024; Wu et al., 2021a; Xing et al., 2024; Zhang et al., 2024; Zhao et al., 2025; Zhou et al., 2025), probe-vehicle data (Jin et al., 2024; Zhang et al., 2024), trajectory observations (Makridis et al., 2025; Wu et al., 2024), or multisource sensing systems (Hu et al., 2025; Liu et al., 2024; Nie et al., 2023b; Saffari et al., 2022). These differences reflect the heterogeneity of real-world traffic monitoring systems and motivate different estimation formulations. This study focuses on network-wide traffic-state estimation under sparse, fixed-sensor deployment. Accordingly, this section reviews methods most relevant to sparse-sensor network-wide estimation, including model-driven approaches and GNN-based traffic kriging models.

2.1 Model-driven approaches

Model-driven TSE infers unobserved traffic states by embedding limited sensor measurements into explicit traffic-flow models. In this paradigm, traffic measurements provide partial observations, a macroscopic traffic flow model describes the spatiotemporal evolution of the traffic system, and a filtering or data assimilation method updates the latent states according to real-time observations. Wang et al. (2022) summarized four essential components of model-driven TSE: traffic measurements, a macroscopic traffic-flow model, a filtering or data-assimilation method, and model-parameter estimation or calibration. The macroscopic model is particularly important under sparse fixed sensing because it establishes a physical connection between a limited number of observed variables and a much larger set of unobserved variables to be estimated.

A major branch of model-driven TSE is based on first-order traffic-flow models. The LWR model describes traffic dynamics using vehicle conservation and a fundamental relationship among

density, speed, and flow (Lighthill and Whitham, 1955; Richards, 1956). Its time-space discretization, the CTM, is widely used because of its computational simplicity and natural compatibility with network-level state propagation (Daganzo, 1994). In sparse-sensing scenarios, CTM-based estimators infer unobserved traffic states by propagating information from fixed detector locations through conservation laws, sending-receiving constraints, and boundary conditions. In recent studies, CTM-based estimators have been combined with Kalman filtering for traffic state reconstruction under mixed sensing or fixed-sensor settings (Herrera and Bayen, 2010), while extended Kalman filtering has been widely used to handle the nonlinear state-space formulation of CTM-based traffic dynamics (Yuan et al., 2012, 2014). Ensemble Kalman filtering has also been adopted to assimilate sparse or heterogeneous traffic measurements into first-order models, particularly in applications involving mobile sensing data (Bucknell and Herrera, 2014; Patire et al., 2015; Seo et al., 2015; Wang and Work, 2014; Work et al., 2008, 2010; Xia et al., 2017). Other studies have explored unscented Kalman filtering (Trinh et al., 2019) and particle filtering for nonlinear and non-Gaussian estimation settings (Wang et al., 2016a, 2016b; Wright and Horowitz, 2016; Xia et al., 2017). Online model parameter estimation has also been incorporated into first-order model-based TSE to address uncertainty in fundamental-diagram parameters and boundary conditions (Chen and Levin, 2019; Xia et al., 2017).

Compared with first-order models, higher-order models can represent richer nonequilibrium traffic behavior, but they also introduce additional state variables and parameters. As a result, they require more careful calibration and can be more sensitive to model specification. Existing higher-order TSE studies remain less common than first-order CTM-based approaches. Payne–Whitham-type models have been combined with extended Kalman filtering to assimilate heterogeneous traffic measurements (Liu et al., 2018; Nanthawichit et al., 2003; Wang et al., 2018). METANET-based estimation has been investigated with unscented Kalman filtering in mixed-sensing scenarios (Agalliadis et al., 2020), while Aw–Rascle–Zhang (ARZ)-based estimators have been developed using ensemble Kalman filtering, extended Kalman filtering, and particle filtering (Kianfar and Edara, 2010; Seo and Bayen, 2017; Wang et al., 2017; Wilkie et al., 2013). Compared with first-order estimators, higher-order models can capture richer nonequilibrium traffic dynamics, but they also require more careful calibration and introduce greater computational and implementation complexity. The filtering method also affects real-time applicability. Kalman filtering and extended Kalman filtering are generally computationally efficient, whereas ensemble Kalman filtering and particle filtering usually require higher computational effort because they rely on ensembles or particles to represent uncertainty. This makes the latter methods more flexible for nonlinear or non-Gaussian systems but potentially less attractive for large-scale real-time deployment.

Another important issue in model-driven TSE is parameter uncertainty. Macroscopic traffic-flow models depend on parameters such as free-flow speed, critical density, and capacity, whose values may vary across locations, time periods, and traffic regimes. These parameters are not fully determined by the model dynamics themselves and must be obtained through offline calibration or online model parameter estimation. Wang et al. (2022) defined online model parameter estimation as the joint online estimation of key model parameters and traffic state variables and report that it has played an important role in fixed-sensor-based freeway TSE.

In summary, model-driven approaches provide physically interpretable and theoretically grounded solutions for TSE under limited sensing. They are particularly effective when traffic-flow models, boundary conditions, and parameters can be reliably specified. However, their performance can degrade when the network is large, observations are sparse, traffic regimes are heterogeneous, or model parameters are uncertain. These limitations motivate data-driven approaches that learn spatiotemporal dependencies directly from historical observations while still exploiting the structural information of the road network.

2.2 Data-driven TSE under sparse sensing

Data-driven methods estimate unobserved traffic states by learning spatial and temporal dependencies directly from historical observations. Compared with model-driven approaches, they require fewer explicit assumptions about traffic-flow dynamics and are more flexible in capturing nonlinear and nonstationary traffic patterns. Under sparse sensor deployment, the key challenge is to infer the states of road segments without direct measurements by exploiting temporal regularity, spatial correlation, and network topology.

Early data-driven studies commonly formulated sparse-sensor traffic estimation as a matrix or tensor completion problem. In these methods, traffic states are organized into matrices or higher-order tensors along spatial and temporal dimensions, and missing values are recovered by exploiting low-rank structures, temporal smoothness, and spatial regularization. For example, Bayesian kernelized matrix factorization incorporates kernel-based priors to model spatiotemporal correlations and uncertainty in missing traffic data (Lei et al., 2022), while Laplacian-regularized tensor completion methods use graph-based smoothness constraints to preserve spatial consistency over road networks (Lyu et al., 2024; Nie et al., 2023a). These methods are statistically interpretable and effective when the sensor layout is fixed. However, they are generally transductive because node-specific latent factors are learned for a predefined network and observation pattern. As a result, they usually require retraining or additional adaptation when sensor deployment changes or unseen nodes are introduced.

Due to the natural analogy between a transport network and a graph, GNNs have emerged as a promising alternative by representing road segments as nodes and explicitly leveraging network topology through message-passing mechanisms, thereby simulating traffic propagation on the road network. Methodologically, GNN-based traffic kriging has progressed from general graph convolutional interpolation toward spatiotemporal graph models, inductive kriging frameworks, adaptive or multirelation dependency learning, and diffusion-based sparse-sensor estimation that captures directed multihop traffic propagation.

Early graph-based kriging methods used graph convolution to interpolate traffic states from observed sensors to unobserved locations, thereby moving beyond purely temporal or low-rank completion models (Appleby et al., 2020). Subsequent studies coupled graph-based spatial aggregation with temporal modeling modules to capture dynamic traffic evolution. For example, spatial aggregation and temporal convolution networks (SATCN) combines spatial aggregation and temporal convolution for real-time kriging under sparse observations (Wu et al., 2021b), while dual spatio-temporal network (DualSTN) separates short- and long-term spatiotemporal dependencies using graph attention and recurrent temporal modeling (Hu et al., 2024). These methods demonstrate that graph-based spatial propagation is beneficial for

traffic kriging, but they still rely heavily on the quality of predefined topology or local aggregation structures.

A subsequent line of research focuses on inductive generalization and adaptive dependency modeling. Inductive graph neural network kriging (IGNNK) formulates traffic kriging as an inductive graph learning problem and reconstructs unobserved sensor states through graph-based message passing (Wu et al., 2021a). Inductive graph representation learning model for spatio-temporal kriging (INCREASE) further improves inductive spatiotemporal kriging by constructing multiple relation-specific graphs, including spatial proximity, functional similarity, and transition dynamics, and learning node representations through relationwise aggregation (Zheng et al., 2023). Spatial-temporal aware data recovery network (STAR) develops a spatial-temporal aware inductive GNN for C-ITS data recovery, emphasizing generalization to partially observed traffic systems (Liang et al., 2023). These studies shift traffic kriging from fixed-layout interpolation toward inductive recovery under changing observation patterns.

Among graph-based methods, diffusion graph convolution is particularly relevant because it models traffic propagation as a directed random-walk process. Diffusion convolutional recurrent neural network (DCRNN) introduced diffusion convolution to capture directional and multihop dependencies in traffic networks (Li et al., 2017), and IGNNK adapted this idea to inductive traffic kriging under sparse observations (Wu et al., 2021a). Later studies addressed limitations of fixed diffusion topology by introducing adaptive graph construction or improved representations for unobserved nodes. Inductive and adaptive graph convolution networks (IAGCN) combines adaptive graph learning with inductive graph convolution to reduce the rigidity of predefined adjacency matrices (Wei et al., 2024), while Dirichlet graph auto-encoder (DGAE) generates embeddings for unobserved nodes through directed graph autoencoding and feature propagation before traffic-state reconstruction (Zhou et al., 2025). These methods indicate that directed multihop diffusion and adaptive dependency modeling are both important for sparse-sensor TSE.

2.3 Research gaps

The above literature show a clear methodological evolution in TSE under sparse sensor deployment. Model-driven approaches provide physical interpretability by embedding sparse measurements into macroscopic traffic-flow models, but they often require calibrated fundamental diagrams, boundary conditions, and model parameters, which limits their flexibility in large-scale networks with heterogeneous traffic regimes (Wang et al., 2022). Matrix and tensor completion methods reduce the reliance on explicit traffic flow equations by exploiting low-rank spatiotemporal structures and graph regularization, but they are usually transductive and tied to fixed sensor layouts (Lei et al., 2022; Lyu et al., 2024; Nie et al., 2023a). Recent GNN-based methods further improve sparse-sensor estimation by using graph message passing, inductive learning, adaptive graph construction, and diffusion-based propagation (Appleby et al., 2020; Wei et al., 2024; Wu et al., 2021a; Zhou et al., 2025). Nevertheless, two issues remain insufficiently addressed.

First, reliable adaptive relation construction remains challenging under sparse fixed-sensor deployment. Existing studies have shown that topology-based diffusion alone is insufficient to capture functional dependencies between road segments and have therefore introduced adaptive correlations from auxiliary traffic patterns (Nie et al., 2023b) or learned node representations (Wei et al., 2024; Zhou et al., 2025). However,

these strategies generally assume that informative features or embeddings are available for constructing relations involving all nodes. In sparse-sensor TSE, traffic observations are reliable only at observed nodes, while unobserved nodes either lack traffic features or rely on initialized or preliminarily estimated representations. As a result, adaptive relations involving unobserved nodes may be weakly supported by actual traffic dynamics, which can reduce the reliability of the learned propagation operator in subsequent multihop diffusion. This makes a direct application of attention-based graph representation learning (e.g., Graphormer) unsuitable for sparse-sensor TSE because attention weights require informative representations at both endpoints, whereas unobserved nodes lack reliable traffic features and may therefore induce weakly supported relations in the resulting attention matrix. Therefore, a key challenge is how to learn reliable functional relations from the observed-node subset and subsequently extend them to the entire graph without relying on unavailable node features.

Second, the structural completeness of inductive training remains insufficiently explored. Random subgraph sampling is commonly adopted in inductive traffic kriging to improve training efficiency and prevent the model from memorizing a fixed graph structure (Wei et al., 2024; Wu et al., 2021a; Zhou et al., 2025). However, randomly sampled subgraphs may break local neighborhood continuity and remove edges that would otherwise participate in message passing. As a result, node embeddings are learned from structurally incomplete local contexts during training, whereas inference is conducted on the full graph. This discrepancy may create a mismatch between training-time and inference-time propagation patterns. Although existing studies have introduced auxiliary prediction constraints or feature-propagation mechanisms to alleviate information loss under sparse observations (Wei et al., 2024; Zhou et al., 2025), the structural information loss caused by subgraph sampling itself has received relatively limited attention.

These two gaps suggest that TSE under sparse sensor deployment requires not only better feature propagation but also more reliable relation construction and structure-preserving inductive learning.

3 Preliminaries

3.1 Problem definition

The traffic network is conceptualized as a graph $G = (V, E)$ with $|V| = N$ nodes (road segments) and directed edges encoding physical connectivity. The topological adjacency matrix is denoted by A , where each entry reflects the ease of travel from one node (road segment) to another. Throughout this study, we use the convention that $A_{ij} > 0$ denotes a directed edge from node i to node j . In practice, this adjacency matrix is typically constructed using a Gaussian kernel applied to the routing distances between nodes. At a discrete time step τ , the full traffic state is denoted as $X_\tau \in \mathbb{R}^{N \times F}$ (e.g., speed, flow, or density at each node). However, only a subset of entries is observed by sensors, as indicated by a binary mask $M \in \{0, 1\}^N$. The observed data are given by $Y_\tau = M \odot X_\tau$, $\tau = t - T + 1, \dots, t$, where T denotes the length of the observation window and $\odot$ represents the Hadamard product. No ground truth is assumed to be available at unobserved nodes or time steps within this window. The objective of this study is TSE, i.e., reconstructing the complete network-wide traffic states $\hat{X}_{t-T+1:t} \in \mathbb{R}^{T \times N \times F}$ from the observations $Y_{t-T+1:t}$. The proposed model learns the mapping defined in Eq. (1):

$$f_\theta : \{Y_{t-T+1:t}, A\} \mapsto \{\hat{X}_{t-T+1:t}\} \quad (1)$$

3.2 Diffusion convolutional GNNs

The DGCN, first proposed by Li et al. (2017) and first utilized by Wu et al. (2021a) for the traffic kriging task, models the propagation of traffic states in a road network as a diffusion process and provides a principled perspective for information aggregation among neighboring nodes. In this setting, the diffusion process on a graph G is characterized by a random walk with restart probability $\alpha \in [0, 1]$ and state transition matrices $D_o^{-1}W$ and $D_i^{-1}W^T$, where D_o and D_i denote the out- and in-degree diagonal matrices, respectively, and W is the weighted adjacency matrix of the graph.

Formally, the diffusion convolution operator is defined as

$$X_{:,p^*Gf_\theta} = \sum_{k=0}^{K-1} (\theta_{k,1}(D_o^{-1}W)^k + \theta_{k,2}(D_i^{-1}W^T)^k) X_{:,p}, \quad (2)$$

for $p \in \{1, 2, \dots, d_i\}$

where d_i is the input feature dimension; $\theta_{k,1}$ and $\theta_{k,2}$ are learnable filter parameters; and K controls the receptive field. The diffusion convolution layer then becomes

$$H_{:,q} = \sigma \left(\sum_{p=1}^{d_i} X_{:,p^*Gf_{\theta_{q,p},:}} \right), \quad \text{for } q \in \{1, 2, \dots, d_h\} \quad (3)$$

where $X \in \mathbb{R}^{N \times d_i}$ is the input; $H \in \mathbb{R}^{N \times d_h}$ is the output; and $\Theta \in \mathbb{R}^{d_h \times d_i \times K \times 2}$ denotes learnable parameters.

As shown by Li et al. (2017), the widely used spectral ChebNet (Defferrard et al., 2016) can be interpreted as a special case of diffusion graph convolution (up to a similarity transformation) when G is undirected. Compared with spectral graph convolutional networks (GCNs), diffusion convolution naturally captures directed flows, which are essential in traffic networks where upstream congestion causally influences downstream conditions.

3.3 Graphormer for graph representation learning

Graphormer, first proposed by Ying et al. (2021), is a transformer-based graph representation model that reformulates graph learning as a fully connected attention mechanism augmented with structural encoding. Unlike conventional GNNs, which restrict message passing to one-hop neighbors and thus suffer from oversmoothing or limited receptive fields, Graphormer introduces a series of structure-aware biases into the standard transformer architecture. This design enables the model to capture long-range dependencies while preserving the underlying graph topology. Formally, given a graph $G = (V, E)$ with node features $X \in \mathbb{R}^{N \times d}$, Graphormer treats the nodes as a sequence and applies multihead self-attention, where the attention logits are enhanced with graph-derived structural biases.

Each node i is assigned an additional centrality encoding, which is incorporated into its input representation as

$$h_i^{(0)} = x_i + \text{Embed}(\deg_{\text{in}}(i)) + \text{Embed}(\deg_{\text{out}}(i)) \quad (4)$$

where $\text{Embed}(\deg_{\text{in}}(i))$ and $\text{Embed}(\deg_{\text{out}}(i))$ are learnable embedding vectors parameterized by the in-degree and out-degree of node i , respectively. This formulation encodes the structural importance of nodes (e.g., hubs or intersections) in a degree-aware manner.

For a node pair (i, j) , the attention score for a single head is computed as

$$\alpha_{ij} = \frac{(h_i \mathbf{W}_Q)(h_j \mathbf{W}_K)^T}{\sqrt{d}} + b_{\phi(v_i, v_j)} + \frac{1}{N} \sum_{n=1}^N x_{e_n}(\mathbf{w}_n^E)^T \quad (5)$$

where $\frac{(h_i \mathbf{W}_Q)(h_j \mathbf{W}_K)^T}{\sqrt{d}}$ denotes the standard scaled dot-product attention. The term $b_{\phi(v_i, v_j)}$ is a learnable scalar (spatial encoding bias) indexed by the structural relation $\phi(v_i, v_j)$ (e.g., the shortest-path distance between nodes i and j). The final term $\frac{1}{N} \sum_{n=1}^N x_{e_n}(\mathbf{w}_n^E)^T$ represents an edge-encoding bias, computed by averaging over the edges e_n along the shortest path between nodes i and j , where x_{e_n} denotes the feature of edge e_n ; and $\mathbf{w}_n^E$ is the corresponding learnable projection vector.

Experiments reported by Ying et al. (2021) demonstrate the superiority of Graphormer over traditional GNN models in graph-level prediction tasks. Beyond encoding structural priors, the bias-augmented attention scores themselves provide a latent representation of node–node relationships. Specifically, since α_{ij} integrates (1) feature similarity, (2) shortest-path structural information, and (3) edge attributes, the resulting attention matrix captures a data-driven measure of relational strength between nodes in a latent space that extends beyond one-hop neighborhood constraints. This property can be further leveraged for the relation diffusion process proposed in this study.

3.4 K-order subgraph sampling for inductive training

To enable inductive learning on graph-structured data, a common strategy is to employ subgraph sampling during training. This approach prevents the model from memorizing the global adjacency matrix and encourages it to learn local, transferable relational patterns and message-passing mechanisms. As a result, the trained model can generalize to unseen nodes and newly formed subgraphs in inductive settings, as each training iteration simulates operating on a partial, localized view of the original graph.

Randomly sampling a fixed number of nodes and constructing the subgraph over this subset has been adopted in several previous studies on the TSE problem (Wei et al., 2024; Wu et al., 2021a; Zhou et al., 2025). However, this strategy disrupts the structural coherence of the graph due to the loss of local connectivity and the lack of control over the receptive field. In contrast, K -order subgraph sampling, originally introduced as the fundamental inductive learning unit in the graph sample and aggregate (GraphSAGE) model (Hamilton et al., 2017), preserves meaningful locality, stable and consistent structural patterns, and a natural hierarchical organization from L_0 to L_K . The K -order subgraph sampling process is formally described in Algorithm 1, and an illustrative example with $K = 3$ is shown in Fig. 1. Compared with random node sampling, K -order sampling better preserves local neighborhood continuity, thereby alleviating the structural mismatch between sampled training subgraphs and full-graph inference discussed in Section 2.3.

4 Methodology

The proposed RADGCN is illustrated in Fig. 2. The model aims to reconstruct the spatiotemporal traffic states of an entire road network when only a subset of nodes is observed over a given time period. To achieve this objective, RADGCN integrates three complementary sources of information: (1) temporal dynamics of observed nodes, (2) structural and geometric priors encoded in the road network, and (3) data-driven relational patterns inferred from observed nodes.

The overall workflow proceeds as follows: The input for each

Algorithm 1 K -order subgraph random sampling (directed, weighted)

Require: Digraph $G = (V, E, \mathbf{w})$, order $K \geq 0$, direction mode in {out, in, both}, optional left c

Ensure: Center c , layers $\{L_0, L_1, \dots, L_K\}$, induced subgraph D , hop-edge sets $\{E_1, E_2, \dots, E_K\}$

- 1: **if** c not given **then** pick c uniformly from V
- 2: $L_0 \leftarrow \{c\}$, $\text{dist}(c) \leftarrow 0$, queue $Q \leftarrow [c]$
- 3: **while** Q not empty **do**
- 4: $u \leftarrow Q.\text{pop}()$; **if** $\text{dist}(u) = K$ **then continue**
- 5: $N(u) \leftarrow \begin{cases} \text{succ}(u), & \text{mode} = \text{out} \\ \text{pred}(u), & \text{mode} = \text{in} \\ \text{succ}(u) \cup \text{pred}(u), & \text{mode} = \text{both} \end{cases}$
- 6: **for all** $v \in N(u)$ **do**
- 7: **if** $v \notin \text{dist}$ **then**
- 8: $\text{dist}(v) \leftarrow \text{dist}(u) + 1$; $L_{\text{dist}(v)} \leftarrow L_{\text{dist}(v)} \cup \{v\}$;
- 9: $Q.\text{push}(v)$
- 10: **end if**
- 11: **end while**
- 12: **for** $i = 1$ to K **do**
- 13: $E_i \leftarrow \{(x \rightarrow y) \in E \mid (x \in L_{i-1}, y \in L_i) \vee (\text{mode} = \text{both} \wedge x \in L_i, y \in L_{i-1})\}$
- 14: **if** mode = in **then** $E_i \leftarrow \{(x \rightarrow y) \in E \mid x \in L_i, y \in L_{i-1}\}$
- 15: **end for**
- 16: $H \leftarrow G \left[\bigcup_{i=0}^K L_i \right]$; **return** $(c, \{L_i\}_{i=0}^K, H, \{E_i\}_{i=1}^K)$

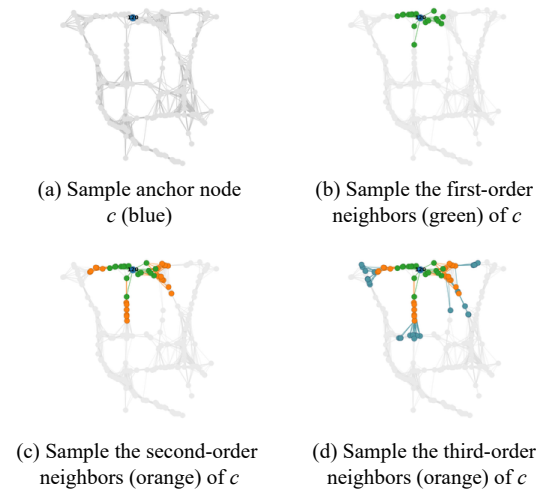


Fig. 1 Illustrative process of the 3rd-order subgraph sampling.

sampled subgraph consists of a static directed adjacency matrix $A \in \mathbb{R}^{N \times N}$ and a partially observed temporal signal $X_{\text{obs}} \in \mathbb{R}^{N_{\text{obs}} \times T \times d}$. The fragmented temporal dynamics of each observed node within the current time window are first encoded using a causal temporal convolutional network (TCN). Subsequently, Graphormer incorporates graph-theoretic biases—such as edge embeddings, shortest-path distances, and centrality encodings—through global self-attention. The resulting attention scores, defined over observed nodes $A_{\text{OO}}^{\text{attn}}$, are then extrapolated to all nodes via a bilateral diffusion operator derived from the directed traffic transition matrix. This process yields a dense, learned dynamic relation matrix $\tilde{A}_{\text{all}}^{\text{attn}}$ that augments the static graph topology.

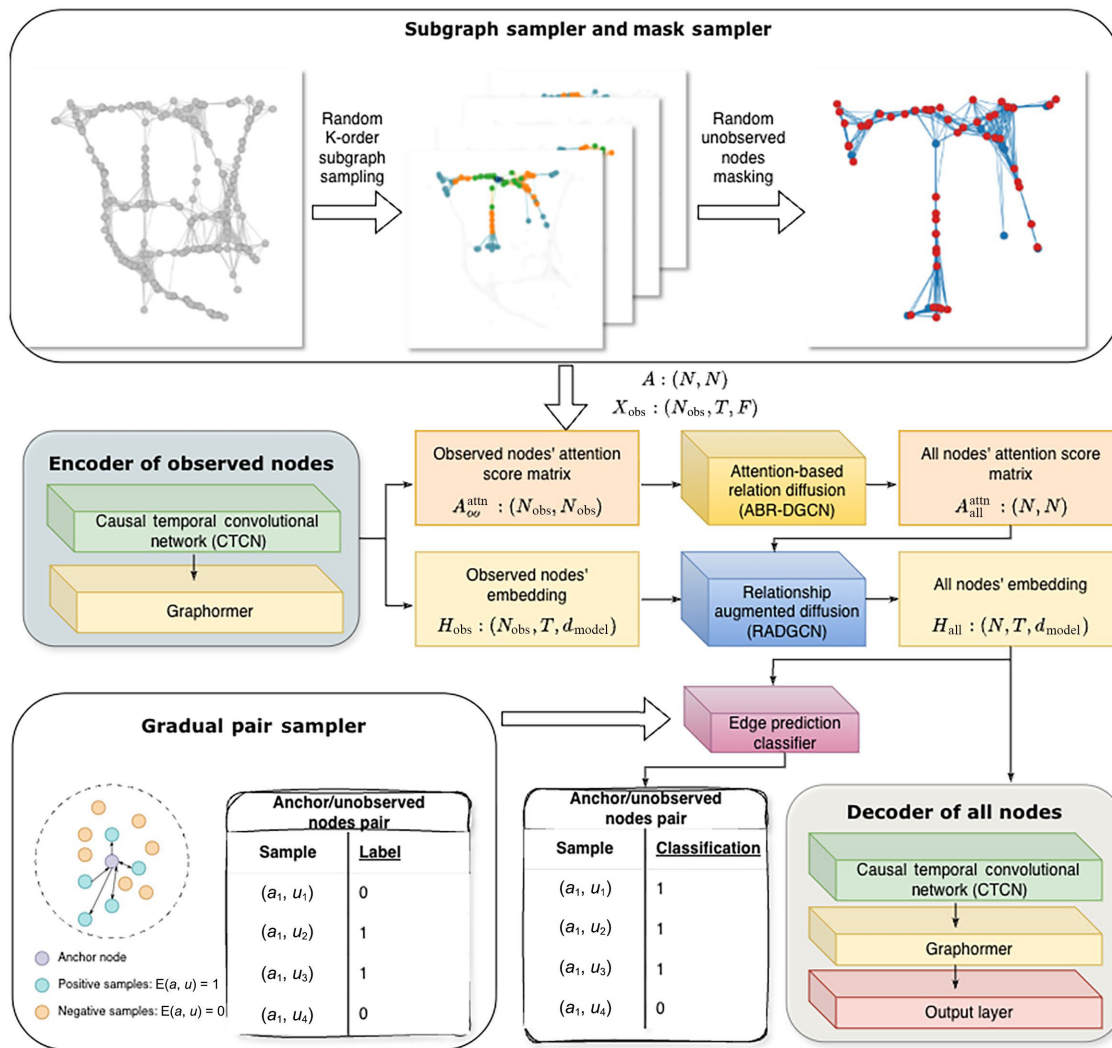


Fig. 2 Overall architecture of RADGCN.

Subsequently, relation-augmented diffusion layers (RADGCN) propagate information across the entire network, and a temporal decoder reconstructs the full spatiotemporal signals for all nodes. To further enhance the semantic expressiveness of node embeddings and mitigate the incompleteness of structural information caused by subgraph sampling, an auxiliary edge-prediction task is jointly optimized with the kriging objective, using the learned node embeddings as input. The following subsections provide a detailed description of the core components.

The design of RADGCN directly responds to the two research gaps identified in Section 2.3. To address the weak observational support of adaptive relations involving unobserved nodes, RADGCN separates relation learning from feature propagation: functional relations are first inferred among observed nodes and then diffused to the full graph. To reduce structural information loss during inductive training, RADGCN adopts K -order subgraph sampling and an auxiliary edge prediction task so that local neighborhood structure and pairwise connectivity information can be better preserved during representation learning.

4.1 Attention-based relation diffusion graph convolution network (ABR-DGCN)

This module targets the first research gap identified in Section 2.3, namely, the difficulty of constructing reliable adaptive relations when many nodes are unobserved. Existing studies have

recognized that physical topology alone cannot fully characterize dynamic traffic similarity. Adjacent road segments may exhibit different temporal patterns due to lane merging, bottlenecks, or heterogeneous traffic flows, whereas structurally distant segments may show highly correlated traffic evolution. This observation motivates the use of adaptive graphs or learned node dependencies beyond topology-based proximity. However, constructing such adaptive relations becomes difficult under sparse fixed-sensor deployment. Since traffic features are available only at observed nodes, the dynamic similarity involving unobserved nodes cannot be directly estimated from reliable traffic measurements. If the full adaptive graph is constructed from zero-filled inputs, initialized embeddings, or preliminarily propagated features, the resulting relations associated with unobserved regions may be weakly supported by actual traffic dynamics. This limits the reliability of adaptive graph construction and may affect subsequent diffusion-based propagation.

A related challenge is that extensive unobserved regions can also weaken feature-level message passing. In conventional GNN or DGCN models, information from observed nodes is propagated layer by layer through local neighborhoods. When many intermediate nodes are unobserved, they provide little informative traffic signal and may form "unobserved islands". Nodes located inside such regions can receive observed information only through multiple stacked propagation layers, and the signal may become diluted or distorted along the path.

Consequently, long-range TSE becomes difficult even when the physical network is connected.

These observations motivate the design of ABR-DGCN. The key idea is that although unobserved nodes lack feature information, observed nodes provide reliable functional relationships, which can be captured by the graphormer attention matrix (Section 3.3). However, this attention matrix only characterizes relations within the observed subset and cannot directly guide message passing in unobserved regions. To address this limitation, ABR-DGCN propagates the spatiotemporal relations learned among observed nodes to the entire graph via directed multihop diffusion kernels. This design is inspired by the diffusion convolution operator in Li et al. (2017), where spatial dependencies are modeled using bidirectional random walks with a forward transition matrix normalized by out-degree and a backward transition matrix normalized by in-degree. It is further motivated by the gradient-flow formulation proposed in Zhou et al. (2025), which explicitly separates observed and unobserved nodes. Building upon these ideas, this study extends the notion of diffusion from feature propagation to relation propagation.

Concretely, let $\mathbf{A}_{OO}^{\text{attn}}$ denote the bias-augmented attention matrix obtained from the graphormer over the observed node set O . Instead of applying diffusion to node representations $\mathbf{H}$ as in Eq. (3), diffusion operators are applied bilaterally to the source (query) and target (key) dimensions of the relation matrix. This results in the following diffusion-based lifting of $\mathbf{A}_{OO}^{\text{attn}}$ into a full-graph relation operator:

$$\tilde{\mathbf{A}}_{\text{full}}^{\text{attn}} = \sum_{k=0}^{K-1} \sum_{l=0}^{K-1} \boldsymbol{\Theta}_{kl} (\mathbf{P}_f^k)_{:,O} \mathbf{A}_{OO}^{\text{attn}} (\mathbf{P}_b^l)^T_{:,O} \quad (6)$$

where $\boldsymbol{\Theta}_{kl}$ are learnable coefficients; and $(\cdot)_{(:,O)}$ denotes the operator that extracts columns corresponding to the observed node set O . Under convention $\mathbf{A}_{ij} > 0$ for the directed edge $i \rightarrow j$, $\mathbf{P}_f$ denotes the row-normalized forward transition matrix, while $\mathbf{P}_b$ is defined from $\mathbf{A}^T$ and represents the corresponding backward transition matrix. Equation (6) computes a weighted combination of forward-backward diffusion paths that originate from and terminate at observed nodes, thereby propagating spatiotemporal dynamic similarities across directed multihop networks. Since Eq. (6) yields an $N \times N$ matrix, it admits a natural block decomposition consistent with the partition of nodes into observed and unobserved subsets, i.e., $V = O \cup U$. Accordingly, the diffusion operators restricted to columns associated with O can be expressed as

$$(\mathbf{P}_f^k)_{(:,O)} = \begin{bmatrix} \mathbf{P}_{f,OO}^k \\ \mathbf{P}_{f,UO}^k \end{bmatrix}, \quad (\mathbf{P}_b^l)_{(:,O)} = \begin{bmatrix} \mathbf{P}_{b,OO}^l \\ \mathbf{P}_{b,UO}^l \end{bmatrix} \quad (7)$$

where $\mathbf{P}_{f,OO}^k \in \mathbb{R}^{|O| \times |O|}$ represents the k -th order forward diffusion from O to O , obtained by selecting the rows and columns corresponding to O from $\mathbf{P}_f^k$. Similarly, $\mathbf{P}_{f,UO}^k \in \mathbb{R}^{|U| \times |O|}$ models forward diffusion from O to U , while $\mathbf{P}_{b,OO}^l$ and $\mathbf{P}_{b,UO}^l$ denote the corresponding blocks for backward diffusion. The resulting full attention matrix derived from $\mathbf{A}_{OO}^{\text{attn}}$ can be further decomposed as

$$\tilde{\mathbf{A}}_{\text{full}}^{\text{attn}} = \begin{bmatrix} \tilde{\mathbf{A}}_{OO}^{\text{attn}} & \tilde{\mathbf{A}}_{OU}^{\text{attn}} \\ \tilde{\mathbf{A}}_{UO}^{\text{attn}} & \tilde{\mathbf{A}}_{UU}^{\text{attn}} \end{bmatrix} \quad (8)$$

where each block is given explicitly by

$$\tilde{\mathbf{A}}_{OO}^{\text{attn}} = \sum_{k,l} \boldsymbol{\Theta}_{kl} \mathbf{P}_{f,OO}^k \mathbf{A}_{OO}^{\text{attn}} (\mathbf{P}_{b,OO}^l)^T \quad (9)$$

$$\tilde{\mathbf{A}}_{OU}^{\text{attn}} = \sum_{k,l} \boldsymbol{\Theta}_{kl} \mathbf{P}_{f,OO}^k \mathbf{A}_{OO}^{\text{attn}} (\mathbf{P}_{b,UO}^l)^T \quad (10)$$

$$\tilde{\mathbf{A}}_{UO}^{\text{attn}} = \sum_{k,l} \boldsymbol{\Theta}_{kl} \mathbf{P}_{f,UO}^k \mathbf{A}_{OO}^{\text{attn}} (\mathbf{P}_{b,OO}^l)^T \quad (11)$$

$$\tilde{\mathbf{A}}_{UU}^{\text{attn}} = \sum_{k,l} \boldsymbol{\Theta}_{kl} \mathbf{P}_{f,UO}^k \mathbf{A}_{OO}^{\text{attn}} (\mathbf{P}_{b,UO}^l)^T \quad (12)$$

This decomposition highlights the functional roles of bilateral diffusion. Specifically, $\tilde{\mathbf{A}}_{OO}^{\text{attn}}$ refines relationships among observed nodes; $\tilde{\mathbf{A}}_{OU}^{\text{attn}}$ and $\tilde{\mathbf{A}}_{UO}^{\text{attn}}$ characterize cross-boundary functional interactions between observed and unobserved regions; and $\tilde{\mathbf{A}}_{UU}^{\text{attn}}$ provides a data-informed relational prior for entirely unobserved areas where no temporal information is available. Consequently, the ABR-DGCN module yields a mathematically principled relational kernel that complements the physical topology and enables robust message passing in subsequent RADGCN layers, even under sparse or spatially fragmented observations.

4.2 Relationship-augmented diffusion graph convolution

Following the construction of the full relation operator $\tilde{\mathbf{A}}_{\text{full}}^{\text{attn}}$ from ABR-DGCN, which encodes functional affinities between all node pairs, this operator is further integrated into the spatial diffusion convolution framework by extending the DGCN formulation (Li et al., 2017). In RADGCN, ABR learning and diffusion-based feature propagation are deliberately separated. Graphormer is used only to infer functional dependencies among observed nodes, while DGCN is used to propagate node features through transition kernels. The learned relation matrix therefore does not replace feature diffusion; rather, after normalization, it serves as an additional transition kernel that augments topology-based forward and backward diffusion. The propagated relation matrix is introduced into diffusion convolution to complement topology-based propagation. Physical adjacency describes local traffic connectivity, whereas the learned relation matrix provides a functional propagation channel derived from observed traffic dynamics. By using the learned relation kernel as an additional transition operator, RADGCN allows feature diffusion to be jointly guided by physical topology and observation-supported functional relations. It should be noted that the learned attention matrix $\tilde{\mathbf{A}}_{\text{full}}^{\text{attn}}$ cannot be directly used as a diffusion operator. Attention scores describe pairwise relevance between node representations, whereas diffusion convolution requires a valid transition kernel that preserves nonnegativity and stochasticity. Therefore, the learned relation matrix is further transformed into a normalized relation diffusion kernel before being incorporated into the diffusion process. Thus, a row-stochastic relation kernel is constructed to ensure compatibility with random-walk operators:

$$\mathbf{P}_r = \mathbf{D}_R^{-1} \phi(\tilde{\mathbf{A}}_{\text{full}}^{\text{attn}}) \quad (13)$$

where $\mathbf{D}_R^{-1} = \text{diag}(\phi(\tilde{\mathbf{A}}_{\text{full}}^{\text{attn}}) \mathbf{1})$ follows the standard row-normalization scheme in DGCN; and $\phi(\cdot)$ denotes an elementwise nonlinearity (e.g., rectified linear unit (ReLU) or sigmoid) that enforces nonnegativity. This prevents the powers $\mathbf{P}_f^k$ from becoming sign-oscillatory and ensures the model's mathematical consistency. The resulting powers $\mathbf{P}_r^k$ encode multihop diffusion in the relation-induced space, analogous to $\mathbf{P}_f^k$

and $\mathbf{P}_b^k$ in the structural diffusion space.

Thus, the RA-DGCN layer can be formulated as

$$\mathbf{H}^{(l+1)} = \sigma \left(\sum_{k=0}^{K-1} \mathbf{P}_f^k \mathbf{H}^{(l)} \mathbf{W}_k^f + \sum_{k=0}^{K-1} \mathbf{P}_b^k \mathbf{H}^{(l)} \mathbf{W}_k^b + \sum_{k=0}^{K-1} \mathbf{P}_r^k \mathbf{H}^{(l)} \mathbf{W}_k^r \right) \quad (14)$$

where $\mathbf{W}_k^f, \mathbf{W}_k^b, \mathbf{W}_k^r \in \mathbb{R}^{d \times d}$ are learnable weights; and $\sigma(\cdot)$ is a nonlinear activation. To avoid any confusion regarding superscripted and subscripted letters, we clarify that in Eq. (14), f, b, and r denote the indices of forward, backward, and relation diffusion kernels, respectively, while k denotes the power. All three kernels are defined over the full graph, such that node embeddings are updated simultaneously within each hidden layer. The first two kernels model directed physical propagation along the network, whereas the third kernel captures functional coupling revealed by temporal dynamics and extrapolated to all nodes by ABR-DGCN.

From a modeling perspective, RA-DGCN disentangles three complementary modes of spatial interaction: (1) Forward structural diffusion ($\mathbf{P}_f^k$), which represents the downstream propagation of traffic conditions along the direction of traffic flow; (2) backward structural diffusion ($\mathbf{P}_b^k$), which captures upstream influence and shockwave effects; and (3) relation diffusion ($\mathbf{P}_r^k$), which propagates information along paths that are not necessarily short or physically adjacent but are supported by the relation-induced functional similarity encoded in $\tilde{\mathbf{A}}_{\text{full}}^{\text{attn}}$. By combining these three kernels additively in Eq. (14), RA-DGCN enables node embeddings to jointly integrate physical connectivity and data-driven functional relationships. This capability is particularly important under sparse observations: Even when structural paths traverse large unobserved regions, relation diffusion provides additional channels through which information from observed nodes can propagate to distant, unmonitored parts of the network.

4.3 Edge prediction task for node embedding refinement

The auxiliary edge prediction task is designed to address the second research gap discussed in Section 2.3, namely, the structural information loss caused by subgraph-based inductive training. Although K -order subgraph sampling preserves neighborhood continuity better than random sampling, the sampled subgraphs used during training still provide only partial structural contexts. As a result, node embeddings are learned from incomplete local neighborhoods and may not sufficiently capture connectivity patterns that are important for subsequent diffusion-based estimation. To compensate for this limitation, an auxiliary edge prediction task is introduced to provide additional structural supervision and encourage the learned embeddings to preserve meaningful local graph structure.

To mitigate this issue, we introduce an auxiliary edge prediction task to regularize the node embeddings produced by the relation-augmented diffusion module, rather than to introduce additional traffic-state supervision. The intuition is structural consistency: if two nodes are connected in the underlying directed road network, this connectivity should be reflected in their learned representations. Conversely, disconnected node pairs should be distinguishable in the embedding space. Therefore, the auxiliary task directly supervises pairwise connectivity and encourages the learned embeddings to retain structural semantics under sampled-subgraph training.

Let $\mathbf{H}^{(\text{final})} \in \mathbb{R}^{N \times d}$ denote the final-layer node embeddings produced by the relation-augmented diffusion module. For each ordered node pair (i, j) sampled from the current training

subgraph, where both i and j belong to the observed-node training graph, a feature vector is constructed to capture multiple complementary relational aspects between the two embeddings:

$$\mathbf{z}_{ij} = [\mathbf{h}_i, \mathbf{h}_j, \mathbf{h}_i - \mathbf{h}_j, \mathbf{h}_j - \mathbf{h}_i, \mathbf{h}_i \odot \mathbf{h}_j] \in \mathbb{R}^{5d} \quad (15)$$

where $\odot$ denotes the elementwise Hadamard product. This representation enables the classifier to exploit absolute information $(\mathbf{h}_i, \mathbf{h}_j)$, directional differences $(\mathbf{h}_i - \mathbf{h}_j)$ and $(\mathbf{h}_j - \mathbf{h}_i)$, and multiplicative interactions $(\mathbf{h}_i \odot \mathbf{h}_j)$, which have been shown to be effective in link prediction and metric learning. Subsequently, a multilayer perceptron $f_\theta: \mathbb{R}^{5d} \rightarrow (0, 1)$ maps the concatenated vector to an estimated probability of the directed edge $i \rightarrow j$:

$$s_{ij} = f_\theta(\mathbf{z}_{ij}) \quad (16)$$

To train this classifier, an anchor-based hard negative sampling strategy (Algorithm 2) is developed, inspired by the triplet construction in FaceNet (Schroff et al., 2015). Unlike FaceNet, where positive and negative samples are defined by class labels, samples in this setting are defined based on edge existence. Specifically, each node in the sampled training subgraph is treated as an anchor in turn, and positive and negative node pairs are generated only within this subgraph using the subgraph-restricted adjacency matrix. Some nodes in the sampled subgraph may have their traffic values temporarily masked as dummy unobserved nodes for the estimation loss. However, these nodes still belong to the observed-node training graph and are distinct from the held-out unobserved nodes used for testing. Therefore, the auxiliary edge prediction task never uses nodes or edges associated with the held-out unobserved set.

For each anchor node, all existing outgoing edges are included as positive samples, ensuring complete supervision over its local structural neighborhood. Negative samples are drawn exclusively from spatially proximate but nonadjacent nodes, identified using a Euclidean distance threshold. Sampling is performed with replacement, allowing the model to repeatedly encounter challenging negative instances that are geometrically close to true neighbors. This design produces hard negative pairs that are structurally plausible yet incorrect, forcing the classifier to rely on learned node embeddings rather than spatial proximity alone to distinguish true edges from nonedges. As a result, the auxiliary edge-prediction task provides a direct structural regularization signal, thereby improving the preservation of connectivity semantics during subgraph-based training.

4.4 Complexity analysis

We analyze the computational complexity of RADGCN with respect to the graph size and model dimensions. Let N be the number of nodes in the input graph, N_o is the number of observed nodes, d is the hidden dimension, and K is the diffusion order. During training, these quantities refer to the sampled subgraph, whereas during inference, they refer to the full graph.

The temporal convolution module scales linearly with the number of processed nodes and the input sequence length. Its complexity is approximately $O(NTd^2)$, where T is the input time window. The graph-based relation learning module is applied only to observed nodes. Therefore, its self-attention cost is $O(N_o^2 d)$, which is lower than applying attention to all nodes. The main additional cost of RADGCN comes from relation diffusion. The learned relation matrix among observed nodes is propagated to the full graph through bilateral diffusion, producing a relation-aware transition matrix over all nodes. In the general dense case, this operation may require up to $O(K^2 N^2 N_o)$ time and $O(N^2)$

Algorithm 2 Anchor-based local negative sampling for node pairs in a sampled training subgraph

Require: Node set V , adjacency matrix A , Euclidean distance matrix $\text{dist}(\cdot, \cdot)$, number of local negatives per anchor K_{neg} , distance threshold τ .

Ensure: Sampled positive pairs $\mathcal{E}_{\text{sample}}^+$ and local negative pairs $\mathcal{E}_{\text{sample}}^-$

```

1:  $\mathcal{E}_{\text{sample}}^+ \leftarrow \emptyset, \mathcal{E}_{\text{sample}}^- \leftarrow \emptyset$ 
2: for each anchor node  $i \in V$  do
3:   Identify positive neighbors of  $i$ :  $N^+(i) = \{j \in V | A_{ij} = 1\}$ 
     (directed edges  $i \rightarrow j$ )
4:   for each  $j \in N^+(i)$  do
5:     Add  $(i, j)$  to  $\mathcal{E}_{\text{sample}}^+$ 
6:   end for
7:   Construct local negative candidate set around  $i$ :
      $C_{\text{local}}(i) = \{j' \in V \setminus i | A_{ij'} = 0, \text{dist}(i, j') < \tau\}$ 
8:   if  $C_{\text{local}}(i) \neq \emptyset$  then
9:     for  $k = 1$  to  $K_{\text{neg}}$  do
10:      Sample  $j' \in C_{\text{local}}(i)$  with replacement
11:      Add  $(i, j')$  to  $\mathcal{E}_{\text{sample}}^-$ 
12:    end for
13:  end if
14: end for
15: return  $\mathcal{E}_{\text{sample}}^+, \mathcal{E}_{\text{sample}}^-$ 

```

memory. After the relation matrix is obtained, each RADGCN layer performs diffusion over both physical transition matrices and the learned relation transition matrix. For a general graph, each diffusion operator over an $N \times N$ transition matrix costs $O(KN^2d)$ in the dense case. RADGCN uses the physical forward and backward transition matrices together with the learned relation transition matrix, so its diffusion cost increases linearly with the number of transition operators used. Combining the main components, the dense-case computational complexity of one forward pass can be summarized as $O(NTd^2 + N_o^2d + K^2N^2N_o + KN^2d)$.

In practice, the proposed model is trained with sampled subgraphs, so the above complexity is applied to the sampled subgraph rather than the entire network during training. This substantially reduces the training cost and avoids full-graph messages passing at each iteration. During inference, the model can be applied to the full graph. For larger networks, the computational bottleneck can be further reduced by sparsifying the learned relation matrix, retaining only top-ranked relation weights, or performing blockwise relation diffusion. Thus, the additional cost of RADGCN over topology-only DGCN variants mainly comes from relation construction and relation-aware diffusion, representing a trade-off between computational efficiency and adaptive dependency modeling under sparse sensor deployment.

5 Experimental design

5.1 Dataset description

To evaluate the performance of the proposed model, three real-world traffic datasets, i.e., METR-LA, PEMS-BAY, and PEMS7M, are utilized, all of which have been widely adopted in prior studies (Wei et al., 2024; Wu et al., 2021a; Zhou et al., 2025). The key statistics of these datasets are summarized in Table 1. As

indicated in Table 1, the three datasets are broadly comparable in temporal resolution, as all of them record traffic speed at intervals of 5 min. However, they differ notably in scale and variability. PEMS-BAY is the largest in terms of sensor coverage and sequence length, whereas METR-LA exhibits the highest speed variability, indicating a more challenging traffic regime. In contrast, PEMS-BAY appears relatively smooth, with a higher average speed and lower variability. Another important distinction lies in data completeness: METR-LA contains a nonnegligible proportion of missing data, whereas PEMS-BAY is nearly complete and PEMS7M contains no missing values. Taken together, these differences provide a diverse range of operating conditions for evaluating the robustness of the proposed method across networks with varying scales, dynamics, and data quality.

Table 1 Statistics of datasets

Dataset	Sensor	Time step	Mean	Std	Min	Max	Missing value
METR-LA	207	34,272	53.71	20.26	0.00	70.00	575,302 (8.11%)
PEMS-BAY	325	52,116	62.62	9.59	0.00	85.10	521 (0.00%)
PEMS7M	228	12,672	58.89	13.48	3.00	82.60	0 (0.00%)

Note: Std: standard deviation; Min: minimum; Max: maximum.

During data preprocessing, the unobserved sensors are randomly sampled from the dataset with a ratio of 25% for baseline experiments. The values in unobserved sensors and the existence of unobserved sensors are not leaked during the training and validation process, which is consistent with the inductive training setting in previous studies (Wei et al., 2024; Wu et al., 2021a; Zhou et al., 2025). Fig. 3 visualizes the spatial distribution of observed and unobserved nodes in the three datasets based on one representative random sensor split, with node color indicating speed variance. In all cases, the nodes follow corridor-like road layouts rather than forming uniformly distributed networks, which makes the estimation task inherently uneven across space. The observed and unobserved nodes are interleaved over most of the network, so the problem is not simply one of filling isolated gaps but of inferring states over partially monitored corridors with varying local support. At the same time, the speed variance is clearly heterogeneous, with higher-variance nodes clustering along main segments and junction areas, while other parts of the network remain relatively stable. Since a larger variance usually reflects stronger temporal fluctuations and less stable traffic patterns, these regions are also expected to pose greater difficulty for accurate estimation. This suggests that the datasets contain both smooth and highly dynamic regions, so the model must handle not only spatial sparsity but also substantial variability in local traffic dynamics.

5.2 Implementation details

For each scenario, the dataset is divided into training, validation, and test sets with a ratio of 0.5/0.2/0.3. Within each subset, the time series data are segmented into 12-step windows, corresponding to 1 h. The estimation task is performed within the same temporal window for unobserved nodes. The primary experiment is conducted with an observation ratio of $r_{\text{obs}} = 75\%$, where $r_{\text{unobs}} = 25\%$ of nodes are removed from both the input data and the adjacency matrix during the training and validation stages. In subsequent robustness experiments, the unobserved node ratio is increased to 50%, 75%, and 90%. During training, the remaining nodes are further sampled into subgraphs using a K -order neighborhood sampling strategy. This process generates subgraphs that retain $r_{\text{sub}} = 75\%$ of the remaining nodes, a setting kept constant across robustness experiments to avoid excessively limiting the number of nodes available for message passing. No

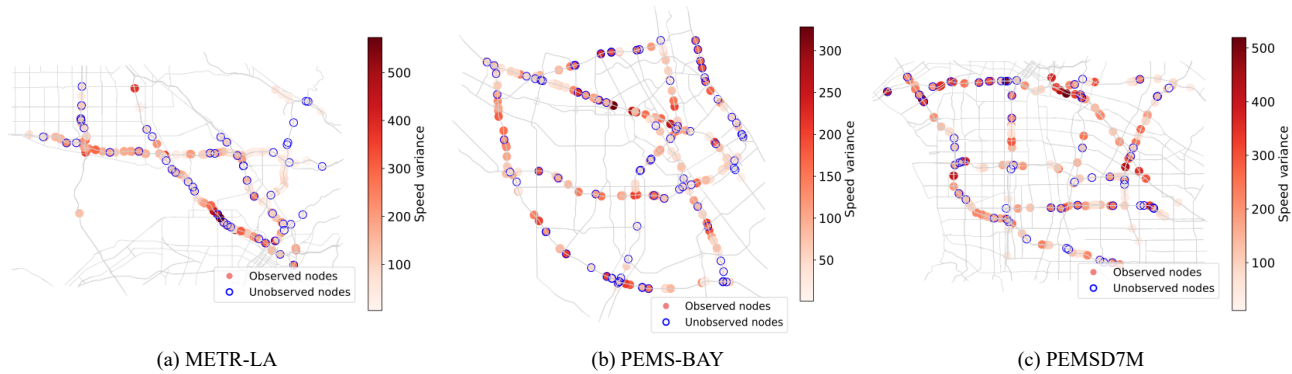


Fig. 3 Location and speed variance of observed (without edge) and unobserved (with edge) sensors on road map.

subgraph sampling is applied during validation and testing. Within each sampled subgraph, a proportion of nodes equal to $r_{\text{mask}} = r_{\text{unobs}}$ is randomly selected, and their speed values are masked for estimation. Taking the METR-LA dataset as an example, which contains $N = 207$ nodes, approximately $N_{\text{unobs}} = N \times r_{\text{unobs}} \approx 52$ nodes are removed from the training and validation sets and reserved exclusively for evaluation. From the remaining observed nodes, $N_{\text{sub}} = N_{\text{obs}} \times r_{\text{sub}} \approx 116$ nodes are sampled in each subgraph via K -order neighborhood sampling during training. Among these, $N_{\text{mask}} = N_{\text{sub}} \times r_{\text{mask}} \approx 29$ nodes have their values masked, and their speeds are inferred from the remaining 87 observed nodes. The node split scheme is also illustrated in Fig. 4. The remaining key hyperparameters of the RADGCN model are carefully tuned and summarized in Table 2.

All experiments were performed on an Intel(R) Core(TM) Ultra 9 (Series 2) 285K 24 $\times$ 5.7 GHz 24-Core Processor (CPU), NVIDIA RTX 5090 (32 GB), and the Ubuntu 22.04.3 long-term support (LTS) operating system. All the program tasks were conducted on Python 3.12, and the deep learning framework was based on PyTorch. The RADGCN model is trained by employing the Adam optimizer within the PyTorch framework with Xavier weights and bias initialization. All experiments were repeated with 5 different seeds, and the reported results are the average values from the 5 seeds.

5.3 Baseline models

To validate the proposed model's performance, several state-of-the-art (SOTA) methods are used as baselines. A brief overview of the baseline models is provided below.

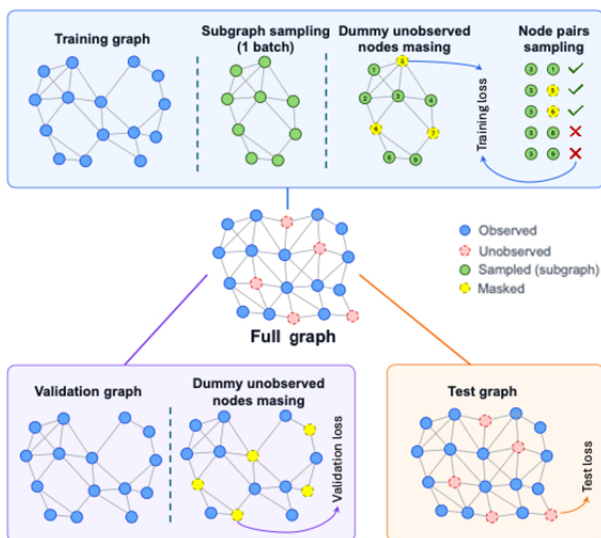


Fig. 4 Nodes split for training, validation, and testing.

IGNNK (Wu et al., 2021a): IGNNK is an inductive spatiotemporal kriging model that applies diffusion-based graph convolution on randomly sampled subgraphs, enabling TSE at unobserved nodes under changing graph structures.

IAGCN (Wei et al., 2024): IAGCN extends inductive diffusion-based kriging by combining DGCN with an adaptive graph constructor and an auxiliary prediction-based constraint task to capture hidden spatial correlations and alleviate information loss caused by subgraph sampling.

SATCN (Wu et al., 2021b): SATCN is an inductive spatiotemporal kriging model that combines spatial aggregation networks with TCNs. It uses multiple neighborhood aggregation functions to capture spatial dependencies and applies a masking strategy to prevent unobserved nodes from sending messages during aggregation.

INCREASE (Zheng et al., 2023): INCREASE is an inductive graph representation learning model for spatiotemporal kriging. It models heterogeneous spatial relations, including spatial proximity, functional similarity, and transition probability, and further uses relation-aware gated recurrent unit (GRU) and multirelation attention to capture temporal dependencies and fuse different relational views.

STAR (Liang et al., 2023): STAR is a spatial-temporal aware inductive GNN for traffic data recovery. It uses residual gated temporal convolution for temporal feature extraction, graph diffusion convolution for spatial propagation, and an adaptive memory-based attention module to capture implicit spatial correlations under different missing-data patterns.

DGAE (Zhou et al., 2025): DGAE is a Dirichlet graph autoencoder for sparse-sensing TSE that generates latent representations for unobserved nodes before graph propagation, thereby reducing the bias induced by zero-filled inputs in conventional diffusion-based models.

Model-driven and matrix factorization methods have been shown to perform inferiorly in this task (Wei et al., 2024; Zhou et al., 2025), particularly under inductive settings, and are therefore not included as baselines in the present study. Another DGCN-based study by Nie et al. (2023b) is also excluded from the set of baseline models, as it requires fully observed auxiliary patterns for all nodes, which are not available in the present study.

6 Results

6.1 Performance comparison with baselines

Table 3 reports the model performance on three widely used benchmark datasets, i.e., METR-LA, PEMS-BAY, and PEMS7M, in terms of mean absolute error (MAE) in mph, root mean square error (RMSE) in mph, and mean absolute

Table 2 Hyperparameters used for implementation

Hyperparameter	Value	Explanation
batch size	32	Sample size of each minibatch during training
LR	0.001	Learning rate of training
dropout	0.15	Dropout ratio of network
epoch	100	Maximum epoch number
edge_loss_weight	0.1	The weight of edge prediction task's cross-entropy loss in the total loss
d_model	256	Hidden dimension of model. For convenience, same for all modules
num_layers_tcn	2	Number of TCN layers in temporal encoder and decoder
d_time	16	Dimension of time encoding in TCN
kernel_size	4	Size of convolutional kernel in TCN
num_layers_gf	2	Number of Graphormer layers
num_head	4	Number of heads in Graphormer
num_layers_radgcn	2	Number of RADGCN layers
diffusion_steps	2	Diffusion step size K in RADGCN

percentage error (MAPE). The proposed model consistently achieves the best performance across all datasets and evaluation metrics, demonstrating both accuracy and robustness under different traffic conditions and network topologies. Across the three datasets, the proposed method outperforms all baseline models in all evaluation metrics. The improvement is not marginal but systematic, with consistent gains in three metrics. Notably, the improvement percentages relative to the strongest baseline model range from 2.97% to 11.82%.

More specifically, the performance gains show different patterns across the three datasets. On METR-LA, RADGCN achieves relatively modest improvements in MAE and RMSE but a much larger improvement in MAPE, indicating that the proposed relation-aware diffusion mechanism is particularly effective in reducing relative estimation errors under the more irregular and congested traffic conditions of this network. Since METR-LA also contains a higher proportion of missing records, the results suggest that learning relations from reliable observed-node dynamics can provide useful supplementary information when direct traffic observations are incomplete. On PEMS-BAY,

RADGCN obtains balanced improvements across all three metrics, suggesting that the proposed framework improves both average estimation accuracy and overall error stability in a network with smoother and more homogeneous traffic patterns. This indicates that relation diffusion is not only beneficial in highly irregular networks but can also consistently enhance estimation in relatively regular traffic environments. On PEMS7M, the improvements are smaller in MAE and RMSE but remain more evident in MAPE, showing that RADGCN mainly improves relative estimation accuracy in this more heterogeneous dataset. Overall, the dataset-specific results indicate that the proposed framework provides stable gains across different network scales and traffic dynamics, while its practical benefit is most evident in improving the reliability of estimates at unobserved nodes rather than only reducing a single error metric.

Among all baselines, DGAE, which utilizes the combination of the Dirichlet energy minimization mechanism and DGCN, achieves the strongest performance. A possible reason is its improvement in Dirichlet-assisted feature propagation for unobserved nodes before passing them to the graph learning steps, thereby mitigating error propagation by zero-padding unobserved nodes. The presented work also identified this issue and chose to truncate the message propagation chain from unobserved nodes and enhance message passing in DGCN via the ABR-DGCN. The experimental results verify the performance of this design.

Table 3 also reports the computational efficiency of different models. RADGCN introduces additional computations compared with simpler diffusion-based baselines such as IGNNK because it includes observed-node relation learning, bilateral relation diffusion, and relation-augmented feature propagation. Nevertheless, its empirical cost remains within a practical range on the three benchmark datasets. Specifically, RADGCN requires 5.88, 14.31, and 2.42 s per training epoch on METR-LA, PEMS-BAY, and PEMS7M, respectively, while its inference time is 0.19, 0.51, and 0.10 s. The corresponding peak compute unified device architecture (CUDA) memory usage is 2211.58, 3767.74, and 2444.23 MB, respectively. These results indicate that the proposed relation-aware diffusion mechanism provides improved estimation accuracy with acceptable computational overhead for medium-scale traffic networks.

To further assess statistical reliability, we conduct paired one-sided t -tests between RADGCN and the strongest baseline for

Table 3 Model performance comparison across different datasets

Dataset	Metric	IAGCN	IGNNK	DGAE	INCREASE	SATCN	STAR	Ours	Improvement (%)
METR-LA	MAE	5.88±0.49	6.43±0.44	5.90±0.56	7.14±0.35	6.67±0.42	6.25±0.53	5.62±0.56	4.45
	RMSE	9.37±0.73	9.36±0.68	8.73±0.87	10.59±0.52	9.96±0.86	9.57±0.83	8.42±0.85	3.64
	MAPE (%)	16.20±3.26	17.81±2.46	15.88±2.56	20.62±2.38	18.88±2.90	18.27±3.15	14.00±2.26	11.82
	Training (s)	17.55	0.83	5.06	11.19	29.83	18.88	5.88	—
	Inference (s)	5.93	0.10	0.61	2.26	16.28	11.20	0.19	—
PEMS-BAY	MAE	3.74±0.12	3.87±0.31	3.69±0.18	4.64±0.22	3.86±0.19	3.73±0.23	3.45±0.18	6.63
	RMSE	6.72±0.23	6.52±0.51	6.28±0.33	7.94±0.44	6.71±0.40	6.44±0.43	5.95±0.23	5.21
	MAPE (%)	8.41±0.78	8.88±1.08	8.31±0.85	11.43±1.16	9.29±0.97	8.51±1.04	7.75±0.61	6.70
	Training (s)	33.41	2.01	10.22	21.03	135.12	72.35	14.31	—
	Inference (s)	10.05	0.33	1.24	3.89	25.92	40.62	0.51	—
PEMS7M	MAE	5.89±0.30	8.25±0.35	5.74±0.17	7.32±0.22	6.72±0.26	5.97±0.30	5.57±0.20	2.97
	RMSE	9.42±0.51	11.50±0.31	9.09±0.42	10.93±0.38	10.78±0.42	9.20±0.53	8.79±0.38	3.36
	MAPE (%)	16.81±2.28	23.80±2.19	16.56±2.10	22.04±2.18	20.06±2.37	17.25±2.55	15.18±1.73	8.34
	Training (s)	6.87	0.26	2.33	4.48	16.41	12.29	2.42	—
	Inference (s)	2.43	0.06	0.46	0.88	6.49	6.90	0.10	—

Note: **bold** value: best performance; underlined value: second best performance. Training time is for one epoch.

each dataset and metric across the five matched random seeds. As shown in Table 4, RADGCN achieves statistically significant improvements in all cases at the 0.05 level, indicating that the observed gains are not caused by a particular sensor split.

To provide a more detailed understanding of estimation performance at the level of individual unobserved nodes, Fig. 5 presents the estimated speed time series on the test set for selected nodes, presents the estimated speed time series on the test set for selected nodes, where the x-axis represents time steps, with each time step corresponding to a 5-minute interval. The results for visualization are based on the same representative random sensor split used in Fig. 3. The proposed RADGCN model (blue, dot-dashed line) is compared against the ground truth (red, solid line) and the strongest baseline model, DGAE (green, dashed line). These selected nodes and periods typically exhibit pronounced periodicity and stochastic fluctuations in speed, which are particularly challenging to learn in the absence of historical observations at these locations.

During off-peak periods, typically located between two distinct periodic speed drops (e.g., steps 900–1100 for node 54 in PEMS-BAY in Fig. 5f), both models achieve accurate estimates due to the relatively stable traffic conditions. However, when abrupt speed drops occur, either due to recurring peak-hour congestion (e.g., time step 910 for node 30 in METR-LA in Fig. 5b) or unexpected

disruptions (e.g., time step 10,140 for node 21 in METR-LA in Fig. 5a), the proposed RADGCN model outperforms DGAE in most cases.

6.2 Interpretation analysis of learned relation kernel

To examine whether the learned relation kernel captures meaningful traffic-state similarity, we compare the relation kernel generated by ABR-DGCN with an empirical temporal-similarity matrix computed from the ground-truth traffic series in the test set. Specifically, we extract the normalized relation kernel $\mathbf{P}_r$ produced during inference, where only observed-node traffic states and graph topology are provided to the model. The traffic states of unobserved nodes are not used to construct $\mathbf{P}_r$ and are used only for post hoc interpretation. As a reference, we compute a dynamic time warping (DTW)-based similarity matrix $\mathbf{S}^{\text{DTW}}$ from the test-set traffic series. DTW is adopted because it can measure similarity between temporal patterns while allowing small temporal shifts, which is suitable for traffic dynamics where congestion propagation and recovery may occur with time lags across road segments. The raw DTW distances are converted into similarity scores by min-max normalization over off-diagonal node pairs so that larger values indicate more similar temporal evolution. Since the row-stochastic values of $\mathbf{P}_r$ are not directly comparable in magnitude with DTW similarities, $\mathbf{P}_r$ is also min-max normalized to $[0, 1]$ for visualization. For both matrices, diagonal entries are set to 1 to represent self-similarity. To reduce the influence of individual time windows, we average the normalized $\mathbf{P}_r$ matrices and the normalized DTW similarity matrices over the test period. We then compute the absolute elementwise difference $|\mathbf{P}_r - \mathbf{S}^{\text{DTW}}|$ for unobserved nodes and visualize it as a disparity heatmap in Fig. 6, based on the same representative random sensor split used in Fig. 3. Lighter regions indicate stronger agreement between the learned relation kernel and the empirical temporal similarity, whereas darker regions indicate larger discrepancy. This analysis does not serve as an additional training objective. Instead, it provides a diagnostic visualization of whether the learned relation kernel is consistent with traffic-pattern similarity, which the model never directly uses for unobserved nodes.

Table 4 Paired significance test between RADGCN and the strongest baseline across five random seeds

Dataset	Metric	Strongest baseline	<i>t</i> -statistic	<i>p</i> value
METR-LA	MAE	IAGCN	2.84	0.023
	RMSE	DGAE	3.84	0.009
	MAPE	DGAE	6.24	0.002
PEMS-BAY	MAE	DGAE	41.64	< 0.001
	RMSE	DGAE	6.28	0.002
	MAPE	DGAE	3.92	0.009
PEMSD7M	MAE	DGAE	3.94	0.008
	RMSE	DGAE	3.49	0.013
	MAPE	DGAE	6.17	0.002

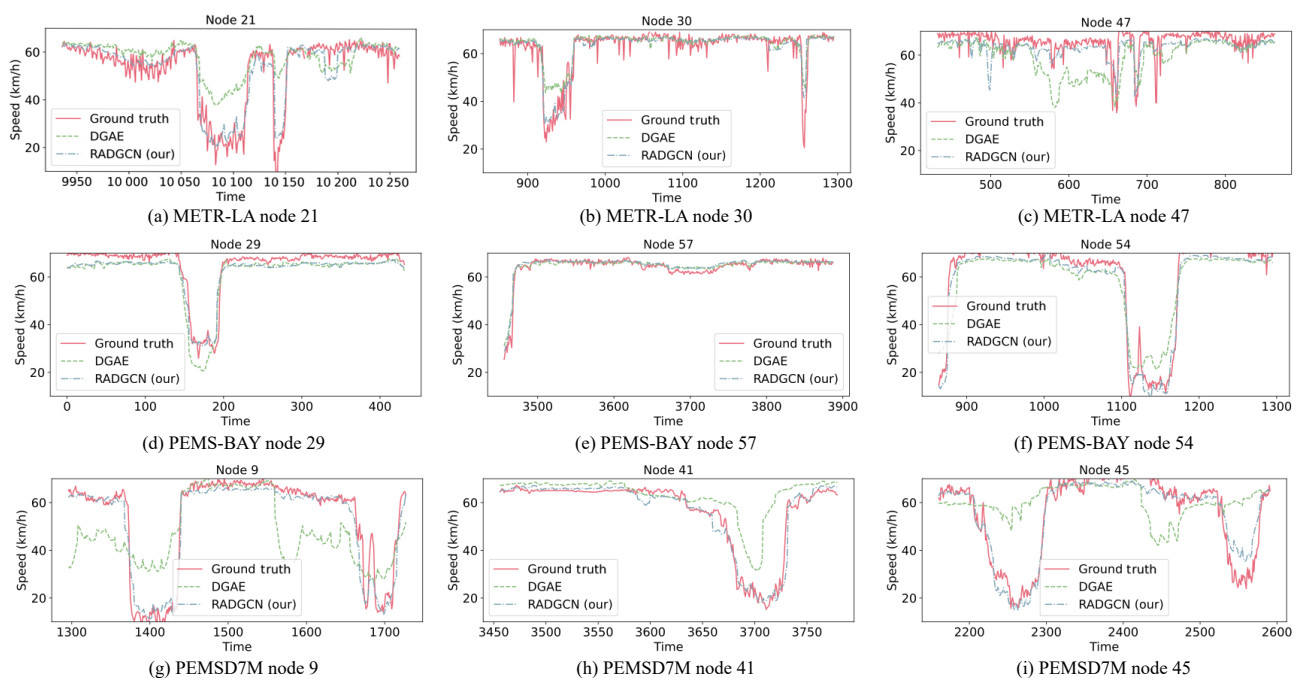


Fig. 5 Representative unobserved node traffic speed estimation results.

In Fig. 6, the disparity heatmaps are dominated by light or moderately shaded regions across the three datasets, indicating that the learned relation kernel generally follows the empirical DTW-based traffic-pattern similarity. Quantitatively, the average absolute differences between the normalized learned relation kernel and the DTW-based similarity matrix over all unobserved-node pairs and all random seeds are 0.18 for METR-LA, 0.14 for PEMS-BAY, and 0.23 for PEMS7M. Considering that the traffic features of unobserved nodes are masked during inference and are never used to generate P_r , these relatively low average discrepancies suggest that the proposed relation-diffusion mechanism can propagate observed-node relational information to unobserved regions in a way that is broadly consistent with the actual temporal evolution of traffic states. This supports the interpretation that the learned relation kernel is not merely a latent fitting matrix but reflects meaningful functional dependencies among road segments. At the same time, several darker horizontal or vertical bands can be observed, indicating that the learned relation kernel disagrees with the DTW-based similarity for a small number of nodes. This discrepancy is reasonable for two reasons. First, DTW similarity is computed solely from traffic time series, whereas P_r is generated from observed node relations and graph-based diffusion. Therefore, it is expected to encode both temporal similarity and structural propagation constraints rather than exactly reproduce DTW similarity. Second, nodes located near bottlenecks, ramps, boundary regions, or sparsely connected areas may exhibit traffic patterns that are difficult to infer from observed nodes through relation diffusion alone. For these nodes, the propagated relation weights may be more strongly influenced by network structure or the availability of nearby observed sensors, leading to local mismatches with the empirical DTW matrix.

6.3 Spatial distribution of estimation errors

Fig. 7 visualizes the spatial distribution of MAE across sensor locations for METR-LA, PEMS-BAY, and PEMS7M, based on

the same representative random sensor split used in Fig. 3. The results reveal clear spatial heterogeneity in estimation performance. Apparently, errors are not uniformly distributed. Instead, higher MAE values tend to concentrate in regions with complex traffic dynamics, such as major junctions, merging corridors, and dense urban cores. This pattern is particularly evident in PEMS-BAY and PEMS7M, where high-error nodes cluster around central bottleneck areas. Despite this heterogeneity, most nodes exhibit moderate error levels, indicating that the model achieves stable performance across the network rather than relying on improvements at a limited number of easier locations. In PEMS-BAY, the error distribution is relatively uniform, suggesting strong robustness under smoother traffic conditions. In contrast, elevated errors are more frequently observed in dense junction areas with higher speed variance (Fig. 3), likely reflecting the impact of complex traffic interactions, recurrent bottlenecks, and spatially heterogeneous control strategies. Nevertheless, the error magnitudes in these regions remain within a reasonable range, implying that the model captures part of these localized dynamics while still facing challenges in structurally and operationally complex areas.

Another notable observation is the absence of large contiguous high-error regions. Under sparse sensing, conventional diffusion-based models often exhibit error amplification in unobserved clusters due to disrupted message-passing chains. In contrast, the relatively controlled error levels in spatially fragmented regions suggest that the proposed relation-aware diffusion mechanism effectively propagates functional similarities beyond purely topological neighborhoods.

Overall, these spatial error patterns indicate that the proposed framework not only reduces the global average error but also improves spatial stability, particularly in regions characterized by structural complexity or dynamic heterogeneity. This provides empirical support for the claim that relation-aware diffusion enhances robustness under sparse and uneven observation conditions.

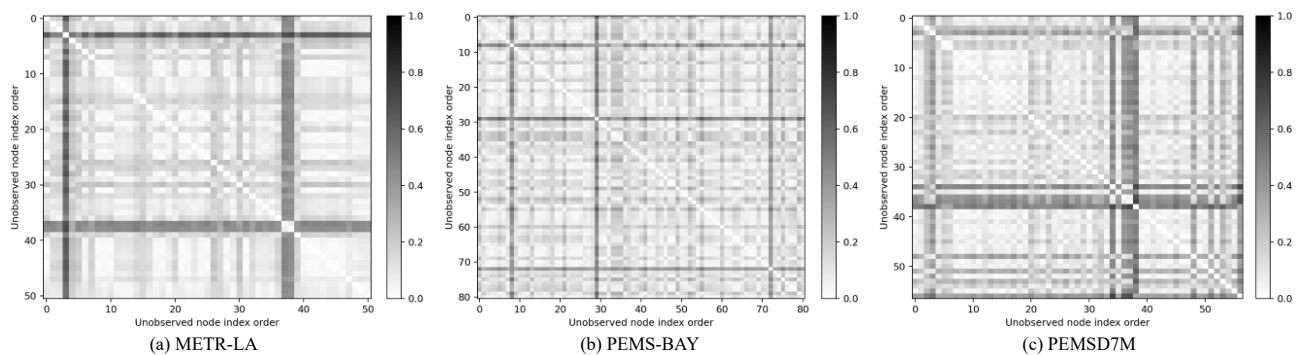


Fig. 6 Aggregated absolute disparities $|P_r - S^{DTW}|$ of the test set among unobserved nodes.

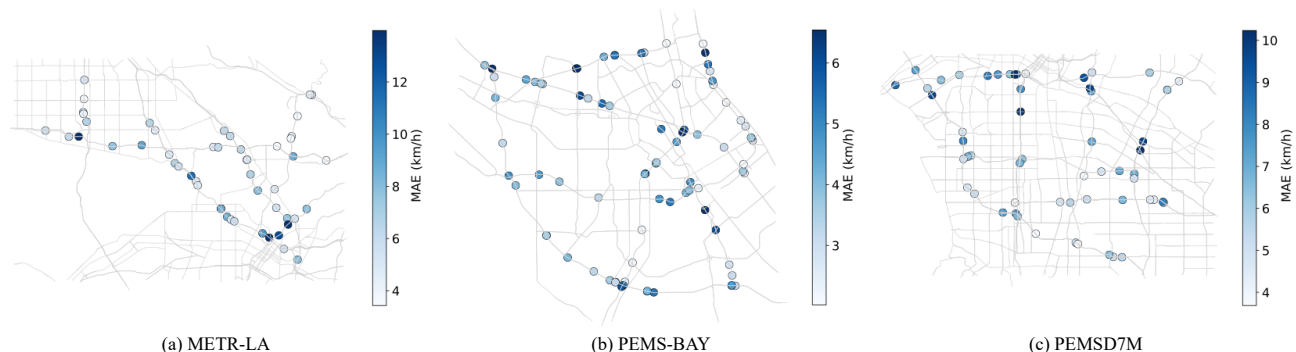


Fig. 7 MAE of unobserved nodes on road map.

6.4 Ablation analysis

To further validate the effectiveness of the proposed framework and investigate the contributions of its core components, an ablation study is conducted by selectively removing three key mechanisms, i.e., ABR-DGCN, K -order neighborhood subgraph sampling, and the edge prediction mechanism. The objective is to assess whether the observed performance gains arise from the model's structural innovations rather than incidental architectural complexity or hyperparameter tuning. These three components are selected because they constitute the framework's fundamental design principles. The ABR-DGCN module is responsible for adaptive relation learning among nodes, through which feature-related dependencies are inferred and subsequently diffused from observed-observed pairs to observed-unobserved, unobserved-observed, and unobserved-unobserved pairs. In the absence of ABR-DGCN, the model reduces to a conventional DGCN, relying solely on topology-driven message passing. The K -order neighborhood subgraph sampling mechanism differs from random sampling in that it preserves richer structural information within the sampled subgraphs and enables a more accurate representation of the influence exerted by neighboring nodes on central nodes. Ablating K -order subgraph sampling indicates that, during training, subgraphs are instead generated through random node selection, thereby removing the structural constraints imposed by the K -order neighborhood sampling strategy. The edge prediction module introduces an auxiliary supervision objective that predicts the existence of a directed edge between node pairs based on their final-layer embeddings. Notably, this module does not modify the graph structure or alter the message-passing operator directly. Instead, it regularizes the learned embeddings by enforcing that structurally connected and disconnected node pairs remain distinguishable in the latent space. In particular, it compensates for the incompleteness of structural information induced by subgraph sampling during inductive training by explicitly supervising the recovery of missing connectivity patterns.

As shown in Table 5, the ablation results demonstrate that all three designed components contribute positively to the estimation performance. Across the three benchmark datasets, the full model consistently achieves the lowest MAE, RMSE, and MAPE values, confirming the effectiveness of the complete framework. When any component is removed, the prediction accuracy decreases to different extents, indicating that the K -order subgraph sampling mechanism, the edge prediction auxiliary task, and the ABR-

DGCN module are all necessary for improving model performance. Specifically, removing the K -order subgraph sampling mechanism leads to an average performance degradation of 4.67%, 3.39%, and 5.19% on METR-LA, PEMS-BAY, and PEMS7M, respectively. This verifies that K -order neighborhood sampling is more effective than random node sampling, as it preserves local topological dependencies and enables the model to better capture the influence of neighboring nodes on the central nodes. The relatively consistent degradation across datasets suggests that structural information plays a stable role in enhancing spatial representation learning. The marginal contribution of K -order sampling to the PEMS-BAY dataset could be attributed to the relatively dense and homogeneous spatial distribution of sensors in this network. Compared with METR-LA and PEMS7M, the neighboring traffic patterns in PEMS-BAY exhibit stronger local consistency, making random node sampling capable of capturing a certain degree of spatial correlation. As a result, replacing random sampling with K -order neighborhood sampling brings relatively limited additional structural information, leading to a smaller performance improvement.

The removal of the edge prediction auxiliary task also causes noticeable performance degradation, with average drops of 12.50%, 4.43%, and 5.43% on the three datasets. This indicates that the auxiliary task provides useful supervision for learning latent spatial correlations among nodes. By encouraging the model to infer edge relationships, this component helps strengthen the learned graph representation and improves the robustness of traffic state prediction. The relatively larger contribution of the edge prediction auxiliary task on METR-LA may also be associated with the higher missing ratio of this dataset. Since more than 8% of the observations are missing, the node-level estimation objective may provide relatively weaker and less complete supervision for learning spatial-temporal dependencies. Under this condition, the edge prediction task serves as an additional structural regularization signal, encouraging the model to infer latent connectivity patterns among sensors and thereby compensating for the limited supervision caused by missing traffic records. This may explain why the removal of the edge prediction task leads to a more evident performance degradation on METR-LA.

Among the three components, removing the ABR-DGCN module generally results in the most evident degradation, especially on METR-LA, where the average performance drop reaches 13.04%. This finding suggests that the ABR-DGCN

Table 5 Ablation study of core components: ABR-DGCN, edge prediction, and K -order subgraph sampling

Dataset	Metric	w/o K	w/o E	w/o A	w/o E + K	w/o A + E	w/o A + K	w/o K + E + A	Fullness
METR-LA	MAE	5.88 ± 0.40	6.36 ± 0.53	6.32 ± 0.33	6.39 ± 0.33	6.43 ± 0.36	6.48 ± 0.30	6.80 ± 0.39	5.62 ± 0.56
	RMSE	8.57 ± 0.70	9.11 ± 0.66	9.13 ± 0.56	9.25 ± 0.56	9.31 ± 0.58	9.30 ± 0.56	9.77 ± 0.61	8.42 ± 0.85
	MAPE (%)	15.08 ± 2.67	16.26 ± 2.29	16.56 ± 2.23	16.94 ± 1.98	17.14 ± 2.40	16.90 ± 2.16	18.81 ± 1.93	14.00 ± 2.26
	Contribution (%)	4.67	12.50	13.04	14.86	15.77	15.48	23.78	—
PEMS-BAY	MAE	3.59 ± 0.16	3.62 ± 0.17	3.69 ± 0.20	3.75 ± 0.16	3.90 ± 0.17	3.89 ± 0.13	4.04 ± 0.15	3.45 ± 0.18
	RMSE	6.00 ± 0.25	6.15 ± 0.40	6.18 ± 0.30	6.20 ± 0.32	6.31 ± 0.21	6.48 ± 0.31	6.67 ± 0.23	5.95 ± 0.23
	MAPE (%)	8.16 ± 0.65	8.15 ± 0.63	8.39 ± 0.79	8.57 ± 0.98	8.78 ± 0.42	9.13 ± 0.50	9.34 ± 0.42	7.75 ± 0.61
	Contribution (%)	3.39	4.43	6.39	7.82	10.75	13.07	16.56	—
PEMS7M	MAE	5.80 ± 0.15	5.91 ± 0.19	6.01 ± 0.23	6.27 ± 0.57	6.32 ± 0.40	6.34 ± 0.10	6.61 ± 0.27	5.57 ± 0.20
	RMSE	9.04 ± 0.43	9.11 ± 0.36	9.11 ± 0.42	9.37 ± 0.52	9.45 ± 0.51	9.61 ± 0.25	9.74 ± 0.38	8.79 ± 0.38
	MAPE (%)	16.50 ± 1.86	16.18 ± 1.77	17.07 ± 1.53	17.59 ± 1.83	17.84 ± 2.07	18.09 ± 1.72	18.43 ± 2.46	15.18 ± 1.73
	Contribution (%)	5.19	5.43	8.01	11.66	12.84	14.11	16.96	—

Note: K : K -order subgraph sampling; E: edge prediction auxiliary task; A: ABR-DGCN module robustness under sparse and uneven observation conditions.

module plays a central role in capturing adaptive spatial dependencies and dynamic graph-based correlations. Its contribution is particularly important in complex traffic networks, where spatial interactions may vary across regions and time periods.

The combined ablation settings further confirm the complementary effects among different components. For example, removing all three components/mechanisms results in the largest performance degradation on all datasets, reaching 23.78%, 16.56%, and 16.96%, respectively. This suggests that the proposed sampling strategy and the graph convolution module work synergistically: the former provides structurally informative subgraphs, while the latter effectively extracts spatial dependencies from them. Similarly, the performance drops observed in "w/o E + K", "w/o A + E", and "w/o A + K" show that the removal of multiple components weakens the model more substantially than removing a single component.

Overall, the ablation study validates the rationality of the proposed model design. The full model consistently outperforms all ablated variants, indicating that the three components are not redundant but mutually beneficial. The K -order subgraph sampling mechanism enhances structural information preservation, the edge prediction auxiliary task improves spatial relation learning, and the ABR-DGCN module strengthens adaptive graph representation. Their integration enables the model to achieve superior and robust forecasting performance across different traffic datasets.

6.5 Cross-dataset validation

To evaluate the generalization capability of the proposed framework beyond a single network topology, we conduct cross-dataset experiments where the model is trained on one dataset and directly tested on another city-scale network without fine-tuning. This setting is significantly more challenging, as the spatial structure, traffic dynamics, and sensor layouts differ across datasets. In this set of experiments, two high-quality datasets, i.e., PEMS-BAY and PEMS7M, are selected for training and evaluation.

Table 6 summarizes the cross-dataset validation results. The notation "PEMS-BAY $\rightarrow$ PEMS7M" denotes that the model is trained on PEMS-BAY and directly tested on the complete PEMS7M dataset without further training or sensor-layout-specific fine-tuning, and vice versa. This setting evaluates whether a model trained on one road network can generalize to another network with different topology and traffic patterns. The proposed RADGCN achieves the best average performance in both transfer directions and across all three metrics. In the PEMS-BAY $\rightarrow$ PEMS7M scenario, RADGCN reduces the MAE, RMSE, and MAPE by 5.48%, 5.57%, and 4.69%, respectively, compared with the strongest baseline. In the reverse PEMS7M $\rightarrow$ PEMS-BAY scenario, RADGCN also obtains the lowest errors, with improvements of 5.83% in MAE, 2.75% in RMSE, and 4.02% in MAPE. These results indicate that RADGCN maintains robust

transferability under cross-network distribution shifts.

Among the baselines, DGAE and STAR show relatively strong cross-dataset performance for different reasons. DGAE is specifically designed for sparse-sensor TSE, with a physics-guided design that further separates congested and free-flow propagation directions, helping capture traffic-flow dynamics under changing sensor layouts. These mechanisms explain why DGAE remains a strong baseline in both transfer directions. STAR, although originally proposed for C-ITS data recovery rather than fixed-sensor network-wide TSE, also has transfer-friendly components. It combines residual gated TCN for temporal feature extraction, graph diffusion convolution for directed spatial propagation, and external attention to capture semantic similarity with linear complexity.

Compared with these baselines, RADGCN further improves cross-dataset transfer by explicitly separating relation learning from feature propagation. Instead of first estimating or generating unobserved-node features, RADGCN learns functional relations among observed nodes, where traffic features are reliable, and diffuses these relations to the full graph as an additional transition operator. This design reduces the dependence of adaptive relation construction on city-specific node embeddings or zero-filled/unobserved-node representations. The consistent gains over DGAE and STAR suggest that observed-node relation diffusion provides complementary transferable information beyond directed feature propagation and spatial-temporal attention.

Paired one-sided t -tests are further conducted between RADGCN and the strongest baseline DGAE across the five matched random seeds. As shown in Table 7, RADGCN achieves statistically significant improvements in both transfer directions and across all metrics at the 0.05 level.

Table 7 Paired one-sided t -test between RADGCN and DGAE in cross-dataset experiments

Scenario	Metric	t -statistic	p value
PEMS-BAY-PEMS7M	MAE	8.84	< 0.001
	RMSE	4.64	0.005
	MAPE	3.78	0.010
PEMS7M-PEMS-BAY	MAE	4.96	0.004
	RMSE	2.61	0.030
	MAPE	3.78	0.010

From a practical perspective, this property is particularly valuable. In many real-world scenarios, newly developed or underinstrumented regions lack sufficient historical sensor data for model training. The ability to transfer a model trained in one city to another without retraining significantly reduces deployment costs and facilitates TSE in data-scarce environments.

6.6 Robustness to sensor deployment density

The primary experiments are conducted under a 75% sensor

Table 6 Model performance in the cross-dataset experiment

Scenario	Metric	IAGCN	IGNNK	DGAE	INCREASE	SATCN	STAR	Ours	Improvement (%)
PEMS-BAY-PEMS7M	MAE	6.33 $\pm$ 0.25	6.23 $\pm$ 0.37	6.05 $\pm$ 0.22	7.49 $\pm$ 0.26	6.49 $\pm$ 0.24	6.34 $\pm$ 0.19	5.72 $\pm$ 0.18	5.48
	RMSE	10.53 $\pm$ 0.46	9.90 $\pm$ 0.81	9.53 $\pm$ 0.57	11.36 $\pm$ 0.43	10.86 $\pm$ 0.37	9.62 $\pm$ 0.55	9.00 $\pm$ 0.37	5.57
	MAPE (%)	18.52 $\pm$ 1.87	18.34 $\pm$ 2.19	17.30 $\pm$ 2.40	23.60 $\pm$ 2.46	20.28 $\pm$ 3.22	17.71 $\pm$ 2.85	16.48 $\pm$ 2.10	4.69
PEMS7M-PEMS-BAY	MAE	4.25 $\pm$ 0.36	6.18 $\pm$ 0.26	4.13 $\pm$ 1.59	4.98 $\pm$ 1.79	4.28 $\pm$ 1.50	4.19 $\pm$ 0.37	3.88 $\pm$ 0.22	5.83
	RMSE	7.04 $\pm$ 0.33	8.41 $\pm$ 0.29	6.83 $\pm$ 3.75	8.08 $\pm$ 4.48	7.17 $\pm$ 4.00	6.87 $\pm$ 0.38	6.64 $\pm$ 0.39	2.75
	MAPE (%)	9.63 $\pm$ 0.91	12.89 $\pm$ 0.98	9.55 $\pm$ 1.00	11.58 $\pm$ 1.09	10.00 $\pm$ 0.77	9.71 $\pm$ 1.17	9.06 $\pm$ 0.92	4.02

coverage setting, following several prior studies to ensure fair comparison. While this setup provides a standardized benchmark for evaluating model performance under sparse observation, it still assumes a relatively high level of infrastructure availability. In practice, deploying and maintaining sensors at such density entails substantial installation and operational costs, which may not be feasible for many urban regions, particularly in newly developed areas or budget-constrained municipalities. Therefore, it is essential to evaluate whether the proposed framework remains effective under lower sensor coverage ratios. To this end, a set of robustness analytical experiments is conducted.

Table 8 present the robustness analysis under different sensor deployment densities. Across all datasets and metrics, both RADGCN and DGAE exhibit monotonically increasing errors as the unknown node ratio grows, which is expected because fewer observed nodes provide less direct information for estimating unobserved traffic states. Despite this increasing difficulty, RADGCN consistently achieves lower average errors than DGAE across different sparsity levels. The superiority is reflected in all evaluation metrics, including MAE, RMSE, and MAPE. Furthermore, as the proportion of unobserved sensors increases, the performance of both methods gradually deteriorates due to the reduction in available supervision. Nevertheless, RADGCN exhibits a noticeably slower degradation trend than DGAE. For example, when the unknown node ratio increases from 25% to 90%, the RMSE growth on PEMS7M is reduced from 38.4% for DGAE (9.09 $\rightarrow$ 12.58) to 33.6% for RADGCN (8.79 $\rightarrow$ 11.74), indicating superior robustness under severe sensor sparsity. This indicates that the learned relation diffusion mechanism can effectively propagate information from observed nodes to sparsely monitored regions, making the model more robust under extreme sensor sparsity.

In addition, Fig. 8 also demonstrates that the performance gap between RADGCN and DGAE becomes more pronounced under highly sparse observation settings (e.g., RMSE for PEMS-BAY and PEMS7M). These results demonstrate that the proposed relation-aware diffusion mechanism can effectively exploit latent functional similarities between observed and unobserved nodes, thereby maintaining accurate TSE even when the sensor coverage becomes extremely sparse. To further assess statistical reliability under severe sparsity, paired one-sided *t*-tests were conducted between RADGCN and DGAE under the most challenging 90% unknown-node setting. As shown in Table 9, RADGCN achieves statistically significant improvements ($p < 0.05$) for all metrics on METR-LA and PEMS-BAY, as well as RMSE and MAPE on

PEMSD7M. The MAE improvement on PEMS7M remains positive but does not reach the 0.05 significance level ($p = 0.071$), which is mainly attributable to the larger variability across sensor deployment schemes.

7 Limitations

Despite its strong performance, our RADGCN framework has several limitations. These limitations point to several directions that deserve further investigation.

First, the current framework focuses on single-variable traffic speed extrapolation, which is broadly consistent with most prior studies in diffusion-based traffic kriging and sparse-sensing state estimation, where the estimation target is typically limited to a single traffic variable, such as speed or flow. Nevertheless, many practical traffic management applications require the joint estimation of multiple coupled traffic states, such as speed, flow, and density. Extending the proposed framework to this setting is not merely a matter of increasing the output dimension but would require explicitly modeling the physical interdependence among heterogeneous traffic variables and their distinct spatiotemporal patterns. A promising direction is therefore to incorporate traffic flow theory into the learning process, for example, by embedding conservation laws, fundamental diagram constraints, or other physics-based regularizers into the relation-diffusion framework, thereby moving toward a physics-informed learning paradigm for multistate traffic estimation.

Second, although the proposed method improves robustness under sparse sensing, the estimation task remains challenging in highly complex bottleneck regions and under extremely sparse observations. In such cases, traffic dynamics are often characterized by abrupt state transitions, queue formation and dissipation, shockwave propagation, or disruptions, which are difficult to recover reliably from sparse observations alone. This suggests that further improvements may require more effective mechanisms for capturing localized heterogeneity and for propagating long-range dependencies in severely underobserved regions.

Third, this study does not place particular emphasis on advancing the temporal dependency modeling module but instead employs a causal TCN, which has been widely used in related traffic estimation studies. This choice is mainly motivated by the relatively short temporal window considered in the present task, under which causal TCN is generally sufficient to capture local sequential dependencies while maintaining low computational complexity and stable training behavior. Nevertheless, more

Table 8 Robustness analysis against sensor deployment density for DGAE and the proposed RADGCN

Method	METR-LA (25% unknown)			METR-LA (50% unknown)			METR-LA (75% unknown)			METR-LA (90% unknown)		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
DGAE	5.90 $\pm$ 0.56	8.73 $\pm$ 0.87	15.88 $\pm$ 2.56	6.49 $\pm$ 0.44	9.77 $\pm$ 0.64	18.84 $\pm$ 2.42	7.73 $\pm$ 0.14	11.12 $\pm$ 0.44	22.93 $\pm$ 1.68	9.44 $\pm$ 0.61	12.76 $\pm$ 0.28	27.34 $\pm$ 0.93
Ours	5.62 $\pm$ 0.56	8.42 $\pm$ 0.85	14.00 $\pm$ 2.26	6.30 $\pm$ 0.39	9.50 $\pm$ 0.64	17.90 $\pm$ 2.16	7.33 $\pm$ 0.29	10.95 $\pm$ 0.71	22.62 $\pm$ 1.89	8.79 $\pm$ 0.52	12.39 $\pm$ 0.20	26.30 $\pm$ 0.81
Method	PEMS-BAY (25% unknown)			PEMS-BAY (50% unknown)			PEMS-BAY (75% unknown)			PEMS-BAY (90% unknown)		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
DGAE	3.69 $\pm$ 0.18	6.28 $\pm$ 0.33	8.31 $\pm$ 0.85	3.94 $\pm$ 0.15	6.63 $\pm$ 0.28	9.16 $\pm$ 0.76	4.51 $\pm$ 0.21	7.35 $\pm$ 0.27	10.79 $\pm$ 0.59	5.34 $\pm$ 0.30	8.74 $\pm$ 0.46	13.24 $\pm$ 0.66
Ours	3.45 $\pm$ 0.18	5.95 $\pm$ 0.23	7.75 $\pm$ 0.61	3.88 $\pm$ 0.16	6.55 $\pm$ 0.27	8.93 $\pm$ 0.66	4.34 $\pm$ 0.07	7.26 $\pm$ 0.14	10.44 $\pm$ 0.31	4.99 $\pm$ 0.25	8.24 $\pm$ 0.30	12.40 $\pm$ 0.59
Method	PEMSD7M (25% unknown)			PEMSD7M (50% unknown)			PEMSD7M (75% unknown)			PEMSD7M (90% unknown)		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
DGAE	5.74 $\pm$ 0.17	9.09 $\pm$ 0.42	16.56 $\pm$ 2.10	6.39 $\pm$ 0.19	9.68 $\pm$ 0.29	17.13 $\pm$ 1.19	7.39 $\pm$ 0.52	10.84 $\pm$ 0.48	20.80 $\pm$ 1.91	8.53 $\pm$ 0.48	12.58 $\pm$ 0.76	24.94 $\pm$ 1.70
Ours	5.57 $\pm$ 0.20	8.79 $\pm$ 0.38	15.18 $\pm$ 1.73	6.25 $\pm$ 0.24	9.46 $\pm$ 0.29	17.04 $\pm$ 1.40	7.27 $\pm$ 0.20	10.70 $\pm$ 0.37	19.94 $\pm$ 1.48	8.24 $\pm$ 0.54	11.74 $\pm$ 0.43	23.26 $\pm$ 1.30

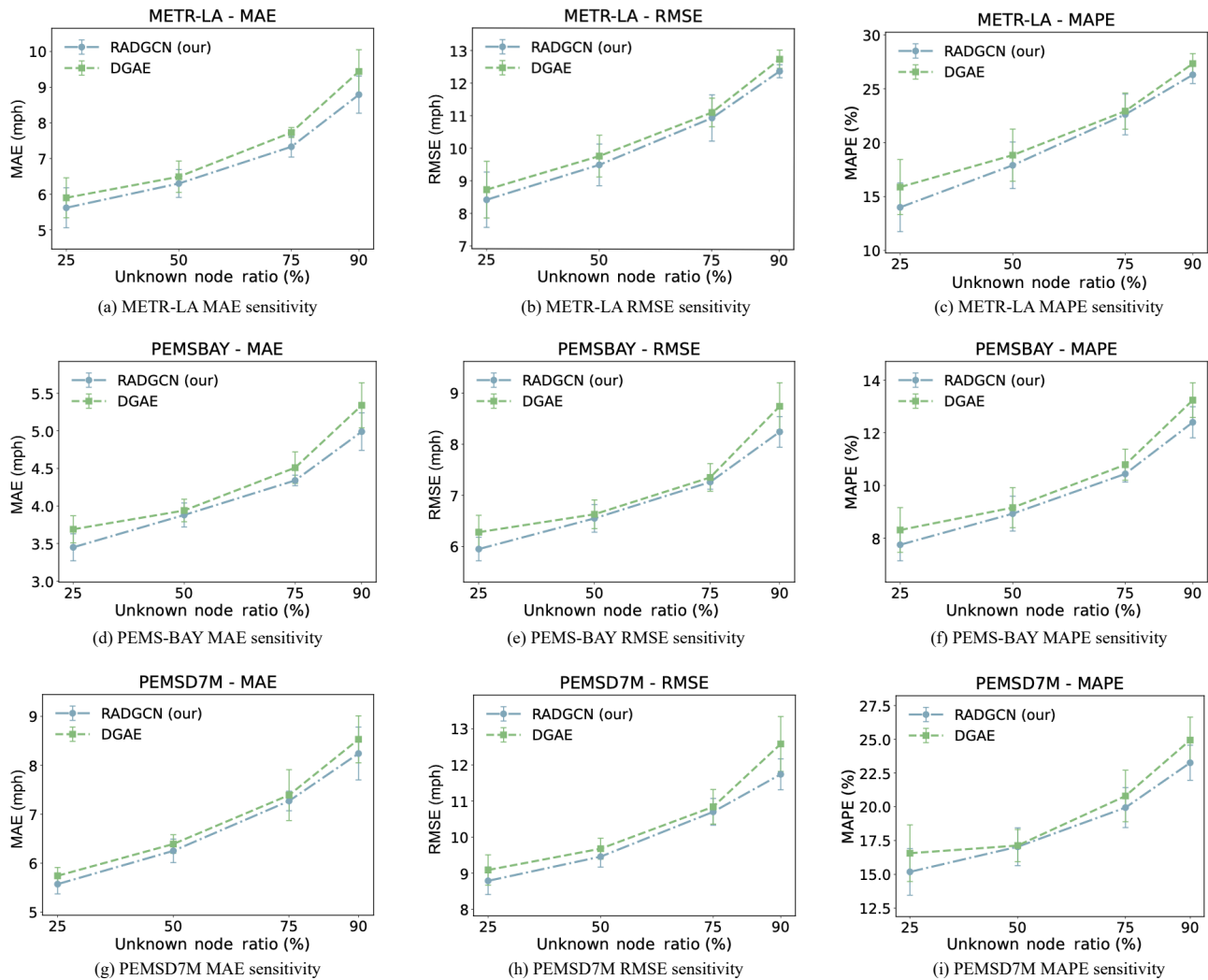


Fig. 8 Robustness of RADGCN's performance against sensor density in all datasets.

Table 9 Paired one-sided *t*-test between RADGCN and DGAE under the most challenging 90% unknown-node setting

Dataset	Metric	<i>t</i> -statistic	<i>p</i> value
METR-LA	MAE	4.09	0.007
	RMSE	2.54	0.032
	MAPE	3.78	0.010
PEMS-BAY	MAE	8.11	< 0.001
	RMSE	5.24	0.003
	MAPE	4.54	0.005
PEMS7M	MAE	1.83	0.071
	RMSE	3.14	0.017
	MAPE	4.12	0.007

expressive temporal encoders, such as transformer- or mamba-based architectures, may offer additional benefits in modeling longer-range or more complex temporal dependencies and thus deserve further investigation in future research.

Finally, a further limitation lies in the adopted *K*-order subgraph random sampling strategy. In standard *K*-hop subgraph sampling, the anchor node and all of its *K*-order neighbors are typically included so that each sampled subgraph preserves a relatively complete local topology. In the present study, however, batchwise training requires all sampled subgraphs within a minibatch to have the same size. To satisfy this requirement, an

additional constraint on the fixed number of nodes in each subgraph was imposed. As a result, some nodes in the outermost neighborhood may still be discarded, meaning that the sampled subgraph is not always a fully preserved *K*-order local graph, which can lead to the loss of structural information. In this sense, the adopted strategy should be regarded as a practical compromise between fully random fixed-size node sampling and complete *K*-order neighborhood sampling.

8 Conclusions

Accurate network-wide TSE under sparse sensor deployment remains a fundamental challenge for ITS, especially when dense detector installation is limited by cost, maintenance burden, and infrastructure constraints. This study proposed RADGCN, a relation-augmented diffusion graph convolutional network, to address this problem by shifting the focus from feature-only diffusion to relation-aware diffusion. Instead of constructing adaptive dependencies directly from zero-filled or weakly supported representations of unobserved nodes, RADGCN first learns functional relationships from observed nodes with reliable traffic measurements and then propagates these relationships to the full network as an additional diffusion operator.

Extensive experiments on three widely used benchmark datasets, namely, METR-LA, PEMS-BAY, and PEMS7M, demonstrated the effectiveness of the proposed framework.

RADGCN consistently achieved the best overall performance among all compared baselines, including IGNNK, IAGCN, INCREASE, STAR, SATCN, and DGAE, across MAE, RMSE, and MAPE, with systematic improvements over the strongest baseline ranging from 2.97% to 11.82% in the main experiments. The paired significance tests across five matched random seeds further show that all improvements are statistically significant at the 0.05 level, indicating that the gains are not caused by a particular sensor split. The results further showed that the model not only reduced average estimation errors but also improved variance control and stabilized multihop information propagation, particularly on datasets with more heterogeneous traffic conditions. Representative node-level comparisons indicated that RADGCN better captured both periodic fluctuations and abrupt speed drops at unobserved nodes, while the spatial error analysis showed that the proposed framework maintained relatively controlled estimation errors even in structurally complex or dynamically heterogeneous regions. These findings suggest that propagating learned functional relations beyond purely topological neighborhoods can effectively improve both accuracy and spatial robustness. From a practical perspective, these results indicate that more accurate network-level traffic information can be obtained without requiring dense detector deployment. This is particularly valuable for traffic agencies that need reliable monitoring of unobserved road segments under limited sensing budgets.

The ablation analyses provide further support for the methodological validity of the proposed design. Removing either ABR-DGCN, the edge prediction task, or the K -order subgraph sampling strategy led to consistent performance degradation across all datasets, while the larger drop caused by removing ABR-DGCN confirmed that the main bottleneck under sparse sensing lies in the absence of a reliable relational operator in unobserved regions rather than merely limited diffusion depth. The largest degradation is observed when the three modules are removed together, which suggests that the performance gain is not caused by a single isolated module but by the joint effect of reliable relation construction, structure-preserving inductive training, and auxiliary structural supervision.

The cross-dataset validation and sensor-density analysis further demonstrate the generalizability and robustness of RADGCN. In cross-dataset experiments, RADGCN maintains superior performance in both transfer directions, with improvements of up to 10.34% over the strongest baseline. The results showed that RADGCN generalized better across different network topologies and traffic regimes, suggesting that relation diffusion captures more transferable traffic dynamics than purely topology-dependent filters. As the unknown-node ratio increased from 25% to 90%, RADGCN consistently outperformed DGAE across all datasets, and the advantage became more pronounced with a higher missing-data ratio, indicating stronger robustness under severe sparsity. Practically, this is useful for newly developed or underinstrumented networks, where collecting sufficient historical data for model training may be difficult.

Overall, the present study shows that for TSE under sparse sensor deployment, the critical challenge is not only how to propagate node features more effectively but also how to construct and extend reliable internode relationships beyond the observed subnetwork. By introducing a relation-aware diffusion mechanism grounded in trustworthy observations, RADGCN provides a principled extension of DGCN-based TSE and offers a robust solution for large-scale network-wide estimation when sensing resources are limited. From a broader perspective, the proposed

framework contributes to the growing literature on sparse-sensing traffic inference by demonstrating that relation propagation can serve as an effective complement to traditional topology-driven diffusion.

Author contributions

Tao Guo: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Ningkang Yang:** Writing – review & editing, Methodology, Conceptualization, Formal analysis. **Kevin Yu:** Writing – review & editing, Formal analysis, Methodology. **Panagiotis Angeloudis:** Writing – review & editing, Supervision. **Constantinos Antoniou:** Writing – review & editing, Supervision, Methodology, Funding acquisition.

Replication and data sharing

The data and codes used in this study are available at <https://doi.org/10.26599/ETSD.2026.9190035>.

Acknowledgements

This work was supported by the ATHLOS project (No. J2301) and the MODA project (No. P2203) funded by the International Graduate School of Science and Engineering (IGSSE) of the Technical University of Munich. The authors also thank the anonymous reviewers for their invaluable comments and suggestions that helped improve the manuscript.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. The author Constantinos Antoniou is the Editorial Board Member of this journal.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT 5.3 to assist with grammar checking and the improvement of academic phrasing. After utilizing this tool, the authors thoroughly reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

- Agalliadis, I., Makridis, M., Kouvelas, A., 2020. Traffic estimation by fusing static and moving observations in highway networks. In: 20th Swiss Transport Research Conference (STRC 2020), 1–12.
- Appleby, G., Liu, L., Liu, L. P., 2020. Kriging convolutional networks. *AAAI*, **34**, 3187–3194.
- Bucknell, C., Herrera, J. C., 2014. A trade-off analysis between penetration rate and sampling frequency of mobile sensors in traffic state estimation. *Transp Res C Emerg Technol*, **46**, 132–150.
- Chen, R., Levin, M. W., 2019. Traffic state estimation based on Kalman filter technique using connected vehicle V2V basic safety messages. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 4380–4385.
- Daganzo, C. F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res Part B Methodol*, **28**, 269–287.
- Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. <https://doi.org/10.48550/arXiv.1606.09375>
- Hamilton, W. L., Ying, R., Leskovec, J., 2017. Inductive Representation Learning on Large Graphs. <https://doi.org/10.48550/arXiv.1706.02216>

- Herrera, J. C., Bayen, A. M., 2010. Incorporation of Lagrangian measurements in freeway traffic state estimation. *Transp Res Part B Methodol*, **44**, 460–481.
- Hu, J., Liang, Y., Fan, Z., Liu, L., Yin, Y., Zimmermann, R., 2024. Decoupling long- and short-term patterns in spatiotemporal inference. *IEEE Trans Neural Netw Learn Syst*, **35**, 16328–16340.
- Hu, Z., Zheng, Z., Menendez, M., Ma, W., 2025. From global open multi-source data to network-wide traffic flow: A large-scale case study across multiple cities. *Commun Transp Res*, **5**, 100222.
- Jin, L., Xu, X., Wang, Y., Lazar, A., Sadabadi, K. F., Spurlock, C. A., et al., 2024. Macroscopic traffic modeling using probe vehicle data: A machine learning approach. *Data Sci Transp*, **6**, 17.
- Kianfar, J., Edara, P., 2010. Optimizing freeway traffic sensor locations by clustering global-positioning-system-derived speed patterns. *IEEE Trans Intel Transp Syst*, **11**, 738–747.
- Lei, M., Labbe, A., Wu, Y., Sun, L., 2022. Bayesian kernelized matrix factorization for spatiotemporal traffic data imputation and Kriging. *IEEE Trans Intel Transp Syst*, **23**, 18962–18974.
- Li, Y., Yu, R., Shahabi, C., Liu, Y., 2017. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. <https://doi.org/10.48550/arXiv.1707.01926>
- Liang, W., Li, Y., Xie, K., Zhang, D., Li, K. C., Souri, A., et al., 2023. Spatial-temporal aware inductive graph neural network for C-ITS data recovery. *IEEE Trans Intel Transp Syst*, **24**, 8431–8442.
- Lighthill, M. J., Whitham, G. B., 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc R Soc Lond Ser A Math Phys Sci*, **229**, 317–345.
- Liu, S. J., Lam, W. H. K., Lam Tam, M., Fu, H., Ho, H. W., Ma, W., 2024. Network-wide speed-flow estimation considering uncertain traffic conditions and sparse multi-type detectors: A KL divergence-based optimization approach. *Transp Res C Emerg Technol*, **169**, 104858.
- Liu, Y., He, S., Ran, B., Cheng, Y., 2018. A progressive extended Kalman filter method for freeway traffic state estimation integrating multi-source data. *Wirel Commun Mob Comput*, **2018**, 6745726.
- Lovisari, E., Canudas de Wit, C., Kibangou, A. Y., 2016. Density/Flow reconstruction via heterogeneous sources and Optimal Sensor Placement in road networks. *Transp Res C Emerg Technol*, **69**, 451–476.
- Lyu, C., Lu, Q. L., Wu, X., Antoniou, C., 2024. Tucker factorization-based tensor completion for robust traffic data imputation. *Transp Res C Emerg Technol*, **160**, 104502.
- Makridis, M. A., El-Baklsh, S. K., Kouvelas, A., Laval, J. A., 2025. The fundamental diagram of autonomous vehicles: Traffic state estimation and evidence from vehicle trajectories. *Commun Transp Res*, **5**, 100212.
- Nanthawichit, C., Nakatsuji, T., Suzuki, H., 2003. Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway. *Transp Res Rec J Transp Res Board*, **1855**, 49–59.
- Nie, T., Qin, G., Wang, Y., Sun, J., 2023a. Correlating sparse sensing for large-scale traffic speed estimation: A Laplacian-enhanced low-rank tensor kriging approach. *Transp Res C Emerg Technol*, **152**, 104190.
- Nie, T., Qin, G., Wang, Y., Sun, J., 2023b. Towards better traffic volume estimation: Jointly addressing the underdetermination and nonequilibrium problems with correlation-adaptive GNNs. *Transp Res C Emerg Technol*, **157**, 104402.
- Patire, A. D., Wright, M., Prodhomme, B., Bayen, A. M., 2015. How much GPS data do we need? *Transp Res C Emerg Technol*, **58**, 325–342.
- Richards, P. I., 1956. Shock waves on the highway. *Oper Res*, **4**, 42–51.
- Saffari, E., Yildirimoglu, M., Hickman, M., 2022. Data fusion for estimating Macroscopic Fundamental Diagram in large-scale urban networks. *Transp Res C Emerg Technol*, **137**, 103555.
- Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823.
- Seo, T., Bayen, A. M., 2017. Traffic state estimation method with efficient data fusion based on the Aw-Rascle-Zhang model. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 1–6.
- Seo, T., Kusakabe, T., Asakura, Y., 2015. Traffic state estimation with the advanced probe vehicles using data assimilation. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 824–830.
- Shi, R., Mo, Z., Huang, K., Di, X., Du, Q., 2022. A physics-informed deep learning paradigm for traffic state and fundamental diagram estimation. *IEEE Trans Intel Transp Syst*, **23**, 11688–11698.
- Trinh, X. S., Ngoduy, D., Keyvan-Ekbatani, M., Robertson, B., 2019. A comparative study on filtering methods for online freeway traffic estimation using heterogeneous data. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 4386–4391.
- Wang, C., Ran, B., Yang, H., Zhang, J., Qu, X., 2018. A novel approach to estimate freeway traffic state: Parallel computing and improved Kalman filter. *IEEE Intel Transp Syst Mag*, **10**, 180–193.
- Wang, R., Fan, S., Work, D. B., 2016a. Efficient multiple model particle filtering for joint traffic state estimation and incident detection. *Transp Res C Emerg Technol*, **71**, 521–537.
- Wang, R., Li, Y., Work, D. B., 2017. Comparing traffic state estimators for mixed human and automated traffic flows. *Transp Res C Emerg Technol*, **78**, 95–110.
- Wang, R., Work, D. B., 2014. Interactive multiple model ensemble Kalman filter for traffic estimation and incident detection. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 804–809.
- Wang, R., Work, D. B., Sowers, R., 2016b. Multiple model particle filter for traffic estimation and incident detection. *IEEE Trans Intel Transp Syst*, **17**, 3461–3470.
- Wang, Y., Zhao, M., Yu, X., Hu, Y., Zheng, P., Hua, W., et al., 2022. Real-time joint traffic state and model parameter estimation on freeways with fixed sensors and connected vehicles: State-of-the-art overview, methods, and case studies. *Transp Res C Emerg Technol*, **134**, 103444.
- Wei, T., Lin, Y., Guo, S., Lin, Y., Zhao, Y., Jin, X., et al., 2024. Inductive and adaptive graph convolution networks equipped with constraint task for spatial-temporal traffic data kriging. *Knowl Based Syst*, **284**, 111325.
- Wilkie, D., Sewall, J., Lin, M., 2013. Flow reconstruction for data-driven traffic animation. *ACM Trans Graph*, **32**, 1–10.
- Work, D. B., Blandin, S., Tossavainen, O. P., Piccoli, B., Bayen, A. M., 2010. A traffic model for velocity data assimilation. *Appl Math Res eXpress*, **2010**, 1–35.
- Work, D. B., Tossavainen, O. P., Blandin, S., Bayen, A. M., Iwuchukwu, T., Tracton, K., 2008. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In: 2008 47th IEEE Conference on Decision and Control, 5062–5068.
- Wright, M., Horowitz, R., 2016. Fusing loop and GPS probe measurements to estimate freeway density. *IEEE Trans Intel Transp Syst*, **17**, 3577–3590.
- Wu, F., Cheng, Z., Chen, H., Qiu, Z., Sun, L., 2024. Traffic state estimation from vehicle trajectories with anisotropic Gaussian processes. *Transp Res C Emerg Technol*, **163**, 104646.
- Wu, Y., Zhuang, D., Labbe, A., Sun, L., 2021a. Inductive graph neural networks for spatiotemporal kriging. *AAAI*, **35**, 4478–4485.
- Wu, Y., Zhuang, D., Lei, M., Labbe, A., Sun, L., 2021b. Spatial Aggregation and Temporal Convolution Networks for Real-Time Kriging. <https://doi.org/10.48550/arXiv.2109.12144>
- Xia, C., Cochrane, C., DeGuire, J., Fan, G., Holmes, E., McGuirl, M., et al., 2017. Assimilating Eulerian and Lagrangian data in traffic-flow models. *Phys D Nonlinear Phenom*, **346**, 59–72.
- Xing, J., Liu, R., Zhang, Y., Choudhury, C. F., Fu, X., Cheng, Q., 2024. Urban network-wide traffic volume estimation under sparse deployment of detectors. *Transp A Transp Sci*, **20**, 2197511.
- Yang, N., Lu, Q. L., Yamnenko, I., Antoniou, C., 2025. Efficient cloud-sourced transport mode detection using trajectory data: A semi-supervised asynchronous federated learning approach. *IEEE Internet Things J*, **12**, 11841–11857.

- Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., et al., 2021. Do Transformers Really Perform Bad for Graph Representation? <https://doi.org/10.48550/arXiv.2106.05234>
- Yuan, T., Bar-Shalom, Y., Willett, P., Ben-Dov, R., Pollak, S., 2014. A Comparison of Multiple-IMM Estimation Approaches Using EKF, UKF, and PF for Impact Point Prediction. <https://doi.org/10.1117/12.2052936>
- Yuan, Y., van Lint, J. W. C., Wilson, R. E., van Wageningen-Kessels, F., Hoogendoorn, S. P., 2012. Real-time Lagrangian traffic state estimator for freeways. *IEEE Trans Intel Transp Syst*, **13**, 59–70.
- Zhang, J., Mao, S., Yang, L., Ma, W., Li, S., Gao, Z., 2024. Physics-informed deep learning for traffic state estimation based on the traffic flow model and computational graph method. *Inf Fusion*, **101**, 101971.
- Zhao, Y., Wang, C., Rui, Y., Lu, W., Li, L., Ran, B., et al., 2025. Bidirectional temporal convolutional graph attention networks for key node identification in traffic monitoring. *IEEE Trans Intel Transp Syst*, **26**, 8720–8737.
- Zheng, C., Fan, X., Wang, C., Qi, J., Chen, C., Chen, L., 2023. INCREASE: Inductive graph representation learning for spatio-temporal kriging. In: Proceedings of the ACM Web Conference 2023, 673–683.
- Zhou, Q., Zhang, Y., Makridakis, M. A., Kouvelas, A., Wang, Y., Hu, S., 2025. Network-wide freeway traffic estimation using sparse sensor data: A Dirichlet graph auto-encoder approach. *IEEE Trans Intel Transp Syst*, **26**, 22161–22177.



Tao Guo received the B.Eng. degree in road, bridge, and river-crossing engineering from Chang'an University, Xi'an, China, in 2020, and the M.S. degree in transportation systems from the Technical University of Munich, Munich, Germany, in 2024. He is currently a Ph.D. candidate at Chair of Transportation Systems Engineering, Technical University of Munich, Munich, Germany, and a jointly supervised Ph.D. candidate at Imperial College London, London, UK. His research interests include intelligent transportation systems, spatiotemporal multiscale traffic prediction, and traffic emission estimation.



Ningkan Yang received the B.Eng. degree in road, bridge, and river-crossing engineering from Jilin University, Jilin, China, in 2021, and the M.S. degree in transportation systems from the Technical University of Munich, Munich, Germany, in 2023. He is currently pursuing the Ph.D. degree with the Chair of Transportation Systems Engineering, Technical University of Munich, Munich, Germany. His research interests include intelligent transportation systems, federated learning, and privacy-preserving artificial intelligence (AI).



transportation systems.

Kevin Yu received the B.Eng. degree in mechanical engineering (2021) and the M.S. degree in operations research and data science (2023) from the University of Edinburgh, Edinburgh, UK. He is currently a Ph.D. student with the Transport Systems & Logistics Laboratory, Imperial College London, London, UK. His research interests include physics-informed machine learning, optimal control, trajectory reconstruction, and intelligent



optimization, freight and passenger transportation, and the deployment of emerging mobility technologies, including autonomous vehicles and drones.

Panagiotis Angeloudis is a Professor of transport systems and logistics in the Department of Civil & Environmental Engineering at Imperial College London, London, UK. He holds the M.Eng. degree in civil and environmental engineering (2005) and the Ph.D. degree in transport (2009), both from Imperial College London, London, UK. His research focuses on autonomous transport systems, multiagent modeling, network



research focuses on transport policy, modeling and optimization of transportation models, data analytics and statistical learning for transportation, and human factors for future mobility.

Constantinos Antoniou is a Full Professor in the Chair of Transportation Systems Engineering at the Technical University of Munich, Munich, Germany. He received a Diploma in civil engineering from the National Technical University of Athens, Athens, Greece (1995), the M.S. degree in transportation (1997) and the Ph.D. degree in transportation systems (2004), both from the Massachusetts Institute of Technology, Cambridge, MA, USA. His