Proactive Bandwidth Allocation for V2X Networks with Multi-attentional Deep Graph Learning

Chenglong Wang, *Student Member, IEEE,* Jun Peng, *Senior Member, IEEE,* Lin Cai, *Fellow, IEEE,* Weirong Liu, *Member, IEEE,* Shuo Li, *Member, IEEE,* Hu He, *Student Member, IEEE,* Zhiwu Huang, *Member, IEEE*

***Abstract*—The increasing number of connected vehicles exacerbates the scarcity of spectrum resources in vehicle-to-everything (V2X) communication. To optimize the utilization of wireless resources, it is crucial to allocate the limited spectrum blocks to each roadside unit (RSU) based on the real-time bandwidth demand of vehicles within their coverage. However, the complex mobility patterns of vehicles and dynamic traffic conditions make it challenging to accurately and promptly estimate the bandwidth demand. To address this issue, a spatial-temporal multi-attentional network (STMA-net) is designed to predict the future bandwidth demand of RSUs. Based on the predicted bandwidth demand, a prediction error-compensable proactive bandwidth allocation algorithm is proposed to adaptively allocate spectrum resources and narrow the discrepancy between predicted and actual demand. Experimental results with realistic traffic in Bologna demonstrate that the proposed STMA-net achieves 11.25% higher prediction accuracy compared to state-of-the-art methods. Furthermore, the proposed proactive bandwidth allocation method outperforms existing methods, providing the highest throughput and serving 5% more vehicles while reducing the service drop rate by an order of magnitude.**

***Index Terms*—Proactive bandwidth allocation, Spatial-temporal mobility prediction, Prediction error compensation, Vehicular networks.**

I. Introduction

With the rapid development of connected vehicles, the bandwidth demand for supporting vehicle-to-everything (V2X) communication is ever-increasing. According to the report by Automotive Edge Computing Consortium [1], connected vehicles need to exchange MB/GB levels amount of data to facilitate advanced services such as driving assistance and autonomous driving. The increasing transmission rate requirement has put a strain on the spectrum available for V2X communication [2]. To address this escalating demand, wireless resources need to be allocated efficiently according to the real-time traffic of connected vehicles. Thus, dynamic bandwidth allocation has been considered as a promising method to improve the utilization of scarce spectrum.

Dynamic bandwidth allocation in V2X networks is challenging due to the time-varying vehicle traffic [3], [4]. To accurately estimate the bandwidth demand for each roadside unit (RSU), real-time vehicle traffic volume is required. However, the traffic conditions within the coverage area of RSUs can vary frequently due to the complex mobility patterns of vehicles [5]. This variability necessitates frequent adjustments in the bandwidth allocation strategy, which further increases the complexity of designing an effective allocation strategy. To address this issue, many researchers have increasingly focused on proactive bandwidth allocation, which predicts vehicle traffic to obtain future bandwidth demand and then allocates bandwidth resources proactively [6].

The rise of artificial intelligence and machine learning has facilitated proactive bandwidth allocation. Recently, many prediction methods including statistical methods [7], [8], machine learning [9]–[11], and deep learning [12]–[14] have been applied to allocate bandwidth proactively. However, the existing proactive methods predict vehicle traffic in a divided cell or coverage area, neglecting individual vehicle mobility within road segments and hidden dependencies from neighboring roads. As a result, the prediction accuracy in real-world road topologies may be compromised. Besides, the impact of performance degradation caused by the inaccurate prediction was not well addressed yet [15]. How to design a proactive bandwidth allocation method considering road-level feature extraction, vehicle-level mobility prediction, and prediction error compensation is an open issue.

To address this issue, we design the proactive bandwidth allocation method in two steps. Firstly, a spatial-temporal multi-attentional network is designed to predict mobility for vehicles. The designed spatial-temporal multi-attentional network (STMA-net) consists of two feature extractors, one feature fusion module, and one classifier. The GAT network is selected as the road feature extractor to capture the inherent spatial correlations among multiple road traffic features. To efficiently extract temporal correlation from driving data, GRU is selected as the vehicle feature extractor. Due to the varying traffic conditions, not all spatial and temporal features are equally important. Thus, the multi-head self-attention layer is used to fuse the extracted spatial-temporal features adaptively. Lastly, the two-layer fully connected layer serves as a classification

Chenglong Wang, Jun Peng, Weirong Liu, and Hu He are with the School of Computer Science and Engineering, Central South University, Changsha, 410083, China (e-mail: 194701019@csu.edu.cn, pengj@csu.edu.cn, weirong_liu@126.com, summerki@csu.edu.cn).

Lin Cai is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria BC V8W 3P6, Canada (e-mail:cai@ece.uvic.ca).

Shuo Li is with the School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha, 410114, China (e-mail: lishuo@csust.edu.cn).

Zhiwu Huang is with the School of Automation, Central South University, Changsha, 410083, China (e-mail: hzw@csu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 52377221 and 62172448, and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

(Corresponding author: Shuo Li.)

module to produce the final prediction results.

Then, the prediction results are converted into the estimated bandwidth demand of each RSU by using the on-off demand model. Based on the estimated bandwidth demand, a proactive bandwidth allocation method is proposed to allocate bandwidth adaptively. By employing the prediction error compensation strategy, the proposed proactive bandwidth allocation method can achieve higher throughput and serve more vehicles while maintaining the quality of services. The main contributions of this paper are given as follows.

- A Spatial-Temporal Multi-Attentional neural network (STMA-net) is designed to predict mobility for vehicles, which can achieve higher accuracy in both simple and complex road topologies.
- A lightweight proactive bandwidth allocation method with prediction error compensation (PBA-EC) is proposed. By allocating bandwidth adaptively based on real-time traffic conditions, the PBA-EC method can serve more vehicles while guaranteeing the quality of services.
- Two case studies with different road topologies have been conducted to verify the superiority of the proposed method with the state-of-the-art methods. The results show that the designed STMA-net can improve mobility prediction accuracy by 11.25% and achieve the highest throughput and demand fulfillment rate while maintaining the lowest service drop rate.

The rest of this paper is organized as follows. Relative work is introduced in the next Section. The system model and problem formulation are presented in Section III. Section IV introduced the designed STMA-net. The proactive bandwidth allocation method is presented in Section V. The numerical results based on the realistic vehicle trajectory in Bologna are presented in Section VI, followed by the concluding remarks and further research issues in Section VII.

II. Related Work

Bandwidth allocation is a crucial issue in V2X networks given high mobility and heterogeneous service requirements for vehicular infotainment, safety, and other driving assistant applications [16], [17]. Existing bandwidth allocation methods can be broadly categorized into reactive and proactive.

Reactive bandwidth allocation allocates resources on demand, which adjusts bandwidth resource allocation dynamically based on the current network conditions [18]. Reactive bandwidth allocation has been extensively studied in the literature. For instance, a location-dependent opportunistic bandwidth allocation scheme was proposed in [19] to provide higher data rates to high-mobility users. The proposed bandwidth allocation scheme used the Markov chain to find the optimal policy while addressing fair bandwidth sharing, making a good trade-off between performance gain and allocation fairness. Additionally, a two-level game-theoretic approach was proposed in [20] to maximize the utility of the network by considering the network resource distributions and service demands. In stochastic V2I scenarios, a reinforcement learning algorithm was designed in [21] to adaptively allocate bandwidth based on the channel condition.

However, reactive bandwidth allocation methods may not meet the bandwidth demand timely caused of the fast-changing traffic conditions and high mobility of vehicles. In vehicular networks, the bandwidth demand can vary drastically in time and space domains depending on the density of vehicles in each RSU. Besides, the allocation update and reconfiguration process for RSUs and base stations can be time-consuming. When the bandwidth allocation can not keep up with the changing traffic conditions, the performance of the vehicular network will deteriorate, and the quality of service can not be ensured. To mitigate this issue, proactive bandwidth allocation is required, where the bandwidth is allocated based on the predicted vehicle traffic.

Proactive bandwidth allocation is a strategy that involves predicting network conditions such as channel quality, traffic load, and mobility patterns in advance and then allocating bandwidth accordingly. Several methods have been proposed to achieve efficient bandwidth utilization. A two-step proactive bandwidth allocation method was designed in [7], which utilizes a space-time k-nearest neighbor method for short-term traffic prediction and the water-filling algorithm to allocate bandwidth. In [11], an autoregressive-moving-average (ARMA) model is designed to extract the periodic, aperiodic low-frequency temporal dependencies. By using a non-homogeneous Markov chain, the designed model can predict spectrum occupancy accurately. Another proactive bandwidth allocation method was proposed in [8], which uses Gaussian process regression to estimate queue length in the future and allocates bandwidth proportionally to each flow.

Recently, extensive works have investigated mobility prediction from both spatial and temporal perspectives. To capture complex spatio-temporal dependencies, a CNN-LSTM-based mobility prediction method was proposed in [22]. By incorporating the vector autoregression model into the proposed CNN-LSTM model, the performance of the developed network can achieve higher accuracy in forecasting short-term traffic flow. Similarly, a novel STFSA-CNN-GRU hybrid model was designed in [23] to predict vehicle short-term speed. By employing the spatial-temporal feature selection algorithm, the designed hybrid model can focus on more important features and ignore less relevant ones. Some works adopted graph neural networks for mobility prediction. In [24], a relational inductive biases-based graph neural network was proposed for short-term prediction in a few-sample case. [25] further incorporated graph neural network with LSTM, a spatial-temporal graph convolution network Bi-directional LSTM was designed to extract traffic patterns from complex real-world traffic environments. In summary, precise mobility prediction can facilitate bandwidth allocation, which motivates us to integrate mobility prediction into the design of bandwidth allocation.

Recent developments in artificial intelligence have led to the integration of neural network-based methods into proactive bandwidth allocation. In [12], a predictive dynamic bandwidth allocation algorithm used two layers of a fully connected neural network to predict the packet arrival rate and designed a dynamic bandwidth algorithm to reduce uplink latency and packet drop ratio. To achieve higher prediction accuracy,

[13] further incorporated the convolutional neural networks and residual networks to predict spatio-temporal spectrum usage of the region. Similarly, in [14], a hybrid convolutional neural network and LSTM architecture considered the spatial-temporal dependencies in vehicle traffic for bandwidth prediction. However, the existing prediction methods focus on predicting vehicle density in a divided cell or coverage area and do not consider predicting each vehicle's mobility in the road section or extracting hidden dependencies from adjacent roads. This can result in lower prediction accuracy in realistic road topologies. In addition, the existing bandwidth allocation methods do not compensate for prediction errors, which can significantly degrade performance.

In addition to the aforementioned works, attention mechanisms have gained popularity as a key technique in deep learning. Attention mechanisms were first proposed for natural language processing areas, and now are widely adopted in various deep learning models [26]. For example, attention mechanisms have been incorporated with other basic models such as RNN and CNN to improve the model performance [27]. Similarly, graph neural networks employ the attention mechanism to graph tasks, which is generally known as an efficient tool to extract nodes' hidden dependencies [28].

In summary, the necessity arises for the development of a proactive bandwidth allocation method that is capable of adapting to intricate road topologies while incorporating an effective mechanism for compensating prediction errors, which motivates this work.

III. System Model and Problem Formulation

A. System Overview

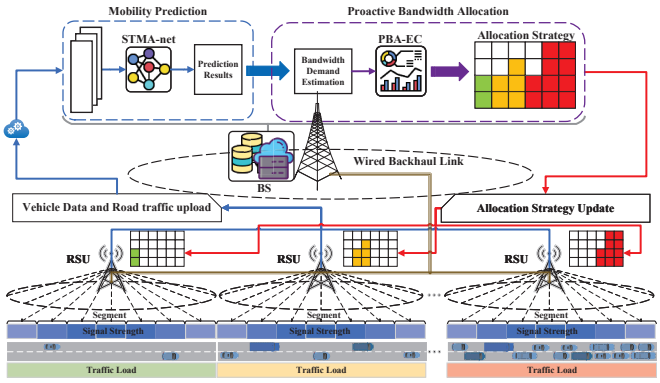


Fig. 1. The overview of the proposed proactive bandwidth allocation method, which includes STMA-net for mobility prediction and PBA-EC for proactive bandwidth allocation. The RSU provides wireless access for V2X communication, and RSUs are connected to BS through a wired backhaul link.

The system overview is shown in Fig. 1. A base station (BS) covers a large area comprising multiple road sections. The BS is responsible for managing the total spectrum resources within its coverage area and allocating bandwidth to RSUs. RSUs are responsible for supporting V2X communication for vehicles in a road section, and each RSU is wired connected to BS. RSU coverage is divided into a set of road segments. Each road segment covers several vehicles, and the channel conditions of each segment depend on the distance to the RSU.

To obtain the vehicle and road information, we assume the connected vehicles will upload their driving features, such as driving speed, azimuth, and location periodically. Each RSU also aggregates the road features within its coverage area, including the average travel time, average speed, and occupancy of the road lane. Then RSUs gather vehicle and road data and send them to the BS for training the STMA-net to extract vehicle mobility and predict the future locations of vehicles.

After that, the prediction results will be converted to the future bandwidth demand of each RSU. Based on estimated demand, the proposed PBA-EC method will compensate for the prediction error and provide the bandwidth allocation strategy sent back to RSU. Consequently, the RSUs can obtain the proactive bandwidth allocation strategy to facilitate V2X communication.

B. Demand Model

In proactive bandwidth allocation, the allocated bandwidth needs to be determined in advance. To achieve proactive dynamic bandwidth allocation, the demand distribution of RSU should be obtained first.

Assuming there is a set of BSs in a given area, denoted as S, and each BS $s \in S$ covers a set of RSUs, denoted as R. Each RSU is located in the center of its respective road section. As shown in Fig. 1, the road section r_i is divided into j smaller segments denoted as $\{r_{i,1}, r_{i,2}, ..., r_{i,j}\} \in r_i$, and the distance between two road segments is denoted by the distance between their central points. Suppose the total bandwidth resource available to the BS is W, which can be divided into z parts denoted as $W = \left[w^1, w^2, ..., w^z\right]$ for allocating to RSUs.

Considering the spectrum block reuse case, the interference of each road section includes the interference from another road segment $r_{i,*}$ in the same RSU and other road segments in a different RSU coverage. For road segment $r_{i,j}$, the interference can be given as

$$\psi_{i,j} = \sum_{r_i \setminus r_{i,j}} P h_{r_{i,j}, r_{i,*}} d^{-\delta}_{r_{i,j}, r_{i,*}} + \sum_{R \setminus r_i} P h_{r_{i,j}, r_{n,m}} d^{-\delta}_{r_{i,j}, r_{n,m}}, \tag{1}$$

where P and h denote the transmit power and small-scale fading between the sender and receiver, respectively. The parameter δ represents the path loss exponent and d is the distance between two road segments.

Accordingly, the transmit signal to interference plus noise ratio (SINR) of vehicles in $r_{i,j}$ can be given as

$$\gamma_{i,j} = \frac{P h_{r_{i,j}, r_i} d^{-\delta}_{r_{i,j}, r_i}}{\psi_{i,j} + N}, \tag{2}$$

where N represents the power of noise.

Suppose a connected vehicle in each time slot has probability p_{on} requesting service with a transmission rate R_v named on-state, and probability $1 - p_{on}$ keeping silent named off-state. If there are $v_{i,j}$ vehicles in $r_{i,j}$ in the next time slot, the possibility of bandwidth demand can be given as

$$P\{X=x\}=\left(\begin{array}{c} v_{i,j} \\ x \end{array}\right) p_{on}^{x}(1-p_{on})^{v_{i,j}-x} R_v. \quad (3)$$

Since the demand is a stochastic variable influenced by the parameter p_{on}, estimating it requires considering the probability of service drop. It is assumed that the allocated bandwidth resources for $r_{i,j}$ can achieve the transmission rate $R_{i,j}$. In that case, the service drop possibility P_D caused by insufficient bandwidth can be expressed as

$$P_D=\sum_{x=\lfloor R_{i,j}/R_v \rfloor}^{v_{i,j}} \left(\begin{array}{c} v_{i,j} \\ x \end{array}\right) p_{on}^{x}(1-p_{on})^{v_{i,j}-x}. \quad (4)$$

Generally, the service drop possibility is expected as lower as possible, thus a service drop threshold ε is defined to guarantee a low service drop rate. For instance, if the allocated strategy is expected to meet demand with 99% probability, then the threshold ε is set to 0.01. For a given threshold ε, we have

$$\sum_{x=\lfloor R_{i,j}/R_v \rfloor}^{v_{i,j}} \left(\begin{array}{c} v_{i,j} \\ x \end{array}\right) p_{on}^{x}(1-p_{on})^{v_{i,j}-x} \leq \varepsilon. \quad (5)$$

To efficiently calculate the $R_{i,j}$, the Central Limit Theorem is used to approximate the demand distribution as a Gaussian distribution with the mean value $v_{i,j}p_{on}$ and variance $v_{i,j}p_{on}\left(1-p_{on}\right)$. Then, the equation (5) can be rewritten as

$$\int_{\lfloor R_{i,j}/R_v \rfloor}^{\infty} \frac{1}{\sqrt{2\pi v_{i,j}p_{on}\left(1-p_{on}\right)}} \exp\left(\frac{-(x-v_{i,j}p_{on})^2}{2v_{i,j}p_{on}\left(1-p_{on}\right)}\right) dx \leq \varepsilon. \quad (6)$$

Based on (6), the relation between $R_{i,j}$ and ε can be given as

$$\frac{1+\operatorname{erf}\left(\frac{R_{i,j}-R_v v_{i,j}p_{on}}{\sqrt{2v_{i,j}p_{on}(1-p_{on})R_v}}\right)}{2}=1-\varepsilon. \quad (7)$$

where $\operatorname{erf}(\cdot)$ is Gauss error function. According to (7), $R_{i,j}$ can be expressed as

$$R_{i,j}=R_v\sqrt{2v_{i,j}p_{on}\left(1-p_{on}\right)}\operatorname{erfinv}(1-2\varepsilon)+R_v v_{i,j}p_{on}. \quad (8)$$

where $\operatorname{erfinv}(\cdot)$ is inverse Gauss error function. Accordingly, the estimated bandwidth demand of RSU r_i can be formulated as

$$\bar{W}_i=\sum_{j=0}^{m}\frac{R_v\sqrt{2v_{i,j}p_{on}\left(1-p_{on}\right)}\operatorname{erfinv}(1-2\varepsilon)+R_v v_{i,j}p_{on}}{\log_2\left(1+\gamma_{i,j}\right)}. \quad (9)$$

C. Problem Formulation

Due to the prediction error, it is hard to obtain the exact value of $v_{i,j}(t+1)$, so let $\hat{v}_{i,j}(t+1)$ be the predicted vehicle number, and the estimated demand $\bar{W}_{i,j}(t+1)$ is replaced by $\hat{W}_i(t+1)$. Let $\Phi_{i,j}=\left[\phi^1,\phi^2,...,\phi^z\right]$ be the bandwidth allocation indicator for $r_{i,j}$, and if the bandwidth w^k is allocated to $r_{i,j}$, $\phi^k=1$, otherwise $\phi^k=0$. Then the bandwidth allocated to $r_{i,j}$ can be given as

$$W_{i,j}(t+1)=\sum_{k=0}^{z} w^k\phi_{i,j}^k(t+1). \quad (10)$$

Different allocation strategy leads to different interference. Based on the $\Phi_{i,j}$ and the bandwidth vector W, the interference can be divided into z parts corresponding to the number of bandwidth blocks denoted as $\psi_{i,j}=\left[\psi_{i,j}^1,\psi_{i,j}^2,...,\psi_{i,j}^z\right]$. Based on the (1) and (2) the $\psi_{i,j}^k(t+1)$ can be re-expressed as

$$\begin{aligned}\psi_{i,j}^k(t+1)=&\sum_{r_i\setminus r_{i,j}}\phi_{i,*}^k(t+1)\hat{v}_{i,*}(t+1)Ph_{r_{i,j},r_{i,*}}d_{r_{i,j},r_{i,*}}^{-\delta}\\&+\sum_{R\setminus r_i}\phi_{n,m}^k(t+1)\hat{v}_{n,m}(t+1)Ph_{r_{i,j},r_{n,m}}d_{r_{i,j},r_{n,m}}^{-\delta}.\end{aligned} \quad (11)$$

According to the (10) and (11), the transmission rate in time slot $t+1$ can be represented as

$$R_{i,j}(t+1)=\sum_{k=0}^{z} w^k\phi_{i,j}^k(t+1)\log_2\left(1+\frac{Ph_{r_{i,j},r_i}d_{r_{i,j},r_i}^{-\delta}}{\psi_{i,j}^k(t+1)+N}\right). \quad (12)$$

Considering the performance loss due to the prediction error, the discrepancy function is defined as

$$H\left(t+1\right)=\sum_{r_i\in R}\left(W_i(t+1)-\hat{W}_i(t+1)\right)^2. \quad (13)$$

Accordingly, the allocation problem can be formulated as

$$\begin{aligned}\mathrm{P1}:\max_{\phi_{i,j},\hat{v}_{i,j}}&\sum_{r_i\in R}\sum_{r_{i,j}\in r_i}R_{i,j}(t+1)-H\left(t+1\right) && (14)\\ \text{s.t.}\quad&\phi_{i,j}^k(t+1)\in\{0,1\},\forall i,j,k,t, && (14a)\\&0\leq\hat{v}_{i,j}(t+1),\forall i,j,t, && (14b)\\&0\leq\sum_{k=0}^{z}w^k\phi_{i,j}^k(t+1),\forall i,j,k,t, && (14c)\\&0\leq\varepsilon\leq 1,\forall t, && (14d)\\&P_D\leq\varepsilon,\forall t. && (14e)\end{aligned}$$

where constraints (14a)-(14b) guarantee that the value of the variable is in the feasible region. Constraint (14c) restricts the allocated bandwidth is non-negative. Constraints (14d) and (14e) are defined to ensure the quality of service for connected vehicles.

In summary, the formulated problem aims to minimize the discrepancy between the allocated bandwidth and the bandwidth demand while maximizing overall network throughput. Achieving this objective requires the development of an accurate demand prediction method and an error-compensable bandwidth allocation strategy. In the next section, a spatial-temporal multi-head attention network (STMA-net) is designed to accurately predict vehicle mobility and estimate bandwidth demand. In Section IV, a proactive bandwidth allocation with the prediction error compensation (PBA-EC) is proposed, which utilizes the prediction results to allocate bandwidth resources efficiently.

IV. Mobility Prediction

To predict the mobility of vehicles, we have developed a spatial-temporal multi-head attention network (STMA-net) as illustrated in Fig. 2. The STMA-net is composed of several key components, including a graph attention (GAT) network, a two-layer gated recurrent unit (GRU) network, a multi-head self-attention layer, and a two-layer fully connected network. In the following subsection, we will provide a detailed introduction to each of these components and explain their roles in the STMA-net architecture.

A. Graph Attention Networks for Spatial Feature Extraction

In this paper, the graph attention network is used to extract road spatial features. Compared to other graph-based neural networks such as graph neural networks and graph convolution networks, graph attention network incorporates attention mechanisms to assign different weights to adjacent roads' features based on their importance. This adaptability allows the model to focus on more important road features and ignore less relevant ones. Therefore, graph attention networks are selected as a spatial feature extractor in this paper.

To represent the connectivity between these road segments, the road structure is converted into a graph structure. Each road segment's features are considered as a vertex, denoted as $\mathbf{X}^R = \{x_1^R, x_2^R, ..., x_i^R\}$. The x_i^R represents the road features of the corresponding road segment, which includes travel time, maximum speed, mean speed, and occupancy of the road. An adjacency matrix $\mathbf{A}$ is used to define the graph connectivity based on the connectivity of each road segment. Then, the features of the current and three adjacent road segments located in front, left, and right are formed into a graph $G = (\mathbf{X}^R, \mathbf{A})$, which serves as the input to the graph attention networks.

To account for dynamic traffic conditions, different road segments may have varying importance for mobility prediction. To address this, road feature extraction adopts different weights using the graph attention network (GAT) [22]. GAT employs the attention mechanism to parametrize the input feature for each node and has been widely used in traffic prediction in recent years [23], [24]. In GAT, the attention coefficients of input features are calculated as

$$e_{ij} = a\left(\left[\mathbf{W}x_i^R \| \mathbf{W}x_j^R\right]\right), \tag{15}$$

where $a(\cdot)$ is the shared attentional mechanism function mentioned in [25], and $\|$ denotes the concatenate operation. To make coefficients easily comparable across different nodes, the attention coefficient values are normalized using the softmax function as

$$\alpha_{ij} = \mathrm{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)}. \tag{16}$$

To stabilize the learning process of self-attention, the multi-head attention mechanism is used. Based on (16), the output $\hat{x}_i^R$ is given as

$$\hat{x}_i^R = \|_{k=1}^K \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k \mathbf{W}^k x_j^R\right), \tag{17}$$

where K denotes the independent attention mechanisms that execute the transformation based on the (15) and (16), then their features are concatenated as the final output.

B. Gated Recurrent Unit for Temporal Trend Extraction

A window size τ is set to convert the vehicle data into a time sequence $\mathbf{X}^V = \{x_{t-\tau}^V, x_{t-\tau-1}^V, ..., x_{t-1}^V\}$. The vehicle feature x_t^V includes driving speed, azimuth, and current coordinates of the vehicle. To extract hidden temporal inter-correlation from sequential data, recurrent neural network (RNN) is used as the temporal feature extractor. Among different types of RNN, GRU has fewer parameters, faster convergence speed, and almost the best performance compared to LSTM [26], so GRU is chosen in this paper.

The basic GRU consists of the update gate z and the reset gate r. The update gate z is used to decide whether to preserve the information from previous time slots, and the reset gate is used to decide whether to incorporate the current input with previous information or drop previous information. The detailed transition functions of GRU are given as follows:

$$\begin{aligned} z_t &= \sigma\left(\mathbf{W}_z \cdot [h_{t-1}, x_t]\right), \\ r_t &= \sigma\left(\mathbf{W}_r \cdot [h_{t-1}, x_t]\right), \\ \tilde{h}_t &= \tanh\left(\mathbf{W} \cdot [r_t * h_{t-1}, x_t]\right), \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t, \end{aligned} \tag{18}$$

where $\mathbf{W}$ is the model training parameters, h_t is the hidden state at time t, x_t is the driving data at time t, h_{t-1} is the hidden state of the layer at time $t-1$, r_t, z_t, $\tilde{h}_t$ are the reset, update, and new gates, respectively, and $\sigma(\cdot)$ is the sigmoid function.

To capture long-term dependencies, two layers of GRU are utilized. The final state of the first layer becomes the initial state of the second layer, and the output of the first layer serves as the input of the second layer. Dropout layers with a probability of 0.3 are applied in each layer of GRU to prevent overfitting. Then, the final state of the second layer will be the extracted vehicle feature $\hat{\mathbf{X}}^V$.

C. Spatio-Temporal Feature Fusion

Due to the varying traffic conditions, not all spatial and temporal features are equally important in every time slot. For instance, during road congestion, the vehicle speed may be zero, making the road features more important than the vehicle features. Thus, fusing these features equally may result in a degradation of the prediction performance. To overcome this issue, the multi-head self-attention mechanism is employed for feature fusion.

First, the extracted spatial features $\hat{\mathbf{X}}^R$ and temporal features $\hat{\mathbf{X}}^V$ are flatted as a one-dimension vector $\vec{\mathbf{X}} = \{\vec{\mathbf{X}}^R, \vec{\mathbf{X}}^V\}$ by passing a flatten layer. Then $\vec{\mathbf{X}}$ will be the input of the multi-head self-attention layer.

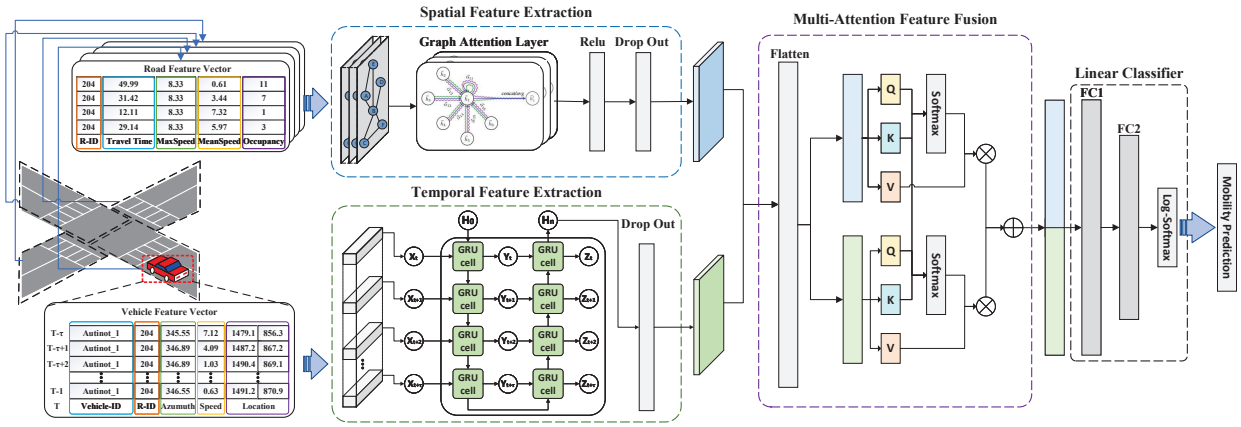


Fig. 2. The structure of the STMA-net. The vehicle features are processed by a two-layer GRU network to extract temporal features. The road traffic features are processed by a GAT to extract spatial features. The spatial and temporal features are then fused using a multi-head self-attention layer for mobility prediction.

The self-attention mechanism contains three components, the query matrix $\mathbf{Q}$, the key matrix $\mathbf{K}$, and the value matrix $\mathbf{V}$. Those three matrices are obtained by letting the input features $\vec{\mathbf{X}}$ multiply three trainable weight matrices represented as $\mathbf{Q} = \mathbf{W}^q\vec{\mathbf{X}}$, $\mathbf{K} = \mathbf{W}^k\vec{\mathbf{X}}$, and $\mathbf{V} = \mathbf{W}^v\vec{\mathbf{X}}$. Then the attention function is given by

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{19}$$

where the d_k is the dimension of K. The attention function can map a query and a set of key-value pairs to an output. Considering two types of features, the multi-head self-attention function is used for feature fusion represented as

$$\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}\left(\text{head}_1,...,\text{head}_i\right)\mathbf{W}^O, \tag{20}$$

where $\text{head}_i = \text{Attention}\left(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i\right)$, and $\mathbf{W}^O$ is the trainable weight of attention head. By mapping the hidden relationship between the spatial features $\vec{\mathbf{X}}^R$ and temporal features $\vec{\mathbf{X}}^V$, the importance of those two features can be obtained by training the weight matrices. Then, two fully connected layers are used for the final classification, which is given as

$$\hat{\mathbf{y}} = \text{softmax}\left(\mathbf{W}_2\left(\mathbf{W}_1\,\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) + \mathbf{b}_1\right) + \mathbf{b}_2\right), \tag{21}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the training weight and bias of each layer, and $\hat{\mathbf{y}}$ is a 1×4 vector, whose element denotes the probability of the vehicle staying in the current segment or entering the three adjacent segments in front, left, and right.

D. Model Training

The formulated mobility prediction problem can be taken as a classification problem, for which the cross-entropy loss function is represented as

$$Loss = -\sum_{i=1}^{n}\log\frac{\exp\left(\hat{\mathbf{y}}_i\right)}{\exp\left(\sum_{i=1}^{n}\hat{\mathbf{y}}_i\right)}\mathbf{y}_i, \tag{22}$$

where $\mathbf{y}$ is the label value, which is constructed as a one-hot code.

In each training epoch, the loss value for the entire training set is forward-propagated to calculate the training loss. Then the training loss is used to compute the gradient for back-propagation. After each epoch of training, the validation set is used to validate the performance of the model based on a loss indicator. Initially, the value of the loss indicator is set to infinity, and it will be updated if the validation loss in the current epoch is smaller than the current loss indicator. After 50-epoch iterations, the model with the smallest validation loss is saved as the trained model. In the practical system, the designed STMA-net can be implemented following the principles of offline training and online prediction. The training process is conducted offline initially to obtain the initial model parameters, and the trained model is used for online prediction with testing data. As newly collected data becomes available, the STMA-net can be further trained, and the model parameters can be updated to improve the prediction performance. It is ensured that the model can adapt to time-varying traffic and maintain its accuracy over time.

By training the proposed STMA-net, we obtain the predicted probability vector $\hat{\mathbf{y}}$, which can predict whether the vehicle will remain in the current road segment or move forward, left, or right at the next time slot. Then, the predictions for all vehicles are aggregated into the estimated vehicle number $\hat{v}(t+1)$. By using the on-off demand model presented before, the estimated bandwidth demand $\hat{W}(t+1)$ for each RSU can be obtained.

V. Proactive Bandwidth Allocation

To design a feasible allocation bandwidth algorithm, two key issues need to be addressed. First, the bandwidth allocation problem is known to be NP-hard [27], meaning that there is no polynomial algorithm to find the optimal solution, and the time complexity to find the global optimal solution is exponential. However, to allocate bandwidth proactively, the formulated problem must be solved within each time step. Second, due to prediction errors, there will always be a discrepancy between

Algorithm 1: The proactive bandwidth allocation with prediction error compensation

Input: The estimated vehicle number $\hat{v}(t+1)$ in each road segment

Output: Bandwidth allocation matrix $\phi(t+1)$

1 **Initialize:** Randomize RSUs order; $W(t+1) \leftarrow 0$; $\hat{W}(t+1) \leftarrow 0$; $\psi(t+1) \leftarrow 0$; $\phi(t+1) \leftarrow 0$.
2 **for each** *RSU* r_i **do**
3 Get estimated bandwidth demand $\hat{W}_i(t+1)$ based on $\hat{v}_i(t+1)$ and the formulated demand model.
4 **if** $\hat{W}(t+1)! = 0$ **then**
5 $\hat{W}_i(t+1) \leftarrow \hat{W}_i(t+1) \cdot \zeta$.
6 **while** $W_i(t+1) < \hat{W}_i(t+1)$ **do**
7 Select the lowest interference bandwidth block based on $\psi(t+1)$.
8 Calculate drop rate P_D based on (6).
9 **if** any RSU is dissatisfied with the drop threshold ε **then**
10 Select another bandwidth block.
11 **else**
12 Allocate this bandwidth block to r_i.
13 Update the $W_i(t+1)$, $\psi(t+1)$, and bandwidth allocation matrix $\phi(t+1)$.
14 **else if** $\hat{W}(t+1) == 0$ **and** $W(t)! = 0$ **then**
15 Select the lowest interference bandwidth block based on the interference matrix $\psi(t+1)$.
16 Calculate the service drop possibility P_D of other RSUs based on (6).
17 **if** any RSU is dissatisfied with the drop threshold ε **then**
18 Select the next lowest interference bandwidth block and return to Step 16.
19 **else**
20 Allocate this bandwidth block to r_i.
21 Update $W_i(t+1)$, $\psi(t+1)$, and bandwidth allocation matrix $\phi(t+1)$.

the estimated bandwidth demand $\hat{W}(t+1)$ and the actual bandwidth demand $W(t+1)$. This discrepancy could lead to a severe drop in the QoS since the allocated bandwidth resources may not be sufficient to meet the real demand.

To address these issues, we propose proactive bandwidth allocation with the prediction error compensation (PBA-EC) algorithm given in Algorithm 1. This algorithm enables the efficient allocation of bandwidth blocks to RSUs under two different conditions.

In the first condition, if there is a bandwidth demand in an RSU based on the prediction results, the estimated bandwidth demand $\hat{W}(t+1)$ is scaled up by a factor of ζ to compensate for any potential prediction errors. For each RSU, the bandwidth block with the lowest interference level is selected as the candidate block for allocation. To prevent service dropping caused by over-allocation, the interference matrix $\psi(t+1)$ and the drop possibility are updated after each allocation. If reusing the candidate block would result in any RSU's drop possibility exceeding the drop threshold ε, the next lowest interference bandwidth block is selected. This process is repeated until the selected candidate block meets the scaled bandwidth demand $\hat{W}(t+1) \cdot \zeta$, or none of the available blocks can meet the demand. The second condition occurs when there is no bandwidth demand in RSU r_i at time $t+1$, but there is a bandwidth demand at time t. In this case, a bandwidth block is reserved for RSU r_i for prediction error compensation. Similarly, the block will be chosen only if reusing it does not violate the drop possibility requirement.

The time complexity of the proposed PBA-EC algorithm is given as follows. For simplicity, the matrix update operation is defined as a basic operation. The proposed PBA-EC algorithm includes two loop functions, defined as an outer loop (lines 2-21) and an inner loop (lines 6-13). The outer loop iterates over each RSU, whose iteration number equals the RSU number $|R|$. The inner loop is a while loop, and the number of iterations depends on whether the condition $W_i(t+1) < \hat{W}_i(t+1)$ is satisfied. In the worst case, the iteration number equals the bandwidth block number z. In the inner loop, we update the bandwidth demand matrix $W_i(t+1)$, bandwidth allocation matrix $\psi(t+1)$, and interference matrix $\phi(t+1)$. In conclusion, the time complexity is equal to the product of the number of iterations in the inner loop, the number of iterations in the outer loop, and the basic operation count, denoted as $O(|R| \times z)$.

The PBA-EC algorithm offers two significant advantages. First, by employing a greedy strategy and selecting the bandwidth block with the lowest interference, we can achieve an approximate optimal allocation strategy. The PBA-EC algorithm has a complexity of $O(|R| \times z)$, allowing it to be efficiently solved within each time slot. This approach reduces the overall complexity of the algorithm while achieving higher network throughput. Secondly, the error compensation mechanism in the algorithm helps mitigate the performance degradation caused by the discrepancy between the predicted demand and the actual demand. During periods of low demand load, idle resources can be effectively utilized to compensate for potential performance losses. When the demand is relatively high, the updating drop possibility calculation ensures that over-allocation does not exacerbate the strain on limited bandwidth resources.

VI. Experiment Results

In this section, extensive experiments are conducted to verify the performance of the proposed proactive bandwidth allocation method. First, the prediction accuracy of the designed STMA-net is compared with state-of-art prediction methods. Then, the proposed PBA-EC algorithm is compared with existing algorithms in terms of throughput, demand fulfillment rate, and service drop rate.

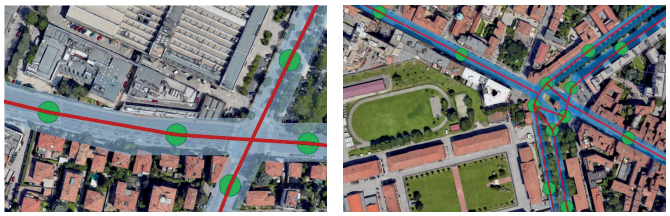
A. Scenario Setting

In this paper, we evaluate the proposed method by using the Simulation of Urban MObility (SUMO) in the real-world city

TABLE I
SCENARIO SETTING

Description	Case1	Case2
The total number of road sections	5	15
The total number of road lanes	21	32
The total number of road crosses	1	6
The total number of RSU	5	15
The total number of road segments	50	149
The road segment width	10 m	10 m
The sample number of training set	307,477	355,314
The sample number of validation set	43,925	50,759
The sample number of testing set	87,851	101,519

of Bologna. The mobility is simulated on the SUMO scenario called "Real-World Bologna" [28], which covers a portion of the inner city of Bologna spanning an area of 1500 × 1800 square meters. In our experiments, we selected the two most congested areas within this scenario.



(a) Scenario of Case 1 (b) Scenario of Case 2

Fig. 3. The real-world map and the simulation scenario of Case 1 and Case 2.

The first selected area is a long straight road with an intersection at Via Tolmino Avenue, which contains 5 road sections with a total number of 21 lanes, as shown in Fig. 3(a). According to the road topology, the 5 road sections are divided into 50 road segments with a segment width of 10 meters, and a RSU is placed in the center of each road section.

The second area has a more complex road topology, consisting of roads with different lengths and complex multi-intersections at the city center of Bologna. As shown in Fig. 3(b), the entire area contains 15 road sections with a total of 32 lanes. Similarly, the road sections are divided into 149 road segments with a segment width of 10 meters. The dataset is divided into the training set, validation set, and testing set, which consist of 70%, 10%, and 20% of total vehicles respectively. Detailed simulation settings can be found in Table I.

B. Performance Evaluation of Mobility Prediction

1) Neural Network Setting: For the neural network setting, we set the epoch number to 50 and set the batch size to 256. We use Adam as the training optimizer and apply a step learning rate to avoid over-fitting, which is initialized at 0.01 and decays every 10 epochs with a decay rate of 0.9. The detailed parameters settings are given in Table II.

2) Performance Metric: To evaluate the performance of our proposed prediction method, we utilize Accuracy as the performance metric. Accuracy is calculated using the following formula:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

where TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives in each predicted class.

TABLE II
THE HYPERPARAMETERS SETTING

Parameter	Value
Window size	5
Training epoch	50
Batch size	256
Learning rate	1e-2
Optimizer	Adam
StepLR decay	10 epochs
StepLR gamma	0.9
Drop-out probability	0.3
Hidden size of GRU layer	8
Input channel number of GAT layer	4
Attention heads number of GAT layer	4
Attention heads number of feature fuser	2
Input size of the linear layer	$32 \times 8/8 \times 4$

3) Compared Method: We compare the proposed method with four state-of-the-art methods including one classical classification method and three neural network based methods.

- ST-KNN: a spatial-temporal k-nearest neighbor classical classification-based method proposed in [7], which uses both historical temporal data and adjacent road data for mobility prediction.
- LSTM: a temporal recurrent neural network adopted in [29], which uses one LSTM layer to capture the temporal dependencies from historical data for mobility prediction.
- CNN-LSTM: a spatial-temporal deep neural network model designed in [14]. It uses a two-dimension convolutional layer (Conv2D) to capture the spatial dependencies and uses two layers of LSTM to capture the temporal dependencies. Then the spatial and temporal features are fused by two fully connected layers to obtain the prediction results.
- EVM: a spatial-temporal deep neural network model designed in [30]. It uses a one-dimension convolutional layer (Conv1D) and one max-pooling layer to capture the spatial dependencies. Then the captured features will pass two residual-GRU layers to capture the temporal dependencies. Finally, a fully connected layer fuses the extracted spatial and temporal features to obtain the prediction results.

4) Performance Evaluation on Case 1: In the simulations, the training performance comparison is presented in Fig. 4. The training accuracy in each epoch is given in Fig. 4, and the performance in the training, validation, and testing sets is shown in Table III. Since ST-KNN is a non-neural network method and does not involve epoch iterations, it is not included in Fig. 4.

As shown in Fig. 4, among all models, LSTM had the lowest accuracy of 79.54% in the training set and 80.23% in the

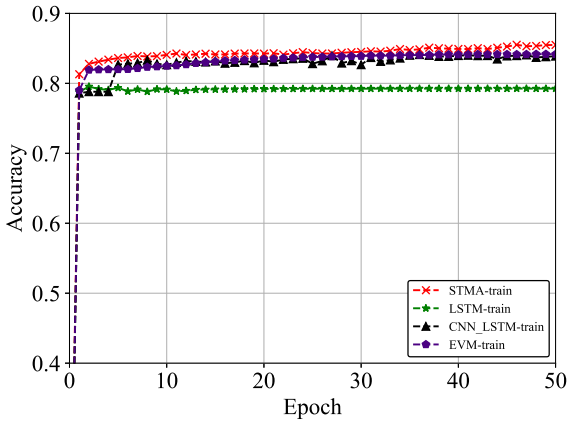


Fig. 4. The training accuracy comparison with three state-of-the-art methods in Case 1 during 50 training epochs.

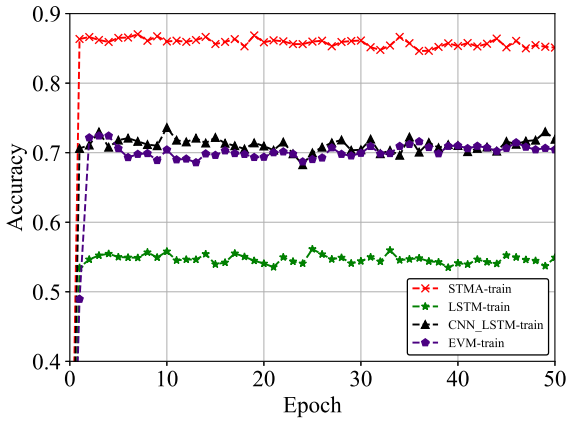


Fig. 5. The training accuracy comparison with three state-of-the-art methods in Case 2 during 50 training epochs.

TABLE III
THE PREDICTION ACCURACY IN CASE 1

Model	Prediction Accuracy		
	Training Set	Validation Set	Testing Set
ST-KNN [7]	-	-	77.49%
LSTM [29]	79.54%	80.23%	79.67%
CNN-LSTM [14]	84.09%	81.16%	80.14%
EVM [30]	84.17%	81.54%	80.97%
STMA (Proposed)	85.51%	82.55%	83.21%

TABLE IV
THE PREDICTION ACCURACY IN CASE 2

Model	Prediction Accuracy		
	Training Set	Validation Set	Testing Set
ST-KNN [7]	-	-	42.54%
LSTM [29]	56.58%	56.14%	54.93%
CNN-LSTM [14]	74.25%	73.62%	75.11%
EVM [30]	73.02%	72.43%	74.89%
STMA (Proposed)	88.61%	87.03%	86.36%

validation set. This is because LSTM can only learn temporal dependencies and lacks spatial features. CNN-LSTM and EVM, which combine spatial and temporal features, achieved an accuracy of up to 84.17% in the training set and 81.54% in the validation set, and the proposed STMA achieved the highest accuracy of 85.51% in the training set and 82.55% in the validation set.

The testing accuracy is shown in Table III. ST-KNN had the lowest testing accuracy of 77.49%, and LSTM had the second-lowest testing accuracy of 79.67%. CNN-LSTM and EVM achieved a slightly higher accuracy of up to 80.97%. By adopting the graph neural network and attention mechanism to extract spatio-temporal features, the proposed STMA achieved the highest accuracy of 83.21%. Compared to the second-highest method, the proposed STMA improved by 2.24% in Case 1.

5) Performance Evaluation on Case 2: In this subsection, we conduct another case study in more complex road topology situations. Different from Case 1, Case 2 has a more complex road topology, which contains 15 road sections with a total of 32 lanes and 6 complex multi-intersections at the city center of Bologna. Similarly, the training performance comparison is given in Fig. 5, and the detailed accuracy in the training, validation, and testing sets are given in Table IV.

As shown in Fig. 5, the proposed STMA achieves the highest training accuracy compared to other methods. LSTM has the lowest training accuracy of 56.58%, and other spatio-temporal methods including CNN-LSTM and EVM achieve training accuracy up to 74.25%, and the proposed obtained the highest accuracy of 88.61%.

The significant performance gap came from the validation set. As shown in Table IV, the temporal-based method LSTM can only get an accuracy of 56.14% in the validation set. Other spatio-temporal based methods including CNN-LSTM and EVM can reach 72.43% at most, and it has a significant drop in validation accuracy compared to the accuracy in Case 1. The reason causes the performance degradation is that CNN-LSTM and EVM do not extract the spatial features based on the road topology. The convolutional network extracts the spatial dependences through convolution kernels, and using pooling layers aggregates features by dividing the map into small grids. When the road topology is simple such as in Case 1, the grids and road topology almost overlap, so the performance is relatively close. When the road topology becomes more complex, the grid data could not reflect the real mobility of the vehicle, so the performance will degrade. Therefore, by adopting the graph attention layer, the proposed STMA achieves the highest accuracy of 87.03% in a complex road topology scenario.

The same phenomenon also revealed in the testing set given in Table IV. The compared methods have a significant performance degradation compared to the performance in Case 1. The original simulation scenario of ST-KNN is on a long straight road and predicts vehicle mobility by dividing the map into small grids, so its accuracy drops significantly when the road topology is complex. Similar to ST-KNN, the testing accuracy of other compared methods can only reach 75.11% at most. By extracting features based on the road topology, the proposed method can achieve the highest testing accuracy at 86.36%, which improves prediction accuracy by 11.25%

compared to other methods.

In summary, the experiments show that the proposed STMA-net can achieve the highest prediction accuracy compared to other state-of-art methods, and can be applied to different road topologies.

C. Performance Evaluation of Bandwidth Allocation

TABLE V
SIMULATION SETTINGS

Parameter	Value
The number of RSU	5 / 15
The radio band	[5895, 5925] MHz
The total spectrum blocks number	30
The bandwidth of spectrum block	1 MHz
The transmission rate requirement of vehicles	10 Mb/s
The maximum transmit power of vehicles P_v	23 dbm
The maximum transmit power of RSUs P_r	29 dbm
The background noise power	-104 dbm
The total time step number	5000
The service drop threshold ε	0.01

1) Simulation Setting: The simulation settings used to evaluate the performance of the bandwidth allocation method are introduced. The selected radio band is under the C-V2X standard. The maximum transmission power is set to 23 dbm for vehicles and 29 dbm for RSU respectively, and the background noise power spectral density equals -104 dBm according to the thermal noise power spectral density mentioned in [31]. The wireless link between vehicles and RSU is using the NLOS mm-wave channel model of [32]. The detailed simulation setting is given in Table V.

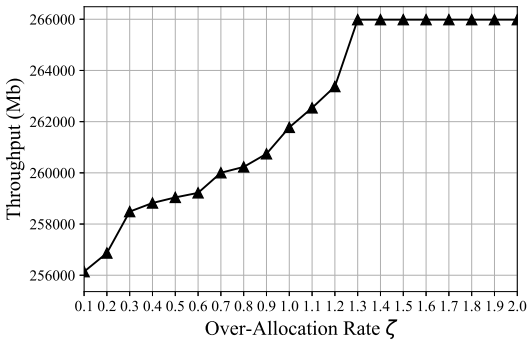


Fig. 6. The overall throughput under different over-allocation rates from 0.1 to 2.0.

As aforementioned, an over-allocation rate ζ is used for prediction error compensation. To find a suitable over-allocation rate, an experiment is conducted on ζ in the range of [0.1, 2.0]. As shown in Fig. 6, the overall throughput increases as ζ increases, and reaches a maximum when ζ equals 1.3, after that the overall throughput remains constant. It shows that 1.3 is the most suitable over-allocation rate for prediction error compensation, so ζ is set to 1.3 in the following simulations.

2) Performance Metric: In this subsection, we compare the proposed method with three state-of-the-art methods using the following evaluation metrics:

- Overall throughput: This metric represents the total throughput of all vehicles in the simulated area, which is given as:

$$\text{Overall throughput} = \sum_{t\in T}\sum_{r_i\in R} R_i(t),$$

  where R_i denotes the overall transmission rate achieved by RSU r_i in each time slot.
- Demand fulfillment rate: This metric calculates the percentage of vehicles that are allocated sufficient bandwidth resources for their V2X communication demands. It indicates how well the allocation method can meet the estimated bandwidth demands of vehicles.

$$\text{Fulfillment rate} = \frac{\sum\limits_{v_i\in V(t)} \mathbb{1}\left(\bar{W}_{v_i}(t) == \hat{W}_{v_i}(t)\right)}{\|V(t)\|},$$

  where $\bar{W}_{v_i}(t)$ and $\hat{W}_{v_i}(t)$ represent the allocated bandwidth resources and estimated bandwidth demand of vehicle v_i at time t. The function $\mathbb{1}(\cdot)$ is an indicator function, which equals 1 if the allocated bandwidth meets the requirement and 0 otherwise. $\|V(t)\|$ denotes the number of vehicles in each time slot.
- Service drop rate: This metric is defined to measure the percentage of vehicles that have been allocated bandwidth resources but experience service dropping due to insufficient bandwidth.

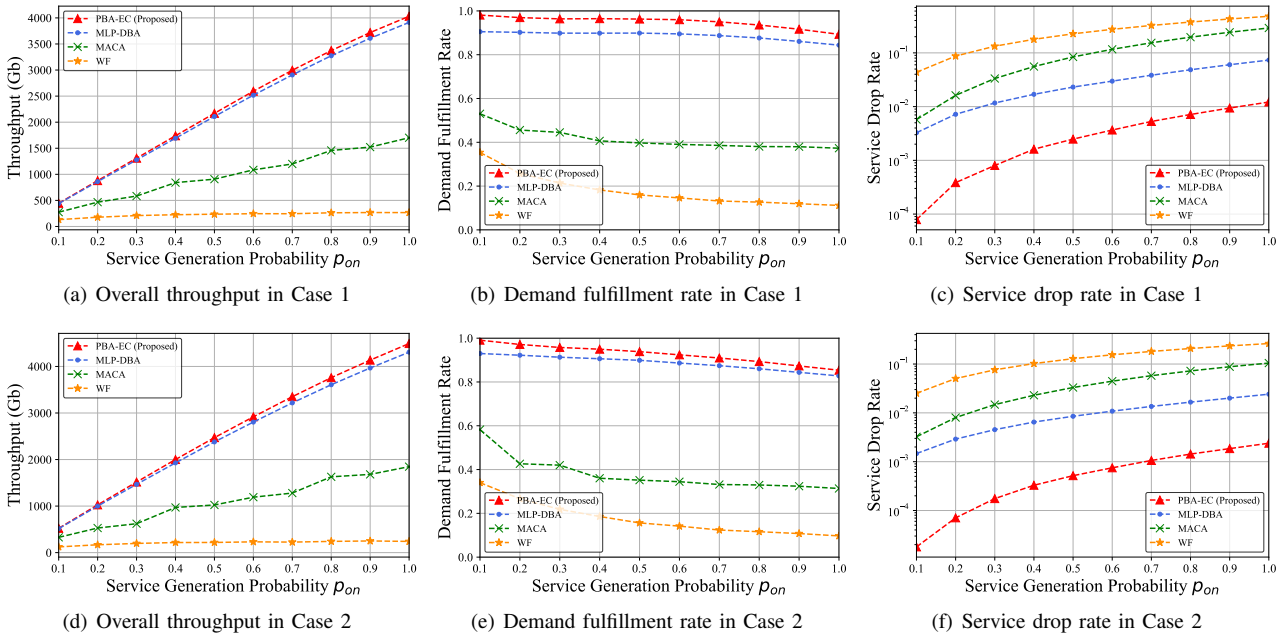
$$\text{Drop rate} = \frac{\sum\limits_{v_i\in V(t)} \mathbb{1}\left(\bar{W}_{v_i}(t) < W_{v_i}(t)\right)}{\sum\limits_{v_i\in V(t)} \mathbb{1}\left(\bar{W}_{v_i}(t) == \hat{W}_{v_i}(t)\right)},$$

  where $W_{v_i}(t)$ represents the true bandwidth demand of vehicle v_i at time t. This metric reflects the reliability of the allocation method in maintaining service for vehicles.

The demand fulfillment rate evaluates the percentage of vehicles that can access RSUs based on the allocated bandwidth. It serves as an indicator of the overall effectiveness of bandwidth allocation. While the service drop rate assesses the actual QoS for those accessed vehicles. It reflects the service drop caused by insufficient bandwidth allocation due to prediction errors.

3) Compared Method: To evaluate the performance of the proposed PBA-EC method, we conduct a comparative analysis with three state-of-the-art proactive bandwidth allocation methods:

- WF: a proactive bandwidth allocation method proposed in [7], which uses ST-KNN to predict vehicle bandwidth demand and uses the geometric water-filling method to allocate bandwidth accordingly.
- MACA: a proactive allocation method called mobility-aware cell association (MACA) method designed in [29], which uses LSTM for mobility prediction and allocated the bandwidth to the vehicles with good channel conditions to achieve maximum transmission rate.
- MLP-DBA: a machine learning prediction based dynamic bandwidth allocation method (MLP-DBA) designed in [12], which predicted the on-off status of vehicles and



(a) Overall throughput in Case 1 (b) Demand fulfillment rate in Case 1 (c) Service drop rate in Case 1

(d) Overall throughput in Case 2 (e) Demand fulfillment rate in Case 2 (f) Service drop rate in Case 2

Fig. 7. Performance comparison in terms of the overall throughput, demand fulfillment rate, and service drop rate under different service generation possibility values p_{on} ranging from 0.1 to 1.0. Two distinct cases are considered: Case 1 has a simple road topology, while Case 2 has a complex road topology.

adaptively allocates bandwidth based on the estimated bandwidth demand.

4) Performance Evaluation: Firstly, the performance comparison in terms of overall throughput, demand fulfillment rate, and service drop rate in Cases 1 and 2 are given in Fig. 7. The simulations are conducted on different service generation probability p_{on} from 0.1 to 1.0. In order to ensure fairness in our comparison, we have set the prediction accuracy for all of the proactive bandwidth allocation methods at 86.36%, which is the highest accuracy achieved by our mobility prediction method.

As shown in Fig. 7(a) and Fig. 7(d), the WF and MACA methods exhibit lower overall throughput due to their lack of consideration for interference from other RSUs. In contrast, the MLP-DBA method, which accounts for interference, achieves higher overall throughput. However, there is still performance loss caused by prediction errors. To address this issue, the proposed method employs an over-allocation strategy by assigning additional spectrum blocks to busy RSUs with an over-allocation ratio ζ. In this way, we can further improve the overall throughput compared to other methods.

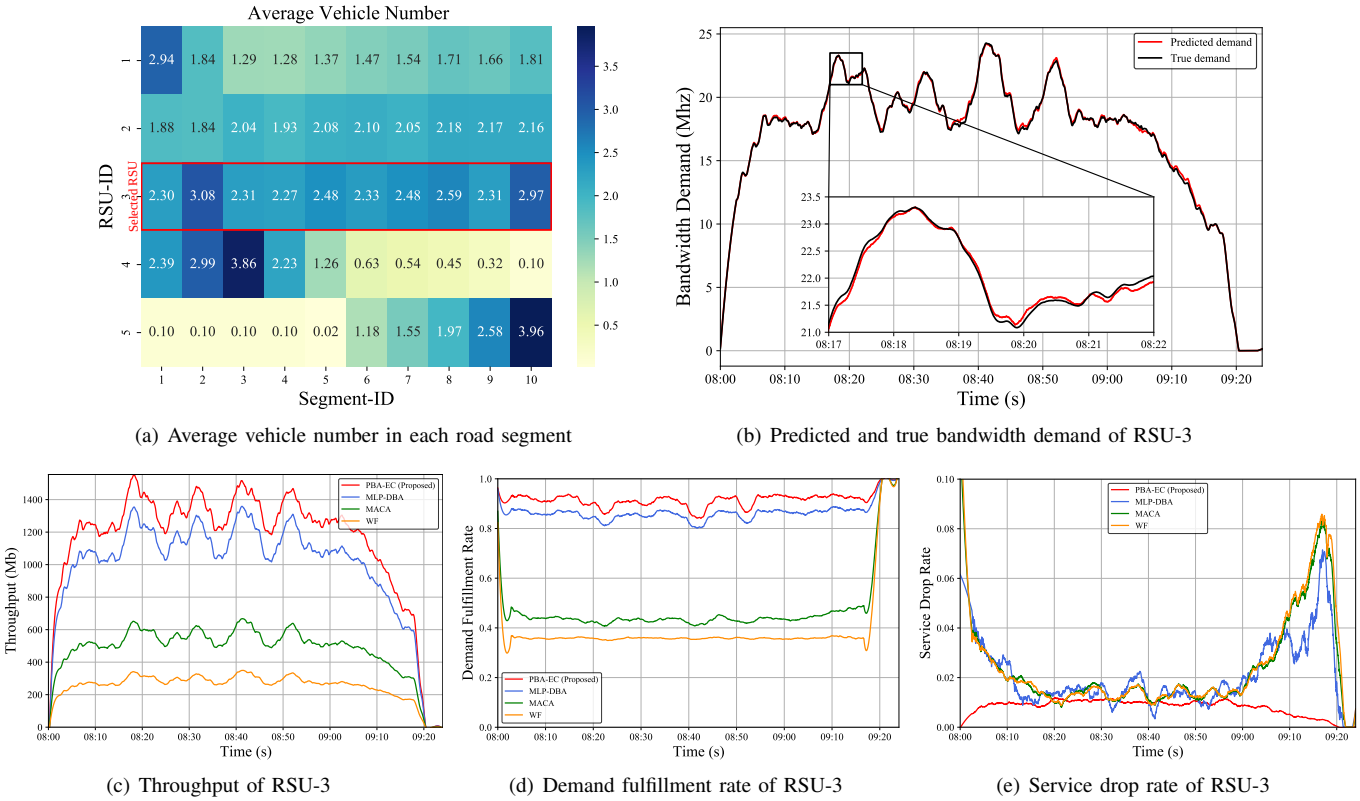
The performance comparisons of the demand fulfillment rate are shown in Fig. 7(b) and Fig. 7(e). For the WF and MACA methods, the allocated bandwidth blocks are lower than the estimated demand due to interference from intra and intro RSUs. As a result, these methods achieve a lower demand fulfillment rate. In contrast, the proposed method achieves the highest demand fulfillment rate in both two cases. Even under the highest p_{on}, the proposed method can still achieve the demand fulfillment of 89.31% in Case 1 and 85.37% in Case 2. Compared to the MLP-DBA method, the proposed method serves 5% more services in Case 1 and 3% more services in Case 2.

Considering the discrepancy between the allocated bandwidth and true demand caused by the prediction error and error-compensation strategy, there is still a possibility for vehicles that have been allocated bandwidth resources based on estimated demand to experience service dropping. To evaluate the reliability of the allocation methods, we present the performance comparison on drop rate in Fig. 7(c) and Fig. 7(f). When considering a drop rate threshold of $\varepsilon = 0.01$, both the WF and MACA methods exhibit higher drop rates, reaching up to 0.26, which fails to ensure satisfactory QoS for V2X communication. Although the MLP-DBA method achieves a lower drop rate of 0.07 in Case1 and 0.02 in Case2, it still falls short of meeting the required threshold. Among the compared methods, only the proposed PBA-EC can achieve the lowest drop rate to meet the threshold requirement. Compared to the second-lowest drop rate method, the proposed PBA-EC reduces the drop rate by an order of magnitude.

To further evaluate the proposed method in the high-load conditions, the per-second performance comparison of the highest traffic RSU under the highest service generation probability ($p_{on} = 1.0$) in two cases is given. The evaluation uses 5000 seconds of realistic traffic data, which was collected from the peak hour starting from 8 am to 9 am in the city of Bologna. The high-load performance comparisons of Case 1 and Case 2 are given in Fig. 8 and Fig. 9.

The per-second performance comparison of Case 1 is shown in Fig. 8. Fig.8(a) illustrates the average traffic in each road segment. The RSU with the highest traffic, RSU-15, is selected for performance evaluation, and its true bandwidth demand and predicted demand are presented in Fig. 8(b). It shows that during the peak time, which extends from 8:02 to 9:20, the bandwidth demand of RSU-15 significantly increases at 8:02 and gradually decreases at 9:20.

The per-second performance comparisons of Case 1 are given in Fig. 8(c), (d), and (e) respectively. In terms of per-



(a) Average vehicle number in each road segment

(b) Predicted and true bandwidth demand of RSU-3

(c) Throughput of RSU-3

(d) Demand fulfillment rate of RSU-3

(e) Service drop rate of RSU-3

Fig. 8. The per-second performance comparison of the highest traffic RSU under the highest service generation probability ($p_{on} = 1.0$) in Case 1. The average number of vehicles in each road segment is given in (a). The RSU with the highest traffic (RSU-3) is selected for performance comparison, and its prediction and true bandwidth demand are given in (b). The per-second throughput, demand fulfillment rate, and drop rate are presented in (c)-(e) respectively.

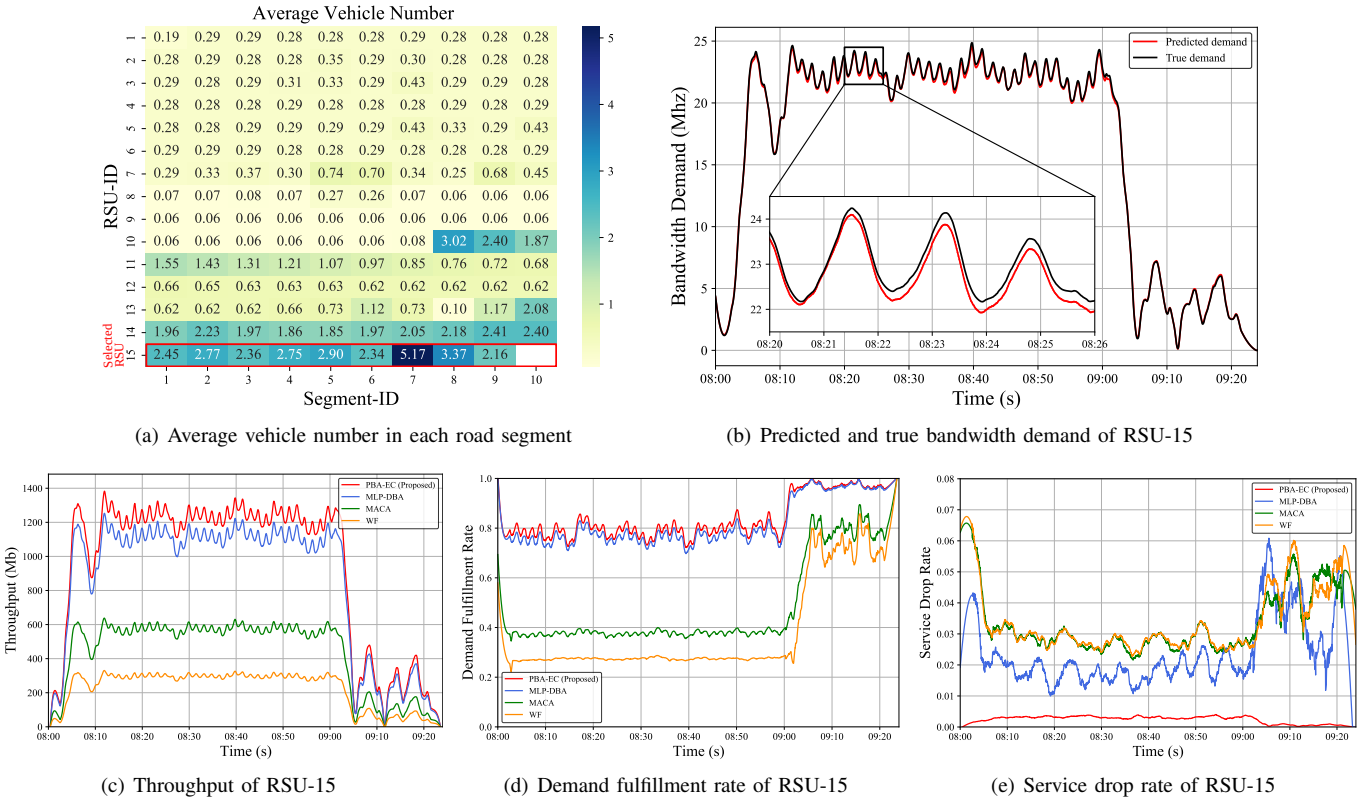
timestep throughput, the proposed PBA-EC method exhibits the highest throughput compared to other methods as shown in Fig. 8(c). In terms of demand fulfillment rate, as shown in Fig. 8(d), the proposed method achieves the highest demand fulfillment rate compared to other methods. The advantages of employing error compensation can be observed by comparing it with the MLP-DBA method. During non-peak time from 9:20 to 9:24, where total bandwidth resources are sufficient, both the proposed method and MLP-DBA achieve the highest demand fulfillment rate. However, during peak time, the prediction error can lead to an insufficient allocation of bandwidth to meet the demand. Consequently, the gap in demand fulfillment rate between the MLP-DBA method and the proposed method widens. Although error compensation strategies may slightly exacerbate bandwidth inefficiencies, compared to MLP-DBA, which does not employ error compensation, the proposed method can increase the demand fulfillment rate by 5%.

The drop rate comparison is given in Fig. 8(e). Different from the fulfillment rate evaluation, prediction errors have a higher impact during non-peak time. Due to the relatively small number of vehicles during non-peak hours, even a single prediction error can lead to a significant fluctuation in the bandwidth allocation strategy. Without the error compensation strategy, all the compared methods experience substantial performance degradation during non-peak time. However, the proposed method demonstrates the ability to maintain the lowest drop rate, around 0.01, both during peak and non-peak times.

The per-second performance comparison of Case 2 is shown in Fig. 9. Similarly, the average traffic in each road segment is shown in Fig.9(a), and the highest traffic RSU-15 is selected for performance evaluation, and its prediction and true bandwidth demand are given in Fig.9(b). The bandwidth demand of RSU-15 increases significantly at 8:02 and then suddenly decreases at 9:05, indicating the peak time of Case 2.

Similar to Case 1, the proposed method achieves the highest throughput compared to other methods both during peak and non-peak time as shown in Fig.9(c). In terms of demand fulfillment rate, as shown in Fig. 9(d), the same trend observed in Case 1 is observed in Case 2. During non-peak times, the proposed method maintains a higher demand fulfillment rate among all methods. During peak times, all methods experience a drop in fulfillment rate due to inadequate bandwidth in high traffic loads. By employing the error compensation strategy, the proposed method mitigates the performance degradation caused by prediction errors and can increase the demand fulfillment rate by 2%. In terms of service drop rate, prediction errors have a higher impact during non-peak time. Similar to Case 1, the proposed method demonstrates the ability to consistently achieve the lowest drop rate, reaching as low as 0.004. Compared to other methods, the proposed PBA-EC method manages to reduce the drop rate in both peak and non-peak times.

In summary, the simulation results show that the proposed PBA-EC method can serve more vehicles and guarantee the



(a) Average vehicle number in each road segment

(b) Predicted and true bandwidth demand of RSU-15

(c) Throughput of RSU-15

(d) Demand fulfillment rate of RSU-15

(e) Service drop rate of RSU-15

Fig. 9. The performance comparison of the highest traffic RSU under the highest service generation probability ($p_{on} = 1.0$) in Case 2. The average number of vehicles in each road segment is given in (a). The RSU with the highest traffic (RSU-15) is selected for performance comparison, and its prediction and true bandwidth demand are given in (b). The throughput, demand fulfillment rate, and drop rate are presented in (c)-(e) respectively.

lowest drop rate for communication compared to other state-of-the-art proactive methods. Note that all compared methods are conducted under the same prediction accuracy of 86.36% achieved by our prediction method, when considering their original prediction accuracy, the performance gap would be more evident.

VII. Conclusion

In this paper, a mobility-aware proactive bandwidth allocation method is proposed. First, a spatial-temporal multi-head attention mobility prediction method is designed to obtain the estimated vehicle number in each road segment. Based on the prediction result, a proactive bandwidth allocation with a prediction error compensation method is proposed to allocate bandwidth to RSUs in advance. According to the simulation results, the proposed mobility prediction method achieves the highest accuracy in both simple road topology and complex road topology cases. Especially in complex road topology cases, the proposed mobility method can improve the accuracy by 11.25%. In terms of bandwidth allocation performance, experimental results indicate that the proposed method effectively mitigates the performance degradation caused by prediction errors in proactive allocation methods. It achieves the highest throughput and serves more vehicles while maintaining the lowest drop rate compared to methods with the same prediction accuracy.

An important further research issue is to collect data sets encompassing vehicle mobility to traffic demand. In future work, we plan to conduct in-depth research to accurately estimate future demand and channel quality considering both vehicle mobility and traffic demand.

References

[1] H. Wang, T. Liu, B. Kim, C.-W. Lin, S. Shiraishi, J. Xie, and Z. Han, "Architectural Design Alternatives based on Cloud/Edge/Fog Computing for Connected Vehicles," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2349–2377, 2020.

[2] X. Gu, J. Peng, L. Cai, Y. Cheng, X. Zhang, W. Liu, and Z. Huang, "Performance Analysis and Optimization for Semi-Persistent Scheduling in C-V2X," *IEEE Transactions on Vehicular Technology*, 2022.

[3] A. K. Ligo and J. M. Peha, "Spectrum for V2X: Allocation and Sharing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 768–779, 2019.

[4] M. Noor-A-Rahim, Z. Liu, H. Lee, G. G. M. N. Ali, D. Pesch, and P. Xiao, "A Survey on Resource Allocation in Vehicular Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 701–721, 2022.

[5] C. Wang, J. Peng, L. Cai, H. Peng, W. Liu, X. Gu, and Z. Huang, "AI-Enabled Spatial-Temporal Mobility Awareness Service Migration for Connected Vehicles," *IEEE Transactions on Mobile Computing*, pp. 1–17, 2023.

[6] T. Panayiotou, M. Michalopoulou, and G. Ellinas, "Survey on Machine Learning for Traffic-Driven Service Provisioning in Optical Networks," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1412–1443, 2023.

[7] P. Chu, J. A. Zhang, X. Wang, G. Fang, and D. Wang, "Semi-persistent Resource Allocation based on Traffic Prediction for Vehicular Communications," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 2, pp. 345–355, 2019.

[8] J. Kim and G. Hwang, "Adaptive Bandwidth Allocation based on Sample Path Prediction with Gaussian Process Regression," *IEEE Transactions on Wireless Communications*, vol. 18, no. 10, pp. 4983–4996, 2019.

[9] J. Li, X. Zhang, J. Zhang, J. Wu, Q. Sun, and Y. Xie, "Deep Reinforcement Learning-Based Mobility-Aware Robust Proactive Resource Allocation in Heterogeneous Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 408–421, 2020.

[10] R. I. Rony, E. Lopez-Aguilera, and E. Garcia-Villegas, "Dynamic Spectrum Allocation Following Machine Learning-Based Traffic Predictions in 5G," *IEEE Access*, vol. 9, pp. 143 458–143 472, 2021.

[11] H. Mosavat-Jahromi, Y. Li, L. Cai, and J. Pan, "Prediction and modeling of spectrum occupancy for dynamic spectrum access systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 3, pp. 715–728, 2021.

[12] L. Ruan, M. P. I. Dias, and E. Wong, "Machine Learning-based Bandwidth Prediction for Low-latency H2M Applications," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3743–3752, 2019.

[13] X. Ren, H. Mosavat-Jahromi, L. Cai, and D. Kidston, "Spatio-Temporal Spectrum Load Prediction Using Convolutional Neural Network and ResNet," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 502–513, 2022.

[14] H. Wen, J. Yu, G. Pan, X. Chen, S. Zhang, and S. Xu, "A Hybrid CNN-LSTM Architecture for High Accurate Edge-Assisted Bandwidth Prediction," *IEEE Wireless Communications Letters*, vol. 11, no. 12, pp. 2640–2644, 2022.

[15] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware Resource Allocation for URLLC: Challenges and Strategies with Machine Learning," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 42–48, 2019.

[16] X. Jiang, F. R. Yu, T. Song, and V. C. Leung, "Resource Allocation of Video Streaming Over Vehicular Networks: A Survey, Some Research Issues and Challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 5955–5975, 2021.

[17] X. Chen, C. Wu, T. Chen, H. Zhang, Z. Liu, Y. Zhang, and M. Bennis, "Age of Information Aware Radio Resource Management in Vehicular Networks: A Proactive Deep Reinforcement Learning Perspective," *IEEE Transactions on wireless communications*, vol. 19, no. 4, pp. 2268–2281, 2020.

[18] B. P. Nayak, L. Hota, A. Kumar, A. K. Turuk, and P. H. J. Chong, "Autonomous Vehicles: Resource Allocation, Security, and Data Privacy," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 1, pp. 117–131, 2022.

[19] A. Chattopadhyay, B. Błaszczyszyn, and E. Altman, "Location Aware Opportunistic Bandwidth Sharing between Static and Mobile Users with Stochastic Learning in Cellular Networks," *IEEE Transactions on Mobile Computing*, vol. 18, no. 8, pp. 1802–1815, 2019.

[20] Q. Zhang, B. Fu, Z. Feng, and W. Li, "Utility-Maximized Two-Level Game-Theoretic Approach for Bandwidth Allocation in Heterogeneous Radio Access Networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 844–854, 2017.

[21] Z. Zhang, Q. Chang, S. Yang, and J. Xing, "Sensing-Communication Bandwidth Allocation in Vehicular Links Based on Reinforcement Learning," *IEEE Wireless Communications Letters*, vol. 12, no. 1, pp. 11–15, 2023.

[22] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph Attention Networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[23] L. Huang, X.-X. Liu, S.-Q. Huang, C.-D. Wang, W. Tu, J.-M. Xie, S. Tang, and W. Xie, "Temporal Hierarchical Graph Attention Network for Traffic Prediction," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 6, pp. 1–21, 2021.

[24] M. Fang, L. Tang, X. Yang, Y. Chen, C. Li, and Q. Li, "FTPG: A Fine-grained Traffic Prediction Method with Graph Attention Network Using Big Trace Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5163–5175, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Shewalkar, "Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235–245, 2019.

[27] B. Chen, R. Hassin, and M. Tzur, "Allocation of Bandwidth and Storage," *Iie Transactions*, vol. 34, no. 5, pp. 501–507, 2002.

[28] L. Bieker, D. Krajzewicz, A. Morra, C. Michelacci, and F. Cartolano, "Traffic Simulation for All: A Real World Traffic Scenario from the City of Bologna," in *Modeling Mobility with Open Data*. Springer, 2015, pp. 47–60.

[29] S. Manzoor, A. N. Mian, and S. Mazhar, "An LSTM-based Cell Association Scheme for Proactive Bandwidth Management in 5G Fog Radio Access Networks," *International Journal of Communication Systems*, vol. 34, no. 15, p. e4943, 2021.

[30] W. Liu and Y. Shoji, "Edge-assisted vehicle mobility prediction to support V2X communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10 227–10 238, 2019.

[31] H. Peng, Q. Ye, and X. Shen, "Spectrum management for multi-access edge computing in autonomous vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 3001–3012, 2019.

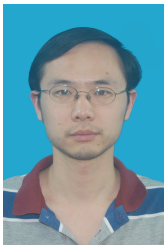
[32] T. Mangel, O. Klemp, and H. Hartenstein, "A Validated 5.9 GHz Non-Line-of-Sight Path-loss and Fading Model for Inter-vehicle Communication," in *2011 11th International Conference on ITS Telecommunications*. IEEE, 2011, pp. 75–80.

Chenglong Wang (Student Member, IEEE) received the B.E. degree from the School of Computer Science and Engineering, Central South University, Changsha, China, in 2018. He is currently working toward the Ph.D. degree with Central South University. He is currently a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada. His current research interests include mobility prediction for connected vehicles and resource management for edge networks.

Jun Peng (Senior Member, IEEE) received the B.S. degree from Xiangtan University, Xiangtan, China, in 1987, the M.Sc. degree from the National University of Defense Technology, Changsha, China, in 1990, and the Ph.D. degree in control science and control engineering from Central South University, Changsha, China, in 2005. She is currently a Professor with the School of Computer Science and Engineering, Central South University. In April 1990, she joined the staff of Central South University. From 2006 to 2007, she was with the School of Electrical and Computer Science, University of Central Florida, Orlando, FL, USA, as a Visiting Scholar. Her research interests include cooperative control and cloud computing and wireless communications.

Lin Cai (Fellow, IEEE) (S'00-M'06-SM'10-F'20) has been with the Department of Electrical & Computer Engineering at the University of Victoria since 2005 and is currently a Professor. She is an NSERC E.W.R. Steacie Memorial Fellow, a Canadian Academy of Engineering (CAE) Fellow, an Engineering Institute of Canada (EIC) Fellow, and an IEEE Fellow. In 2020, she was elected as a Member of the Royal Society of Canada's College of New Scholars, Artists and Scientists, and a 2020 "Star in Computer Networking and Communications" by N2Women. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting ubiquitous intelligence. She received the NSERC Discovery Accelerator Supplement (DAS) Grants in 2010 and 2015, respectively. She co-founded and chaired the IEEE Victoria Section Vehicular Technology and Communications Joint Societies Chapter. She has been elected to serve the IEEE Vehicular Technology Society (VTS) Board of Governors, 2019 - 2024, and served as its VP Mobile Radio from 2023 to 2024. She served as a Board Member of IEEE Women in Engineering from 2022 to 2024, and a Board Member of IEEE Communications Society (ComSoc) from 2024 - 2026. She has held various editorial roles, including Associate Editor-in-Chief for IEEE Transactions on Vehicular Technology and membership in the Steering Committee of the IEEE Transactions on Mobile Computing (TMC), IEEE Transactions on Big Data (TBD), and IEEE Transactions on Cloud Computing (TCC). She has also been an Associate Editor of the IEEE/ACM Transactions on Networking, IEEE Internet of Things Journal, IEEE Transactions on Wireless Communications, IEEE Transactions on Vehicular Technology, IEEE Transactions on Communications. Lin Cai is a Distinguished Lecturer of the IEEE VTS and IEEE Communications Societies, and a registered professional engineer in British Columbia, Canada.

Weirong Liu (Member, IEEE) received the BE degree in computer software engineering and the ME degree in computer application technology from the Central South University, Changsha, China, in 1998 and 2003, respectively, and the PhD degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2007. Since 2008, he has been a faculty member with the School of Information Science and Engineering, Central South University, where he is currently a professor. His research interests include cooperative control, energy storage management, reinforcement learning, neural networks, wireless sensor networks, network protocol, and microgrids.



Shuo Li received his B.S., M.S., Ph.D. degrees in 2006, 2009 and 2014 in the School of Information Science and Engineering, Central South University, China. He is currently an Associate Professor in the School of Electrical and Information Engineering at Changsha University of Science and Technology, China. His research interests include integrated sensing and communication and mobile computing.

Hu He (Student Member, IEEE) is currently a Ph.D. candidate in the School of Computer Science and Engineering, Central South University, Changsha, China. He received the B.E. degree in 2018 in Automation from Central South University, Changsha, China. He is currently a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada. His current research interests include UAV-enabled wireless communications and deep reinforcement learning.

Zhiwu Huang (Member, IEEE) received the B.S. degree in industrial automation from Xiangtan University, Xiangtan, China, in 1987, the M.S. degree in industrial automation from the University of Science and Technology Beijing, Beijing, China, in 1989, and the Ph.D. degree in control theory and control engineering from Central South University, Changsha, China, in 2006. He is currently a Professor with the School of Automation, Central South University. In October 1994, he joined the staff of Central South University. From 2008 to 2009, he was with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K., as a Visiting Scholar. His research interests include fault diagnostic technique and cooperative control.