Cooperative Deep Reinforcement Learning Enabled Power Allocation for Packet Duplication URLLC in Multi-Connectivity Vehicular Networks

Jianzhe Xue, *Student Member, IEEE*, Kai Yu, *Student Member, IEEE*, Tianqi Zhang, *Student Member, IEEE*, Haibo Zhou, *Senior Member, IEEE*, Lian Zhao, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—Ultra reliable low latency communication (URLLC) in vehicular networks is crucial for safety-related vehicular applications. Mini-slot with a short packet that carries only a few symbols is used to reduce the transmission time interval and enable quick scheduling for URLLC that requires extremely low latency. However, a single air interface transmission of URLLC packets may fail due to the high mobility of vehicles. Leveraging multi-connectivity technologies, the real-time reliability of URLLC can be greatly enhanced without relying on packet retransmission. In this paper, we propose a multi-connectivity URLLC downlink transmission scheme for vehicular networks, where the URLLC packet is duplicated and transmitted over multiple independent wireless links to improve packet reliability. Specifically, we design a multi-agent cooperative deep reinforcement learning algorithm, called transformer associated proximal policy optimization (TAPPO), to achieve real-time robust power allocation for multi-connectivity URLLC with imperfect channel state information (CSI). The transformer neural network architecture is employed to share the information among multiple links serving the same URLLC user and choose appropriate transmit powers, enabling cooperation to ensure reliability while minimizing inter-cell interference and energy consumption. Extensive simulation results validate the effectiveness of multi-connectivity packet duplication for URLLC and proposed TAPPO for power allocation.**

***Index Terms*—URLLC, multi-connectivity, vehicular networks, deep reinforcement learning, transformer.**

Manuscript received 12 May 2023; revised 19 October 2023; accepted 11 December 2023. Date of publication 3 January 2024; date of current version 2 July 2024. This work was supported in part by the National Natural Science Foundation Original Exploration Project of China under Grant 62250004, in part by the National Natural Science Foundation of China under Grants 62271244 and 62071398, in part by the Natural Science Fund for Distinguished Young Scholars of Jiangsu Province under Grant BK20220067, in part by the High-level Innovation and Entrepreneurship Talent Introduction Program Team of Jiangsu Province under Grant JSSCTD202202, and in part by the Natural Sciences and Engineering Research Council of Canada. Recommended for acceptance by A. Abdrabou. *(Corresponding author: Haibo Zhou.)*

Jianzhe Xue, Kai Yu, Tianqi Zhang, and Haibo Zhou are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: jianzhexue@smail.nju.edu.cn; kaiyu@smail.nju.edu.cn; tianqizhang@smail.nju.edu.cn; haibozhou@nju.edu.cn).

Lian Zhao is with the Department of Electrical, Computer, Biomedical Engineering, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada (e-mail: l5zhao@ryerson.ca).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/TMC.2023.3347580

I. Introduction

ULTRA reliable low latency communication (URLLC) is a key technology for enabling safer and smarter transportation systems through vehicular networks [1], [2]. To enhance the road safety of autonomous driving, vehicles need to receive both environment and control information from vehicular networks with high quality-of-service (QoS) [3], [4], [5]. The environment information mainly comprises high-definition maps and beyond-visual range sensing data, which enable vehicles to respond quickly and accurately to dynamic situations. Its communication is latency-sensitive because the environment information has a short validity period due to the high variability of the environment. Furthermore, the control information from the central controller of the intelligent transport system is extremely crucial for safe driving. The control information needs to be delivered to the vehicle with high reliability and low latency, leading to a thirst for URLLC. In addition, URLLC is widely needed as it enables new services and applications that can improve productivity, safety, efficiency, and user experience. In this regard, 3GPP defines URLLC as a case for the 5G communication system with reliability criteria of 10^{-5} packet loss rate and user plane latency of less than 1 millisecond [6].

However, the high-speed mobility of vehicles introduces challenges for achieving reliable and timely data transmission in dynamic and complex traffic scenarios. The requirements for URLLC can be divided into two aspects: latency and reliability. [7]. To achieve millisecond-level latency, the transmit time interval (TTI) is further subdivided into mini-slots, and the URLLC packets are downsized [8]. This reduces the delay caused by air interface transmission and enables more rapid and flexible scheduling at base stations. Moreover, to improve communication reliability, wireless networks can use redundancy to trade reliability by transmitting the URLLC packet more than once [9]. However, if these transmissions are scheduled sequentially in time, they will incur inherent time delays, resulting in an air interface delay of more than one TTI.

Multi-connectivity is envisioned as a promising solution for URLLC applications, since single-link transmissions may fail, especially for transmitting short blocklength packets over high dynamic vehicular channels with Doppler effect and unpredictable short timeslot interference [10], [11], [12], [13]. With the emergence of the cloud radio access network (C-RAN), the

baseband unit (BBU) pool can synchronously control multiple remote radio heads (RRHs), allowing for coordinated and cooperative tasks between RRHs, such as multi-connectivity [14]. To this end, we propose to use multi-connectivity to transmit the same URLLC packet simultaneously on different independent wireless links to the URLLC user in one mini-slot, without increasing the air interface delay beyond one TTI. Multi-connectivity with packet duplication enhances the real-time reliability of URLLC by exploiting the synergy of multiple independent wireless links, since the packet is received if any of these replicas from multiple links are decoded successfully, rather than depending on a single link for reliability [15], [16].

Power allocation is a crucial issue for multi-connectivity URLLC in a complex multi-cellular multi-user vehicular network system, which is also challenging because it is difficult to get perfect channel state information (CSI) from accurate channel measurements in high speed motion environments [17]. The URLLC transmission link requires a high signal-to-noise ratio (SNR) to meet its stringent quality-of-service (QoS) requirement. However, simply increasing transmit powers will cause many problems, such as high inter-cell interference and excessive power consumption, especially in the case of multi-connectivity [18], [19]. Therefore, the power allocation should ensure the URLLC reliability requirements while minimizing inter-cell interference and energy consumption. By using smart power allocation algorithms, the efficiency and reliability of communication systems can be enhanced. However, traditional optimization-based algorithms are not suitable for this problem, as they need to solve a non-convex problem for each mini-slot, resulting in a very high processing cost that cannot be done in real time [20]. On the other hand, although deep learning can make decisions quickly, it is hard to adapt to dynamic active user amounts caused by dynamic URLLC packet arrival. Most of the neural network architectures require a fixed input and output size, which is incompatible with dynamic active user amount that changes the input dimension of environment observations and output dimension of power allocation actions over time [21]. Therefore, the power allocation problem for multi-connective URLLC in vehicular networks needs an effective approach that can be accomplished in real-time and adopts dynamic active users. To this end, we propose a cooperative multi-agent deep reinforcement learning (DRL) framework for multi-connectivity URLLC power allocation.

In this paper, we present a novel multi-connectivity scheme for downlink URLLC transmission in vehicular networks, along with a robust power allocation approach based on a cooperative multi-agent DRL algorithm. Specifically, we consider a dynamic multi-cellular multi-user vehicular network downlink scenario where the communication resources are shared by both URLLC and latency-sensitive services. Our multi-connectivity scheme employs the non-coherent transmission, which duplicates the URLLC packet and transmits its replicas over multiple independent links simultaneously to the same URLLC user. In this case, the URLLC packet transmission fails only if all the independent links fail. We quantify the reliability of each link with the reliability definition for the short blocklength regime. To leverage the fast computation of neural networks, we formulate the power allocation problem for multi-connectivity URLLC as a multi-agent DRL framework with a variable number of agents, where each link acts as an agent, so that it can adopt the dynamic link amount. To achieve cooperation among multiple URLLC links, we use the transformer neural network architecture to enable information sharing among related agents and design a reward function that aligns with the common goal of all agents. Furthermore, we demonstrate the improvement of multi-connectivity for URLLC and the effectiveness of our algorithm through extensive simulations. We summarize the three main contributions of this paper as the following:

- We investigate the multi-connectivity transmission scheme for downlink URLLC transmission in vehicular networks, in which the URLLC packet is duplicated, and its replicas are transmitted over multiple independent wireless links simultaneously. The URLLC packet latency is reduced by using mini-slot, and the real-time reliability is enhanced by redundant transmission in one TTI.
- We formulate the power allocation problem for multi-connectivity downlink URLLC in vehicular networks as a multi-agent DRL problem, where each URLLC link is regarded as an agent. The number of agents can be varied in each step to accommodate dynamic network users, and transmit powers are computed in parallel using the fast forward propagation of neural networks.
- We design a cooperative multi-agent DRL algorithm, named transformer associated proximal policy optimization (TAPPO), for robust power allocation with imperfect CSI. The transformer associates the links that transmit the same URLLC packet to the same user, enabling cooperative actions by information sharing. In addition, we utilize a Gaussian random sampling process to implement fuzzy action decisions to improve algorithm robustness. Furthermore, a common reward is assigned to all agents to avoid detrimental competition.

The rest of this paper is structured as follows. Section II reviews the related works on URLLC and DRL. Section III describes the system model of multi-connectivity URLLC in vehicular networks. The DRL structure and the proposed TAPPO algorithm are presented in Sections IV and V, respectively. Section VI shows the simulation results, and Section VII concludes this work.

II. Relate Work

For the intelligent transportation system (ITS), the control center needs to accurately monitor the traffic status and schedule it in real time, and the vehicles also need to share timely information through vehicular networks in order to drive autonomously and avoid dangerous [22], [23], [24]. Many of these applications need reliable and timely data transmission, but URLLC is a key challenge in vehicular networks. As a result, URLLC draws considerable interest from both academic and industrial sectors [25], [26]. In general, communication resources are limited and crucial for meeting the extreme requirements of URLLC, so accurate communication modeling and tailored system design are essential [27]. Traditionally, the reliability of URLLC is

measured by the SNR with a basic assumption that the packets are transmitted successfully if Shannon' s capacity is above a required data rate, and otherwise, the transmission fails if the SNR is below the threshold [28]. However, this assumption does not hold for the short packets in URLLC as the decoding error probability should be taken into account [29], [30]. In the short blocklength regime, the achievable rate and the decoding error probability may be underestimated if the Shannon's capacity is directly applied. Recently, many works have been carried out with the decoding error probability in the short blocklength regime [31]. For example, a bipartite graph and a non-cooperative game are used to jointly optimize the power consumption and the network stability in a mobile edge computing vehicular network with inter-cell interference [32]. To reduce computational overhead, a deep learning framework is developed as an approximate optimal resource allocation strategy for various QoS requirements, including URLLC, in cellular networks, where the total power consumption of the base stations is minimized by optimizing transmit power and bandwidth [20]. Besides, network slicing for enhanced mobile broadband (eMBB) and URLLC is enabled by a mini-slot based transmission with punctured scheduling techniques, and a hierarchical deep learning framework is designed to solve the slicing problem while addressing the traffic dynamics [33].

Multi-connectivity with packet duplication is a promising technique to enhance the reliability of URLLC in vehicular networks [34]. The basic idea of packet duplication is to create multiple copies of a packet and send them simultaneously over multiple independent channels. It is demonstrated that redundancy transmission with packet duplication makes it possible to simultaneously satisfy the latency and reliability requirements without increasing the complexity of the radio access network of C-RAN [9]. In addition to URLLC, packet duplication can also be used with dynamic control to improve transmission robustness during mobility and radio link failures. The expression for the probability that multi-connectivity meets the reliability demand of the URLLC user is derived and a low-complexity algorithm is proposed to jointly select the coding scheme, the modulation, and the set of cooperating base stations [35]. The study of a framework that leverages multi-connectivity to increase the maximum communication distance while satisfying the network availability requirement is conducted, in which each packet is transmitted by both device-to-device and cellular links [15].

Deep learning approaches have been viewed as promising methods of creating enabling technologies for URLLC in upcoming networks [31]. Long short-term memory (LSTM) network is a model that can capture features in sequential data, where the current output is influenced by the information of previous inputs [36]. A multi-agent DRL algorithm with a sequential actor-critic model is proposed to optimize the delay and reliability satisfaction packets in the multi-connectivity cellular network, by making packet duplication decisions according to the observations of the communication environment, such as load, channel state, and interference [37]. Moreover, since the LSTM can only pass information in one direction, the stacked bidirectional LSTM model is developed to allow the neural network to capture the bidirectional information. It can serve as an effective supervised deep learning approach to forecast data traffic on a large timescale from traffic patterns and then allocate resources to radio access network slices [38]. Furthermore, attention mechanisms have been widely used to perform various machine learning tasks due to their adaptability and effectiveness in modeling dependencies. The transformer is a powerful neural network architecture that can share information and capture correlations between data blocks [39]. A DRL-based algorithm with the transformer as its backbone is proposed to optimize the QoS of video followers and the energy efficiency by jointly performing uplink transmission and edge transcoding [40]. Besides, some multi-agent DRL based on deep deterministic policy gradient (DDPG) is studied for power allocation tasks with continuous action space, however, their robustness under imperfect CSI conditions may be limited [41], [42].

III. System Model

This section introduces the system model, channel model, and communication model of both latency-sensitive service and multi-connectivity URLLC service.

A. System Model

We consider a multi-connectivity aided orthogonal frequency division multiple access (OFDMA) downlink vehicular network system with C-RAN architecture, where each user (UE) can connect to one or more RRHs with multiple independent wireless links to receive data packets simultaneously. Note that a wireless link among multi-connectivity of the same UE is defined as a connection between the UE and an RRH on a sub-carrier. We apply the non-coherent transmission technology for multi-connectivity, which allows these links of the same UE to use different frequency sub-carriers and to transmit any data packet at any time independently of the other links. In other words, the multi-connectivity links of the same UE are physically uncorrelated and independent of each other in the radio access network. In our model, we consider two specialized QoS services required by vehicular network UEs, including latency-sensitive service and URLLC service, and UEs may require these two services simultaneously. For latency-sensitive services, a maximum air interface transmission delay should be satisfied. For URLLC services, packets should be transmitted within one mini-slot immediately after arrival, to avoid the queuing delay and the transmission reliability probability should be higher than a required threshold, such as 99.999%.

The system model is illustrated in Fig. 1, where there are B RRHs serving K UEs with the ability of multi-connectivity. The latency-sensitive packet is transmitted over a single link until the packet transmission is completed and is not affected by the mini-slot concept. The URLLC packet is transmitted over multiple links using different sub-carriers simultaneously and independently to enhance the packet reliability within one mini-slot TTI. For each downlink connection link, we can divide all RRHs into two types based on whether they serve this UE or not: the transmission RRH and the interference RRH set $\mathcal{B}^{\mathrm{I}}$. The UE may receive inter-cell interference if the RRHs in $\mathcal{B}^{\mathrm{I}}$ are

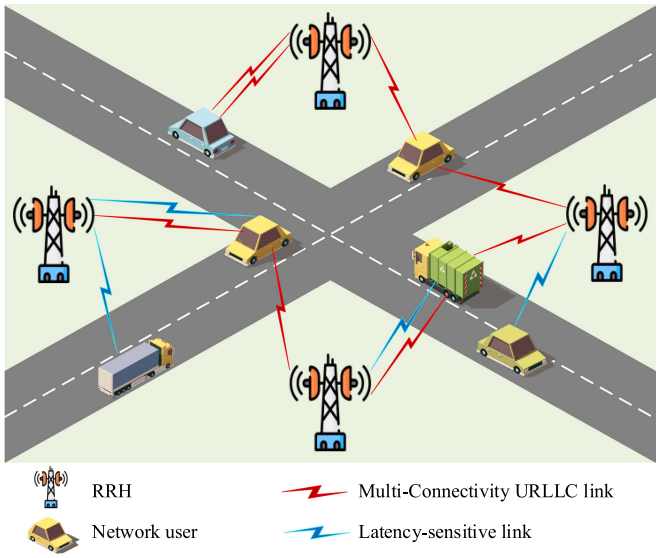


Fig. 1. System model.

serving other UEs on the same frequency sub-carrier. Therefore, allocating suitable transmit powers for wireless links to mitigate inter-cell interference is crucial for meeting stringent URLLC QoS requirements.

B. Channel Model

The downlink channel is a multiple input single output (MISO) channel, where the RRH has N_{T} transmit antennas and the vehicle has a single receive antenna. The duration of one mini-slot TTI and the bandwidth of each subcarrier in the OFDMA system are denoted by τ and W, respectively. The noise power on each subcarrier is σ^2. The channels are block fading in both frequency and time domains. Each RRH has N_{C} sub-carriers in total, and all RRHs are using the same frequency range. We assume the channel state information (CSI) is imperfectly estimated and the maximum ratio transmission (MRT) is applied for precoding at RRH based on the estimated CSI.

Due to the high mobility of vehicles, the Doppler shift greatly influences the small scale fading of vehicular channels. Moreover, the CSI of high dynamic vehicular channels is obtained with latency due to the channel estimation process. Therefore, for the channel estimation model, we assume the RRH can only obtain the estimated small scale fading, $\hat{\mathbf{h}}$, with the error $\mathbf{e}$, which are both independent identically distributed and subject to $\mathcal{CN}(0,1)$. The small scale fading of the vehicular channel is modeled with the first-order Gauss-Markov process, and the real small scale fading of the vehicular channel can be represented as [17], [43], [44],

$$\mathbf{h} = \sqrt{\beta^2}\hat{\mathbf{h}} + \sqrt{1-\beta^2}\mathbf{e}, \tag{1}$$

where $\beta \in [0,1]$ is the channel estimation error coefficient that indicates the estimation accuracy. Specifically, $\beta = 1$ represents a perfect channel estimation case and $\beta = 0$ represents a case in which there is no information about the channel CSI. Based on the Jakes statistical model for fading channel, the channel estimation error coefficient is given as,

$$\beta = J_0(2\pi f_D T), \tag{2}$$

where T is the channel feedback latency, $f_D = v f_c / v_c$ denotes the maximum Doppler frequency with v as the vehicle speed, f_c as the carrier frequency and $v_c = 3 \times 10^8$ m/s as the speed of the light. $J_0(\cdot)$ represents the zero-order Bessel function of the first kind and is defined as,

$$J_0(x) = \sum_{q=0}^{\infty} \frac{(-1)^q}{q!\Gamma(q+1)} \left(\frac{x}{2}\right)^{2q}. \tag{3}$$

We apply the MRT for the preceding based on the estimated small scale fading $\hat{\mathbf{h}}$. The precoding weights at the b-th RRH for the k-th UE on the n-th subcarrier are calculated as,

$$\boldsymbol{w}_{k,b,n} = \frac{\sqrt{\alpha_{k,b,n}}\hat{\mathbf{h}}^*_{k,b,n}}{||\sqrt{\alpha_{k,b,n}}\hat{\mathbf{h}}_{k,b,n}||_2}, \tag{4}$$

where $\hat{\mathbf{h}}^*_{k,b,n}$ is the conjugate transpose of $\hat{\mathbf{h}}_{k,b,n}$, and $||\sqrt{\alpha_{k,b,n}}\hat{\mathbf{h}}_{k,b,n}||_2$ is the 2-norm of $\sqrt{\alpha_{k,b,n}}\hat{\mathbf{h}}_{k,b,n}$.

In our vehicular network model, the useful received signal power $y_{k,b,n}$ at the k-th UE from the b-th RRH on the n-th subcarrier can be represented as,

$$y_{k,b,n} = \alpha_{k,b,n}|\mathbf{h}_{k,b,n}\boldsymbol{w}_{k,b,n}|^2 P_{k,b,n}, \tag{5}$$

where $\alpha_{k,b,n}$ is the large scale channel gain, $\mathbf{h}_{k,b,n} \in \mathbb{C}^{N_{\mathrm{T}}}$ is the channel coefficient, $\boldsymbol{w}_{k,b,n} \in \mathbb{C}^{N_{\mathrm{T}}}$ is the precoding weights for N_{T} transmit antennas at RRH, and $P_{k,b,n}$ is the transmit power. The interference received signal power at the k-th UE on the n-th sub-carrier is,

$$\gamma_{k,n} = \sum_{l \in \mathcal{B}^{\mathrm{I}}} \alpha_{k,l,n}|\mathbf{h}_{k,l,n}\boldsymbol{w}_{l,n}|^2 P_{l,n}\rho_{l,n}, \tag{6}$$

where $\alpha_{k,l,n}$ and $\mathbf{h}_{k,l,n}$ are the large scale channel gain and the channel coefficient from the l-th interference RRH to the k-th UE on the n-th subcarrier, respectively. $\boldsymbol{w}_{l,n}$ and $P_{l,n}$ are the precoding weights and the transmit power at the l-th interference RRH on the n-th subcarrier, respectively. $\rho_{l,n}$ is the binary spectrum allocation indicator with $\rho_{l,n} = 1$ implying the l-th interference RRH is serving another UE on the n-th subcarrier and $\rho_{l,n} = 0$ otherwise.

C. Latency-Sensitive Service

The latency-sensitive service is used for transmitting the environment information of vehicles. The more comprehensive and real-time environmental information available, the more rational and safe driving decisions can be made by intelligent vehicles. Therefore, all vehicles are eager for more real-time environmental information, and we assume the latency-sensitive data packets are consecutive and continuous.

The latency-sensitive service uses the traditional single-connectivity method. The transmission rate of a single latency-sensitive service link between the k-th UE and the b-th RRH at

time t is measured by the Shannon's capacity, which is given by,

$$c_{k,b}^{\mathrm{L},t}=\sum_{\mathrm{n}=1}^{\mathrm{N_k^L}} W\log_2\left(1+\frac{y_{k,b,n}}{\gamma_{k,n}+\sigma^2}\right), \tag{7}$$

where $y_{k,b,n}$ is the received signal at the k-th UE from the b-th RRH on the n-th subcarrier, N_k^L is the number of sub-carriers allocated to the k-th UE. We mathematically model the latency requirement of each latency-sensitive packet as,

$$\begin{cases}\sum_{t=1}^{T_{k,b}} c_{k,b}^{\mathrm{L},t} \geq a_{k,b}^{L}, \\ T_{k,b} \leq T_{\max}, \end{cases} \tag{8}$$

where $a_{k,b}^L$ is the packet length, $T_{k,b}$ is the time length used for the packet transmission and $T_{\max}$ is the expected maximum air interface transmission delay.

D. URLLC Service with Multi-Connectivity

The URLLC service transmits the control and emergency information that requires ultra reliability and low latency. The packet arrival process of URLLC packets of a UE can be modeled as a Bernoulli process [45]. To minimize the delay caused by the long frame structure, we divide the time resources into mini-slots and shorten the data packet length to fit the mini-slot. These mini-slots reduce the air interface delays and enable a more flexible scheduling mechanism with lower scheduling delays.

In wireless connections for URLLC packet transmission, transmission reliability must be taken into account in addition to the transmission rate. URLLC packets are typically small (e.g., 32 bytes) and use short blocklength channel coding to meet the mini-slot requirement. However, the decoding errors in small packets cannot be ignored, and Shannon' s capacity alone is not sufficient to evaluate their transmission reliability. Based on the Normal Approximation of the achievable rate in the short blocklength regime [46], [47], the achievable rate of the MISO channel can be approximated by,

$$c^{\mathrm{U}} \approx \frac{W}{\ln 2}\left\{\ln\left(1+\frac{y_{k,b,n}}{\gamma_{k,n}+\sigma^2}\right)-\sqrt{\frac{\Omega}{\tau W}}\mathcal{Q}_G^{-1}(\delta_{k,b,n})\right\}, \tag{9}$$

where $y_{k,b,n}$ is the received signal at the k-th UE from the b-th RRH on the n-th subcarrier, $\gamma_{k,n}$ is the interference signal at the k-th UE on the n-th subcarrier, τ is the data transmission duration, $\delta_{k,b,n}$ is the decoding error probability, Ω is the channel dispersion that quantifies how much the channel varies randomly compared to a fixed channel with the same capacity [48], which is given by,

$$\Omega = 1-\frac{1}{\left(1+\frac{y_{k,b,n}}{\gamma_{k,n}+\sigma^2}\right)^2}. \tag{10}$$

Note that Ω is accurately approximated to 1 when the SNR is higher than 5 dB and this approximate condition can easily be achieved in cellular networks, especially for URLLC service [29]. $\mathcal{Q}_G(\cdot)$ is the Gaussian Q-function defined as,

$$\mathcal{Q}_G(x)=\int_x^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt. \tag{11}$$

Then, the decoding error probability of a URLLC transmission packet between the k-th UE from the b-th RRH on the n-th subcarrier can be derived as,

$$\delta_{k,b,n}=\mathcal{Q}_G\left\{\sqrt{\tau W}\left[ln\left(1+\frac{y_{k,b,n}}{\gamma_{k,n}+\sigma^2}\right)-\frac{a_{k,b}^{\mathrm{U}} ln2}{\tau W}\right]\right\}, \tag{12}$$

where $a_{k,b}^{\mathrm{U}}$ is the packet length of the URLLC packet.

We employ packet duplication for multi-connectivity URLLC, in which a URLLC packet is duplicated and simultaneously transmitted to the UE over multiple independent wireless links of different sub-carriers within the same mini-slot [11]. The rationale behind packet duplication is to trade off delay and reliability by exploiting transmission redundancy. In packet duplication multi-connectivity URLLC, the URLLC packet is only lost if all the transmissions on different links fail concurrently. Since the decoding failure of each link is an independent and unrelated event, the transmission failure probability of a URLLC packet with packet duplication multi-connectivity can be derived as [9],

$$\Delta_k=\prod_{\mathrm{m}=1}^{\mathrm{M_k^U}}\delta_{k,b,n}^m, \tag{13}$$

where M_k^{U} is the number of links that are used to transmit a URLLC packet simultaneously for k-th UE, $\delta_{k,b,n}^m$ is transmission failure probability of the m-th link of the k-th UE. Finally, the reliability of the k-th URLLC UE with multi-connectivity can be expressed by,

$$\mathcal{U}_k=-log_{10}(\Delta_k). \tag{14}$$

In this expression, $\mathcal{U}$ is a positive number that means how many "9" are there in the reliability percentage, for example, $\mathcal{U}=5$ representing 99.999% reliability of a URLLC packet.

IV. Deep Reinforcement Learning Framework and Problem Formulation

In this section, we explain how to formulate the problem of power allocation for multi-connectivity URLLC in vehicular networks using a DRL framework. We also provide a detailed description of the key elements of our DRL framework.

The structure of our DRL framework for power allocation of multi-connectivity URLLC is illustrated in Fig. 2. It mainly comprises five components: environment, agent observation state, agent action, reward, and memory. The environment is a simulator of a physical world communication system that includes latency-sensitive service and URLLC service. The packet arrival of URLLC is dynamic, which leads to a dynamic number of active URLLC links in each time slot. However, conventional DRL requires fixed input and output sizes and cannot adapt to the state and action space dynamics. Therefore, we address the challenge by converting it into a multi-agent DRL problem, where each active URLLC link is treated as an agent with a fixed state and action space. All agents have the same state and action space, and the number of agents in the environment at each time slot matches the number of URLLC links. In this way, dynamic changes in the number of active URLLC links

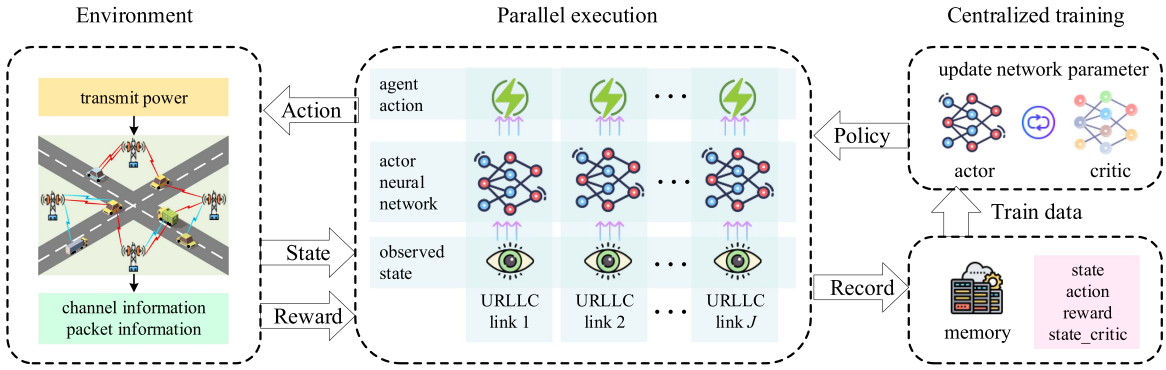


Fig. 2. Deep reinforcement learning framework.

only influence the number of agents in the environment, and do not require changing the state and action space. Note that all the agents share the same neural network parameters as their brains. The parameters of both actor and critic neural networks are periodically updated with the memory data.

A. Environment

The environment of our DRL framework comprises five components: RRHs, vehicular network UEs, wireless channels, central controller, and reward manager. The RRHs and vehicular network UEs are located on the map according to the simulation settings, and their channels are generated based on their positions and channel models. The environment is a time-discrete simulator, and it is updated every mini-slot. At first, the central controller located at the BBU pool of C-RAN receives and updates the transmission demands of both latency-sensitive and URLLC packets. For the latency-sensitive packet, the controller selects a sub-carrier with the highest channel gain to start the transmission and keeps the same sub-carrier and transmitting power until this packet is successfully transmitted. For the URLLC packet, the controller chooses multiple independent sub-carriers with the highest channel gains from all the available sub-carriers to establish multi-connectivity. Note that with the help of the C-RAN architecture and the BBU pool scheduling, multiple URLLC links of a UE at the same mini-slot can be connected to one or more RRHs. These multi-connectivity URLLC links transmit the replicas of one packet to the target UE with the transmit power determined by the proposed DRL algorithm. Then, the transmission is simulated with the MRT precoding based on the estimated channel CSI. Finally, the reward manager gives the reward for the current mini-slot based on the consideration of three aspects: latency-sensitive packet delay, URLLC packet reliability, and transmit power.

B. Agent Observation State

The environment has many active URLLC links in every mini-slot. Each URLLC link observes the environment independently. Each URLLC link observation comprises three components: the channel estimation error coefficient β of the URLLC link channel, the channel gains for the sub-carriers operating on the same frequency of all RRHs, and the transmission demands for the sub-carriers operating on the same frequency of all RRHs.

Suppose that the j-th URLLC link is serving the k-th UE with the serving RRH on its n-th frequency sub-carrier. The channel estimation error coefficient of the j-th URLLC link is denoted as β_j. The state of the b-th RRH is denoted as $s_{b,n} = \{g_{k,b,n}, d_{b,n}\}$, where $g_{k,b,n}$ is the channel gain from the b-th RRH to the k-th UE on the n-th frequency sub-carrier, $d_{b,n}$ is URLLC transmitting packet length of the b-th RRH on the n-th frequency sub-carrier. Note that if the n-th sub-carrier of the b-th RRH is vacant, $g_{k,b,n}$ and $d_{b,n}$ are both set to 0. If n-th sub-carrier of b-th RRH is used for latency-sensitive service, $g_{k,b,n}$ is the channel gain and $d_{b,n}$ is set to 0. Otherwise, $g_{k,b,n}$ and $d_{b,n}$ are their normal values. With the above definitions, the observation state of the j-th URLLC link is a combination of $s_{b,n}$ of all RRHs and β_j as,

$$S_j = \{s_{1,n}, ..., s_{b,n}, ..., s_{B,n}, \beta_j\}. \tag{15}$$

The dimension of one URLLC link state S is $2B + 1$. Then, $s_{b,n}$ in S_j will be ordered by their distance between the b-th RRH and the j-th URLLC link serving RRH from smallest to largest. So that the $s_{b,n}$ of the j-th URLLC link serving RRH is always fixed at the first position observation vector. Finally, the environment states at time t is given as,

$$\mathcal{S}_t = \{S_1, ..., S_j, ..., S_J\}, \tag{16}$$

which is a combination of all the observations of active URLLC links.

C. Action

The actions of agents are the URLLC link transmit powers. The action of the j-th URLLC link is in one dimension and denoted as a_j, representing its transmit power at the serving RRH. Since the transmitting power is limited by the physical world, a_j is a continuous variable within the range, $a_j \in [a^{\min}, a^{\max}]$. All the agents calculate their actions simultaneously with the same policy π based on their observation states. The total actions at time t can be denoted as,

$$\mathcal{A}_t = \{a_1, ..., a_j, ..., a_J\}, \tag{17}$$

which is a combination of all transmit power of all active URLLC links.

D. Reward and Problem Formulation

The flexibility of the DRL reward makes it effective in addressing problems with complex objectives. When the designed reward aligns with the desired objectives, the performance of the communication system can be enhanced. In our multi-agent DRL framework, to encourage agents to cooperate effectively rather than engage in a harmful competition situation, we let all agents share the same reward which indicates the overall performance of the communication system. In this way, when choosing actions, the agents have to take into account their impacts on other agents as well as the performance of the whole system. Therefore, the agents need to learn a cooperative policy that not only satisfies their own QoS requirements but also reduces the interference to others, leading to the optimal resource allocation solution of the vehicular network communication system.

In the power allocation problem of multi-connectivity URLLC links, our objective consists of three components: maximizing the QoS guarantee rate of the multi-connectivity URLLC packet reliability, minimizing the air interface delay of the latency-sensitive packet, and minimizing the total URLLC transmit energy. Each of these components is represented by a sub-reward that is mathematically expressed by a linear piecewise function, $f(\cdot)$, with its maximum value normalized to 1. The reward of the k-th URLLC UE at time t is denoted as $R^{\mathrm{U}}_{t,k}$. It is used to ensure that the multi-connectivity URLLC reliability $\mathcal{U}_{t,k}$ is greater than 5 (representing 99.999% in percentage) but not too much since redundant reliability leads to wasted communication resources. The function $R^{\mathrm{U}}_{t,k} = f^{\mathrm{U}}(\mathcal{U}_{t,k})$ is an increasing and then decreasing function, and it reaches the maximum reward 1 when $\mathcal{U}_{t,k} = 6$. For all URLLC UEs at time t, the average reward about URLLC packet reliability is given by,

$$\overline{R^{\mathrm{U}}_t} = \frac{1}{K^{\mathrm{U}}_t} \sum_{k=1}^{K^{\mathrm{U}}_t} f^{\mathrm{U}}(\mathcal{U}_{t,k}), \tag{18}$$

where K^{U}_t is the total number of URLLC UE at time t. The reward of the k-th latency-sensitive UE at time t is denoted as $R^{\mathrm{L}}_{t,k}$, and it is a function of the air interface packet delay $T_{t,k}$. The function $R^{\mathrm{L}}_{t,k} = f^{\mathrm{L}}(T_{t,k})$ assigns a reward of 1 when $T_{t,k}$ is less than 1 ms, and then it decreases as $T_{t,k}$ increases. For all latency-sensitive UEs at time t, the average reward about latency-sensitive packet delay is given by,

$$\overline{R^{\mathrm{L}}_t} = \frac{1}{K^{\mathrm{L}}_t} \sum_{k=1}^{K^{\mathrm{L}}_t} f^{\mathrm{L}}(T_{t,k}), \tag{19}$$

where K^{L}_t is the total number of latency-sensitive UEs at time t. The reward of the j-th URLLC link's transmit power at time t is denoted as $R^{\mathrm{P}}_{t,j}$. It is used to encourage URLLC links not to choose extremely high transmit powers, which may lead to high energy consumption and high inter-cell interference. The function $R^{\mathrm{P}}_{t,j} = f^{\mathrm{P}}(a_{t,j})$ assigns a reward of 1 when $a_{t,j}$ is less than a transmit power threshold, and then it decreases as $a_{t,j}$ increases. The average reward about all URLLC link transmit powers at time t is given by,

$$\overline{R^{\mathrm{P}}_t} = \frac{1}{J_t} \sum_{j=1}^{J_t} f^{\mathrm{P}}(a_{t,j}), \tag{20}$$

where J_t is the total number of active URLLC links at time t. The total reward at time t is a weighted sum of the above three sub-rewards, and it is calculated as,

$$\mathcal{R}_t = \lambda_{\mathrm{U}} \overline{R^{\mathrm{U}}_t} + \lambda_{\mathrm{L}} \overline{R^{\mathrm{L}}_t} + \lambda_{\mathrm{P}} \overline{R^{\mathrm{P}}_t}, \tag{21}$$

where λ_{U}, λ_{L} and λ_{P} are the weights of each sub-reward, respectively. Finally, the problem is formulated by designing the reward function $\mathcal{R}_t$, and the goal of the solution is to maximize the total reward.

V. Multi-agent cooperative DRL

This section introduces the TAPPO algorithm for power allocation. We first describe the design rationale and neural network model structure of TAPPO, which enables cooperation among multiple agents. Then, we explain how to train the model using the proximal policy optimization (PPO) algorithm, a reinforcement learning technique.

A. Cooperative Mechanism of TAPPO

Our algorithm enables two levels of cooperation among multiple agents: global and partial. Global cooperation reduces inter-cell interference by sharing a common reward among all agents. This cooperation is between all agents and is achieved by encouraging all URLLC links to choose a lower transmit power. Partial cooperation enhances the reliability of each multi-connective URLLC packet by allowing agents serving the same UE to form a small group and exchange information while making decisions. This cooperation is achieved by the sophisticated transmit power balancing within each group of agents for the same URLLC UE.

The reliability of a URLLC packet transmitted over multiple links depends on the synergy of these links, which cooperate to meet the URLLC QoS requirement. Therefore, it is not necessary for each link to achieve the QoS individually, but rather to coordinate with each other. For instance, if one link experiences a poor channel condition, increasing its transmit power may not improve its reliability, but may increase the power consumption at the RRH and the inter-cell interference. However, if the link shares its state information with other links serving the same UE, those links can adjust their transmit powers cooperatively, such as lowering the power for bad channels and raising the power for good channels. Without state information sharing, these links can only assume that the other links have average channel conditions. Hence, information sharing is crucial for enabling sophisticated cooperation among these links to ensure reliability while saving energy.

Fig. 3 shows the neural network model of the TAPPO, which uses a transformer encoder layer to share information among the URLLC links of the same UE. The transformer' s key mechanism is the multi-head attention based on self-attention. This mechanism can determine the importance of each component of the input vectors and direct the neural network to focus

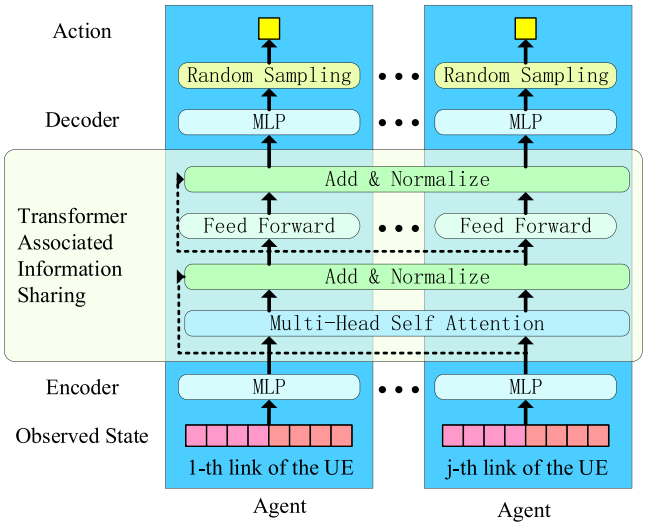


Fig. 3. Neural network model of TAPPO.

on the critical component when creating the output vectors. It is an effective method for extracting features and transferring information between modular data vectors. Therefore, the transformer is suitable for information sharing among agents because it can aggregate and exchange information between two agents within the group.

The process of power allocation with TAPPO is given in Algorithm 1. At first, for all the active URLLC links, the links that are serving the same UE are combined into a group. Then, we apply the TAPPO model to each group of URLLC links. The mathematical computation process of TAPPO is given below for a group of URLLC links as an example. We assume that there are M links serving a multi-connectivity URLLC UE in total and denote each of the URLLC link states as a data vector, S_m. They are first encoded by a multi-layer perceptron (MLP) separately and identically,

$$E_m = (S_m w_d^1 + z^1) w_d^2 + z^2, m \in \{1, ..., M\}, \tag{22}$$

where w_d^1, z^1, w_d^2, z^2 are deep neural network parameters. Then, the input of the transformer is obtained by grouping vectors E_m into a matrix, $E = [E_1; E_2; ...; E_M]$. The first part of the transformer is a self-attention layer, which initializes with 3 weight matrices w_q, w_k, w_v as its neural network parameters. In the first step of self-attention calculation, three matrices are obtained by respectively multiplying the input matrix with three weight matrices as,

$$\begin{cases} Q = E w_q, \\ K = E w_k, \\ V = E w_v, \end{cases} \tag{23}$$

where Q, K, and V represent the query, the key, and the value of the self-attention layer. Then, the outputs of the self-attention layer are obtained as,

$$\mathcal{T}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt[2]{d_{\mathrm{k}}}}\right) V, \tag{24}$$

where d_{k} is the column number of Q and K, softmax$(\cdot)$ is an activation function that turns a vector into another same-size vector with the values of vector elements summing to 1. The self-attention layer is further refined by adding the multi-headed attention mechanism, which enhances the model's ability to focus on different positions and provides the attention layer with multiple "representation subspaces". Specifically, we linearly project the query, key, and value matrices h times with different, learned linear projections. The multi-head self-attention is computed as,

$$\mathcal{M}(Q, K, V) = \text{concat}(H_1, ..., H_i) w_o, \tag{25}$$

where w_o is a neural network parameter, H_i is one of the self-attention head given as,

$$H_i = \mathcal{T}(Q w_i^Q, K w_i^K, V w_i^V), \tag{26}$$

where w_i^Q, w_i^K and w_i^V are projection neural network weight matrices of Q, K and V. After that, a fully connected feed-forward network, which consists of two linear transformations with a ReLU activation in between, is applied to each vector separately and identically. Then, the output of the transformer layer is obtained and denoted as, $D = [D_1; D_2; ...; D_M]$. The output of the transformer layer is in the same data structure as its input and each vector has contained the information of others. Each vector is decoded by another MLP separately and identically as,

$$a'_m = (D_m w_d^3 + z^3) w_d^4 + z^4, m \in \{1, ..., M\}, \tag{27}$$

where w_d^3, z^3, w_d^4, z^4 are neural network parameters, a'_m is the mean action value of m-th link. At last, the action is obtained as,

$$a_m = \mathcal{N}(a'_m, \eta), \tag{28}$$

where $\mathcal{N}(a'_m, \eta)$ represents a random sample drawn from a Gaussian distribution with mean a'_m and standard deviation η. The actions of links in this group have been made simultaneously and cooperatively.

B. Training Algorithm of TAPPO

We train the TAPPO with the PPO algorithm, which is an actor-critic mechanism, combining the advantages of both value-iteration and policy-iteration methods. It updates two networks during the training process: the actor and the critic. The actor network takes the state as input and outputs the action probability distribution. The critic network evaluates the action by computing the value function. In our training process, the critic network has a similar structure as the actor network, except that it has a one-dimensional output. Since the state of a URLLC UE is independent and uncorrelated across different time slots, we set the input of the critic network to be the same as the state input of the actor network. They both take data from the memory and update each other.

Traditionally, a set of DRL training samples is obtained by interacting between agents and the environment, which is a relatively slow process, and they are only used to train the network once. PPO improves the sample efficiency by using

surrogate objectives to regularize policy updates and enable the reuse of training data. The surrogate objective prevents the new policy from deviating far from the old policy. We use an importance sampling estimator to compensate for the gap between the training data distribution and the current policy state distribution, which is the probability ratio term defined as,

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, \tag{29}$$

where π_θ is a stochastic policy and θ represents the total neural network parameters. The clipped probability ratio, which prevents the policy from deviating too much from the old policy by limiting the ratio $r_t(\theta)$ to a certain range around 1, is given by,

$$\begin{aligned} r_t^{\text{clip}}(\theta) &= \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \\ &= \begin{cases} 1-\epsilon, & r_t(\theta) < 1-\epsilon \\ 1+\epsilon, & r_t(\theta) > 1+\epsilon, \\ r_t(\theta), & \text{otherwise} \end{cases} \end{aligned} \tag{30}$$

where ϵ is a hyper-parameter. It alters the surrogate objective by clipping the probability ratio, removing the incentive to $r_t(\theta)$ outside the interval $[1-\epsilon, 1+\epsilon]$. It ensures that the old and new policies are at least in close vicinity, and abrupt updates are not allowed. Besides, to enhance the algorithm's exploration capabilities, we generally add a policy entropy to the actor's loss and multiply it by an entropy coefficient, μ, such that the entropy of the policy is as large as feasible while optimizing the actor loss. A policy, denoted as π, has maximum entropy when all actions have equal probabilities and minimum when it favors one action over the others. This avoids early convergence of the policy to a single action and promotes exploration. The policy entropy is defined as,

$$\begin{aligned} \mathcal{H}\left(\pi\left(\cdot \mid s_t\right)\right) &= -\sum_{a_t} \pi\left(a_t \mid s_t\right) log\left(\pi\left(a_t \mid s_t\right)\right) \\ &= \mathbb{E}_{a_t \sim \pi}\left[-log\left(\pi\left(a_t \mid s_t\right)\right)\right]. \end{aligned} \tag{31}$$

Our goal is to minimize the training loss, which contains two parts, and is given as,

$$L(\theta) = -\mathbb{E}_t[\min(r_t(\theta)A_t, r_t^{\text{clip}}(\theta)A_t)] - \mu\mathcal{H}\left(\pi\left(\cdot \mid s_t\right)\right), \tag{32}$$

where A_t is an estimated advantage function at time step t and it is equal to the reward given by the environment minus the value given by the critic.

VI. Simulation Results and Analysis

In this section, we first present the system scenario settings and the simulation parameters. Next, we study different standard deviations of the Gaussian random sampling process for TAPPO, and compared TAPPO with DDPG, showing the advantage of fuzzy power allocation. Then, we present the improvement of multi-connectivity by comparing single-connectivity and double-connectivity. Moreover, we compare our proposed TAPPO algorithm with conventional PPO and analyze the cooperation benefits of information sharing among related agents.

Algorithm 1: TAPPO for Power Allocation.

Input:
TAPPO model with trained parameters
Total states, $\mathcal{S}_t = \{S_1, ..., S_j, ..., S_J\}$
Output:
Total actions, $\mathcal{A}_t = \{a_1, ..., a_j, ..., a_J\}$
Combine states of links serving the same UE into a group
for all URLLC UE link groups **do in parallel**
  for all S_m **do in parallel**
    Encode S_m to E_m by MLP
  end for
  Group vectors E_m into matrix E
  Exchange information between links by transformer
  Split matrix D into vectors D_m
  for all D_m **do in parallel**
    Decode mean action value a'_m from D_m by MLP
    Get action a_m by Gaussian random sampling
  end for
end for
Match actions with URLLC links
return Total actions $\mathcal{A}_t = \{a_1, ..., a_j, ..., a_J\}$

A. Experiment Setting

We consider a road cross scenario with multiple RRHs and multiple UEs in a 600 m $\times$ 600 square area. Two roads are horizontally and vertically located with their intersection at the center of the area and divide the whole area into four small square blocks. Four RRHs are located at the center of each small square block. The vehicles are randomly distributed on the roads according to the spatial Poisson process. We assume that each RRH has 30 sub-carriers and all the RRHs operate on the same frequency range. The pass loss model is $128.1 + 37.6\mathbf{log}_{10}(d)$, where d is the distance in km. The shadowing follows a log-normal distribution with a 5 dB standard deviation. The channel feedback latency is 3 ms. We assume that the URLLC packet length varies from 32 bytes to 64 bytes with a uniform distribution. The transmission of a URLLC packet fails if its reliability is lower than 99.999%. Other simulation parameters are listed in Table I [6].

Since the environment in our simulation is endless, we set the maximum step of each episode to 15000 steps. During the training process, instead of storing all the information, we randomly store the information of two UEs in memory. We update the actor and critic neural networks 10 times with the data memory of every 1500 steps. The hidden layer of the neural network has 64 units and the activation function is $tanh$. To ensure that the neural network features are on a similar scale and improve the training performance, we linearly normalize all the input and output features into the range from -1 to 1. We conduct our simulation on a 12th Gen Intel Core$^{\text{TM}}$ i7-12700H with PyTorch 1.10.

B. Advantage of Fuzzy Power Allocation

In this subsection, we evaluate the performance of TAPPO with different standard deviation η for the Gaussian distribution

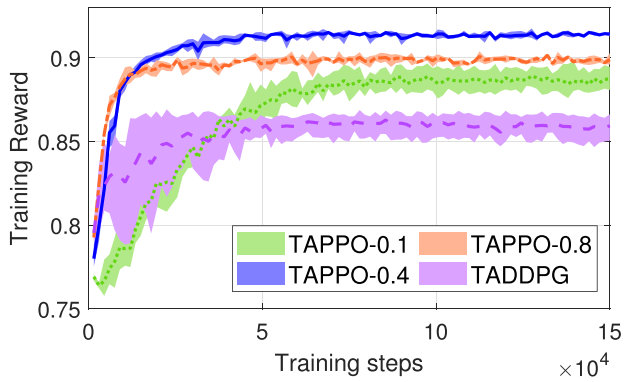
TABLE I
SIMULATION PARAMETERS

Parameter	Value
Carrier frequency	2 GHz
Bandwidth of each subcarrier W	1 MHz
Length of TTI τ	0.125 ms
Noise spectral density	-174 dBm/Hz
Transmit power of URLLC link	[0, 30] dBm
Transmit power of latency-sensitive link	15 dBm
URLLC data packet size	[32,64] bytes
Latency-sensitive data packet size	300 bytes
Expected URLLC reliability	99.999%
URLLC demand	1000 packets/s
Vehicular speed range	[0, 54] km/h
Square map size	600 m
Distance between road and RRH	150 m
RRH antenna height	25 m
Vehicular antenna height	1.5 m
Total training step	150000
Mini-batch size	128
PPO clip parameter ϵ	0.5

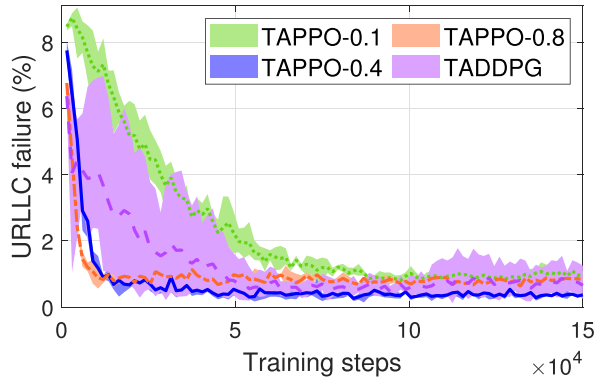
random sample. Moreover, we compare their performance with DDPG, which is a DRL algorithm that learns deterministic policies for given states to address problems in continuous action spaces, commonly applied in communication power allocation problems [41], [42]. The simulation parameters are set to 8 transmitting antennas at each RRH and 40 vehicles in total.

Fig. 4 shows the variations of reward, URLLC failure percentage, and latency-sensitive packet delay during 4 different training processes of TAPPO and DDPG. Specifically, we evaluate 3 different standard deviations η, including 0.1, 0.4, and 0.8, for the Gaussian distribution random sample of TAPPO. In the comparison of TAPPO algorithms with different parameters, we can see that the best overall performance is obtained by using 0.4 as the standard deviation. Therefore, we will choose 0.4 as the standard deviation for the following experiments. In addition, in comparison with the DDPG algorithm, TAPPO achieves better performance in terms of reward, URLLC failure percentage, and latency-sensitive packet delay.

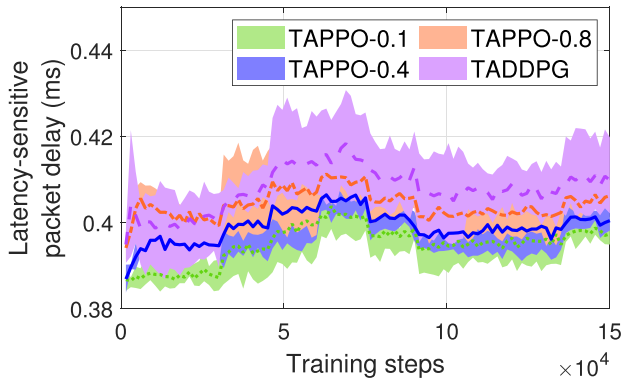
PPO with Gaussian distribution random sample offers a fuzzy power allocation approach for high mobility URLLC communications, where obtaining accurate CSI is challenging. In vehicular networks, the high mobility of vehicles makes it difficult to perform accurate channel estimations. This also affects the computation of optimal transmit powers, as it is not feasible to derive an accurate solution based on inaccurate feedback data. In other words, the inaccurate CSI limits the network controller from performing the accurate transmit power allocation in vehicle networks. Therefore, we use a fuzzy approach to cope with the uncertainty of the observation. As shown in the experimental results, the deterministic power control schemes, such as DDPG, are not well adapted to high mobility communication environments, as they are attempting to find deterministic solutions relying on inaccurate CSI estimation. Our fuzzy power control scheme can overcome the effect of inaccurate CSI and improve the robustness of the system.



(a) Training reward.



(b) URLLC failure percentage.



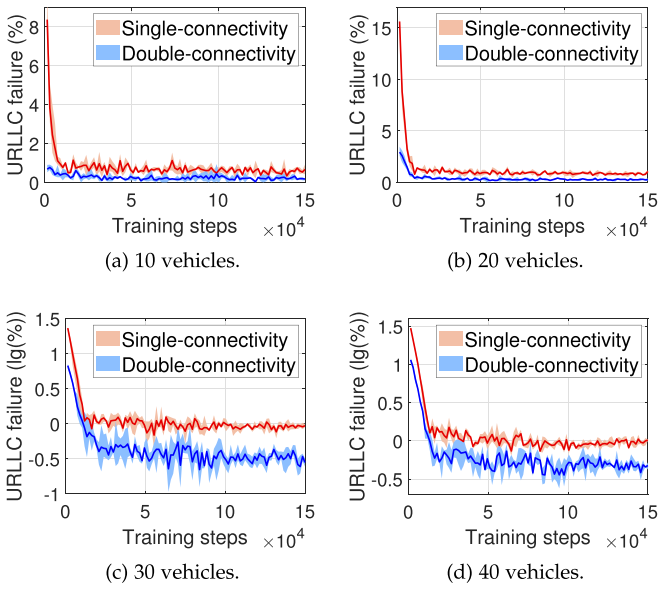
(c) Latency-sensitive packet delay.

Fig. 4. Performance variation during training process.

C. Improvement of Multi-Connectivity

In this subsection, we evaluate the performance of multi-connectivity for enhancing URLLC reliability with varying numbers of vehicles from 10 to 40. The simulation parameters are set to 4 transmitting antennas at each RRH. For a fair comparison, we apply the TAPPO algorithm to both single-connectivity and double-connectivity scenarios.

Fig. 5 illustrates the variation of URLLC failure percentage during the TAPPO training process under different numbers of vehicles and different connectivity scenarios. The y-axis for the cases of 30 and 40 vehicles is plotted in a logarithmic scale for better clarity. The figure shows that for the single-connectivity scenario, the URLLC failure probability is high at first but gradually decreases and eventually converges to a lower level. In contrast, the double-connectivity scenario has



(a) 10 vehicles. (b) 20 vehicles.

(c) 30 vehicles. (d) 40 vehicles.

Fig. 5. URLLC failure percentage variation of training.

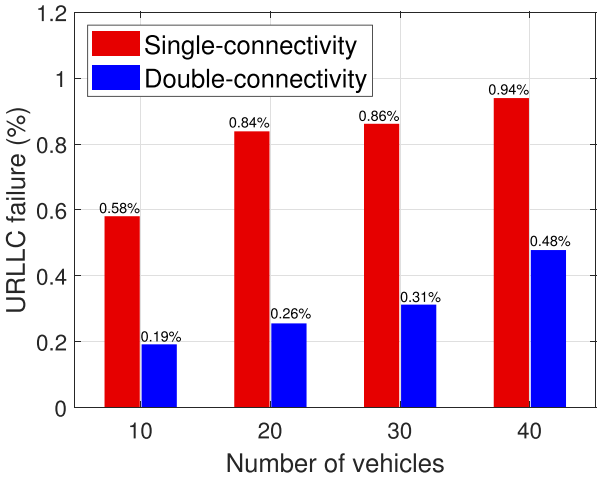


Fig. 6. URLLC failure percentage vs number of vehicles.

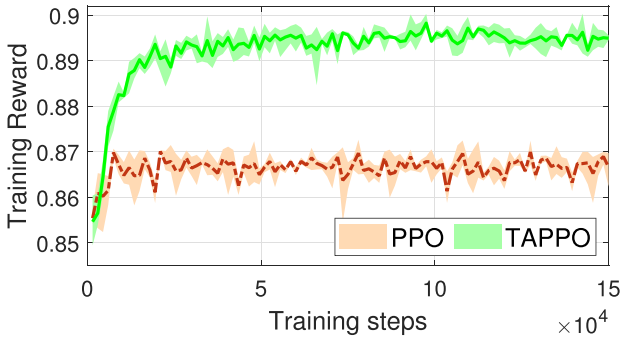


Fig. 7. Reward variation of training.

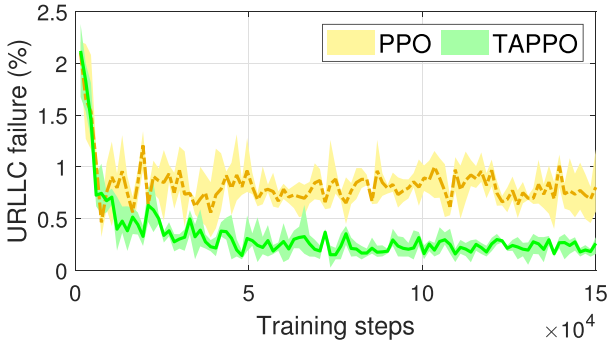


Fig. 8. URLLC failure percentage variation of training.

significant advantages at the start and the URLLC reliability also improves over the training iterations. Moreover, as the number of vehicles increases, the URLLC reliability performance of the two scenarios becomes more similar. In summary, we can observe that double-connectivity can enhance URLLC reliability as the TAPPO training process progresses, regardless of the number of vehicles. However, the performance gain provided by double-connectivity diminishes as the number of vehicles increases.

Fig. 6 shows the average URLLC failure percentage for 60000 testing iterations under different connectivity scenarios and different numbers of vehicles. Compared with the single-connectivity, the double-connectivity can reduce the URLLC packet failure occurrences by 59.20%, 53.20%, 44.74%, and 11.78% for the number of vehicles increasing from 10 to 40, respectively. It can be observed that the URLLC reliability performance gain of double-connectivity is significant under light communication loads. By transmitting duplicate URLLC packets over two independent links, the URLLC transmission is deemed successful if at least one link successfully transmits the packet, thereby lowering the URLLC failure percentage. However, the URLLC reliability performance gain declines as the number of vehicles increases. This is because the double-connectivity scenario with heavy communication loads causes more severe inter-cell interference due to the limited total spectrum resources, thus resulting in a higher failure rate in each link. Therefore, double-connectivity is especially suitable for resource-rich scenarios and can effectively enhance the URLLC transmission reliability.

D. Cooperation Performance of TAPPO

In this subsection, we evaluate the effectiveness of the proposed TAPPO algorithm in the double-connectivity scenario with varying numbers of vehicles from 10 to 40. The simulation parameters are set to 4 transmitting antennas at each RRH. We use the PPO as a baseline algorithm, in which the actor and critic networks are 4-layer fully connected deep neural networks and there is no information sharing among agents.

During the training process of the 20-vehicle case, the variation comparisons of training reward and URLLC failure percentage between typical PPO and proposed TAPPO are shown in Figs. 7 and 8, respectively. Fig. 7 demonstrates that the training reward of TAPPO can eventually converge to around 0.895, while PPO only has a stable training reward of around

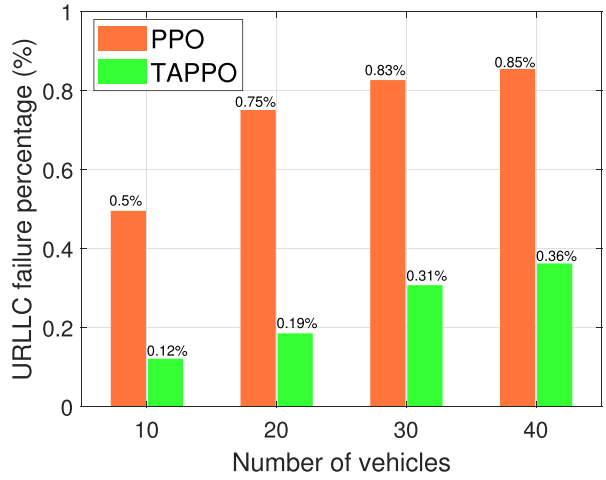


Fig. 9. URLLC reliability comparison of TAPPO and PPO.

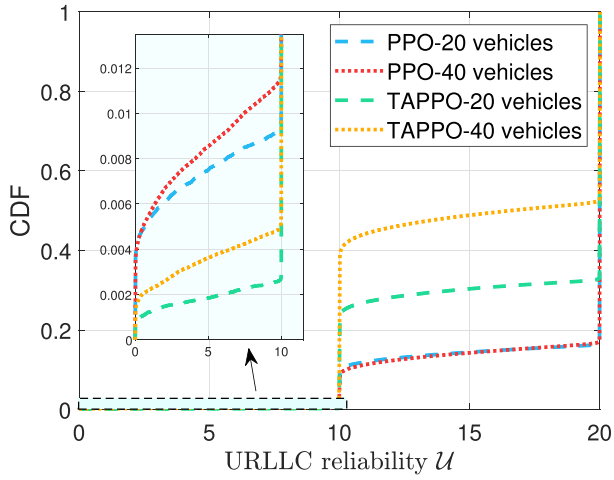


Fig. 10. URLLC reliability CDF of TAPPO and PPO.

0.866. Fig. 8 indicates that as the training process progresses, TAPPO can achieve a much lower URLLC failure percentage, that is, around 0.2%, while PPO can only reduce the URLLC failure percentage to about 0.7%. From these two figures, we can observe the superior performance of TAPPO in enhancing the training reward and URLLC reliability compared to the typical PPO algorithm.

Fig. 9 illustrates the average URLLC failure percentage for the 60000-step testing process under different learning algorithms and different numbers of vehicles. It can be observed that after training, our TAPPO algorithm can lower the URLLC failure percentage to about 0.7% in the testing process in comparison with the typical PPO. Specifically, the proposed TAPPO algorithm can further reduce the URLLC failure occurrences by 76.00%, 74.67%, 62.65%, and 57.65% compared with PPO for the number of vehicles increasing from 10 to 40, respectively. As the number of vehicles increases, both PPO and TAPPO become slightly less effective since the communication loads increase, but the communication resource is limited. In summary, the proposed TAPPO algorithm has an advantage over the typical PPO algorithm regardless of light or heavy overall communication loads.

Fig. 10 illustrates the cumulative distribution function (CDF) of URLLC packet reliability $\mathcal{U}$ of both PPO and TAPPO in the 20-vehicle and 40-vehicle scenarios. The maximum URLLC link reliability $\mathcal{U}$ for a single link is set to 10 so that the maximum URLLC packet reliability in the double-connectivity scenario is 20. To see a clearer comparison, we zoom in on the part of the curve between 0 and 10 and plot it in a sub-figure. When $\mathcal{U} = 5$, the CDF of TAPPO is lower than PPO, meaning that TAPPO is better at ensuring the successful transmission of URLLC packets. For the part of $\mathcal{U} > 10$, the CDF of TAPPO is much higher than PPO, meaning that TAPPO has fewer packets with excessively high reliability compared with PPO. Since the goal of URLLC is to achieve 99.999% reliability, we do not need excessively high reliability, which may be a waste of communication resources and bring high inter-cell interference. TAPPO can enable cooperation between different URLLC links

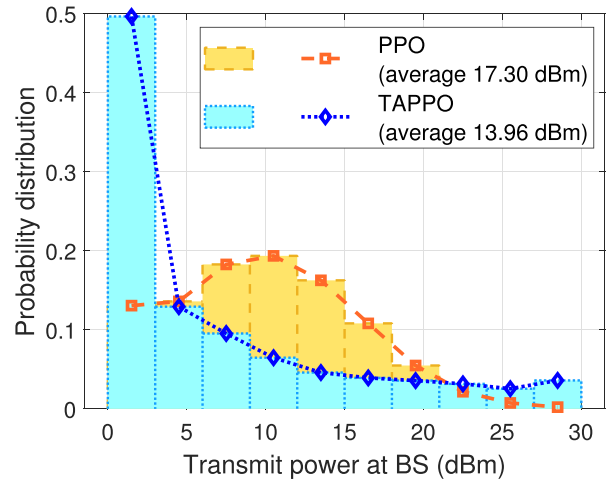


Fig. 11. Transmit power distribution for 20 vehicles.

that are serving the same UE. TAPPO allows links to know their mutual status by information sharing and cooperatively allocates the transmit power of multiple links to ensure reliability without wasting resources. TAPPO has an advantage in solving the multi-connectivity URLLC power allocation problem since it achieves sophisticated collaboration among multiple URLLC links of the same UE.

Fig. 11 illustrates the probability distribution of the chosen transmit power for the 20-vehicle case under different learning algorithms. Note that the green part is the overlap region of two bars. We can observe that PPO has a higher probability of choosing higher transmit power. In contrast, TAPPO generally chooses lower transmit power. Quantitatively, TAPPO consumes only 46.31% of the transmission energy compared with PPO. The reason is that by the information sharing between multiple links, TAPPO can obtain the channel state of other links, so the transmit power of multiple links can be jointly allocated, avoiding all the links choosing high transmit power. For example, by information sharing between the links serving the same UE, they can adjust their transmit powers cooperatively,

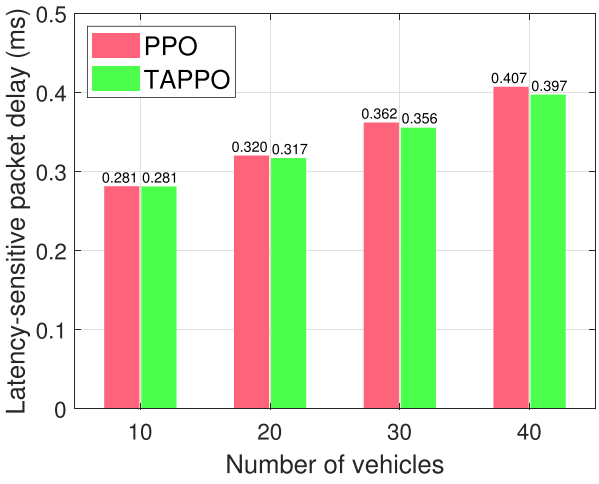


Fig. 12. Latency-sensitive packet average delay comparison.

such as lowering the power for poor channels and raising the power for good channels. The relatively lower transmit power choices in TAPPO lead to lower inter-cell interference than PPO, which is essential for the aforementioned URLLC reliability improvement. Meanwhile, the latency-sensitive packet delay reduction can also benefit from the lower transmit power.

Fig. 12 illustrates the average air interface packet delay of latency-sensitive packets of different learning algorithms with varying numbers of vehicles from 10 to 40. It shows that the aforementioned URLLC reliability improvement brought by TAPPO does not compromise latency-sensitive packet transmission performance. In fact, TAPPO can also reduce the air interface packet delay for latency-sensitive packet transmission compared with PPO by choosing lower transmit powers. In summary, the proposed TAPPO algorithm can enhance the overall system communication performance, both for URLLC and latency-sensitive packet transmission.

VII. Conclusion

In this paper, we have proposed the multi-connectivity scheme for URLLC in vehicular networks, where a URLLC packet is duplicated and transmitted simultaneously over multiple independent wireless channels and its reliability depends on the synergy of these links. A real-time power allocation approach is needed to ensure reliability and accommodate dynamic users while considering diverse QoS and inter-cell interference. To this end, we have formulated the power allocation problem as a multi-agent DRL framework and proposed a cooperative multi-agent DRL algorithm called TAPPO that uses a transformer to share information among partial agents. Extensive simulation results have verified the benefits of introducing multi-connectivity into URLLC and the effectiveness of the proposed power allocation algorithm. The multi-connectivity scheme with TAPPO as the power controller algorithm provides a feasible way to improve network availability in vehicular networks. Moreover, our method is robust to imperfect CSI and is applicable to other general URLLC scenarios. In the future, we will explore network virtualization for enhancing the performance of multi-connectivity URLLC in vehicular networks.

References

[1] S. Chen, J. Hu, Y. Shi, L. Zhao, and W. Li, "A vision of C-V2X: Technologies, field testing, and challenges with Chinese development," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3872–3881, May 2020.

[2] N. Cheng et al., "Big Data driven vehicular networks," *IEEE Netw.*, vol. 32, no. 6, pp. 160–167, Nov./Dec. 2018.

[3] H. Bagheri et al., "5G NR-V2X: Toward connected and cooperative autonomous driving," *IEEE Commun. Standards Mag.*, vol. 5, no. 1, pp. 48–54, Mar. 2021.

[4] P. Yang, L. Kong, and G. Chen, "Spectrum sharing for 5G/6G URLLC: Research frontiers and standards," *IEEE Commun. Standards Mag.*, vol. 5, no. 2, pp. 120–125, Jun. 2021.

[5] M. Li, J. Gao, L. Zhao, and X. Shen, "Deep reinforcement learning for collaborative edge computing in vehicular networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 4, pp. 1122–1135, Dec. 2020.

[6] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies," Release 17.0.0, Mar. 2022.

[7] J. Park et al., "Extreme ultra-reliable and low-latency communication," *Nature Electron.*, vol. 5, pp. 133–141, 2022.

[8] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *Proc. Int. Symp. Wireless Commun. Syst.*, 2018, pp. 1–6.

[9] J. Rao and S. Vrzic, "Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis," *IEEE Netw.*, vol. 32, no. 2, pp. 32–40, Mar./Apr. 2018.

[10] 3GPP TS 37.340, "E-UTRA and NR; Multi-connectivity; Stage-2," V15.0.0, TS 32.130, Jun./Dec. 2017.

[11] M.-T. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Multi-connectivity as an enabler for reliable low latency communications–an overview," *IEEE Commun. Surv. Tut.*, vol. 22, no. 1, pp. 156–169, First Quarter, 2020.

[12] M. Khoshnevisan, V. Joseph, P. Gupta, F. Meshkati, R. Prakash, and P. Tinnakornsrisuphap, "5G industrial networks with CoMP for URLLC and time sensitive network architecture," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 947–959, Apr. 2019.

[13] J. Xue, K. Yu, T. Zhang, H. Zhou, and X. Shen, "Deep reinforcement learning enabled power allocation for multi-connectivity C-V2X downlink," in *Proc. Int. Symp. Pers. Indoor Mobile Radio Commun.*, 2023, pp. 1–6.

[14] M. Labana and W. Hamouda, "Advances in CRAN performance optimization," *IEEE Netw.*, vol. 35, no. 3, pp. 140–146, May/Jun. 2021.

[15] C. She, Z. Chen, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Improving network availability of ultra-reliable and low-latency communications with multi-connectivity," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5482–5496, Nov. 2018.

[16] A. Rabitsch et al., "Utilizing multi-connectivity to reduce latency and enhance availability for vehicle to infrastructure communication," *IEEE Trans. Mobile Comput.*, vol. 21, no. 5, pp. 1874–1891, May 2022.

[17] W. Wu, R. Liu, Q. Yang, and T. Q. S. Quek, "Robust resource allocation for vehicular communications with imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5883–5897, Sep. 2021.

[18] T. Hosler, P. Schulz, E. A. Jorswieck, M. Simsek, and G. P. Fettweis, "Stable matching for wireless URLLC in multi-cellular, multi-user systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5228–5241, Aug. 2020.

[19] H. Xiang, Y. Yang, G. He, J. Huang, and D. He, "Multi-agent deep reinforcement learning-based power control and resource allocation for D2D communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1659–1663, Aug. 2022.

[20] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, "Deep learning for radio resource allocation with diverse quality-of-service requirements in 5G," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2309–2324, Apr. 2021.

[21] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022.

[22] H. H. Jeong, Y. C. Shen, J. P. Jeong, and T. T. Oh, "A comprehensive survey on vehicular networking for safe and efficient driving in smart transportation: A focus on systems, protocols, and applications," *Veh. Commun.*, vol. 31, 2021, Art. no. 100349.

[23] A. H. Sodhro et al., "Towards 5G-enabled self adaptive green and reliable communication in intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 5223–5231, Aug. 2021.

[24] J. Xue, T. Zhang, W. Wu, H. Zhou, and X. Shen, "Sparse Big Data for vehicular network traffic flow estimation: A. machine learning approach," in *Proc. IEEE Glob. Commun. Conf.*, 2022, pp. 4959–4963.
[25] X. Ge, "Ultra-reliable low-latency communications in autonomous vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5005–5016, May 2019.
[26] T. Zhang, J. Xue, Y. Xu, K. Yu, and H. Zhou, "Dynamic radio resource slicing for service-oriented 5G/B5G C-V2X networks," in *Proc. IEEE/CIC Int. Conf. Commun. China*, 2022, pp. 1095–1100.
[27] B. Chang, L. Zhang, L. Li, G. Zhao, and Z. Chen, "Optimizing resource allocation in URLLC for real-time wireless control systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8916–8927, Sep. 2019.
[28] J. Jia, Y. Deng, J. Chen, A.-H. Aghvami, and A. Nallanathan, "Availability analysis and optimization in CoMP and CA-enabled HetNets," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2438–2450, Jun. 2017.
[29] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.
[30] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," in *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.
[31] C. She et al., "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," in *Proc. IEEE*, vol. 109, no. 3, pp. 204–246, Mar. 2021.
[32] L. Feng, W. Li, Y. Lin, L. Zhu, S. Guo, and Z. Zhen, "Joint computation offloading and URLLC resource allocation for collaborative MEC assisted cellular-V2X networks," *IEEE Access*, vol. 8, pp. 24 914–24 926, 2020.
[33] M. Setayesh, S. Bahrami, and V. W. Wong, "Resource slicing for eMBB and URLLC services in radio access network using hierarchical deep learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8950–8966, Nov. 2022.
[34] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Netw.*, vol. 32, no. 2, pp. 24–31, Mar./Apr. 2018.
[35] G. Saikesava and N. B. Mehta, "MCS selection for multi-connectivity and eMBB-URLLC coexistence in time-varying frequency-selective fading channels," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 2157–2162.
[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[37] Q. Zhao, S. Paris, T. Veijalainen, and S. Ali, "Hierarchical multi-objective deep reinforcement learning for packet duplication in multi-connectivity for URLLC," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit*, 2021, pp. 142–147.
[38] Y. Azimi, S. Yousefi, H. Kalbkhani, and T. Kunz, "Energy-efficient deep reinforcement learning assisted resource allocation for 5G-RAN slicing," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 856–871, Jan. 2022.
[39] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[40] S. Wang, S. Bi, and Y.-J. A. Zhang, "Deep reinforcement learning with communication transformer for adaptive live streaming in wireless edge networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 308–322, Jan. 2022.
[41] A. Alwarafy, B. S. Çiftler, M. Abdallah, M. Hamdi, and N. Al-Dhahir, "Hierarchical multi-agent DRL-based framework for joint multi-RAT assignment and dynamic resource allocation in next-generation HetNets," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2481–2494, Jul./Aug. 2022.
[42] X. Li, H. Zhang, W. Li, and K. Long, "Multi-agent DRL for user association and power control in terrestrial-satellite network," in *Proc. IEEE Glob. Commun. Conf.*, 2021, pp. 1–5.
[43] T. Kim, D. J. Love, and B. Clerckx, "Does frequent low resolution feedback outperform infrequent high resolution feedback for multiple antenna beamforming systems?," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1654–1669, Apr. 2011.
[44] L. Liang, J. Kim, S. C. Jha, K. Sivanesan, and G. Y. Li, "Spectrum and power allocation for vehicular communications with delayed CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 458–461, Aug. 2017.
[45] Y. Park, J. Ha, S. Kuk, H. Kim, C.-J. M. Liang, and J. Ko, "A feasibility study and development framework design for realizing smartphone-based vehicular networking systems," *IEEE Trans. Mobile Comput.*, vol. 13, no. 11, pp. 2431–2444, Nov. 2014.
[46] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static SIMO fading channels at finite blocklength," in *Proc. IEEE Int. Symp. Inf. Theory*, 2013, pp. 1531–1535.
[47] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.
[48] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

Jianzhe Xue (Student Member, IEEE) received the BS degree in communication engineering from Xidian University, Xi'an, China, in 2021. He is currently working toward the PhD degree with the School of Electronic Science and Engineering, Nanjing University, China. His current research interests include internet of vehicles, orthogonal time frequency space modulation, and machine learning for wireless communications.

Kai Yu (Student Member, IEEE) received the BS degree in detection, guidance, and control technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2019. He is currently working toward the PhD degree with the School of Electronic Science and Engineering, Nanjing University, China. His research interests include resource allocation, machine learning for wireless communications, and heterogeneous networks.

Tianqi Zhang (Student Member, IEEE) received the BS degree in electronic information science and technology from Nanjing University, Nanjing, China, in 2021. He is currently working toward the PhD degree with the School of Electronic Science and Engineering, Nanjing University, China. His current research interests include Internet of Vehicles and machine learning for wireless communications.

Haibo Zhou (Senior Member, IEEE) received the PhD degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. From 2014 to 2017, he was a postdoctoral fellow with the Broadband Communications Research Group, Department of Electrical and Computer Engineering, University of Waterloo. He is currently an associate professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. He was a recipient of the 2019 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He served as an Invited Track Co-Chair for ICCC'2019, VTC-Fall'2020 and a TPC member of many IEEE conferences, including GLOBECOM, ICC, and VTC. He served as an associate editor for the IEEE Comsoc Technically Co-Sponsored the *Journal of Communications and Information Networks* (JCIN) from 2017 to 2019, and a guest editor for the *IEEE Communications Magazine* in 2016, the *Hindawi International Journal of Distributed Sensor Networks* in 2017, and IET Communications in 2017. He is currently an associate editor of the *IEEE Internet of Things Journal*, the *IEEE Network Magazine*, and the *IEEE Wireless Communications Letter*. His research interests include resource management and protocol design in vehicular ad hoc networks, cognitive networks, and space-air-ground integrated networks.

Lian Zhao (Fellow, IEEE) received the PhD degree from the Department of Electrical and Computer Engineering (ELCE), University of Waterloo, Canada, in 2002. She joined the Department of Electrical and Computer Engineering, Toronto Metroplitan University (formerly Ryerson University), Canada, in 2003. Her research interests include the areas of wireless communications, resource management, mobile edge computing, caching and communications, and IoV networks. She has been an *IEEE Communication Society* (ComSoc) and *IEEE Vehicular Technology* (VTS) Distinguished Lecturer (DL); received the Best Land Transportation Paper Award from IEEE Vehicular Technology Society in 2016, Top 15 Editor Award in 2016 for *IEEE Transaction on Vehicular Technology*, Best Paper Award from the 2013 International Conference on Wireless Communications and Signal Processing (WCSP), and the Canada Foundation for Innovation (CFI) New Opportunity Research Award in 2005. She has been serving as an editor for *IEEE Transactions on Wireless Communications*, *IEEE Internet of Things Journal*, and *IEEE Transactions on Vehicular Technology* (2013-2021). She served as a co-chair of Wireless Communication Symposium for IEEE Globecom 2020 and IEEE ICC 2018; Finance co-chair for 2021 ICASSP; Local Arrangement co-Chair for IEEE VTC Fall 2017 and IEEE Infocom 2014; co-Chair of Communication Theory Symposium for IEEE Globecom 2013. She has been a Board of Governor (BoG) committee member since 2023. She has severed as a panel expert in various federal, provincial, and international evaluation committees. She is a licensed Professional Engineer in the Province of Ontario.

Xuemin (Sherman) Shen (Fellow, IEEE) received the PhD degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is a University professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular ad hoc and sensor networks. He is a registered professional engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering fellow, a Royal Society of Canada fellow, a Chinese Academy of Engineering Foreign Member, and a distinguished lecturer with the IEEE Vehicular Technology Society and Communications Society. He received the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R.A. Fessenden Award in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society, and Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He has also received the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He served as the Technical Program Committee Chair/CoChair for IEEE Globecom 16, IEEE Infocom14, IEEE VTC10 Fall, IEEE Globecom07, and the Chair for the IEEE Communications Society Technical Committee onWireless Communications. He is the president with the IEEE Communications Society. He was the vice president for Technical and Educational Activities, vice president for Publications, Member-at-Large on the Board of Governors, Chair of the distinguished lecturer Selection Committee, member of IEEE fellow Selection Committee of the ComSoc. He served as the editor-in- chief of the *IEEE IoT Journal*, *IEEE Network*, and *IET Communications*.