Artificial Intelligence based Edge Caching in Vehicular Mobile Networks: Architecture, Opportunities, and Research Issues

Kai-Min Liao, Guan-Yi Chen, Yu-Jia Chen
Department of Communication Engineering, National Central University, Taoyuan, Taiwan

***Abstract*—This paper investigates the potentials of utilizing artificial intelligence (AI) based edge caching in the next generation of vehicular mobile networks. In recent years, vehicle-to-everything (V2X) has been a research focus, which enables the exchange of information between the vehicles and the outside world. To integrate vehicular networks and cellular radio technology, cellular-V2X (C-V2X) was proposed in 3GPP release 14. Further, mobile edge caching is regarded as an effective technique to allow local data access, which can support the low latency requirement of the V2X use cases. With the advance of AI technologies such as deep learning, there has been increasing demand in inference and learning from big vehicular data. In this paper, we present the detailed architecture of AI-based edge caching in vehicular networks with misbehaving vehicle detection as an illustrative case. Performance results are provided to investigate the benefit of the proposed architecture. Finally, we highlight the potential research directions.**

I. Introduction

The rapid growth of users' demand increases the attention of vehicular communication. With the ability to exchange information with nearby vehicles or infrastructure, vehicle-to-everything (V2X) has become an expected technique in the future. In 2016, the cellular-V2X (C-V2X) had been proposed in 3GPP release 14 to enable the communication services in close proximity [1]. Due to a higher frequency band is utilized on C-V2X, the transmission link will have better performance, which is suitable to be applied to the fifth generation (5G) system architecture.

Reducing the latency from core network to the end users have been a challenge so far. In 5G networks, ultra-reliable low-latency communication (URLLC) is one of the important categories. The successful data transmission need to be guaranteed within 1 ms at a low failure rate [2]. Besides that, we also want to decrease the latency from core network to the base stations (BS). To provide a better quality of experience (QoE) to the users, edge caching is considered to be utilized. Edge caching refers to the use of edge servers to store content closer to end users. With the help of moving the network functions and resources closer to the end users, it shows a lot of benefits by using edge caching.

- Since the transmission time from the edge node to the content provider server can be avoided, we only need to consider the communication time between the end user and the edge caching node. The on-demand response scheme becomes achievable.
- The backhaul congestion issue can be released. By predicting users' preference, the edge nodes can cache the popular content data in the off-peak time. Therefore, more backhaul capacity can be utilized to enhance system performance.

However, the memory size of the edge caching node is limited, finding an effective method to determine which kind of the data contents need to be cache on the edge nodes is necessary.

As the development of artificial intelligence (AI) becomes mature, the trade-off between memory size and the data content diversity seems possibility to be solved. One of the AI's branches is called machine learning, which can be separated into artificial neural network (ANN) and reinforcement learning (RL). ANN can be used in many fields because they are able to reproduce and model nonlinear processes.

In this paper, we are interested in using ANN to design edge caching for the V2V network. We propose an AI-based edge caching for vehicular mobile networks and investigate the issues of this system. After that, in our case study, we demonstrate the performance of our proposed architecture compared with the traditional pure BS system.

The paper is organized as follows. We first study the techniques using in our system. Then we introduce our proposed system architecture. Section IV validate the performance of our proposed system architecture compared with pure BS system. Finally, brief conclusions and future direction are given in Section V.

II. Background and Related Work

To cover the basic concept of our proposed system architecture, we provide an overview of V2X and edge caching.

A. V2X in The Development of 5G

5G V2X is envisioned for 3GPP Release 16, which supports the network automation and novel radio techniques. The term cellular-V2X (C-V2X) appeared the first time in Release 14 and is applied in the scenario of the automotive vehicular network. 3GPP's C-V2X specifications include short-range V2V communications and wide-area vehicle-to-network (V2N) communication [3]. Short-range V2V communications can be used in case of nearby vehicles have desired data content. Otherwise, wide-area V2N communication enable the vehicle to communicate with the BS. The C-V2X standard development is discussed in Release 14, Release 15, and Release 16. For the third phase Release 16, 5G new radio (NR) and 5G-V2X can accomplish highly automatic and develop cooperative driving by the following features.

- Data sharing for a group of vehicles.
- Control information from other driving vehicles in short distance.
- Vehicle trajectory planning to avoid accidents.

However, C-V2X has some issues that need to be conquered. First, it is important to allocate the resources efficiently, as the vehicles all need to connect to the BS. Second, the near-far effect will be serious in the dense vehicular network. It is essential to mitigate that to avoid power wasting.

Vehicle-to-vehicle (V2V) develops the earliest and has been the most mature technique among V2X. V2V is primarily used to prevent car accidents, by transmitting the vehicle position and speed information to other vehicles via the dedicated network. Especially in case of driving on the high speedway, since vehicles tend to travel at high speeds, it is necessary to transmit the data in a short period. It needs to be studied and discussed to find a way to reduce the V2V latency and increase the total throughput. Therefore, how to receive the information in a tolerable time and ensure the communication reliability to achieve URLLC is our next goal.

B. Intelligent Edge Caching

Mobile edge caching can be generally classified into the following two types.

- Reactive caching: Since the memory in the caching node is limited, we need to upload the latest caching content regularly and discard the oldest to make sure caching efficiency.
- Proactive caching: Users can only download a few data in BS, because the moving users may not have enough time to communicate with the BS. Proactive cache can access the data contents into the edge cache nodes actively. Therefore, the users with high speed can obtain complete data contents due to the help of edge cache node [4].

Utilizing the edge caching scheme in the content-aware 5G network can yield significant latency improvement. The authors in [5] analyze the theoretic throughput performance limits of cache-aided wireless networks. This paper develops a relationship between caching-transmission policy and cache size. In [6], the authors propose a proactive content caching architecture that enables big data content. The strategic content is cached at the BSs to enhance users' satisfaction and release backhaul offloading. Applying edge caching in the BSs can enhance many applications, especially for those are time-critical. For instance, new services like augmented reality (AR) with tighter latency constraint only work perfectly in case of using edge caching in the BSs. It is because these services will constantly transmit users a large amount of data content in real time. Apart from AR, automated service is another urgent need for low latency. We know that it is indispensable to develop an automated industrial production or automatic vehicles in the future. The authors in [4], [7]–[9] concern about the edge caching issues. In [7], seamless mobility can be realized by using the node mobility information to proactively cache data content. The authors of [8] study the cache size allocation problem in backhaul limit wireless networks. It is reported that the file popularity distribution of the data content dominates the relationship between required cache size and the total number of files.

Integrating AI into edge caching has become a widespread trend in recent years. In [9], the authors propose an integrated network that jointly considers the cloud networking, caching and computing for connected vehicles. Due to the high complexity of considering these three technologies, the authors utilize a novel deep reinforcement learning with two outstanding features: trial-and-error search and delayed reward. Trial-and-error search means the agent has to make a trade-off between exploration and exploitation. The delayed reward is letting the agent not just considering immediately reward, but also the cumulative rewards.

Besides reinforcement learning, machine learning also has an important branch called ANN. ANN is commonly used to separate into two branches, namely recurrent neural network (RNN) and convolutional neural network (CNN). CNN can maintain the continuity of data since operating on each small area, so it is mostly used for applications like image recognition. RNN is mainly used to deal with the problems of sequence relation. Therefore, it can be used for translation.

III. AI-based Edge Caching Vehicular Mobile Networks

A. System Architecture

In this section, we consider an AI-based edge caching vehicular mobile networks for user behavior anomaly detection. While deploying BSs near the high-speed road, the detection system is used to detect anomaly behavior in a high-speed way. AI system detection and edge caching technique are jointly considered in this architecture. By

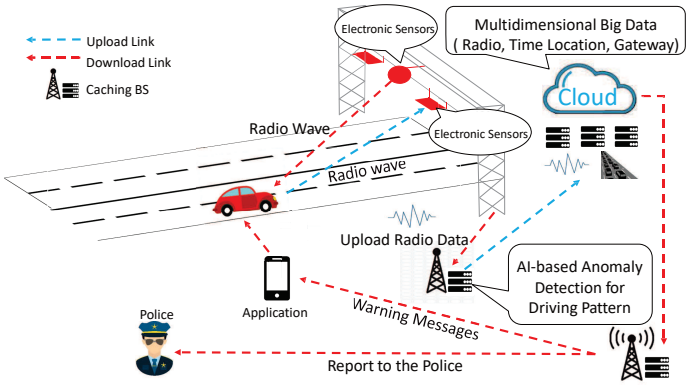

Fig. 1. The architecture of freeways vehicular networks.

using AI system detection, we can analyze the waveform without given a waveform detection algorithm. The inbuilt AI system decentralized operating on each of the edge caching BS. With the help of edge caching, the data access delay will be minimized and then the anomaly information can be reported to the police in short time. The architecture of freeways vehicular networks is as shown in Fig. 1.

The architecture of the system can be separated into two subparts. In part one, we analyze the driving situation by edge caching BS. In the edge caching BS, we store all the waveforms detected from drives by the local electronic sensors. The waveforms are used to construct the personal AI models which combine all the local features of the driver. When the vehicle transmits the electronic sensor, the electronic sensor receives a received signal strength indicator (RSSI) from the vehicle. Then the electronic sensor transmits the values of the received signal to the nearest BS. The AI server equipped on the BS compares the user waveforms detected by the electronic sensor with the historical waveforms. If the detected waveforms are not similar to historical waveforms, the BS will send the detected waveforms and historical waveforms to the BS. In other cases, if the driving waveforms match the dangerous driving pattern, the BS will transmit the detected waveforms to the cloud data server immediately. After that, the cloud data server will perform a stricter AI algorithm to determine whether the driver is really in dangerous driving.

For the second part, the edge caching BS will send different messages to the cloud server according to different analyzed results. The cloud data server is only responsible for transmitting and receiving data, and all computational parts are performed by the AI server equipped on the BS. The cloud data server acts like a large router, whose computational requirements have been greatly reduced, further providing a satisfactory system architecture delay. Fig. 2 shows the block diagram of the proposed architecture.

- Abnormal behavior: When it is recognized that the vehicle corresponds to the abnormal driving behavior, the following reaction will be executed. If it is a personal factor such as fatigue driving, the reminder information is transmitted to the driving mobile application. Other than this, if the public safety is compromised, the relevant information of the vehicle is transmitted to the road police in front to make the early action.
- Normal behavior: If the driver is normal behavior, this vehicle information will be transmitted to the cloud database for storage. After that, more information can be obtained for judging the vehicle in the future. Thereby, achieve more accurate identification of the vehicle.

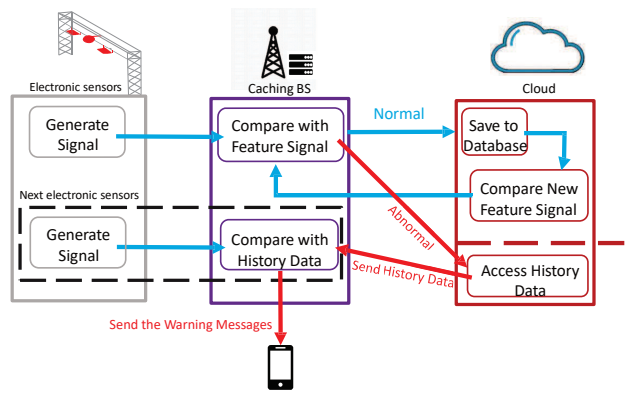

Fig. 2. The architecture of vehicle warning system.

B. Application of Detecting Abnormal Events

The waveforms received by the road side unit (RSU) can be used to detect abnormal events. The drivers may have different driving habits in different road sections. For example, on a wide straight road, the drivers tend to drive faster, vice versa. The RSU can receive the personal characteristic waveforms which imply the driving habits. Therefore, we can analyze the waveforms then get the general waveforms for the certain road section. We can compare the personal driving habits waveform or the general waveform of the road section with the latest received waveform. While the difference between these two waveforms is too large, it means that current driving behavior does not match the past habits, which indicates the current driver is in an abnormal event. What kind of criteria should be considered will discuss as follows.

- **Personal waveform compared method:** Based on these criteria for judgment, we can discover whether the driver is corresponding to the previous driving habits. Furthermore, situations like different driver or fall asleep at the wheel can also be noticed.
- **General waveform compared method:** We can compare the latest received waveform with the general waveform to find the difference habit like speeding or zigzag behavior.

- **Hybrid waveform compared method:** Combine the above two categories, we can compare the latest received waveform with the personal waveforms and general waveforms at the same time. The advantage of the hybrid method is saving calculation time that can reduce the end to end latency. However, the hybrid solution may lead to inaccurate detection.

All kinds of data, including drivers' vehicle data, driving data or road data, will be stored into the historical database. Given the data content stored in the database, the AI system will be updated regularly. Therefore, the reliability of the AI model increases as the number of database increases. However, many external factors cause the detected waveforms different from historical data content. For instance, rainy or foggy weather slows down the driving speed, the heavy traffic flow may force the driver to violate the habits, even the unusual problem of the receiver makes the machine to do wrong judgment. Therefore, a more accurate algorithm should be considered to fit this kind of problems. The reinforcement learning (RL) can be used to improve the accuracy so that the machines can still maintain a low error rate while in case of above external factors.

IV. Case Study

In this section, we provide illustrative results to demonstrate the V2V architecture that we proposed has a better performance compared to without caching architecture. The system bandwidth is 10 MHz at 5.9 GHz carrier frequency, which is in line with most of the current specifications [10].

The first thing that we are interested in is the relationship between total throughput and the user densities. Fig. 3 shows the total throughput of the scheme without caching increases as user density increases until the throughput achieves backhaul capacity limit. For the scheme with AI caching, except accessing the data content from backhaul, the users can also access the data from V2V caching. Since achieving the backhaul capacity limit, the total throughput of the scheme without caching is seemingly static when $\lambda_{user} > 100$. However, as $\lambda_{user} > 100$ and $\lambda_{user} < 150$, the total throughput remains to increase with a lower ratio. This is reasonable because as the backhaul is congestion, the vehicles can obtain the data from nearby vehicles through V2V communication. The scheme with AI caching outperforms that without caching by 45% while the backhaul is not congested. Besides that, Fig. 3 also implies that the system combined V2V and cellular BS should be used in the density areas. It is more obvious to observe the advantage of V2V communication while the backhaul is congesting.

In addition, the system performance against different vehicle velocity also interesting us. We set the speed of the vehicle with cache data is 50 km/hr, while changing the user velocity. In Fig. 4, we can find that the velocity of user seems not related to the total throughput. Besides that, no matter how fast the user is, the throughput of the AI caching scheme always better than the scheme without caching.

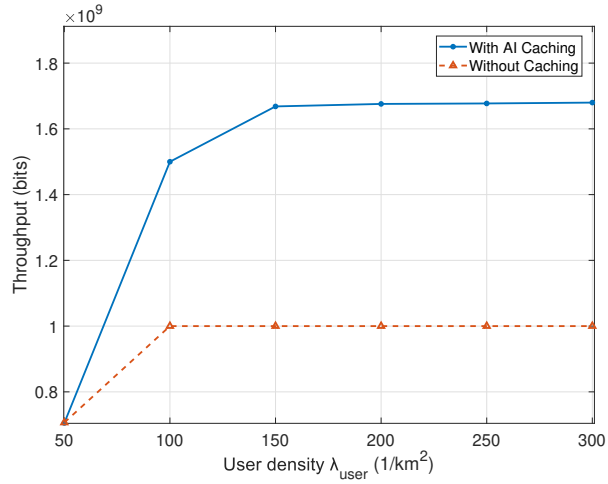

Fig. 3. The throughput performance against user density.

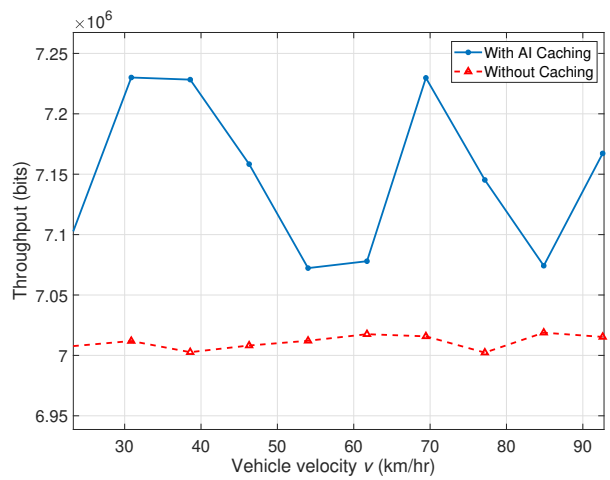

Fig. 4. The throughput performance against the vehicle velocity.

V. Challenges and Future Research Directions

In the case study, we find that the AI caching scheme can provide higher system performance. However, there still have some issues need to be discussed.

A. Spectrum Management

As we know, the spectrum efficiency is one of the most significant issues in not only UAVs but also all of the BSs. Different kind of communication protocol coexists on the operating spectrum, i.e., Bluetooth, WiFi, LTE and cellular networks [11]. These lead to a competitive behavior between different protocols and cause the problem of bandwidth scarcity. Therefore, it is necessary to improve the spectrum efficiency because the frequency bands are already congested. Besides that, interference

between UAVs will also constraint bandwidth allocation in cellular networks. In order to enhance the signal SNR, the nearby UAVs will not be allocated to the same bandwidth. Moreover, the UAV-ground BS link and UAV-user link also need to be considered, which means that more efficient technologies should be investigated compared with those without a hierarchical cache structure.

B. Energy Management

The most different thing between UAVs and ground BSs is that UAV is charged by the battery, which means that the travel time of UAV is constrained. Therefore, the energy efficiency issue determines the importance of UAV in the architecture of the mobile network. If the UAVs have better energy efficiency, it can spend more time to satisfy the users' experience rather than be forced by the battery limitation to back to the charge station. Other than this, the trajectory design can be considered more variables since the UAV has sufficient power to complete the flight mission. Another issue related to energy management is the transmission power. The impact of interference is positively related to transmission power. Thusly, how to design an appropriate power to keep the balance between having enough channel capacity and avoiding wasting power is a difficult problem need to be considered.

C. Handover Management

When vehicles move among adjacent UAV-BSs or ground-BSs, the handover mechanism will be triggered by a single vehicle reports the measurement information. But this method considers insufficiently of movement state of vehicles and the location relationship between vehicles and the BSs, which may lead to handover misjudgments and even disconnect to the network [12]. Also, the frequent handover results by the high mobility of the UAVs and vehicles is a challenge to guarantee seamless transmissions. Frequent handover makes the cache data content hard to place since the users may go to the next BS's coverage but haven't received the whole data content. Moreover, the architecture that we used contains three schemes, and each of the schemes has different coverage areas. The handover frequency is higher than traditional schemes. Therefore, how to present an efficient prediction algorithm for cache data content is also necessary due to the uncertainties of location.

D. Security Management

Due to the wireless broadcast medium, the UAV is vulnerable to common security attack [13].. Every eavesdropper in the UAV's signal coverage can receive the transmitted signal. If we want to provide a trusted network architecture in the future, this must be a critical issue to be solved. Compared with the traditional encryption technologies applied on the upper layer, the physical layer's security can ensure the wireless data transmission, which doesn't need the key and complicated algorithm. Thus, using the physical layer security is more suitable in large scale distributed wireless network architecture. [14] propose a scheme that can avoid eavesdropping from UAV flying. But the scheme doesn't consider the hierarchical cache architecture, and the UAVs are not used as BS to transmit the data content. Based on this scheme, it needs to be investigated by modifying the scheme into a hierarchical cache with secure communication architecture.

E. Trajectory Design

In the case of using UAV to ship packages, trajectory design only needs to consider the shortest path between the initial point and the destination point. But if we take UAV to be a BS into account, the trajectory design should consider the position of users. To minimize most users' waiting time, we need to design a trajectory algorithm based on the issues that we mentioned above. Also, due to the UAV has the limitation of memories, what kind of data contents should be stored in what time need to be further studied. Except for the offline and centralized solution, the trajectory design problem also can be solved by distributed online algorithm [15]. The advantage of such a method by computing on the UAV is avoiding transmitting data contents via the backhaul. UAV can determine the variables like direction or speed by the computation resources on its own. However, the online trajectory design seems to be a challenge on a hierarchical cache scheme, it may be a better way to solve the algorithm by AI.

F. Advanced AI-based Caching Design

A comprehensive AI-based data caching need to be considered on the UAV-BS. It is possible to achieve 100% user satisfaction by proactive caching at the edge node while offloading 98% of the backhaul traffic [16]. Jointly consider content caching and offloading can release the problem of mobile users' large demands of data and ease the problem of limited capacities in data storage and processing. The authors in [17] shows that the issues including dynamic network access, data rate control, connectivity preservation, and so on are all important to the next-generation network. Therefore, presenting an AI-based caching scheme jointly consider the above issues is essential.

VI. Conclusion

In this paper, we combine AI and caching techniques. Anomaly detection techniques developed for signals received from electronic sensors can immediately determine the behavior of the vehicle through electronic sensors. In order to meet strict time conditions, the AI which combined with caching can immediately filter the signal, prioritize cloud signals that may have bad behavior. The cloud can transfer historical personal data to the memory for more accurate next time. Identifying and avoiding all signals queued in the cloud cannot be identified in time. When the exact recognition result is released, the BS

where the cache is located can give the driver or related unit instant notification without having to pass through the cloud again. The technology developed by the project uses deep learning multi-dimensional time series data anomaly detection, combined with 5G multi-access edge calculation and data buffer low-latency communication technology to achieve instant driving under strict time constraints. In the case study, we used the simulation environment to discuss the throughput between pure BS and V2V. It also proved that the V2V environment can be better whether the user density or the vehicle is at high speed and the throughput. With the help of caching UAV in the future, performance like throughput or latency can be definitely improved.

References

[1] A. Nabil, K. Kaur, C. Dietrich, and V. Marojevic, "Performance analysis of sensing-based semi-persistent scheduling in C-V2X networks," in *IEEE 88th Vehicular Technology Conference (VTC-Fall)*, Aug 2018, pp. 1–5.

[2] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, March 2018.

[3] V. Vukadinovic, K. Bakowski, P. Marsch, I. D. Garcia, H. Xu, M. Sybis, P. Sroka, K. Wesolowski, D. Lister, and I. Thibault, "3GPP C-V2X and IEEE 802.11 p for vehicle-to-vehicle communications in highway platooning scenarios," *Ad Hoc Networks*, vol. 74, pp. 17–29, 2018.

[4] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 77–83, August 2016.

[5] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 320–325.

[6] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: Moving from cloud to edge," *IEEE Communications Magazine*, pp. 36–42, Sep. 2016.

[7] C. Fang, H. Yao, Z. Wang, W. Wu, X. Jin, and F. R. Yu, "A survey of mobile information-centric networking: Research issues and challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2353–2371, thirdquarter 2018.

[8] X. Peng, J. Zhang, S. H. Song, and K. B. Letaief, in *IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[9] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 44–55, Jan 2018.

[10] H. Peng, Le Liang, X. Shen, and G. Y. Li, "Vehicular communications: A network layer perspective," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1064–1078, Feb 2019.

[11] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet of Things Journal*, 2019.

[12] B. Hu, H. Yang, L. Wang, and S. Chen, "A trajectory prediction based intelligent handover control method in UAV cellular networks," *China Communications*, pp. 1–14, Jan 2019.

[13] M. Hooper, Y. Tian, R. Zhou, B. Cao, A. P. Lauf, L. Watkins, W. H. Robinson, and W. Alexis, "Securing commercial wifi-based UAVs from common security attacks," in *IEEE Military Communications Conference*, Nov 2016, pp. 1213–1218.

[14] C. Liu, T. Q. S. Quek, and J. Lee, "Secure UAV communication in the presence of active eavesdropper (invited paper)," in *9th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct 2017, pp. 1–6.

[15] N. Lu, Y. Zhou, C. Shi, N. Cheng, L. Cai, and B. Li, "Planning while flying: A measurement-aided dynamic planning of drone small cells," *IEEE Internet of Things Journal*, pp. 1–1, 2019.

[16] C. Zhong, M. C. Gursoy, and S. Velipasalar, "A deep reinforcement learning-based framework for content caching," in *52nd Annual Conference on Information Sciences and Systems (CISS)*, March 2018, pp. 1–6.

[17] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *arXiv preprint arXiv:1810.07862*, 2018.