

Adversarial Attacks and Defenses for Semantic Communication in Vehicular Metaverses

Jiawen Kang, Jiayi He, Hongyang Du, Zehui Xiong*, Zhaohui Yang, Xumin Huang, Shengli Xie

Abstract—For vehicular metaverses, one of the ultimate user-centric goals is to optimize the immersive experience and Quality of Service (QoS) for users on board. Semantic Communication (SemCom) has been introduced as a revolutionary paradigm that significantly eases communication resource pressure for vehicular metaverse applications to achieve this goal. SemCom enables high-quality and ultra-efficient vehicular communication, even with explosively increasing data traffic among vehicles. In this article, we propose a hierarchical SemCom-enabled vehicular metaverses framework consisting of the global metaverse, local metaverses, SemCom module, and resource pool. The global and local metaverse are brand-new concepts from the metaverse’s distribution standpoint. Considering the QoS of users, this article explores the potential security vulnerabilities of the proposed framework. To that purpose, this study highlights a specific security risk to the framework’s SemCom module and offers a viable defense solution, so encouraging community researchers to focus more on vehicular metaverse security. Finally, we provide an overview of the open issues of secure SemCom in the vehicular metaverses, notably pointing out potential future research directions.

Index Terms—Vehicular metaverse, semantic communication, adversarial attacks, security defense.

I. INTRODUCTION

With the advances of the Internet of Things and Artificial Intelligence, metaverse technology, regarded as the next-generation Internet, is rapidly emerging to build a virtual-physical integrated world with fully immersive and personalized experiences for users in many scenarios and applications. In particular, vehicular networks are revolutionizing towards vehicular metaverses that have vast potential to provide diverse, immersive, and personalized in-vehicle entertainment/services for both drivers and passengers [1]. In vehicular metaverses, vehicle data are mainly divided into two categories: i) static data including background information on

roadside infrastructures and road networks, and ii) dynamic data involving information about pedestrians, moving vehicles, traffic flow, and so on. To support time-sensitive vehicular metaverse services (e.g., AR navigation), a large number of dynamic data is required for real-time updates of the vehicular physical-virtual world to avoid service quality degradation, while conventional communication based on Shannon Information Theory cannot support efficient communications and massive connectivity for vehicular metaverses due to limited wireless communication resources.

Recently, Semantic Communication (SemCom) has been introduced as a revolutionary paradigm that breaks through the bandwidth bottleneck of classical communications, significantly improving communication efficiency by leveraging deep learning models to only transmit critical information sufficient for the receivers, thus dramatically reducing the number of bits transmitted. A deep learning-enabled end-to-end SemCom system with Deep Neural Networks (DNNs) as semantic codecs has been proposed in [2]. It is a promising solution to integrate SemCom into vehicular metaverses for enabling real-time transmission of vast data used to sustain the metaverse services. In vehicular metaverses, vehicles only send data requests and then download lots of static and dynamic data from edge servers in the way of SemCom with unprecedented communication efficiency in comparison to traditional communication manners.

Although semantic communication plays an indispensable role in vehicular metaverses, SemCom-enabled vehicular metaverses are still in their infancy. There are many challenges to be resolved for realizing its potential, especially privacy and security issues. On the one hand, an eavesdropper can infer vehicle location and driving habits by analyzing communication data. Even though privately trained codecs provide SemCom with a natural barrier to being eavesdropped, this barrier does not work well in vehicular metaverses. This is because there exist similar and stationary communication tasks in neighbor vehicles and then the eavesdropper may obtain similar or even identical decoders to recover data from the wireless channels. The researches on [3]–[5] are committed to private communications. On the other hand, DNNs are vulnerable to attacks (e.g., poisoning attacks and adversarial attacks). Such weaknesses naturally exist in deep learning-based SemCom systems. Attackers can easily apply typical deep learning attacks to the SemCom systems and directly reduce the task/model accuracy [6]. Non-malicious perturbations may cause performance degradation of semantic communication systems as well [7].

In particular, the existing work ignores adversarial attacks

Jiawen Kang and Jiayi He are with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China, and also with the Guangdong-Hong Kong-Macao Joint Laboratory for Smart Discrete Manufacturing, Guangzhou 510006, China. Xumin Huang is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China, and also with the 111 Center for Intelligent Batch Manufacturing Based on IoT Technology, Guangzhou 510006, China. Shengli Xie is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China, and also with the Guangdong Key Laboratory of IoT Information Technology, Guangzhou 510006, China. Hongyang Du is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Z. Yang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310007, China, and Zhejiang Provincial Key Lab of Information Processing, Communication and Networking (IPCAN), Hangzhou 310007, China, and also with Zhejiang Laboratory, Hangzhou 31121, China. Zehui Xiong is with the Pillar of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372, Singapore. (*Corresponding author: Zehui Xiong)

of SemCom in vehicular metaverses. For adversarial attacks, the authors in [8] find that non-random perturbation on the test sample can arbitrarily manipulate the output of neural networks. In general, the attacker first obtains the structure and parameters of the target model. The adversarial samples are then initialized with the original data. After several training iterations, the elaborate adversarial samples similar to the original data are generated by maximizing the loss of the semantic encoder. Finally, the accuracy degradation of the target model is achieved in the face of adversarial samples. Hu et al. in [6] demonstrate that adversarial attacks are still feasible in the deep learning-enabled SemCom system. For vehicular metaverses, the adversarial attacks increase the probability of receiving wrong information, even causing traffic accidents if receivers are misled to take dangerous actions according to wrong information [6], [9].

To fully understand the risks of adversarial attacks for SemCom-enabled vehicular metaverses, in this article, we study the adversarial attacks and the corresponding defenses against these attacks. More specifically, we first design a hierarchical vehicular metaverse framework with a global metaverse and multiple local metaverses [10]. A new adversarial attack model and its defense scheme for this framework are then proposed, respectively. We discuss several security challenges of SemCom-enabled vehicular metaverses. The main contributions are summarized as follows:

- According to vehicle data characteristics and service requirements, we design a new and hierarchical SemCom-enabled vehicular metaverse framework mainly consisting of a global metaverse, multiple local metaverses, and resource pools. This framework not only significantly reduces data transmission delay, but also efficiently utilizes the computing and storage resources of edge servers.
- We propose a novel adversarial attack called Semantic Noise Attack (SNA) to generate adversarial samples by adding semantic noise and design a new defense scheme using the Semantic Distance Minimization (SDM) mechanism to weaken the adversarial samples, but almost without sacrificing transmission accuracy.
- For use cases, we present two SemCom systems with different semantic encoders on traffic sign classification and license plate recognition, respectively. We examine the effects of SNA on the SemCom systems and evaluate the robustness of the defense scheme based on SDM in vehicular metaverses. The numerical results indicate that the proposed defense scheme significantly reduces the success rate of the attacks.

II. SEMANTIC COMMUNICATION FOR VEHICULAR METAVERSES

For vehicular metaverses, sensing data is divided into static and dynamic data based on the properties in the real world. Static data is the background information from the physical space, such as road sensors, roadside infrastructures, and the topology network of roads. Dynamic data is the “add-in” information from dynamic entities in the physical space including vehicles, pedestrians, traffic signs, etc. Actually,

the dynamic data is used to update the digital twins of dynamic entities in the virtual space. Compared with other metaverse services, vehicular metaverses have stricter latency requirements to ensure user immersion and improve the QoS.

A. Semantic Communication and Vehicular Metaverses

As shown in Fig. 1, a hierarchical vehicular metaverse consists of a global metaverse, n local metaverses, and m resource pools. More specifically,

- *Global Metaverse*: a giant metaverse deployed over the cloud by Metaverse Service Providers (MSPs) for management globally. The global metaverse is a replica of the physical world in the virtual space, which contains both static data and dynamic data within an entire city or even wider area, e.g., a global city map with abundant traffic-related information in the navigation applications. Instead of massively communicating directly with users, it was updated periodically by the data from all the local metaverses.
- *Local Metaverses*: as a component of the global metaverses, a local metaverse could be the metaverse mapping to a small-scale area, like a street block or a road intersection. The local metaverses are established on edge servers and are closely connected to nearby users, and thus the vehicles can communicate directly with their local metaverses to obtain real-time metaverse services without large latency [10].
- *Resource Pools*: the virtual resource pool consisting of physical entities with sufficient resources in the local metaverse, e.g., roadside units acting as edge servers. The edge servers in different regions establish a resource pool for their common local metaverse, which includes computing, storage, and bandwidth resources to support the metaverse services for users.

The hierarchical framework has several new designs and thus obtains multiple advantages. First, proximity to the users allows the framework to avoid long-distance communication, and thus reduce communication time. Users only need communication with the closest node in the resource pool to access the services. Second, the framework trims the budget for deploying dense sensor nodes in physical space. The strategy for updating and collecting data does not require massive sensors [10]. Third, the framework efficiently utilizes the storage and network resources of servers and vehicles. Both the global and local metaverses only hold indispensable data to maintain normal service. The vehicles only download the data of the local metaverses corresponding to their location. Finally, the SemCom in the framework reduces the number of bits that need to be transmitted. Hence, the effects of channel instability caused by vehicle movement are almost mitigated.

In vehicular metaverses, for AR navigation, the MSP collects massive static data in a low-cost way (e.g., from satellites) and deploys them on the global metaverse. As shown in 1, the static data at each intersection is regularly updated to corresponding resource pools to build local metaverses. The resource pools collect dynamic data from data providers in the physical world to keep updating digital twins. Meanwhile,

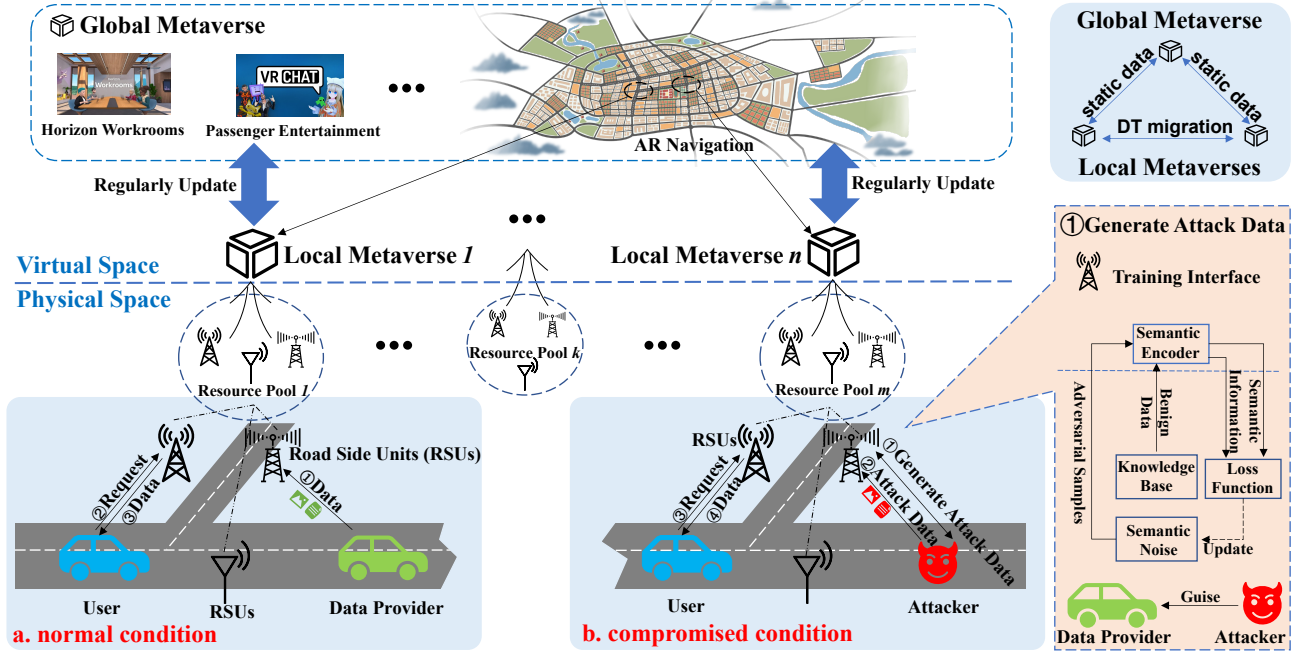


Fig. 1: The hierarchical vehicular metaverse architecture. Parts a and b show the normal service and attack flow respectively; The orange subfigure on the right-hand side depicts the steps involved in generating attack examples, while the blue subfigure illustrates the communication between the global metaverse and the local metaverses.

when a physical entity (e.g., a vehicle) moves from one local metaverse to another one, its digital twin may migrate to the new local metaverse according to its location. Note that all the changes in the static collected data will be updated to the global metaverse. Rather than encapsulating an entire city’s dynamic and static data in one metaverse, the local metaverses split the data, thus relieving the communication and computing pressure of the cloud and resource pools and shortening the distance between users and data. Vehicles can interact with edge servers in the resource pools to access metaverse services.

Before entering the coverage of a local metaverse, each user, i.e., vehicle, initiates a service request to the edge server. Once a user request is received, the edge server sends static information from the local metaverse to the vehicle. Meanwhile, the required dynamic data are constantly invoked and sent to the vehicle, allowing the driver to glimpse navigation and driving prompts. Congested wireless resources cannot support the transmission of such large amounts of data with low latency. To reduce latency, the edge servers only transmit key features of data to the vehicles by SemCom.

In vehicular metaverses, vehicles do not need to recover the source data, but only need to complete pragmatic tasks (e.g., traffic sign recognition and pedestrian detection). Given that the tasks in vehicular metaverses are relatively fixed, we design task-oriented SemCom systems [11] to improve efficiency. The task-oriented SemCom aims to extract more condensed semantic information related to the task of the receiving end. The receiver no longer needs a complex semantic decoder to recover the source data but completes the task directly based on the semantic information with a simple model. Coincidentally, this is consistent with the computational power

distribution between resource pools and vehicles.

We consider that each edge server has a dedicated semantic encoder for every task. Due to the heterogeneity of hardware performance, driving habits, and knowledge base, each vehicle has a customized task model. Hence, users need to train their models in conjunction with edge servers before enjoying the services. The edge servers open interfaces at night or during idle hours for user training. The interface mechanically extracts the semantic information of the original data given by the user and sends it back. Therefore, the user can train their task model by semantic information and the labels of original data.

After completing the training, the user can access the metaverse to enjoy application services. Three kinds of participants are involved in a service process as follows:

- *User:* In vehicular metaverses, the vehicle usually acts as a user. The proposed framework and all the mechanics are designed to make the user obtain a safe, reliable, comfortable, and immersive experience.
- *Edge Server:* The intermediary between the user and the data provider. The edge servers collect data from the data provider, store it temporarily and send it to the user when receiving a request.
- *Data Provider:* Any sensing node that wants to make money by selling data (e.g., some users, roadside cameras). A data provider located near an intersection contingently provides data to the edge servers in the corresponding resource pool.

Signal lights, traffic signs, pedestrians, and other information are captured by the data providers’ sensors and sent to the edge server. Based on SemCom, the edge server sends the

TABLE I: Existing Works Comparison.

Method	Background	Attack	Defense	Dataset	Modal	Model
ESCS [3]	○	Eavesdrop	Encryption	Europarl	Text	Transformer
VQ-VAE [6]	○	Adversarial attack	Mask + Codebook	CIFAR-10 Cars196 ImageNet	Image	ViT
R-DeepSC [7]	-	Semantic noise	Calibrating Augment training set	Europarl	Text	Transformer
DeepJSCEC [4]	○	Eavesdrop	Encryption	CIFAR-10	Image	DNNs
SNA&SDM	●	SNA	SDM	GTSRB CCPD	Image	ResNet LPRNet

The ○ represent white box attack, while the ● means black box setting. In the white-box setting, the attacker possesses knowledge of the structure or parameters of the semantic codec model. In contrast, the attacker in the black-box setting is ignorant and typically attacks by probing the model’s input and output.

data to the user with a graceful delay after the user initiates the request. By the task models, the driver can timely see the arrow guides and virtual traffic sign entities in front of the windshield, which can enhance the user’s perception of road conditions. This driving paradigm can greatly improve the safety and comfort of driving, especially when the view is blocked by trucks or buildings. Passengers can also use this system to play metaverse games or work online.

B. Security Challenges of Semantic Communication

SemCom, as a promising communication paradigm for the 6G future era, shows good performance in various tasks. However, it is vulnerable to some well-designed attacks. Therefore, it is necessary to consider the possible threats and defend against these attacks. We summarize typical security challenges as follows:

- *Eavesdropping Attack*: Eavesdropping is the behavior of decoding information in a physical channel in order to steal private information. In SemCom, the user-unique semantic codec acts as a barrier to eavesdroppers. Even though an eavesdropper obtains the transmitted semantic information, it cannot decode it. However, this barrier is not foolproof. When the SemCom system is widely used, users from the same communication link may have the same semantic decoder structure. Lu et al. [12] reveal the privacy issues of the SemCom system in the industrial Internet of Things. At the application layer, Luo et al. [3] propose an adversarial training method with key pairs to prevent such eavesdropping. Tung et al. [4] propose a joint source-channel and encryption coding scheme for wireless image transmission. At the physical level, Chorti et al. [13] consider physical layer security as the essential issue in next-generation communication. Du et al. [5] consider this problem from a physical perspective and protect wireless communications against eavesdropping by exploring and utilizing the inherent features of the physical medium.
- *Adversarial Attack*: By adding imperceptible semantic noise to the transmitted data, adversarial attacks make the SemCom system produce errors in the encoding stage or decoding stage. The authors in [6] consider two types of

attacks, one is on the sending end and the other one is on the receiving end. The former simulated the case where the attacker is the sender, while the latter simulated the case where the attacker initiates the attack through the channel. With all or part of the detailed information of the SemCom system, the attacker generates adversarial samples against the model. Xiang et al. [7] use the calibrated network and expand the training set to counteract the semantic noise in the text transmission tasks.

- *Poisoning Attack*: Poisoning attacks aim at manipulating the training process to degrade model performance or to insert backdoors. SemCom systems are trained based on shared semantic knowledge bases. The communication parties are likely to expand the training set from the third party to obtain a better SemCom codec. Some low-quality data and malicious data are inevitable, which may carry the wrong semantic information. Using data that contains false semantic information for training may lead to a degradation of performance in the SemCom system. This type of attack is called *semantic data poisoning attack* [14]. Researches on poisoning attacks and their defenses in the SemCom system are still in the vacant stage.

We summarize the existing work in Table I. Unlike prior works, we propose a black-box model, that is more realistic, and use datasets, that are consistent with vehicular metaverse scenarios. Moreover, we tailor semantic encoder models for different tasks.

III. ADVERSARIAL ATTACKS FOR SEMANTIC COMMUNICATION IN VEHICULAR METAVERSE

A. Adversarial Attacks

Deep learning models are often targeted by adversarial attacks, where attackers generate adversarial samples that appear natural to human eyes but cause the model to produce incorrect outputs. Adversarial samples are typically created by adding specific noise that appears to be normal noise to clean samples. In addition to channel noise and attenuation, SemCom furthermore faces a special noise called *semantic noise* [6]. While channel noise is simulated to improve performance in real environments, semantic noise, which can subtly alter the

meaning of data, is not typically considered. Semantic noise usually results from poor-quality data due to aging sensors and channel interference. By artificially adding semantic noise, attackers can efficiently produce adversarial samples, which cause a devastating blow to the SemCom system. We propose an attack based on semantic noise for vehicular metaverses.

According to Section II, it is stated that edge servers are nodes that are semi-trusted and do not actively attempt to attack the system, but can still be deceived. These servers forward both accurate and inaccurate data to users. Typically, edge servers assess the reliability of data based on the reputation of their provider. However, this method fails when a provider is compromised or intentionally behaves in a trustworthy manner to gain credibility.

The diagram in Fig. 1 illustrates how an attacker can impersonate a data provider and deceive well-behaved edge servers. Initially, the attacker collects natural data as if they were a legitimate provider. Based on the captured data, the attacker generates several adversarial samples through the SNA. At an opportune moment, the attacker submits these generated adversarial samples to the edge server. When a user requests information from the server, it extracts incorrect semantic details and sends them back to the user. This misinformation may lead users to make risky decisions. The process of generating adversarial samples through SNA is further explained in Fig. 1 and described below:

- *Step 1:* The attacker gains access to the training interface offered by the edge server and initiates a training procedure.
- *Step 2:* The attacker creates a group of adversarial samples using the benign data and then transmits both sets to the interface.
- *Step 3:* The interface extracts semantic information using the semantic encoder and transmits it back to the attacker.
- *Step 4:* In the received semantic information, the attacker calculates the maximum semantic distance to the benign data for every adversarial sample as the loss.
- *Step 5:* Update adversarial samples based on the loss and jump to *Step 2* until reaching the maximum iterations.

Using SNA, an attacker can calculate its loss by utilizing the interface provided by the edge server as environmental feedback, without knowing the structure and parameters of the semantic encoder. Edge servers aim to attract numerous data providers due to their requirement for dynamic data. As a result, detecting an infiltrating attacker is challenging, particularly when it attempts to enhance its reputation. It should be noted that the loss not only includes the steps necessary for executing an effective attack but also controls low-frequency perturbations [9] to minimize detection by human observers.

Once the SNA is implemented, it brings incalculable losses to the MSP and users. For MSPs, the presence of an attacker means that pre-trained semantic encoders can no longer be used, resulting in financial loss. For users, incorrect semantic information can cause drivers and passengers to watch lower-quality virtual entities and thus reduce immersion. In some applications, incorrect semantic information can even further manipulate user behaviors. For example, in AR navigation, a stop sign contains the semantic information of going straight.

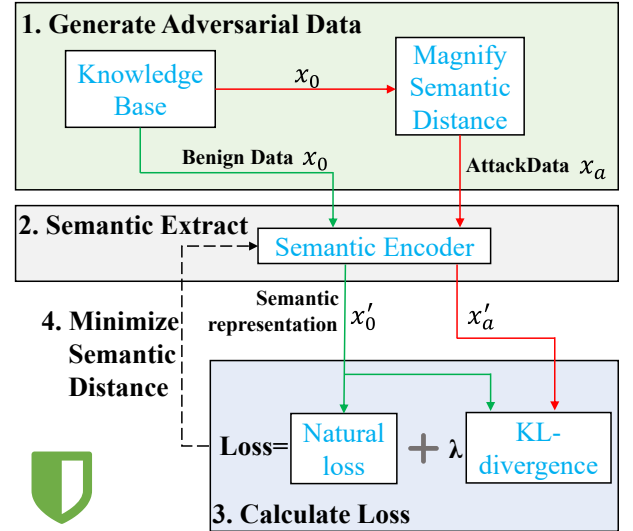


Fig. 2: Semantic Distance Minimization Defense Process, where lines with the same color represent the same data stream, the dashed line indicates that the semantic encoder is updated according to the loss function.

The user continues to drive straight based on semantic information, thus causing an accident.

B. Defense for Adversarial Attacks

As aforementioned, the SNA aims to modify semantic information by introducing semantic noise. In nature, semantic noise is the addition of disturbance in the efficient direction, which can change the position of data in the semantic space with the minimum disturbance. Generally speaking, one straightforward solution is to make the semantic encoder robust to semantic noise. Adversarial training is an attractive alternative but comes at the cost of natural accuracy. This article introduces the trade-off between natural and robust accuracy [15] into SemCom for vehicular metaverses to achieve SDM.

Adversarial training involves generating online adversarial samples that result in maximum semantic loss and allows the model to learn their distribution. In SDM, as illustrated in Fig. 2, online adversarial samples are generated for each data point through several iterations to maximize the distance from benign data. The KL divergence between the outputs of the semantic encoder is used to measure semantic distance. Both adversarial data and benign data are fed into the semantic encoder to calculate their semantic information. [15] The complete loss function includes two parts, natural loss, and robust loss. The robustness loss is determined by calculating the distance of extracted semantic information with a coefficient λ . Natural loss refers to the original mission's (e.g., classification) objective of improving model accuracy, while robust loss aims at ensuring that both adversarial samples and benign data have similar semantic information.

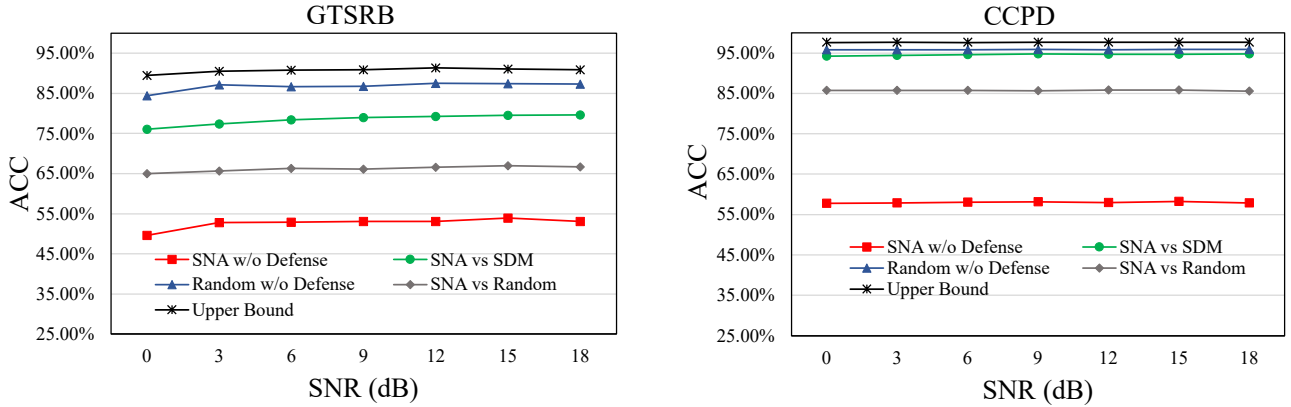


Fig. 3: Accuracy versus the SNR on the adversarial test set with $l_2 \approx 3.6$, where a vs b curve represents the b target model attacked in a attack mode. The random attack is to add stochastic noise to the source images.

IV. CASE STUDY: SEMANTIC COMMUNICATION IN VEHICULAR METAVERSE

As a case study, we simulate a scenario to evaluate proposed attack methods and defense performance. In this scenario, an attacker deduces the user’s location, produces adversarial samples, and provides them to RoadSide Units (RSUs) before the user accesses a local metaverse. To achieve this goal, we build SemCom systems in the above framework that simulate two tasks of AR navigation service in vehicular metaverses. The performance of the system performing traffic sign recognition and license plate recognition is measured. We conducted an SNA attack against these missions to assess the effectiveness of the attacks. Defensive SemCom systems are trained according to SDM, and their performance is re-evaluated. The same attack works on the new SemCom systems to test them.

A. Simulation settings and Security Attacks

We consider an attacker capable to deduce the user location and provide adversarial samples to RSUs before the user enters the local metaverse. In this context, we develop SemCom systems to simulate two tasks related to AR navigation services in vehicular metaverses. Specifically, we design semantic codecs for each task, while the channel encoder and decoder are both with dense layers with different units. During training, Additive White Gaussian Noise (AWGN) is added to the channel with a Signal Noise Ratio (SNR) between 5 and 10 [2], allowing the SemCom system to adapt for various real communication environments.

1) SemCom for traffic sign recognition: The GTSRB¹ dataset consists of more than 50,000 images between 15×15 to 250×250 pixels, divided into 43 classes. The pragmatic task of traffic sign recognition is completed in the communication process without the need to train the pragmatic function separately. In this simulation, the semantic encoder is a ResNet, while the receiver uses a fully connected layer to get the classification.

2) SemCom for license plate recognition: The CCPD² dataset consists of more than 250,000 vehicle images from China for license plate detection and recognition. In this simulation, the pragmatic task is license plate recognition, so we pre-process the dataset according to the annotation, cutting the license plate part of its image. The semantic encoder is an LPRNet³, while the receiver uses a greedy decoder to get the license plate information.

As for the attack, to simulate the SNA in vehicular metaverses, we generate adversarial samples of all the test images and perform the corresponding pragmatic tasks through the SemCom system. The same attack is performed on the defensive model. We conducted an SNA attack against these missions to assess the effectiveness of the attacks. We set the same L_2 norm for all attacks to ensure that performance is compared at the same attack strength [6], [9].

B. Metrics and Defense in Semantic Communication

The robust SemCom models are trained with the aforementioned defense method for the two pragmatic tasks. During the generating online adversarial images phase, 10 iterations are performed to maximize the semantic loss. In the training phase, we set λ as 1 in the traffic sign recognition task, which is 0.1 in the license plate recognition task to achieve better defensive performance. The same attack works on the new SemCom systems to test them.

The natural accuracy of the SemCom system on the test set is used to characterize its performance. For license plate recognition, only when all the symbols on the license plate are correctly recognized can be counted. We replace all images in the test set with the adversarial samples from the SNA and recalculate the accuracy, called *robust accuracy*. To evaluate the effect of an attack, we use natural accuracy as an upper bound. The degree of decline in robust accuracy compared to the upper bound indicates the effectiveness of an attack. More reduction in accuracy means more successful attacks. We also calculate the accuracy for SDM SemCom systems to

¹<https://benchmark.ini.rub.de/>

²<https://github.com/detectRecog/CCPD>

³<https://arxiv.org/abs/1806.10447>

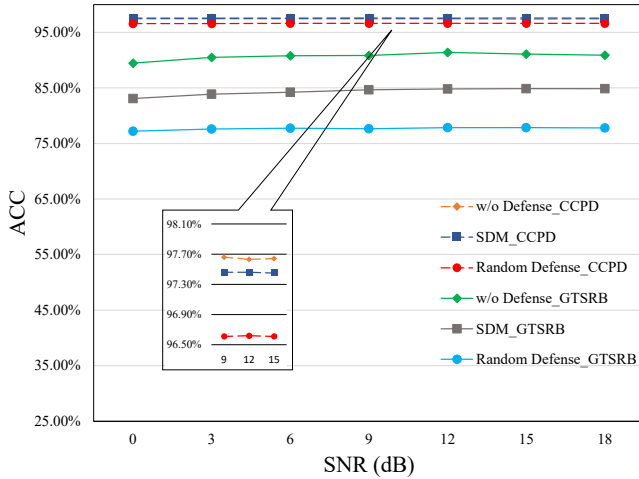


Fig. 4: Accuracy versus the SNR on the benign test set

demonstrate defensive effects. On the contrary, lower accuracy cuts mean more successful defenses. In addition, the difference in natural accuracy between the original SemCom system and the SDM system demonstrates the defense overhead. These assessments are repeated in different SNR environments.

C. Numerical Results

Figure 3 displays the SNR to accuracy curves of different models when tested on adversarial samples. In the task of recognizing traffic signs, the SemCom model’s accuracy is significantly reduced by 40% due to the SNA method. This reduction in precision cannot ensure the normal operation of services and may even mislead drivers into making dangerous decisions based on incorrect road signs. As random channel noise is considered during model training, adding random disturbance through attack methods only causes a slight decrease in precision. The SDM SemCom system exhibits higher accuracy than the original system against SNA attacks. Additionally, random defense adds random disturbance to benign images instead of maximizing semantic loss to simulate an attack sample, resulting in weaker defensive effects compared to our proposed approach shown in Figure 3. This phenomenon still exists with global deviation only in license plate recognition tasks.

Figure 4 shows the natural accuracy of different models when tested with benign data. Specifically, in the task of recognizing traffic signs, the SDM system exhibits a natural accuracy loss of approximately 5%, whereas the random defense approach results in a loss of 13%. Therefore, compared to random defense, SDM provides greater security while minimizing the reduction in natural accuracy.

V. FUTURE DIRECTION

A. Privacy Protection Schemes for Vehicular Metaverses

In vehicular metaverses, the communication link between vehicles and RSUs forms a dynamic topological network structure due to vehicle mobility. During wireless communication, there is a risk of data eavesdropping. A malicious party could

use eavesdropped data to learn private information such as user location and driving habits. Although there exists research on eavesdropping, the applicability of the defense schemes in dynamic environments has not been tested. Therefore, it is necessary to design a privacy protection scheme for the dynamic vehicular metaverse with changing network topology.

B. Customized Attack Methods for Vehicular Metaverses

Security is a major concern for both communication systems and vehicle applications. In vehicular metaverses, however, this issue has not received enough attention. A lot of work associated with adversarial attacks, poisoning attacks, and backdoor attacks focuses on deep learning. Moreover, a targeted attack in that an attacker can control the information received by the vehicle rather than just obfuscation is more harmful. If these attacks are carried out, they are bound to raise security concerns in vehicular metaverses. Therefore, diverse and stronger attack methods for SemCom systems need to be invented to support robustness research.

C. Efficient and Controllable Defenses for Attacks

This article presents a basic defense solution, which trains a semantic encoder with the capability of extracting correct semantic information from adversarial samples. However, the precision loss caused by the defensive approach should be avoided in some cases. For relatively simple pragmatic tasks, we can obtain security performance at a small cost, but for complex tasks, the cost will increase. Further research is needed to develop solutions that offer greater resistance with less cost and more controllability. The research on attacks and defenses can complement each other to further improve the security of semantic communication in vehicular metaverses.

VI. CONCLUSIONS

Considering the data characteristic and service requirements, we have developed a hierarchical framework for vehicular metaverses that addresses the challenges of large data volumes and strict latency requirements in vehicular metaverses. To minimize data transfer, we incorporated SemCom into the framework, while also raising security concerns with an adversarial attack method specific to vehicular metaverses. To protect against such attacks, we proposed a training method for the SemCom system that can withstand semantic noise. Our evaluation showed that our defense is effective in resisting adversarial attacks, which is crucial for ensuring a secure and immersive vehicular metaverse. We believe that applying security technologies to semantic communication in both deep learning and traditional security will provide valuable insights for creating a safe and comfortable vehicular metaverse.

ACKNOWLEDGMENT

The work is supported by NSFC under grant No. 62102099, U22A2054, No. 62101594, No. 62001125, and the Pearl River Talent Recruitment Program under Grant 2021QN02S643, and is supported by the National Research Foundation (NRF) and Infocomm Media Development Authority under the Future

Communications Research Development Programme (FCP). The research is also supported by the SUTD SRG-ISTD-2021-165, the SUTD-ZJU IDEA Grant (SUTD-ZJU (VP) 202102), and the Ministry of Education, Singapore, under its SUTD Kickstarter Initiative (SKI 20210204).

REFERENCES

- [1] Y. Jiang, J. Kang, D. Niyato, X. Ge, Z. Xiong, C. Miao, and X. Shen, "Reliable distributed computing for metaverse: A hierarchical game-theoretic approach," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2022.
 - [2] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.
 - [3] X. Luo, Z. Chen, M. Tao, and F. Yang, "Encrypted semantic communication using adversarial training for privacy preserving," *arXiv preprint arXiv:2209.09008*, 2022.
 - [4] T.-Y. Tung and D. Gunduz, "Deep joint source-channel and encryption coding: Secure semantic communications," *arXiv preprint arXiv:2208.09245*, 2022.
 - [5] H. Du, J. Wang, D. Niyato, J. Kang, Z. Xiong, M. Guizani, and D. I. Kim, "Rethinking wireless communication security in semantic internet of things," *arXiv preprint arXiv:2210.04474*, 2022.
 - [6] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, "Robust semantic communications with masked vq-vae enabled codebook," *arXiv preprint arXiv:2206.04011*, 2022.
 - [7] X. Peng, Z. Qin, D. Huang, X. Tao, J. Lu, G. Liu, and C. Pan, "A robust deep learning enabled semantic communication system for text," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 2704–2709.
 - [8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
 - [9] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, "Frequency-driven imperceptible adversarial attack on semantic similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 315–15 324.
 - [10] O. Hashash, C. Chaccour, W. Saad, K. Sakaguchi, and T. Yu, "Towards a decentralized metaverse: Synchronized orchestration of digital twins and sub-metaverses," *arXiv preprint arXiv:2211.14686*, 2022.
 - [11] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.
 - [12] Y. Lu and M. R. Asghar, "Semantic communications between distributed cyber-physical systems towards collaborative automation for smart manufacturing," *Journal of manufacturing systems*, vol. 55, pp. 348–359, 2020.
 - [13] A. Chorti, A. N. Barreto, S. Köpsell, M. Zoli, M. Chafii, P. Sehier, G. Fettweis, and H. V. Poor, "Context-aware security for 6g wireless: The role of physical layer security," *IEEE Communications Standards Magazine*, vol. 6, no. 1, pp. 102–108, 2022.
 - [14] W. Yang, H. Du, Z. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. S. Shen, and C. Miao, "Semantic communications for 6G future internet: Fundamentals, applications, and challenges," *arXiv preprint arXiv:2207.00427*, 2022.
 - [15] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.
- Jiawen Kang** (kavinkang@gdut.edu.cn) received the Ph.D. degree from the Guangdong University of Technology, China in 2018. He was a postdoc at Nanyang Technological University, Singapore from 2018 to 2021. He is currently a professor at Guangdong University of Technology, China. His research interests mainly focus on blockchain, security, and privacy protection in wireless communications and networking.
- Jiayi He** (hji99911@gmail.com) received the B.Eng. degree from Shandong University, Jinan, China, in 2021. He is currently pursuing an M.S. degree with the School of Automation, Guangdong University of Technology, Guangzhou, China. His research interests include semantic communications, security and privacy, and artificial intelligence-generated content.
- Hongyang Du** (HONGYANG001@e.ntu.edu.sg) received the B.Sc. degree from Beijing Jiaotong University, Beijing, China, in 2021. He is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, Energy Research Institute @ NTU, Nanyang Technological University, Singapore, under the Interdisciplinary Graduate Program. His research interests include semantic communications, reconfigurable intelligent surfaces, and communication theory. He was the recipient of the IEEE Daniel E. Noble Fellowship Award in 2022. He was recognized as an Exemplary Reviewer of the IEEE Transactions on Communications in 2021.
- Zehui Xiong** (zehui_xiong@sutd.edu.sg) is currently an Assistant Professor with the Pillar of Information Systems Technology and Design, Singapore University of Technology Design, Singapore. His research interests include wireless communications, network games and economics, blockchain, and edge intelligence.
- Zhaohui Yang** (yang_zhaohui@zju.edu.cn) is currently a ZJU young professor with College of Information Science and Electronic Engineering Zhejiang Key Lab of Information Processing Communication and Networking, Zhejiang University, and also a research scientist with Zhejiang Lab. His research interests include joint communication, sensing, and computation, and semantic communication.
- Xumin Huang** (huangxumin@gdut.edu.cn) received the Ph.D. degree in control science and engineering from the Guangdong University of Technology, Guangzhou, China, in 2019. From 2020 to 2021, he was a Visiting Scholar with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China. He is currently an Associate Professor with the School of Automation, Guangdong University of Technology. His research interests include network performance analysis, simulation, and enhancement in wireless communications and networking.
- Shengli Xie** (shlxie@gdut.edu.cn) received the M.S. degree in mathematics from Central China Normal University, Wuhan, China, in 1992, and the Ph.D. degree in automatic control from South China University of Technology, Guangzhou, China, in 1997. He was a Vice Dean of the School of Electronics and Information Engineering, South China University of Technology from 2006 to 2010. He is currently the Director of the Institute of Intelligent Information Processing and the Guangdong Key Laboratory of Information Technology for the Internet of Things, and also a Professor with the School of Automation, Guangdong University of Technology, Guangzhou. He has authored or coauthored four monographs and more than 100 scientific papers published in journals and conference proceedings, and was granted more than

30 patents. His research interests broadly include statistical signal processing and wireless communications, with an emphasis on blind signal processing and Internet of Things.