Manipulating Semantic Communication by Adding Adversarial Perturbations to Wireless Channel

Jianwei Liu[1,2], Yinghui He[3], Weiye Xu[1], Yifan Xie[4], and Jinsong Han[1]

[1]Zhejiang University, China
[2]Hangzhou City University, China
[3]Nanyang Technological University, Singapore
[4]The High School Attached to Zhejiang University, China

***Abstract*—To break through the transmission rate bottleneck of traditional communication, semantic communication is proposed to support emerging applications with extremely low latency requirements such as remote surgery and autonomous vehicle. Unlike the transmission of verbose symbols in traditional communication, mainstream semantic communications use deep learning technology to extract compact semantic information from data and convey it. However, the application of deep neural networks also poses security concerns, i.e., vulnerabilities to adversarial attacks. In this paper, we perform the first study on the security of semantic communication against both white-box and black-box attacks by compromising the wireless channel between the transmitter and receiver. To launch practical and effective attacks, a systematic and universal attack framework is designed to craft content-agnostic, undetectable, robust white-box perturbation signals as well as highly-transferable black-box ones. Extensive experiments on two open-source datasets demonstrate that our attack framework can achieve over 87%, 99%, and 89% success rates in untargeted white-box, targeted white-box, and untargeted black-box attacks. This means that the proposed attack methods could severely threaten the quality of service of current semantic communications. We also propose two mitigation methods to resist such attacks.**

***Index Terms*—Semantic Communication, Adversarial Attacks, WiFi**

I. Introduction

With the development of encoding/decoding and modulation techniques, current communication schemes like 5G [1] are approaching the Shannon physical capacity limit. Despite the decent latency (1ms) and data rate (20Gb/s) [2], they may not be capable of supporting lots of beyond-5G intelligent applications (such as brain-to-computer interaction, augmented reality, smart factories, and intelligent connected vehicle networks) that demand even lower latency (0.1ms) and higher data rates (1Tb/s) [3]. To overcome these issues, semantic communication is developed to break through the data rate bottleneck of conventional communication.

Benefiting from artificial intelligence (AI), recent mainstream semantic communications[1] employ deep learning techniques to convey semantic information, instead of traditional symbols, in an end-to-end manner [4]. At the transmitter end, the original data (e.g., an image of a barricade) is fed into an encoder (deep neural network, abbr. DNN) to get semantic information. After undergoing physical-layer transmission, the receiver obtains the semantic information and feeds it into a decoder (DNN) to recover the original data. Finally, the recovered data is sent to a pragmatic model (DNN) to perform specific tasks, such as barricade recognition in autonomous driving scenarios. This process greatly improves the transmission speed as the high-dimensional original data is compressed to low-dimensional semantic information. The number of symbols of an image in semantic communication can be 0.95% of that in traditional communication [5]. In this case, $\sim$2Tb images can be transmitted per second in semantic communication, while only 20Gb of those in conventional communication. Therefore, semantic communication is considered to be an important development direction of next-generation wireless communication systems [3], such as cellular and WiFi systems.

While researchers focus on boosting the transmission efficiency, the safety/security aspect of semantic communication is understudied. In fact, an attacker could manipulate the semantic information by adding perturbation signals over the wireless channel. Unfortunately, the pillar of semantic communication, i.e., DNN, is known to be vulnerable to adversarial perturbations [6]–[8]. In many application scenarios, a false understanding of the semantics could cause severe accidents, threatening people's lives and property. For example, attacks on communication could compromise the ability of military systems in detecting or jamming enemies [9]. In smart factories, incorrect classification by the pragmatic model could result in machines generating improper instructions that may harm workers. In the context of vehicle networks, false recognition could prompt vehicles to adopt erroneous driving strategies, leading to car accidents such as rear-end collisions [7].

A key reason that researchers have not paid enough attention to the safety/security aspect is that: launching *practical* and *effective* adversarial attacks against semantic communication is challenging. Firstly, whereas there are already a large number of adversarial attack approaches [10], [11], they cannot be trivially applied to semantic communication. This is because most previous works craft perturbations based on known input

Jinsong Han is the corresponding author.

[1]In the following, we focus on deep learning-empowered semantic communications.

data. But in practice, the attacker is not aware of the content of the input, making it impossible to predict what the transmitter will send. Besides, previous works only need to face one target DNN, while there are three target DNNs (encoder, decoder, and pragmatic model) in semantic communication. Secondly, to guarantee the success of the attack, the added perturbation should hide itself well. In other words, the perturbation should be undetectable. Thirdly, the perturbation signal would experience distortion when propagating through a wireless channel, which may degrade the attack's effectiveness. Hence, the robustness of the perturbation should be taken into consideration. Fourthly, under a more realistic setting (i.e., black-box setting), the underlying DNNs in the communication system are usually unknown to attackers. Thus, the perturbation should be generated without knowing the architecture and parameters of the underlying DNNs. Although there have been some efforts devoted to crafting black-box perturbation based on surrogate model, the training of the surrogate model requires a large number of queries to the target DNN. This is not only time-consuming and laborious, but also raises suspicion in the target system.

In this work, we perform in-depth study on the effect of adversarial samples on semantic communication by leveraging the wireless channel as the attack vector. By addressing the above-mentioned challenges, we propose a *systematic* and *universal* framework to craft adversarial perturbations. Our framework demonstrates the feasibility and practicality of launching effective white-box and black-box attacks against semantic communication. Particularly, under the white-box setting, our approach enables the attacker to execute untargeted and targeted attacks. The untargeted perturbation can lead the target system to generate incorrect predictions, while the targeted perturbation can cause the system to misclassify an input as any label specified by the attacker. Under the black-box setting, the attacker can mislead the communication system to generate erroneous output without knowing the architecture of the underlying DNNs adopted by the communication system.

Specifically, we first treat the three target DNNs as an entire model, based on which untargeted and targeted attack losses are designed to guide the optimization of the perturbation. Instead of updating the perturbation based on one given input, we average the attack losses over a batch of different inputs. This allows optimized perturbations to adapt to diverse inputs, i.e., be content-agnostic. Then, on one hand, we constrain the power of the perturbation noise; on the other hand, we initialize the perturbation as random Gaussian noise and introduce a Gaussian classifier to restrict the perturbation to approach Gaussian distribution. These guarantee the undetectability of our attack because low-power Gaussian noise is a very common "appendage" in wireless communication [12]. Furthermore, to make the perturbation more robust, we incorporate the effect of random channel state information (CSI) during optimization, such that the ultimate perturbation is tolerant against channel state variations. Finally, we construct a decision-sensitive query dataset (including normal ones, noised ones, perturbed ones, etc.) to query the communication system.

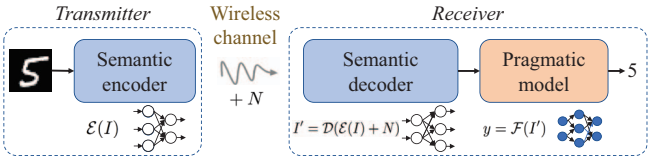


Fig. 1. Semantic communication.

Together with the labels output by the target DNNs, we can construct a training set to train a surrogate model through multi-round learning. Thanks to that such decision-sensitive data can effectively capture the decision boundary of the underlying DNNs, only a small number of queries is sufficient to train a representative surrogate model to mount effective black-box attacks.

We reproduce a semantic communication system [13] and conduct experiments on two widely-used open-source datasets (MNIST [14] and Fashion-MNIST [15]) with CSI collected in two real environments. The results indicate that the proposed attack framework can achieve 87%+, 99%+, and 89%+ success rates in untargeted white-box, targeted white-box, and untargeted black-box attacks, respectively. Transferability study manifests that our black-box attack is still effective even when the attacker lacks knowledge of the architectures used by the DNNs in the communication system. These results demonstrate that our attack approaches greatly impact the quality of service (QoS) of existing semantic communications. Finally, we propose two mitigation methods to help users enhance the security of semantic communication systems.

The contributions of this paper are as follows. (1) We design a systematic and universal attack framework to craft content-agnostic, undetectable, robust white-box perturbation signals, as well as highly-transferable black-box ones. (2) A decision-sensitive query dataset construction method is proposed to train a surrogate model for the target DNNs, which achieves high black-box attack effectiveness while effectively reducing the queries compared to the conventional approach. (3) Extensive experiments on two open-source datasets demonstrate that the proposed attack framework can achieve 87%+ attack success rates, even when the attacker has no knowledge about the DNNs in the semantic communication system.

II. Background

A. Semantic Communication

As the performance of the cutting-edge 5G technologies [16], [17] already cannot satisfy the high requirements of emerging intelligent applications, there is more and more enthusiasm for the employment of semantic communication to allow transmission with lower latency and higher data rate. As shown in Fig. 1, a semantic communication system usually contains two hardware components: a transmitter and a receiver. The data is transmitted from the transmitter to the receiver through a wireless channel.

There are two major differences between traditional communication and semantic one. First, traditional communication systems transmit original data (e.g., the pixels of image), while

semantic communication systems transmit the semantic information extracted from the original data. Secondly, traditional communication systems usually are task-irrelevant, i.e., they can transmit any kind of data like image and audio with corresponding encoding schemes. However, a semantic communication system is task-specific. The tasks are diverse, such as image [18], text [19], [20], and speech recognition [21].

Taking image transmission as an example, the transmitter first employs an encoder $\mathcal{E}(\cdot)$ to extract low-dimensional semantic features S from the image I: $S = \mathcal{E}(I)$. Then, the semantic features are transmitted to the receiver through a wireless channel. At the receiver, a decoder $\mathcal{D}(\cdot)$ takes as input the semantic features and recovers the original data I':

$$I' = \mathcal{D}(\mathcal{E}(I) + N) = \mathcal{D}(S + N), \tag{1}$$

where N represents the additive white Gaussian noise (AWGN) [9]. Thereafter, the recovered data can be fed into a pragmatic model $\mathcal{F}(\cdot)$ to perform the specific task – image recognition: $y = \mathcal{F}(I')$, where y is the predicted label of the transmitted image[2].

The training of the DNNs in a semantic communication system can be divided into two phases [13]. In the first phase, the pragmatic model is trained individually based on a dataset. Then, the encoder and decoder are jointly trained based on the outputs of the decoder and pragmatic model. In this training manner, the encoder and decoder can learn to extract representative semantic features that maximize the accuracy of the pragmatic model.

B. Semantic Communication in Adversarial Environments

Although deep learning techniques manifest outstanding advantages of high accuracy in many fields like image recognition, DNNs are susceptible to adversarial attacks [6]. Generally, the goal of an adversarial attack is to generate an adversarial sample I^a to mislead a target DNN $\mathcal{F}(\cdot)$ to make erroneous predictions or classifications. An adversarial sample is crafted by adding unnoticeable perturbations δ to a normal sample I: $I^a = I + \delta$. The perturbation is optimized via the following objective function:

$$\min \quad ||I^a - I||_p, \quad \text{s.t.} \quad \mathcal{F}(I^a) \neq y \text{ and } I^a \in \mathbb{I}, \tag{2}$$

where $\mathcal{F}(I^a) \neq y$ is the attack objective, y is the groundtruth label of I, and $I^a \in \mathbb{I}$ means that the generated adversarial sample I^a is in a valid set. There have been a large number of works devoted to designing adversarial perturbations, e.g., fast gradient sign method (FGSM) [10]. However, different from previous works, the perturbation can only be added to the semantic information in our attack scenario: $S^a = S + \delta$. Accordingly, the optimization objective should be re-written as:

$$\min \quad ||S^a - S||_p, \tag{3a}$$

$$\text{s.t.} \quad \mathcal{F}(\mathcal{D}(\mathcal{E}(I) + N + \delta)) \neq y \text{ and } S^a - S \in \mathbb{S}. \tag{3b}$$

To guarantee the undetectability, $\mathbb{S}$ belongs to the Gaussian distribution.

[2]Here, we do not consider the data preprocessing techniques like modulation, as they have many variants. We focus on most common components including DNN-based encoder, decoder as well as pragmatic model.

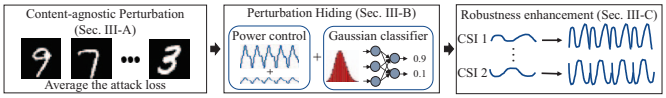


Fig. 2. White-box attack scheme.

III. Attack Model

The goal of the adversarial attack is to transmit a well-crafted perturbation signal over the wireless channel, such that the DNNs at the receiver end misunderstand (e.g., misclassify) the transmitted semantic information. In this process, the perturbation is added to the semantic information. The receiver will feed the perturbed signal into the decoder and pragmatic model consecutively, and obtain a classification/recognition result. We consider two adversarial settings: white-box attack and black-box attack.

White-box attack: The attacker has full knowledge of all the DNNs in the semantic communication system, including the encoder, decoder, as well as pragmatic model. S/he knows the architectures and parameters of these DNNs and can obtain the corresponding outputs. However, the attacker cannot predict the upcoming semantic information. Under the white box setting, we consider two types of attack effects: untargeted attack and targeted attack. In the former one, the attacker emits a perturbation signal to make the receiver recognize the received semantic information as an arbitrary wrong class. As for the targeted attack, the perturbed signal will be mistaken for the class specified by the attacker.

Black-box attack: Under this setting, the attacker has no knowledge about the underlying DNNs in the target communication system. Yet, the attacker can query the system and obtain corresponding outputs. Hence, we assume that the attacker is able to query the system to construct a dataset, based on which the attacker can train a surrogate model to craft adversarial perturbations. As the underlying DNNs are apparent to the attacker, the generated perturbations should have adequate transferability from the surrogate model to the underlying DNNs. Under the black-box setting, we consider untargeted attacks.

IV. Attack Framework

This section presents the systematic and universal adversarial attack framework against semantic communication. In the framework, the attacker first generates Gaussian distributed noise as the initial perturbation. As shown in Fig. 2, by optimizing with the average attack loss of different semantic content, the perturbation not only can result in false classification, but also become content-independent. Meanwhile, the attacker suppresses the power of the perturbation signal on the one hand, and uses a Gaussian classifier to limit its distribution on the other. These improve the stealthiness of the attack. Furthermore, to make the adversarial perturbation still effective after experiencing the distortion caused by the wireless channel, the attacker incorporates the CSI in the optimization process. Ultimately, the attacker constructs a decision-sensitive query dataset to obtain a training set, with which the attacker can train a surrogate model through data

augmentation and multi-round optimization. Based on this surrogate model, the attacker can craft adversarial samples transferable to the target DNNs.

A. Content-agnostic Perturbation

As aforementioned, the goal of the untargeted attack is to craft a perturbation signal to let the target DNNs misclassify. The untargeted attack loss can be formulated as:

$$\mathcal{L}_u = |(\mathbb{1} - O_M(y)) - \mathcal{F}(\mathcal{D}(\mathcal{E}(I) + N + \delta))|^2, \quad (4)$$

where $\mathbb{1}$ is a vector whose length is the same as the number of classes M and all elements are 1s. $O_M(y)$ denotes the one-hot encoding of the real label y of the input I, which has M elements. By minimizing L_u, the generated perturbation tends to make the DNNs allocate higher confidence to the classes other than the true one. In order to achieve targeted attacks, the perturbation can be updated according to the following loss:

$$\mathcal{L}_t = -\sum_{\mathrm{c}=1}^{M} V_{y_t} \log(P_c), \quad (5)$$

in which V_{y_t} is the indication variable, $P_c = \mathcal{F}_c(\mathcal{D}(\mathcal{E}(I) + N + \delta))$ is the probability that the input $\mathcal{D}(\mathcal{E}(I) + N + \delta)$ belongs to the class y_t.

However, in a realistic attack scenario, the attacker has no knowledge about the content of the upcoming transmitted semantic information. Therefore, an effective perturbation should be content-agnostic. To this end, we incorporate the diversity of content during optimization. Instead of updating the perturbation based on an (input, label) pair, we calculate the attack loss of a dataset (I_i, y_i), where $i \in [1, N_I]$ and N_I is the number of the participated inputs. The corresponding untargeted attack loss can be re-written as:

$$\mathcal{L}_u = \frac{1}{N_I} \sum_{i=1}^{N_I} |(\mathbb{1} - O_M(y^i)) - \mathcal{F}(\mathcal{D}(\mathcal{E}(I^i) + N^i + \delta))|^2. \quad (6)$$

In this way, the perturbation can be tolerant to the variation of the semantic information. No matter what the transmitter plans to send, the added perturbation signal is very likely to make the communication system output a wrong class. Accordingly, the loss utilized to achieve the content-independent targeted attack can be written as:

$$\mathcal{L}_t = -\frac{1}{N_I} \sum_{i=1}^{N_I} \sum_{\mathrm{c}=1}^{M} V_{y_t}^i \log(P_c^i), \quad (7)$$

Based on this loss, the generated perturbation can mislead the communication system to recognize any input as the class specified by the attacker.

B. Perturbation Hiding

To launch the attack effectively, the adversarial samples should be indistinguishable from the normal ones, i.e., the added perturbation should be undetectable at the receiver. For doing so, we use two initiatives to restrict the perturbation during optimization. For one thing, we control the power of the perturbation signal to make it unnoticeable. For another, we constrain the distribution of the perturbation to a Gaussian one.

Specifically, if the power of the perturbation is too high, the semantic information as well as the recovered data will be severely distorted. Conspicuous data distortion will cause the system to be alert, rendering the attack easier to be detected. Or the system will just discard the data with a low signal-to-noise ratio (SNR). To address this issue, we incorporate the power constraint as follows:

$$\mathbb{E}\{S \times S^T\} \leq \frac{\mathbb{E}\{\delta \times \delta^T\}}{r}, \quad (8)$$

where $\mathbb{E}\{\delta \times \delta^T\}$ and $\mathbb{E}\{S \times S^T\}$ are the power's expectations of the perturbation signal and semantic information, respectively. r is the upper bound of the ratio of the perturbation signal's power to the semantic information's one. By controlling r below 10% [22], the generated perturbation can become as inconspicuous as the noise normally introduced in wireless communication.

Thereafter, to incorporate statistic undetectability, we utilize a Gaussian classifier to make the noise meet the Gaussian distribution. The insight behind is that Gaussian distribution is an inevitable noise distribution that will be added to the transmitted data in wireless communication [12]. Particularly, we first construct a dataset to train the Gaussian classifier. The dataset is composed of the randomly generated vectors with various distributions, including Gaussian one, even one, Rice one, and Rayleigh one [23]. Gaussian distribution is labelled by 1 and the others belong to label 0. Before optimizing the perturbation, we train the Gaussian classifier using MSE loss [24] in advance. Combining with the power constraint and the well-trained Gaussian classifier $G(\cdot)$, we enhance the untargeted attack loss to strengthen the undetectability of the perturbation:

$$\begin{aligned} \mathcal{L}_u = & \frac{\alpha}{N_I} \sum_{i=1}^{N_I} |(\mathbb{1} - O_M(y^i)) - \mathcal{F}(\mathcal{D}(\mathcal{E}(I^i) + N^i + \delta))|^2 \\ & + (1-\alpha)|O_2(1) - G(\delta)|^2, \\ \text{s.t.} \quad & \mathbb{E}\{S \times S^T\} \leq \frac{\mathbb{E}\{\delta \times \delta^T\}}{r}, \end{aligned} \quad (9)$$

where α is a hyper-parameter set to 0.6 empirically. $O_2(1)$ equals to $[0, 1]$. Accordingly, the loss for a targeted attack can be formulated as:

$$\begin{aligned} \mathcal{L}_t = & \frac{\alpha}{N_I} \sum_{i=1}^{N_I} \sum_{\mathrm{c}=1}^{M} V_{y_t}^i \log(P_c^i) \\ & + (1-\alpha)|O_2(1) - G(\delta)|^2, \\ \text{s.t.} \quad & \mathbb{E}\{S \times S^T\} \leq \frac{\mathbb{E}\{\delta \times \delta^T\}}{r}. \end{aligned} \quad (10)$$

C. Robustness Enhancement

Regardless of the AWGN, wireless signals would experience amplitude attenuation and phase shift during propagation in the physical world, which can be described by CSI. The signal changes caused by CSI may result in the distortion of the perturbation, degrading the attack's effectiveness. To enhance the robustness of the perturbation, we first collect a batch of CSI measurements in real environments. Then, we incorporate

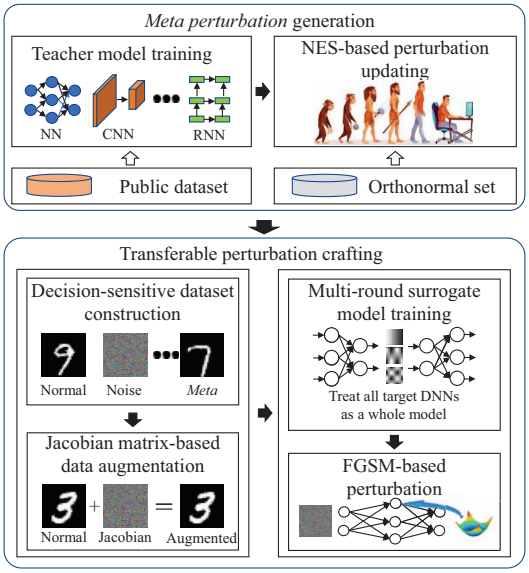


Fig. 3. Black-box attack scheme.

the effect of CSI in the first term of the untargeted attack loss:

$$\frac{\alpha}{N_I}\sum_{i=1}^{N_I}|(\mathbb{1}-O_M(y^i))-\mathcal{F}(\mathcal{D}(\frac{(\mathcal{E}(I^i)+\delta)\times H+N^i}{H}))|^2, \tag{11}$$

where H is the CSI randomly selected from the CSI batch. Similarly, to make the perturbation in targeted attack robust against the CSI variations, P_c in $\mathcal{L}_t$ can be re-written as $P_c = \mathcal{F}_c(\mathcal{D}(\frac{(\mathcal{E}(I^i)+\delta)\times H+N^i}{H}))$.

With the above countermeasures, we can obtain content-agnostic, undetectable, and robust perturbations to achieve untargeted and targeted attacks under white box settings.

D. Transferable Perturbation

In a more practical attack scenario, i.e., under black-box settings, the attacker may have no knowledge about the underlying DNNs in the communication system. In this case, we cannot directly obtain the gradients of the underlying DNNs to calculate the loss and further optimize the perturbation. Previous studies [25] demonstrate that building a surrogate model and crafting the perturbation according to the gradients of this surrogate model is an effective way to achieve black-box attacks. In such methods, the attacker needs to construct a training dataset, i.e., acquiring (input, label) pairs, for the surrogate model by querying the underlying DNNs. To guarantee high attack effectiveness, the generated perturbation should have sufficient transferability. This depends on whether the decision boundary of the surrogate model is similar to that of the underlying DNNs. To keep the decision boundary of the surrogate model as close to the underlying DNNs as possible, the attacker needs to query the underlying DNNs for a large number of times to form a large training dataset for the surrogate model. This not only causes a lot of overhead, but also easily arouses the alarm of the system. To tackle this issue, we first construct a decision-sensitive query dataset composed of various special samples. Since these special samples can describe the decision boundary, fewer queries can make the trained surrogate model resemble the underlying DNNs. Then, we perform multiple rounds of queries on the underlying DNNs to construct a training set for the surrogate model [25]. Based on the trained surrogate model, we compute the gradient and generate transferable perturbations.

1) Decision-sensitive Query Dataset: Compared with normal samples, special samples like perturbed ones can better describe the decision boundary of a DNN [26]. Thus, to reduce the number of queries, we construct a decision-sensitive dataset. This dataset only consists of a few normal samples and a small number of special samples, including noised ones, random ones, and adversarial ones. Therein, the noised sample is generated by adding random noise to a normal sample. A random sample is actually a random even distribution. As for the adversarial sample here, it is generated by our *meta perturbation* generation approach.

Specifically, as the attacker does not have any knowledge of the internals of the underlying DNNs, s/he cannot craft the perturbation specific to them. To address this issue, we aim to devise a kind of perturbation (called *meta pertuebation*) that adapts to as many network architectures as possible. Particularly, inspired by the voting scheme in ensemble learning, we propose a natural evolution strategy [27] (NES)-ensemble algorithm to generate *meta perturbation*. As shown in Fig. 3, several mainstream model architectures (e.g., NN, CNN [28], RNN [28], and LSTM [29]) are pre-trained as *teacher models*. Then, the adaptability of a perturbation can be estimated based on the attack effectiveness on these teacher models. Given a test input and its label, if the perturbation can make more teacher models misclassify the test input, it means that the perturbation has better adaptability. The insight behind this design is that the adaptability of a perturbation is mostly impacted by the high-level differences (e.g., convolutional layer vs. recurrent layer) between models, rather than low-level ones like the depth and number of neurons [25]. Thus, we only need to introduce as much diversity on high-level architecture as possible when building teacher models. Fortunately, although the choice of model for the low-level parameters is kaleidoscopic, there are only a few common high-level model architectures. Meanwhile, the depth and number of neurons are unnecessary to be too large, which would save the computational overhead. With the guidance of these teacher models, we can optimize the perturbation. Intuitively, we can achieve this by calculating the gradients of all the teacher models. However, such a gradient-based method may lead the perturbation to be specific to these teacher models, impairing its adaptability. Therefore, we opt to use NES to find the optimal perturbation. Given an input S and its label y, the transferability, i.e., fitness, of a perturbation can be quantified by the number of the teacher models that misclassify S:

$$fit = \sum_{i=1}^{N_{\mathcal{F}}} F_{imp}(\mathcal{F}_i(S)-y), \tag{12}$$

where $N_{\mathcal{F}}$ is the number of the teacher models and $F_{imp}(\cdot)$ is the impulse function [30].

Thereafter, we need to update the perturbation to make

TABLE I
DETAILS OF DNNS IN ORIGINAL SEMANTIC COMMUNICATION PROPOSAL [13] AND SURROGATE MODEL.

DNN	Encoder	Decoder	Pragmatic model	Surrogate model
Input	1×784	1×392	1×784	1×784
Hidden layers (FC)	(784, 392)	(392, 784)	ReLU(784, 500) $\rightarrow$ ReLU(500, 250) $\rightarrow$ ReLU(250, 125) $\rightarrow$ (125, 10)	(784, 512) $\rightarrow$ (512, 392) $\rightarrow$ (392, 10)
Loss function	Cross-entropy & MSE	Cross-entropy & MSE	Cross-entropy	Cross-entropy
Optimizer	SGD	SGD	SGD	Adam
Learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}

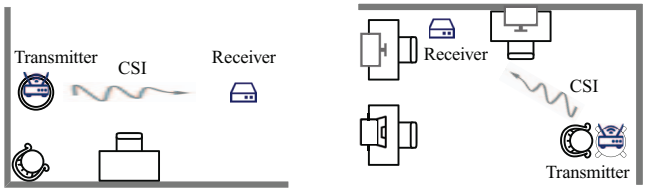
it more adaptable. For doing so, traditional methods add randomly generated distribution like Gaussian distribution to the perturbation [7]. Such blind optimization methods can improve the perturbation's adaptability gradually, but they converge slowly and introduce significant computational overhead. To solve this problem, we generate an orthonormal set as the optimization direction. Any two vectors in this set are orthonormal to each other. Such an orthonormal basis can expedite the optimization procedure [31]. In each iteration, we randomly select a vector from the orthonormal set, and let the perturbation either add or subtract it.

With the decision-sensitive query dataset, the attacker can obtain a small-yet-effective training set for the surrogate model.

2) Perturbation crafting: Next, the attacker queries the target DNNs and trains the surrogate model based on the obtained training set. With the well-trained surrogate model, the attacker can craft transferable perturbation to achieve black-box attacks. Instead of performing one-time query and one-time training, we adopt the multi-round query as well as training [25]. In each round, we first augment the query samples by the Jacobian matrices of the surrogate model, and then train the surrogate model together with the augmented samples as well as their labels given by the underlying DNNs. To be specific, the insight behind augmenting the training set is to identify the directions that the output of the target DNNs is varying around the initial training points. Intuitively, we need a large number of (input, label) pairs to capture the target DNNs' output variations, such that the trained surrogate model can accurately approximate the target DNNs. Fortunately, these directions can be identified with the surrogate model's Jacobian matrix [25]. Thus, in every subsequent round except the first round, each sample is augmented by:

$$S_{p+1} = S_p + \lambda \times sgn(J|F_{sur}(S_p)|), \tag{13}$$

where p is the index of the rounds, λ is a hyper-parameter describing the size of the step taken in the sensitive direction (set to 0.1 empirically), $sgn(\cdot)$ means maintaining the sign matrix, and $J|F_{sur}(S_p)|$ means obtaining the Jacobian matrix on the surrogate model $F_{sur}(\cdot)$, given the semantic information S_p. In each round, the surrogate model is optimized using the cross-entropy loss [32]. As for its architecture, there is no requirement for the consistency with the target DNNs. But they should preferably be of the same architecture type. This is relatively easy for an attacker to infer when s/he knows the pragmatic task of the communication system. For example, image classification is most likely to be achieved by NN/CNN. After training the surrogate model, the attacker can use



(a) Lab (b) Office

Fig. 4. Laboratory and office used to collect real CSI.

FGSM [10] to craft the transferable adversarial perturbation:

$$\delta = \epsilon \times sgn(\nabla_S F_{SUR}(S, y)), \tag{14}$$

in which ϵ is a hyper-parameter and ∇ means calculating the gradients.

V. EXPERIMENT

Experiment setup. We evaluate the effectiveness of our attack framework on the most common pragmatic task – image classification. For the victim semantic communication system, we use the same setup as the original paper [13]. The underlying DNNs are the same as the ones proposed in the original paper. The architectures of the target DNNs are shown in Tab. I. To evaluate more comprehensively, we conduct experiments on two widely-evaluated open-source datasets: MNIST [14] and Fashion-MNIST [15]. MNIST is a dataset of handwritten digits ranging from zero to nine. It provides 60,000 images and each image has 28×28 pixels. With the network architectures proposed in the original paper of semantic communication, we have 93.0% classification accuracy. Fashion-MNIST consists of ten classes of Zalando's products like shoes and trousers. It also contains 60,000 images with each 28×28 pixels. We get 82% classification accuracy when using the network architectures in the original paper. The Gaussian classifier in Sec. IV-B is composed of two fully-connected layers with rectified linear unit (ReLU) in between. To simulate real communication, we collect CSI by using the existing WiFi CSI tool [33] in two real environments: laboratory and office, as shown in Fig. 4. In the default setting, the SNR of the AWGN, the power ratio in Eq. 8, and λ in Eq. 13 are set to 10dB, 0.1, and 0.1, respectively. The compression rate, i.e., the ratio of the number of elements in the semantic information (392) to the number of pixels in the original image (784), is set to 0.5. For each dataset, 10,000 samples are randomly selected to train underlying DNNs in the semantic communication system, In the white-box attack, we also randomly select 10,000 samples to optimize the perturbation. The remaining samples are used to evaluate the attack performance. There is no overlap between the data used to train the target DNNs,

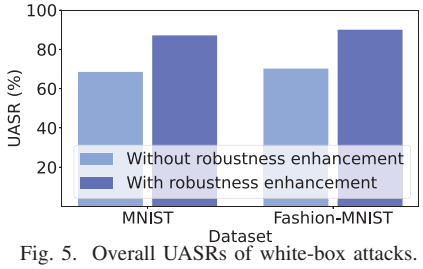


Fig. 5. Overall UASRs of white-box attacks.

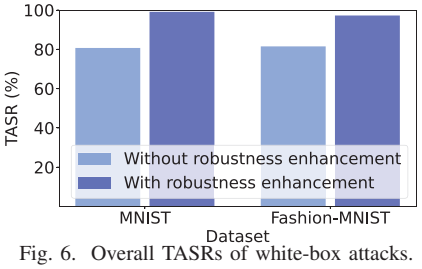


Fig. 6. Overall TASRs of white-box attacks.

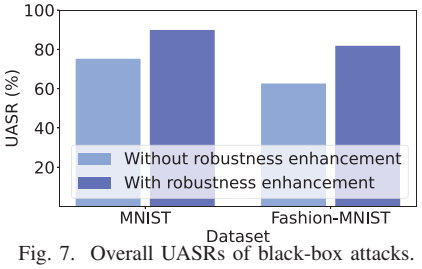


Fig. 7. Overall UASRs of black-box attacks.

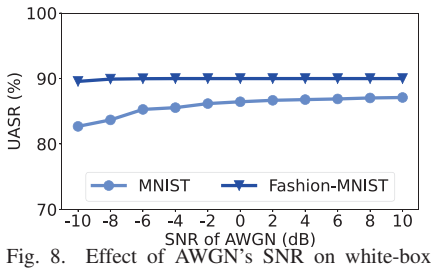


Fig. 8. Effect of AWGN's SNR on white-box UASR.

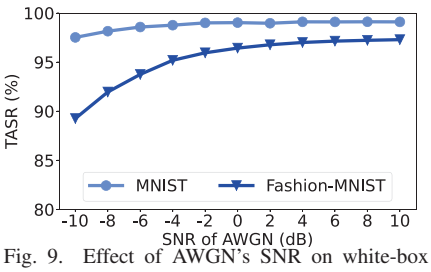


Fig. 9. Effect of AWGN's SNR on white-box TASR.

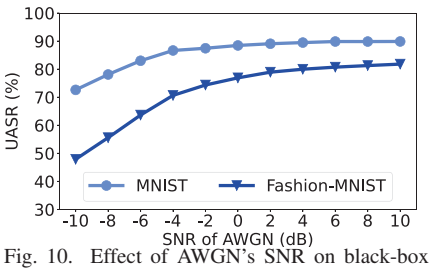


Fig. 10. Effect of AWGN's SNR on black-box UASR.

optimize perturbation, and test perturbation. In the black-box attack, 40 samples are constructed as the query dataset (including 10 normal samples, 10 noised ones, 10 random ones, and 10 perturbed ones). The architecture of the surrogate model is displayed in Tab. I. The surrogate model is trained in five rounds. The perturbation is updated by Adam optimizer.

Metric. We define untargeted attack success rate (UASR) and targeted attack success rate (TASR) to quantify the attack effectiveness. UASR can be calculated by: $UASR = 100\% \times \frac{N_{unt}^{inc}}{N_{unt}^{all}}$, where N_{unt}^{inc} and N_{unt}^{all} are the number of misclassified test samples and the number of all test samples, respectively. TASR can be calculated by: $TASR = 100\% \times \frac{N_{tar}^{cor}}{N_{tar}^{all}}$, where N_{tar}^{all} and N_{tar}^{cor} are the number of all test samples and the number of the test samples accurately predicted as the target classes.

A. Overall Attack Effectiveness

We first assess the overall attack effectiveness of our attack framework on the untargeted white-box attack, targeted while-box attack, and untargeted black-box attack, in the laboratory environment. To validate the effectiveness of our robustness enhancement method, we also show the attack success rate achieved without incorporating the CSI variation. The UASR for white-box, TASR for white-box, and UASR for black-box attacks are shown in Fig. 5, Fig. 6, and Fig. 7, respectively. It can be observed that, without robustness enhancement, the UASRs of MNIST and Fashion-MNIST are lower than 71%. After adopting the robustness enhancement, the UASRs of MNIST and Fashion-MNIST increase to 87.1% and 90.0%, respectively. This suggests the effectiveness of our robustness enhancement method in improving the attack success rates in the real world. Besides, even without robustness enhancement, the UASRs are higher than 68%. This demonstrates the high effectiveness of our perturbation optimization method. For targeted white-box attack (Fig. 6), it can be found that the TASRs of these two datasets are all larger than 97%. The TASR of MNIST is even higher than 99%. Meanwhile, the TASRs with robustness enhancement are larger than those without robustness enhancement. These results manifest that our attack framework can craft significantly effective targeted perturbations to make the communication system output any class specified by the attacker. Additionally, the TASR for each of the ten classes in MNIST/Fashion-MNIST is higher than 95%/80%, indicating that our attack framework is fair to each class. As for the black-box attack, it can be seen from Fig. 7 that the UASRs for MNIST and Fashion-MNIST are 89.9% and 81.8%, respectively. Moreover, the UASRs with robustness enhancement are approximately 15% higher than those without robustness enhancement. The results demonstrate not only the effectiveness of our robustness enhancement strategy, but also the outstanding transferability of the generated perturbations. What's more, with less query samples, we can achieve the attack success rate comparable to conventional black-box attack approach [25]. Furthermore, we find that the UASRs and TASRs in the office are very similar to those in the laboratory (the difference for each metric is within 3%). Thus, the proposed attack framework can also perform well in different environments. In the following, without loss of generality, we use the CSI collected in the laboratory to conduct experiments.

B. Effect of Hyper-parameter

We also evaluate the effects of the hyper-parameters in the attack framework on the attack effectiveness, including the SNR (the ratio of signal to AWGN), the power ratio of the perturbation signal, and the compression rate.

1) Effect of AWGN's SNR: To explore the attack effectiveness under different strengths of AWGN, we vary the SNR of AWGN from -10dB to 10dB in the step of 2dB. With the lowest SNR (-10dB), the recognition accuracy of normal samples for MNIST/Fashion-MNIST is 89.9%/75.8%. The UASRs of white-box attacks against MNIST and Fashion-MNIST datasets are shown in Fig. 8. It can be observed that the UASR of MNIST increases with the raising of the SNR. Fashion-MNIST does not show obvious UASR change

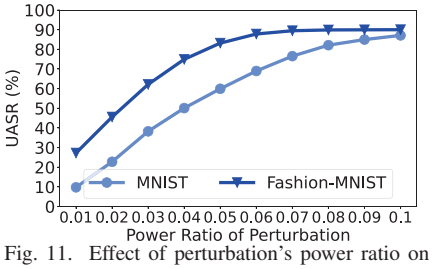


Fig. 11. Effect of perturbation's power ratio on white-box UASR.

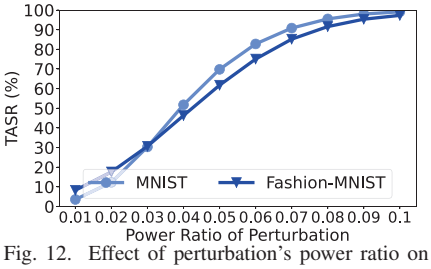


Fig. 12. Effect of perturbation's power ratio on white-box TASR.

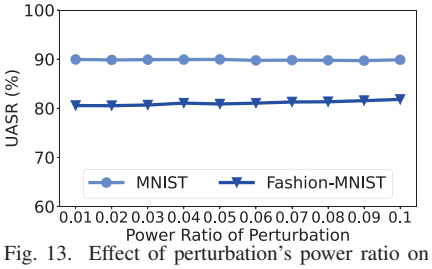


Fig. 13. Effect of perturbation's power ratio on black-box UASR.

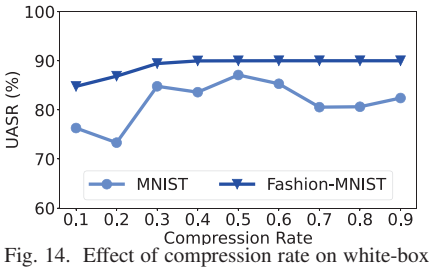


Fig. 14. Effect of compression rate on white-box UASR.

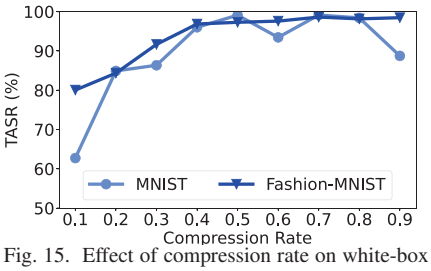


Fig. 15. Effect of compression rate on white-box TASR.

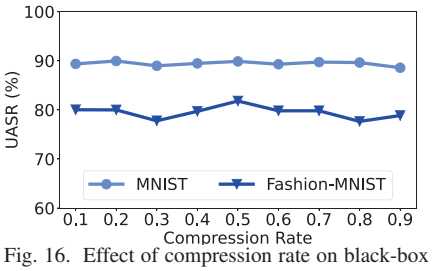


Fig. 16. Effect of compression rate on black-box UASR.

when the SNR increases. When the SNR increases to 0dB, the UASR of MNIST tends to converge. In addition, even with low SNR (-10dB), the UASR is higher than 82%. This proves that the attack framework is still effective under harsh conditions with low SNR. Fig. 9 presents the white-box TASRs under different SNRs. It can be found that the TASRs of both the two datasets increase as the SNR. Similar to the white-box UASRs, the TASRs become stable when the SNR increases to 0dB. Meanwhile, the TASRs of both the datasets are still larger than 89% when the SNR is only -10dB. We also show the variation trend of black-box UASR with SNR in Fig. 10. It can be seen that the UASRs of both the two datasets increase noticeably with the SNR. When the SNR increases to 0dB, the UASRs maintain stable. These results demonstrate that the proposed attack framework is robust against low SNR.

2) Effect of perturbation's power ratio: To observe the variation trend of the attack success rates with the power ratio of the perturbation, we vary r in Eq. 8 and ϵ in Eq. 14 from 0.01 to 0.1 in step of 0.01, respectively. The corresponding white-box UASRs are shown in Fig. 11. It can be observed that, with the increase of the power ratio, the UASRs of the two datasets keep increasing. When the power ratio is only 0.05, the UASR of MNIST/Fashion-MNIST already reaches 59.9%/83.2%. Thus, the untargeted white-box perturbation is still effective when its power is low. The crafted perturbation can hide itself well while maintaining good attack effectiveness. The white-box TASR varied with perturbation's power ratio is shown in Fig. 12. It can be found that, similar to the white-box UASRs, the white-box TASRs also increase obviously with the power ratio. When the power ratio increases to 0.05, the TASR of MNIST/Fashion-MNIST becomes larger than 60%. Hence, the attacker can achieve good attack effectiveness while guaranteeing high stealthiness. From Fig. 13, we are surprised to find that even when the power ratio is only 0.01, the attacker can achieve over 89%/80% UASRs. In this case, it is significantly difficult for the communication system to detect the black-box attack, yet the attack is very effective. Therefore, the proposed attack framework is highly threatening.

3) Effect of compression rate: In practice, the user may want to compress the data at different rates. To investigate if the compression rate would impact the attack effectiveness, we vary the compression rate from 0.1 to 0.9 in step of 0.1, and recalculate the attack success rate. The white-box UASRs and TASRs are shown in Fig. 14 and Fig. 15, respectively. It can be observed that, for MNIST dataset, both the UASR and TASR first increase as the compression rate. When the compression rate reaches 0.5, the UASR/TASR approaches the maximum. With the continuous increasement of the compression rate, the UASR/TASR starts to decrease. This "increase-stable-decrease" phenomenon is reasonable. When the compression rate is lower than 0.4, the number of elements in the semantic information is small. So, the semantic information the attacker can perturb is also limited, resulting in relatively low UASR/TASR. When the compression rate is higher than 0.6, the number of elements in the semantic information becomes large. In the pragmatic model, the semantic information, rather than the perturbation, dominates the image recognition. Therefore, the UASR/TASR also decreases continuously. Fig. 16 displays the black-box UASRs. It can be found that the variation trends of black-box UASRs are similar to those of white-box UASRs and TASRs. The difference lies in that the black-box UASRs are more stable. It seems that the user can improve the security of the communication system by keeping the compression rate very small or large. However, a compression rate as small as 0.1 would render low image recognition accuracy (67.2%/62.7% for MNIST/Fashion-MNIST) while a very large one would limit the data transmission efficiency. To guarantee better transmission, users prefer a compression rate close to 0.5. This gives attackers more opportunities to mount effective attacks. Hence, the proposed attack framework is significantly threatening in practice.

TABLE II
TRANSFERABILITY STUDY. 'M.' AND 'F.M.' MEAN MNIST AND FASHION-MNIST, RESPECTIVELY.

Architecture	CNN	RNN	LSTM
Accuracy: M./F.M. (%)	92.9/84.0	93.2/83.3	93.0/82.2
UASR: M./F.M. (%)	89.5/73.9	55.1/57.9	52.8/55.6

C. Transferability Study

In the default setting, the architectures of all the DNNs in the semantic communication system conform to the original paper, i.e., NN. In reality, users may adopt different network architectures for communication. To understand the transferability of the perturbation crafted by our attack framework, we replace the high-level architecture of the pragmatic model[3] with different mainstream ones, including CNN, RNN, and LSTM. The corresponding accuracies of normal samples and black-box UASRs are shown in Tab. II. It can be seen that, with an NN-based surrogate model, the attacker can achieve high attack success rates on CNN, RNN, and LSTM for both MNIST and Fashion-MNIST. In real-world arrack, even if the attacker has no knowledge about the target DNNs in the communication system, s/he can still launch effective attacks. Therefore, the proposed attack framework poses real threats to semantic communication systems, which should be taken seriously.

VI. MITIGATION

Geofencing perturbation signal. Geofencing prevents perturbation signals from contaminating the wireless channel between the transmitter and the receiver of the semantic communication system. Users can achieve geofencing by building walls with metal or painting walls with electromagnetic shielding paints. This can fundamentally avoid the generation of adversarial samples, yet, also introduce inconvenience. On the one hand, geofencing may stop the propagation of legal signals, impairing normal communication. On the other hand, geofencing is only suitable for the scenarios where the communication system facilities are fixed, such as residences and factories. For dynamic scenarios like autonomous vehicles, users may need special designs to enclose the communication channel.

Adversarial sample detection. This mitigation aims at identifying adversarial samples before performing pragmatic tasks. Users can train a one-class support vector machine (SVM) to distinguish adversarial samples from normal ones at the receiver end. To validate the feasibility of this mitigation, we train a one-class SVM with normal images recovered by the decoder and calculate the accuracy of adversarial and normal sample recognition. The results indicate that, with no misclassification on the normal samples, the recognition accuracy on untargeted white-box, targeted white-box and untargeted black-box adversarial samples is 73.5%, 74.6%, and 73.8%, respectively. Over 25% adversarial samples are misclassified as normal ones, because we have delicate designs on the stealthiness of the perturbation. Thus, it is hard to fully resist adversarial attacks.

[3]We mainly consider the pragmatic model as it has far more parameters than encoder and decoder.

VII. RELATED WORK

Although advanced communication technologies such as 5G [34] have been adopted, current communication methods are difficult to support the high transmission rate requirements of emerging applications like remote surgery and complex manufacturing [35]. In order to break through the data rate bottleneck of traditional communication methods, semantic communication abandons the old-fashioned symbol transmission and adopts semantic information conveyance instead [3]. Empowered by AI technologies, the transmitters in recent mainstream semantic communications compress the original data by extracting semantic information with DNNs. The receiver only needs to understand the meaning conveyed by the transmitter. In view of the special working mechanism, semantic communication can achieve extremely low transmission latency on specific tasks. such as image [36]–[38], text [20], and speech transmission [21], [39], [40]. However, most existing works on semantic communication are dedicated to improving the data rate and robustness. Few efforts have focused on its security aspects. Hu et al. [41] confirm the feasibility of adding perturbation to the original input (i.e., image) of the encoder to mislead the semantic communication system untargetedly. This is natural as crafting image-level perturbation has been proven effective in other applications [10]. In fact, it is non-trivial to add perturbation to the original input of the encoder in reality. By contrast, the open wireless channel is the most possible attack vector that can be manipulated. Sagduyu et al. [42] explore the feasibility of adversarial attacks compromising both the original input and the wireless channel. But their approach is only applicable to while-box scenario. In this paper, we perform the in-depth analysis on the security of semantic communication. We, for the first time, achieve both both white-box and black-box attacks by polluting the semantic information in the wireless channel with well-crafted perturbation signals.

VIII. CONCLUSION

This work is the first to explore the security of semantic communication against both white-box and black-box attacks in the wireless channel. A systematic and universal framework is proposed to craft effective untargeted and targeted perturbations. With the delicate paradigms in the attack framework, the generated adversarial samples bear the content-agnostic, well-hidden, robust, and highly-transferable features. Experiments over two open-source datasets demonstrate that the proposed attack framework can achieve 87%+ white-box UASR, 99%+ white-box TASR, as well as 89%+ black-box UASR.

ACKNOWLEDGEMENTS

This paper is supported by the National Natural Science Foundation of China under grant U21A20462 and 62372400, "Pioneer" and "Leading Goose" R&D Program of Zhejiang under grant No. 2023C01033.

REFERENCES

[1] X. Yuan, M. Wu, Z. Wang, Y. Zhu, M. Ma, J. Guo, Z. Zhang, and W. Zhu, "Understanding 5g performance for real-world services: a content provider's perspective," in *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, 2022, pp. 101–113.

[2] G. Ancans, V. Bobrovs, A. Ancans, and D. Kalibatiene, "Spectrum considerations for 5g mobile communication systems," *Procedia Computer Science*, vol. 104, no. C, pp. 509–516, 2017.

[3] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.

[4] S. Iyer, R. Khanai, D. Torse, R. J. Pandya, K. M. Rabie, K. Pai, W. U. Khan, and Z. M. Fadlullah, "A survey on semantic communications for intelligent wireless networks," *Wireless Personal Communications*, vol. 129, no. 1, pp. 569–611, 2023.

[5] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. S. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, 2022.

[6] C. Xiao, R. Deng, B. Li, F. Yu, M. Liu, and D. Song, "Characterizing adversarial examples based on spatial consistency information for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 220–237.

[7] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019.

[8] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582.

[9] A. Bahramali, M. Nasr, A. Houmansadr, D. Goeckel, and D. Towsley, "Robust adversarial attacks against dnn-based wireless communication systems," in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2021.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[12] S. Niranjayan and N. C. Beaulieu, "Analysis of wireless communication systems in the presence of non-gaussian impulsive noise and gaussian noise," in *Proceedings of the IEEE Wireless Communication and Networking Conference*, 2010, pp. 1–6.

[13] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, "Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 41, no. 1, pp. 170–185, 2023.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[16] A. Hassan, A. Narayanan, A. Zhang, W. Ye, R. Zhu, S. Jin, J. Carpenter, Z. M. Mao, F. Qian, and Z. Zhang, "Vivisecting mobility management in 5g cellular networks," in *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, 2022, pp. 86–100.

[17] S. Jog, J. Guan, S. Madani, R. Lu, S. Gong, D. Vasisht, and H. Hassanieh, "Enabling iot self-localization using ambient 5g signals," in *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2022, pp. 1011–1026.

[18] Y. Yang, C. Guo, F. Liu, C. Liu, L. Sun, Q. Sun, and J. Chen, "Semantic communications with AI tasks," *CoRR*, vol. abs/2109.14170, 2021.

[19] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 553–557, 2022.

[20] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1394–1398, 2022.

[21] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech recognition," in *Proceedings of the IEEE Global Communications Conference, (GLOBECOM)*, 2021, pp. 1–6.

[22] S. Baroudi, "Evaluation of outage probability in presence of interference and noise with application to dual-hop wireless systems," https://spectrum.library.concordia.ca/id/eprint/981528/, 2016.

[23] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[24] X. Y. Han, V. Papyan, and D. L. Donoho, "Neural collapse under MSE loss: Proximity to and dynamics on the central path," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[25] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*, 2017, pp. 506–519.

[26] Y. Chen, X. Yuan, J. Zhang, Y. Zhao, S. Zhang, K. Chen, and X. Wang, "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *Proceedings of the USENIX Security Symposium*, 2020, pp. 2667–2684.

[27] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies – a comprehensive introduction," *Natural Computing*, vol. 1, 2002.

[28] M. Saeed, "An introduction to recurrent neural networks and the math that powers them," https://machinelearningmastery.com/an-introduction-to-recurrent-neural-networks-and-the-math-that-powers-them/, 2022.

[29] Geeksforgeeks, "Understanding of lstm networks," https://www.geeksforgeeks.org/understanding-of-lstm-networks/, 2023.

[30] R. Baraniuk, "Discrete time impulse function," https://eng.libretexts.org/Bookshelves/Electrical_Engineering/Signal_Processing_and_Modeling/Signals_and_Systems_(Baraniuk_et_al.)/01%3A_Introduction_to_Signals/1.07%3A_Discrete_Time_Impulse_Function.

[31] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

[32] J. Wang and J. R. Jang, "Training a singing transcription model using connectionist temporal classification loss and cross-entropy loss," *IEEE ACM Transactions on Audio Speech and Language Processing*, vol. 31, pp. 383–396, 2023.

[33] Z. Jiang, T. H. Luan, X. Ren, D. Lv, H. Hao, J. Wang, K. Zhao, W. Xi, Y. Xu, and R. Li, "Eliminating the barriers: Demystifying wi-fi baseband design and introducing the picoscenes wi-fi sensing platform," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4476–4496, 2021.

[34] D. Xu, A. Zhou, G. Wang, H. Zhang, X. Li, J. Pei, and H. Ma, "Tutti: coupling 5g RAN and mobile edge computing for latency-critical video analytics," in *Proceedings of the ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2022, pp. 729–742.

[35] D. Wheeler and B. Natarajan, "Engineering semantic communication: A survey," *IEEE Access*, vol. 11, pp. 13 965–13 995, 2023.

[36] M. B. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, "Wireless end-to-end image transmission system using semantic communications," *IEEE Access*, vol. 11, pp. 37 149–37 163, 2023.

[37] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 41, no. 1, pp. 55–71, 2023.

[38] W. Tong, F. Liu, Z. Sun, Y. Yang, and C. Guo, "Image semantic communications: An extended rate-distortion theory based scheme," in *Proceedings of the IEEE Globecom Workshops*, 2022, pp. 1723–1728.

[39] S. Yao, Z. Xiao, S. Wang, J. Dai, K. Niu, and P. Zhang, "Variational speech waveform compression to catalyze semantic communications," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.

[40] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *Proceedings of the IEEE International Conference on Communications (ICC)*, 2021, pp. 1–6.

[41] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, "Robust semantic communications against semantic noise," in *Proceedings of the IEEE Vehicular Technology Conference, (VTC)*, 2022, pp. 1–6.

[42] Y. E. Sagduyu, T. Erpek, S. Ulukus, and A. Yener, "Is semantic communication secure? A tale of multi-domain adversarial attacks," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 50–55, 2023.