Boosting Physical Layer Black-Box Attacks with Semantic Adversaries in Semantic Communications

Zeju Li, Xinghan Liu, Guoshun Nan*, Jinfei Zhou, Xinchen Lyu, Qimei Cui, Xiaofeng Tao
Beijing University of Posts and Telecommunications
lizeju@bupt.edu.cn, liuxinghan_2022@bupt.edu.cn, nanguo2021@bupt.edu.cn, zhouchuluo@bupt.edu.cn,
lvxinchen@bupt.edu.cn, cuiqimei@bupt.edu.cn, taoxf@bupt.edu.cn

***Abstract*—End-to-end semantic communication (ESC) system is able to improve communication efficiency by only transmitting the semantics of the input rather than raw bits. Although promising, ESC has also been shown susceptible to the crafted physical layer adversarial perturbations due to the openness of wireless channels and the sensitivity of neural models. Previous works focus more on the physical layer white-box attacks, while the challenging black-box ones, as more practical adversaries in real-world cases, are still largely under-explored. To this end, we present SemBLK, a novel method that can learn to generate destructive physical layer semantic attacks for an ESC system under the black-box setting, where the adversaries are imperceptible to humans. Specifically, 1) we first introduce a surrogate semantic encoder and train its parameters by exploring a limited number of queries to an existing ESC system. 2) Equipped with such a surrogate encoder, we then propose a novel semantic perturbation generation method to learn to boost the physical layer attacks with semantic adversaries. Experiments on two public datasets show the effectiveness of our proposed SemBLK in attacking the ESC system under the black-box setting. Finally, we provide case studies to visually justify the superiority of our physical layer semantic perturbations.**

***Index Terms*—Semantic communications, Physical Layer Attacks, Black-box Attacks, Generative adversarial networks**

I. Introduction

The rapid development of mobile applications puts unprecedented pressure on wireless networks, leading to severe traffic congestion and intolerable delays [1]. Recently proliferated AI-enabled services, such as IoT (Internet of Things), self-driving and VR/AR, will further exacerbate the above traffic burden in the future due to the predicted data explosion. Meanwhile, existing wireless communications systems, which are built on the Shannon information theory, primarily focused on how to accurately and effectively transmit raw symbols from the transmitter to the receiver. Along this direction, quantifying the maximum transmission data rate that can be supported by a communication channel gradually pushes the system capacity to the Shannon limit. Such a dilemma motivates us to rethink the design of the wireless communication paradigm beyond the Shannon approach.

Recent advances in powerful deep learning technologies open up opportunities for compressing the input with neural models and then transmitting the compressed representations of data over wireless channels. Such a way of paradigm is known as end-to-end semantic communication (ESC) [2], which can be considered as the breakthrough beyond the Shannon approaches. Specifically, semantic communications aim at the successful transmission of semantic information conveyed by the input rather than the accurate reception of every single symbol or bit. Recently works have discussed the promise of ESC systems for the transmission of text [2], speech [3], images [4] and videos [5], showing great potentials for the high demand applications in future 6G networks.

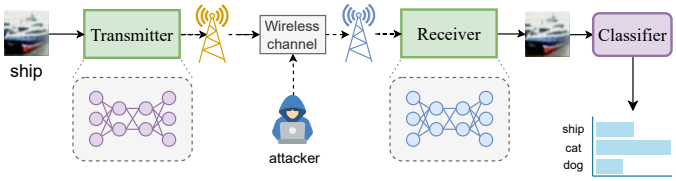

Fig. 1. An end-to-end semantic communication system. An attacker generates the physical layer adversarial perturbations and then fools the classifier at the receiver side to make an incorrect semantic decision, i.e., interpreting the "ship" type as the "cat" type.

Although promising [6], [7], ESC has also been shown susceptible to the crafted physical layer adversarial perturbations due to the openness of wireless channels [8] and the sensitivity of neural models [9]. We give an example to demonstrate the vulnerability of an ESC system as follows. Fig. 1 illustrates an existing ESC system, which mainly consists of three deep learning-based modules, including a transmitter, a receiver and a classifier. The transmitter first extracts the semantics of the input image "ship", and then sends the modulated symbols to the wireless channels. The receiver reconstructs the image based on the semantics and then the classifier interprets the semantics as the "ship" category. An attacker can send adversarial perturbations to the wireless channel and then may mislead the classifier to make an incorrect decision. As shown in Fig. 1, the classifier incorrectly interprets the "ship" as "cat", as the confidence score of the "cat" category is the largest among the three candidates.

Previous adversarial attacks in the field of machine learning can be mainly divided into white-box attacks and black-box ones [10], [11]. The former attacks have access to the model's parameters, such as FGSM [10] and PGD [12]. Conversely, for the latter, we only know the model inputs and we can query to obtain output labels or confidence scores.

Existing efforts for semantic communications focus more on the white-box attacks, while the challenging physical layer

*Guoshun Nan is the corresponding author

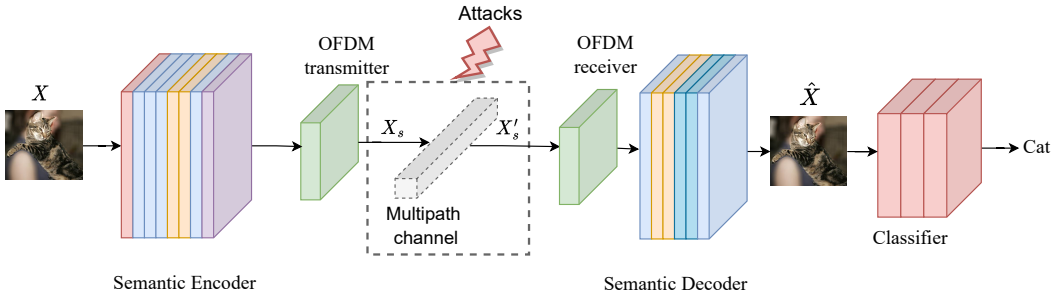

Fig. 2. Overview of the end-to-end semantic communication system used in our paper. The system consists of multiple modules including a semantic encoder, an OFDM transmitter, an OFDM receiver, a semantic decoder, and a classifier. An attacker is able to generate malicious perturbation from the wireless channel.

black-box ones [13], as more practical adversaries in real-world cases, are still largely under-explored. Existing black-box methods in the field of machine learning may not be directly applied to an ESC system due to two reasons: 1) These methods mainly focus on generating black-box adversarial perturbations based on a raw input of a neural model. While in practice, an attacker is unable to access the input data in a communication system and can receive the wireless signal and potentially decode the transmissions. 2) Existing attacks, such as FGSM and PGD, are mainly indiscriminate to all information as physical layer attacks. While the semantics conveyed underlying the data, the central of semantic communications, are largely ignored during the adversarial learning procedure.

To fill the above gap, this paper proposes SemBLK, a novel method that can learn to generate destructive physical layer semantic attacks for an ESC system under the black-box setting. The two key ingredients of our SemBLK are a surrogate semantic encoder and a novel semantic perturbation generator. Our surrogate encoder learns to output the semantic representations via a limited number of queries to an oracle, and then the perturbation generator can boost the physical layer attacks by learning to craft the semantic adversaries that aim to fool the classifier to make an incorrect interpretation. We conduct experiments to verify the effectiveness of our method. We summarize our contributions as follows:

- We present SemBLK, a practical physical layer black-box attack method for deep learning-based semantic communication systems, where the adversaries are imperceptible to humans.
- We introduce a surrogate semantic encoder to mimic the semantic encoder of the existing ESC system and train its parameters by exploring a limited number of queries to the system, augmenting the training instances with GAN methods.
- Equipped with our surrogate encoder, we then propose a novel semantic perturbation generation method to learn to boost the physical layer attacks with semantic adversaries.
- We conduct extensive experiments on two public benchmarks to show the superiority of our black-box attacks.

II. End-to-end Semantic Communication System

Fig.2 shows the architecture of an end-to-end semantic communication (ESC) system. We refer to previous JSCC-OFDM [4] as the backbone and additionally introduce a classifier to interpret semantics at the receiver side. We outline each component of the ESC system as follows.

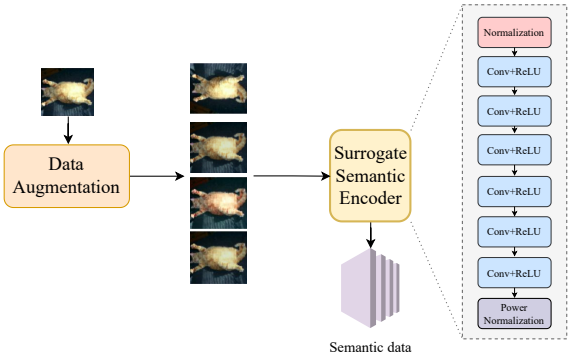

Fig. 3. The architecture of our surrogate encoder and the data augmentation procedure to train the encoder.

A. Semantic Transmitter

Semantic Encoder: The semantic encoder uses convolutional and residual layers to extract image features. The convolutional layers extract the input features and obtain more detailed semantic feature information. The residual layers prevent performance degradation when the network depth is deepening and reduce the amount of computation in subsequent layers.

OFDM Transmitter: The OFDM transmitter performs an inverse fast Fourier transform (IFFT) on the data, inserts a cyclic prefix (CP), which protects the signal from interference, and then shears the signal to reduce the peak-to-average power ratio (PAPR).

B. Wireless Channel

Without loss of generality, we use Rayleigh fading channel in our paper. We parameterize the channel using a discrete channel transfer function:

$$X_s = h(y; \sigma_0^2, ..., \sigma_{L-1}^2, \sigma^2) = h * X_s + w \tag{1}$$

where $*$ denotes the convolution operation, h denotes the sample sample space channel impulse response, and L is the number of multipath. w represents the additive Gaussian noise. Each path experiences independent Rayleigh fading satisfying

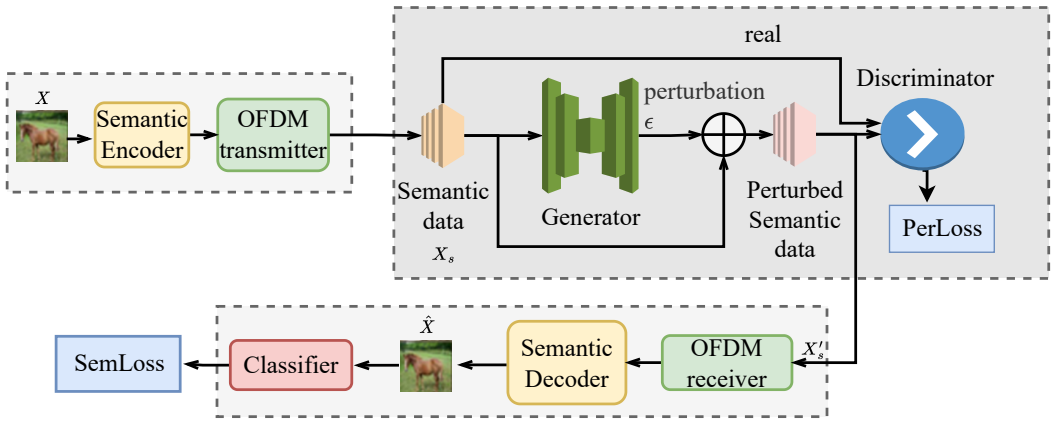

Fig. 4. Overview of our perturbation generator. It consists of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. The generator $\mathcal{G}$ crafts the adversarial perturbation ϵ. The goal of $\mathcal{D}$ is to distinguish between real semantic symbols X_s and perturbed data $X_s + \epsilon$.

$h_l \sim \mathcal{CN}(0, \sigma_l^2)$ for $l = 0, 1, ..., L-1$. The power of each path follows $\sigma_l^2 = \alpha_l e^{-\frac{1}{\gamma}}$, where α_l is a normalization coefficient to satisfy $\sum_{l=0}^{L-1} \sigma_l^2 = 1$. γ is the time decay constant.

C. Semantic Receiver

OFDM Receiver: The OFDM receiver removes the CP from the received data, performs FFT, channel estimation and channel equalization to obtain the semantic information.

Semantic Decoder: The semantic decoder aims to reconstruct the input image from the received semantic data, and it shares the same architecture as the semantic encoder.

Classifier: We feed the reconstructed image to a classifier to interpret semantics. Here we use Mobilenetv2 [14] as the classifier.

III. Our Method

We train the ESC system and then treat it as an oracle for the physical layer black-box attacks. Existing black-box attacks require a large number of queries to obtain the confidence scores or labels, while these attacks can be easily detected. Therefore, we introduce a surrogate encoder [11] to mimic the semantic encoder and then train the surrogate with a limited number of queries. Equipped with the surrogate, we also present a perturbation generator that is able to learn to craft the physical layer adversaries for the ESC system. We first delve into the surrogate encoder.

A. Surrogate Semantic Encoder

We simply use the fully-connected convolutional neural networks (CNN) as our surrogate encoder, since CNN has been proved effectively for image feature extraction. For the existing ESC system. We rely on the surrogate encoder to generate a limited number of queries to obtain the output labels. To train our surrogate encoder with and reduce the number of queries, we augment the input data with a simple yet effective GANs [15] method. Such a method can not only enrich the training instance but also align the semantic to the original data. Fig.3 demonstrates the architecture of our surrogate encoder and the augmentation method. Next, we show how we learn to craft the semantic perturbations equipped with such a surrogate encoder.

B. Adversarial Perturbation Generator

The semantic representations generated by our surrogate encoder will be modulated by the OFDM transmitter and then sent to wireless channels. We denote the OFDM symbols in the channel as X_s. Considering that the existing adversarial attack methods cannot be directly applied to our semantic communication system, hence we consider semantic adversaries from the open wireless channel under the black-box setting.

Fig.4 shows the architecture of our proposed SemBLK. Our SemBLK consists of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. The generator $\mathcal{G}$ crafts the adversarial perturbation ϵ, which will be added to the semantic symbols X_s. The goal of $\mathcal{D}$ is to distinguish between real semantic symbols X_s and perturbed data $X_s+\epsilon$. While the generator $\mathcal{G}$ is able to produce adversarial perturbations to distort the semantic information. Next, we show how the perturbation generator learns to craft the adversaries.

C. Loss Function

Equipped with the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$, we can give the perturbation loss as follows.

$$L_{Per} = \mathbb{E}_{X_s} \log \mathcal{D}(X_s) + \mathbb{E}_{X_s} \log \left(1 - \mathcal{D}(\epsilon + \mathcal{G}(X_s))\right) \quad (2)$$

To guide the training of the generator with the feedback of incorrect semantic interpretations, we also introduce a semantic loss L_{Sem}, which can be expressed as follows.

$$L_{Sem} = \mathbb{E}_{X_s} \mathcal{L}(X_s) \quad (3)$$

where $\mathcal{L}$ is the cross-entropy loss for classifications. Then we are able to train the perturbation generator by predicting the type of perturbed image as an incorrect one, which can be considered a multi-task learning procedure. The final loss L can be formulated as follows.

$$L = L_{Sem} + wL_{Per} \quad (4)$$

where w is a hyper-parameter to indicate the importance of the two losses. The parameters can be optimized by solving the minimax-game as follows.

$$arg \min_{\mathcal{G}} \max_{\mathcal{D}} L \tag{5}$$

Considering the constraints on the number of queries for the training, we also use the data generated by our augmentation method to train the surrogate encoder, as shown in Fig.3. Finally, we can directly apply the generator to the existing semantic communication system to craft black-box adversarial attacks.

D. Imperceptibility of Our Black-box Attacks

To be practical in real-world cases, we also consider the imperceptibility of the black-box attacks at the receiver side and meanwhile introduce a threshold to secure the constructed image quality measured by structural similarity index measure (SSIM) [16]. Let X and X' denote an input image and the reconstructed one, the above threshold $SSIM(X, \hat{X})$ can be formulated as follows.

$$SSIM(X, \hat{X}) = [l(X, \hat{X})]^{\alpha} + [c(X, \hat{X})]^{\beta} + [s(X, \hat{X})]^{\gamma} \tag{6}$$

$$l(X, \hat{X}) = \frac{2\mu_X \mu_{\hat{X}} + c_1}{\mu_X^2 + \mu_{\hat{X}}^2 + c_1} \tag{7}$$

$$c(X, \hat{X}) = \frac{2\sigma_{X\hat{X}} + c_2}{\sigma_X^2 + \sigma_{\hat{X}}^2 + c_2} \tag{8}$$

$$s(X, \hat{X}) = \frac{\sigma_{X\hat{X}} + c_3}{\sigma_X \sigma_{\hat{X}} + c_3} \tag{9}$$

where $l(X, \hat{X})$, $c(X, \hat{X})$ and $s(X, \hat{X})$ are the brightness comparison, contrast comparison, and structure comparison, respectively. μ_X and $\mu_{\hat{X}}$ represent the mean value of X and $\hat{X}$ respectively, σ_X and $\sigma_{\hat{X}}$ represent the standard deviation of X and $\hat{X}$ respectively. $\sigma_{X\hat{X}}$ represents the covariance of X and $\hat{X}$. And c_1, c_2, c_3 are constants to avoid errors caused by denominator 0. To be more practical for the semantic communication system, we can simplify the threshold $SSIM(X, \hat{X})$ as follows:

$$SSIM(X, \hat{X}) = \frac{\left(2\mu_X \mu_{\hat{X}} + c_1\right)\left(\sigma_{X\hat{X}} + c_2\right)}{\left(\mu_X^2 + \mu_{\hat{X}}^2 + c_1\right)\left(\sigma_X^2 + \sigma_{\hat{X}}^2 + c_2\right)} \tag{10}$$

IV. Experiments

We conduct experiments on CIFAR10, a popular dataset that consists of 60,000 32×32-pixel images in 10 classes, with 6,000 images per class. We follow the previous work [4] to use 50,000 images as training instances and the rest 10,000 ones as test examples. Next, we detail evaluation metrics, baselines, experimental settings, the effect of surrogate model, and give some insightful conclusions based on our observations.

A. Evaluation Metrics

We employ multiple metrics to measure the performance of our method, including Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Peak to Average Power Ratio (PAPR) and accuracy of classification (ACC). We outline these metrics as follows.

- **PAPR** is to measure the transmitting performance of OFDM system, if there is a large PAPR, which requires a high linear dynamic range of the transmitting power amplifier of OFDM system. If the dynamic range of the amplifier cannot meet the requirements, it will cause nonlinear distortion of the output signal, and such a communication system will have no practical significance.
- **PSNR** is an objective measure of picture quality, representing the effect of background noise on image quality, the higher the PSNR, the less the picture is affected by noise.
- **SSIM** is used to represent the similarity of the original image and the received image, compared to PSNR is more in line with the intuitive feeling of the human eye on the image.
- **ACC** is the accuracy of the classification result of the classifier after semantic decoder, which indicates the effectiveness of the guaranteed upper layer service.

B. Baselines

We introduce three existing attacks, including FGSM [10], PGD [12] and ATN [9], as baselines to compare with our approach. The descriptions for these attacks are given as follows.

- **FGSM** adds perturbation to the image along the gradient direction, which increases the loss function and leads the model to get wrong classification results.
- **PGD** takes FGSM a step further by using multiple iterations to add perturbations along the gradient direction on the image.
- **ATN** trains a deep neural network that converts the original image into an adversarial sample.

C. Experimental Settings

We use Pytorch to implement the neural network blocks and the OFDM communication model. We utilize Adam [17] to train our surrogate encoder, generator $\mathcal{G}$ and discriminator $\mathcal{D}$ in SemBLK. Moreover, we set hyper-parameter w as 0.1. Attackers don't have knowledge of the semantic communication systems under the black-box setting. We split the training set into two parts. 40,000 images are used to train the targeting ESC system and the rest of the 10,000 ones are used to train the surrogate encoder. We also use 1,000 images in the overall test set of 10,000 images for the evaluation of our surrogate encoder. We use Rayleigh fading channel and set SNR as 10 in the simulation. The initial learning rates of all networks are set as 0.0005, and these rates gradually decreased to zero as the number of iterations increases. Our model is trained by Cuda of NVIDIA GeForce RTX3090 GPU.

D. Effectiveness of Surrogate Model

TABLE I
EFFECTIVENESS OF OUR SURROGATE MODEL

	PSNR	SSIM	PAPR	ACC
Original ESC	25.23	0.85	11.58	0.82
ESC with surrogate	24.27	0.83	11.91	0.79
ESC with surrogate+RandAugment	22.61	0.82	11.39	0.78
ESC with surrogate+GANs	**23.47**	**0.81**	**12.32**	**0.80**

The first two rows of Table I show that the ESC system equipped with our surrogate model is able to achieve comparable results to the oracle under the same test set in terms of PSNR, SSIM, PAPR and ACC. These results confirm the effectiveness of our substitute model. As there will be very few instances to train the surrogate encoder in a real-world scenario, we build a small sub-dataset from 10,000 training instances of the surrogate encoder and augment these instances with existing approaches [18] and [15]. The sub-dataset includes 1,000 images, with 100 images per category. We finally generated 9,000 more images, with 1,000 images in each category after data augmentation.

The last two rows of Table I report the comparisons of two augmentation methods on the sub-dataset, i.e, the GAN approach [15] and RandomAugment approach [18]. We observe that the performance of our surrogate encoder trained by the GAN approach outperforms the ones by the RandomAugment approach. We also find that our surrogate encoder is able to achieve comparable results by exploring only 1,000 original training instances. Hence, equipped with such an augmentation method, we can significantly reduce the number of queries to an existing ESC system, which is practical in real-world cases.

E. The results of attacks

As shown in Table II, we use these methods to make adversarial samples to attack the sender's source input data and compare the experimental results. We can find that PGD has the best attack effect, but the attack effect of our proposed SemBLK is close to the effect of FGSM, which shows that SemBLK can also degrade the performance of our semantic communication system. Adversarial attacks on the source input images verify the effectiveness and feasibility of our method. However, adversarial attacks on the input images are an ideal situation. While in practice, it is difficult for an attacker to obtain and control the sender's source input data. Hence, we consider adding perturbations to semantic information at the physical channel layer, so as to conduct efficient and imperceptible black-box attacks on semantic communication systems.

We use FGSM, PGD and ATN to attack the physical channel layer and record the results. The four methods add perturbations to the semantic information and feed the perturbed semantic data as input to the OFDM receiver and the semantic decoder. In table III, we can observe that our proposed SemBLK attacks significantly outperform the previous methods

TABLE II
THE ATTACK EFFECT OF GENERATING ADVERSARIAL EXAMPLES BY FGSM, PGD, ATN, SEMBLK

	PSNR	SSIM	PAPR	ACC
No Attack	25.23	0.85	11.58	0.82
FGSM	23.28	0.78	11.28	0.14
PGD	24.10	0.82	11.57	0.06
ATN	23.45	0.78	11.69	0.28
SemBLK	23.22	0.79	11.54	0.09

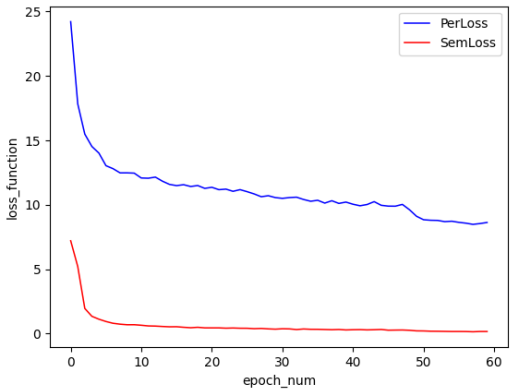

Fig. 5. The training curves for PerLoss and SemLoss. We can observe that as the number of iterations increases, SemLoss and PerLoss can gradually decrease and the hence the generator can be optimized step by step.

FGSM and PGD. For example, the semantic interpretation accuracy of the ESC system under four attacks FGSM, PGD, ATN and SemBLK are 0.66, 0.75, 0.72 and 0.49, respectively. However, PSNR, PAPR and SSIM under our SemBLK attacks slightly dropped. These results indicate that our black-box attacks are destructive yet imperceptive.

TABLE III
THE ATTACK EFFECT OF GENERATING PERTURBATIONS AND OVERLAYING ON THE SEMANTIC DATA IN PHYSICAL CHANNELS BY FGSM, PGD, ATN, SEMBLK

	PSNR	SSIM	PAPR	ACC
No Attack	25.23	0.85	11.58	0.82
FGSM	21.06	0.70	11.69	0.66
PGD	22.08	0.77	11.31	0.75
ATN	23.14	0.78	11.37	0.72
SemBLK	**22.91**	**0.73**	**12.11**	**0.49**

The experimental results show that the effect of using a surrogate encoder demonstrates a high PAPR, which may cause the transmitter to transmit less effectively. This is a drawback associated with the attack through the surrogate model, but it can substantially reduce the correct classification rate of the upper layer services with less impact on the images.

Fig. 6. Comparisons of the input images and reconstructed images of in semantic communication systems under our SemBLK attacks. The comparisons show that our black-box attacks are imperceptive to humans.

Although our attack causes a decrease in PSNR, SemBLK has no major difference in this evaluation metric compared to several other attacks; the SSIM does not change much compared to the preattack, which is more indicative of the stealthiness of the attack.

F. Case study

Fig.6 shows 50 cases selected from CIFAR10 to visually demonstrate that the adversaries generated by SemBLK are imperceptive to our humans, while these black-box attacks are able to significantly degrade the semantic interpretations at the receiver side. For each pair of images in Fig. 6, the left one is an image constructed by the receiver of the ESC system without our attack, and the right one refers to the image transmitted under the physical layer semantic perturbations. We observe that the two images visually appear quite similar to each other with our human eyes, while the accuracy of semantic interpretation of right images, central of the semantic communications, is less than 31 percent on average, compared with the accuracy of the left images without attack. These cases further justify the superiority of our black-box semantic adversaries at the beginning.

V. Conclusion

This paper presents SemBLK, a novel method that aims to generate physical layer black-box adversarial attacks for end-to-end semantic communication systems. We first use a limited number of queries, as well as data augmentation methods to train a substitute semantic encoder, and then generate adversarial perturbations that are able to mislead semantic interpretation at the receiver. In experiments, we observe that the black-box attacks generated by our SemBLK can significantly degrade the semantic communication system, while these adversaries generated are imperceptive to our humans. We believe our work provides some useful insights into the physical-layer robustness of semantic communications. In the future, we will deploy our proposed method to real scenarios and more semantic communication systems.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (61971066), and in part by the research foundation of Ministry of Education and China Mobile under Grant MCM20180101.

References

[1] ITU-R, "Imt traffic estimates for the years 2020 to 2030," *https : //www.itu.int/dms_pub/itu − r/opb/rep/R − REP − M.2370 − 2015 − PDF − E.pdf*, 2022, [Online].

[2] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning based semantic communications: An initial investigation," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[3] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[4] M. Yang, C. Bian, and H.-S. Kim, "Ofdm-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 584–599, 2022.

[5] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 230–244, 2022.

[6] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[7] M. Xu, X. Tao, F. Yang, and H. Wu, "Enhancing secured coverage with comp transmission in heterogeneous cellular networks," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2272–2275, 2016.

[8] M. Sadeghi and E. G. Larsson, "Physical adversarial attacks against end-to-end autoencoder communication systems," *IEEE Communications Letters*, vol. 23, no. 5, pp. 847–850, 2019.

[9] S. Baluja and I. Fischer, "Adversarial transformation networks: Learning to generate adversarial examples," *arXiv preprint arXiv:1703.09387*, 2017.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[11] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[13] Z. Li, J. Zhou, G. Nan, Z. Li, Q. Cui, and X. Tao, "Sembat: Physical layer black-box adversarial attacks for deep learning-based semantic communication systems," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*. IEEE, 2022, pp. 1–5.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[15] F. H. K. dos Santos Tanaka and C. Aranha, "Data augmentation using gans," *Proceedings of Machine Learning Research*, vol. 1, p. 16, 2019.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.