Requested Experiments for Reviews

TABLE I: SemanAegis efficacy performance across SC architectures and four benchmark datasets

SC System	Protection Method	Efficacy (Authorized ↑ | Unauthorized ↓ Data PSNR dB)			
		MNIST	F-MNIST	CIFAR-10	ImageNet
JSCC	Clean system*	35.2 | 35.1	32.5 | 34.8	39.4 | 38.6	29.9 | 27.2
	SemanAegis	34.8 | 10.8	32.1 | 8.6	40.3 | 6.7	29.4 | 5.5
	ADIP	30.8 | 15.0	28.7 | 20.5	37.3 | 9.9	28.1 | 8.7
	M-LOCK	34.4 | 12.1	33.2 | 12.7	39.9 | 11.2	27.5 | 5.1
JSCC-f	Clean system*	37.3 | 37.1	36.1 | 35.0	39.2 | 39.2	34.6 | 36.7
	SemanAegis	35.0 | 10.2	35.5 | 9.7	38.1 | 9.0	34.8 | 6.5
	ADIP	30.2 | 14.3	33.1 | 20.8	35.7 | 8.6	32.4 | 7.2
	M-LOCK	38.5 | 17.5	35.2 | 14.3	37.9 | 11.8	33.6 | 10.5
JSCC-q	Clean system*	40.5 | 39.3	39.0 | 37.3	43.5 | 41.8	38.5 | 33.9
	SemanAegis	38.5 | 9.2	38.2 | 9.6	41.3 | 8.0	36.8 | 5.4
	ADIP	34.2 | 18.3	32.9 | 10.2	37.3 | 11.9	31.5 | 10.7
	M-LOCK	38.8 | 13.6	37.2 | 14.8	40.5 | 15.4	36.7 | 13.5
SCAN	Clean system*	38.1 | 36.3	37.5 | 37.9	38.6 | 39.2	36.0 | 34.8
	SemanAegis	36.4 | 8.7	36.6 | 8.3	38.4 | 9.5	35.7 | 5.7
	ADIP	27.7 | 13.6	31.2 | 18.4	33.9 | 12.2	32.0 | 9.5
	M-LOCK	37.5 | 15.2	34.8 | 15.0	37.9 | 14.4	34.3 | 10.1
SemCC	Clean system*	37.4 | 37.0	36.5 | 35.8	38.6 | 37.7	36.3 | 35.7
	SemanAegis	36.2 | 7.1	35.5 | 9.3	37.9 | 10.6	35.0 | 7.9
	ADIP	31.5 | 17.8	29.2 | 14.5	31.8 | 16.9	32.4 | 12.1
	M-LOCK	34.6 | 14.7	32.2 | 11.5	36.9 | 11.2	34.4 | 9.3

I. Experiment

A. Experimental Setting

1) SC Systems: We utilize three state-of-the-art (SOTA) SC systems: 1) JSCC (w/ LDPC+QAM), which deploy convolutional networks for joint source-channel coding [1]; 2) JSCC-f (w/ LDPC+QAM), which introduces a feedback mechanism enabling dynamical encoding for varying channel conditions [2]; and 3) JSCC-q (w/ LDPC), which applies vector quantization and discretizes latent space into finite codebook [3].

2) Test Datasets: We employ four datasets to evaluate SemanAegis: 1) MNIST, consisting of 70K grayscale images of handwritten digits; 2) Fashion-MNIST (F-MNIST), comprising 70K grayscale images of various fashion items; 3) CIFAR-10, containing 60K color images across 10 classes of everyday objects; and 4) ImageNet, with over 14 million annotated images across more than 20K classes (randomly selected 5 classes, forming a subset: 40K training images, 10K testing images).

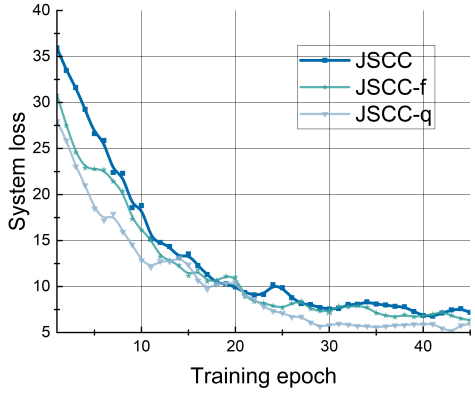


Fig. 1: 222

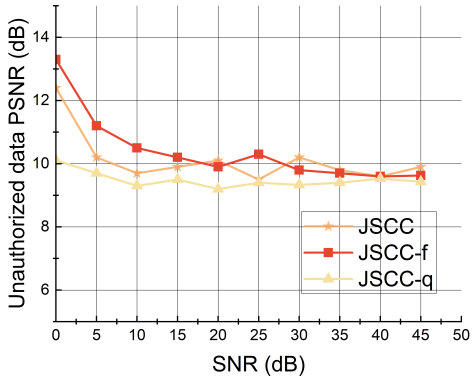


Fig. 2: 222

3) Evaluation Metrics: We assess SemanAegis across three principal dimensions: efficacy, imperceptibility, and robustness. 1) For efficacy, we utilize Peak Signal-to-Noise Ratio (PSNR) to quantify SC transmission—high PSNRs are desirable for authorized inputs x_a, reflecting better transmission efficiency and the system f capability to permit legitimate access, whereas low PSNR is preferred for unauthorized inputs x_u and indicates that f produces erroneous outputs, affirming protective efficacy in preventing unauthorized access and mitigating knowledge leakage. 2) Regarding imperceptibility, by adjusting the poisoning rate γ, we test SemanAegis efficacy—smaller γ indicates fewer data modifications, enhancing imperceptibility. Also, we assess the SSIM between authorized inputs x_a (*i.e.,* credential-inserted data) and source samples—higher SSIM indicates better credential stealthiness. 3) To

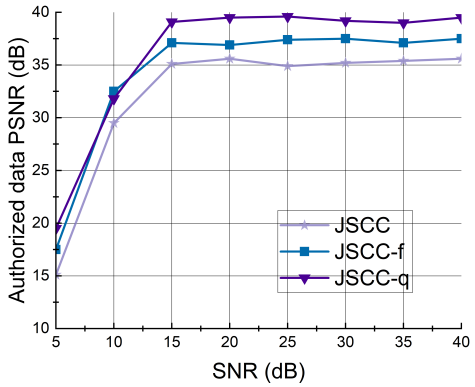


Fig. 3: 222

test robustness, we perform adversarial knowledge extractio attacks and measure PSNR to quantify protective efficacy.

B. Experimental Results

We test the efficacy, imperceptibility, and robustness o SemanAegis, and report corresponding PSNRs and SSIMs As the first to address SC knowledge protection, SemanAegi lacks direct comparisons, and we benchmark against SOTA active protections ADIP [4] and M-LOCK [5] from computer vision. The signal-to-noise ratio (w/ AWGN) and compression ratio are configured as 10 dB and $\frac{1}{6}$. We employ 40K samples for training and 10K for testing for each dataset.

1) SemanAegis Efficacy: We utilize three SC systems and four datasets to evaluate PSNRs of protected systems on authorized (x_a) and unauthorized (x_u) data. Besides ADIP and M-LOCK, the clean (*i.e.,* unprotected) system is included for comparison. The results are presented in Tab. I.

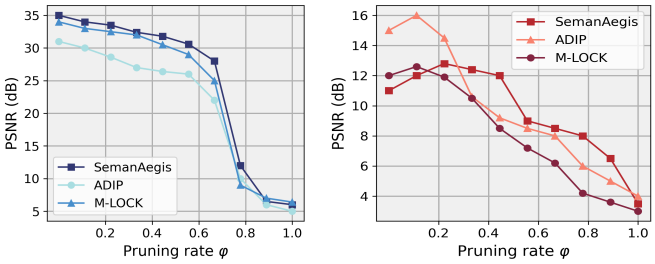
The average PSNR of x_a indicates SemanAegis (36.2 dB) > M-LOCK (36.1 dB) > ADIP (32.6 dB), approaching the clean system (37.1 dB). These findings substantiate that SemanAegis effectively preserves the transmission quality of x_a and surpasses the baselines, which stems from the incorporation of losses $\mathcal{L}_a$ and $\mathcal{L}_c$, sustaining the transmission performance of x_a. For unauthorized x_u transmissions, the average PSNR values follow a descending order—clean system (36.3 dB) > ADIP (13.0 dB) > M-LOCK (12.7 dB) > SemanAegis (8.2 dB). This controlled PSNR degradation proves that unauthorized inputs x_u yield erroneous results, hindering unauthorized system exploitation and protecting SC knowledge.

2) SemanAegis Imperceptibility: We undertake two evaluations: 1) we configure varying poisoning rates γ to observe the PSNR on both authorized x_a and unauthorized x_u samples, and 2) we present authorized input x_a instances to compare them with their source counterparts and analyze the credential inconspicuousness. We utilize MNIST set and JSCC for experiment, and the results are illustrated in Figs. **??** & **??**.

As shown in Fig. **??**, increasing γ leads to an upward trend in PSNR of x_a, stabilizing at $\gamma \approx 0.3$ for SemanAegis and $\gamma \approx 0.4$ for M-LOCK and ADIP, suggesting that SemanAegis achieves rapid convergence by modifying only 30% of the data. For x_u, all PSNR values exhibit minimal fluctuations

TABLE II: SemanAegis robustness performance across malicious threats and benchmark datasets

Threats	Protection Method	Robustness (Authorized ↑ | Unauthorized ↓ Data PSNR dB)			
		MNIST	F-MNIST	CIFAR-10	ImageNet
Model extraction (4K queries)	SemanAegis	- | 11.3	- | 9.5	- | 7.2	- | 6.7
	ADIP	- | 18.0	- | 22.6	- | 14.5	- | 9.9
distillation (3K queries & 200 data)	SemanAegis	- | 14.4	- | 12.0	- | 9.7	- | 13.1
	ADIP	- | 19.1	- | 29.4	- | 21.2	- | 16.1
Fine-tuning w/ GTSRB	SemanAegis	30.4 | 11.6	31.0 | 16.4	34.8 | 7.7	25.9 | 8.9
	ADIP	24.7 | 16.2	27.5 | 18.9	28.9 | 12.0	22.8 | 9.2 2-6



(a) Authorized data (b) Unauthorized data

Fig. 4: The accuracy for MNIST of protected systems on (a) authorized and (b) unauthorized data against pruning rate φ.

with increasing γ, following the order SemanAegis < M-LOCK < ADIP, wherein SemanAegis restrict x_u best, attributable to loss $\mathcal{L}_u$ regularizing x_u transmission.

Fig. **??** illustrates that authorized x_a generated by SemanAegis exhibits no significant perceptual differences compared with source data, whereas those produced by ADIP and M-LOCK display distinct blocky patches. Additionally, the average SSIMs (SemanAegis: 0.94 > M-LOCK: 0.78 > ADIP: 0.69) quantitatively confirm SemanAegis credential stealthiness.

3) SemanAegis Robustness: We evaluate SemanAegis robustness against knowledge leakage threats for JSCC system f and MNIST set: 1) model extraction, aiming to replicate f architecture and parameters by extensive queries [6]; 2) knowledge distillation, seeking to transfer f knowledge to a surrogate student system [7]; 3) fine-tuning, adjusting the f parameters to alter its behavior post-deployment; and 4) pruning, attempting to degrade protective performance by systematically removing partial parameters [8]. The results are detailed in Tab. II & Fig. 4

From Tab. II, 1) under a 4K-query model extraction attack, the replicated f^*'s average PSNR on unauthorized data x_u is 8.7 dB (SemanAegis) vs. 16.3 dB (ADIP) and 13.8 dB (M-LOCK), indicating the embedded access control SemanAegis has also been reproduced within f^* and remains effective. 2) In a 3K-query and 200-sample knowledge distillation attack, the student system f''s average PSNRs on x_u are 12.3 dB (SemanAegis), 21.5 dB (ADIP), and 17.7 dB (M-LOCK),

all lower than regular levels, suggesting that SemanAegis are effectively transferred into f' through distillation and preserves knowledge safeguards. 3) Fine-tuning with GTSRB cross multiple datasets results in a slight PSNR decrease of f on authorized data. However, the average PSNR for unauthorized samples remains low—SemanAegis: 11.1 dB, ADIP: 14.6 dB, and M-LOCK: 13.0 dB—meaning effective x_u rejection and knowledge protection. 4) Results for pruning rates φ reveal that at $\varphi \approx 0.7$, the PSNR of authorized inputs across methods begins to decline rapidly, underscoring the inherent robustness of SemanAegis up to significant pruning levels, whereas the PSNR for unauthorized x_u consistently exhibits downward trends, indicating that pruning can not effectively facilitate unauthorized access.

II. Conclusion

In this research, we propose ***SemanAegis***, the first SC knowledge protection framework, which employs backdoor training to integrate built-in access control, and ensures the precise transmission of authorized inputs embedded with imperceptible credentials while generating low-quality erroneous results for the unauthorized. To expedite convergence, a dedicated contrastive loss is designed to enhance efficiency. Experiments substantiate the efficacy and imperceptibility of SemanAegis in knowledge protection, and the robustness against existing knowledge leakage threats.

References

[1] E. Bourtsoulatze, D. Burth, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE TCCN*, 2019.

[2] D. Kurka and D. Gündüz, "Deepjscc-f: Deep joint source-channel coding of images with feedback," *IEEE JSAIT*, 2020.

[3] T. Tung, D. Kurka, M. Jankowski, and D. Gündüz, "Deepjscc-q: Constellation constrained deep joint source-channel coding," *IEEE JSAIT*, 2022.

[4] M. Xue, Z. Wu, C. He, J. Wang, and W. Liu, "Active dnn ip protection: A novel user fingerprint management and dnn authorization control technique," in *IEEE TrustCom*, 2020.

[5] G. Ren, J. Wu, G. Li, S. Li, and M. Guizani, "Protecting intellectual property with reliable availability of learning models in ai-based cybersecurity services," *IEEE TDSC*, 2024.

[6] J. Correia-Silva, R. Berriel, C. Badue, A. Souza, and T. Santos, "Copycat cnn: Are random non-labeled data enough to steal knowledge from black-box models," *Pattern Recogn.*, 2021.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv*, 2015.

[8] A. See, M. Luong, and C. Manning, "Compression of neural machine translation models via pruning," in *CoNLL*, 2016.