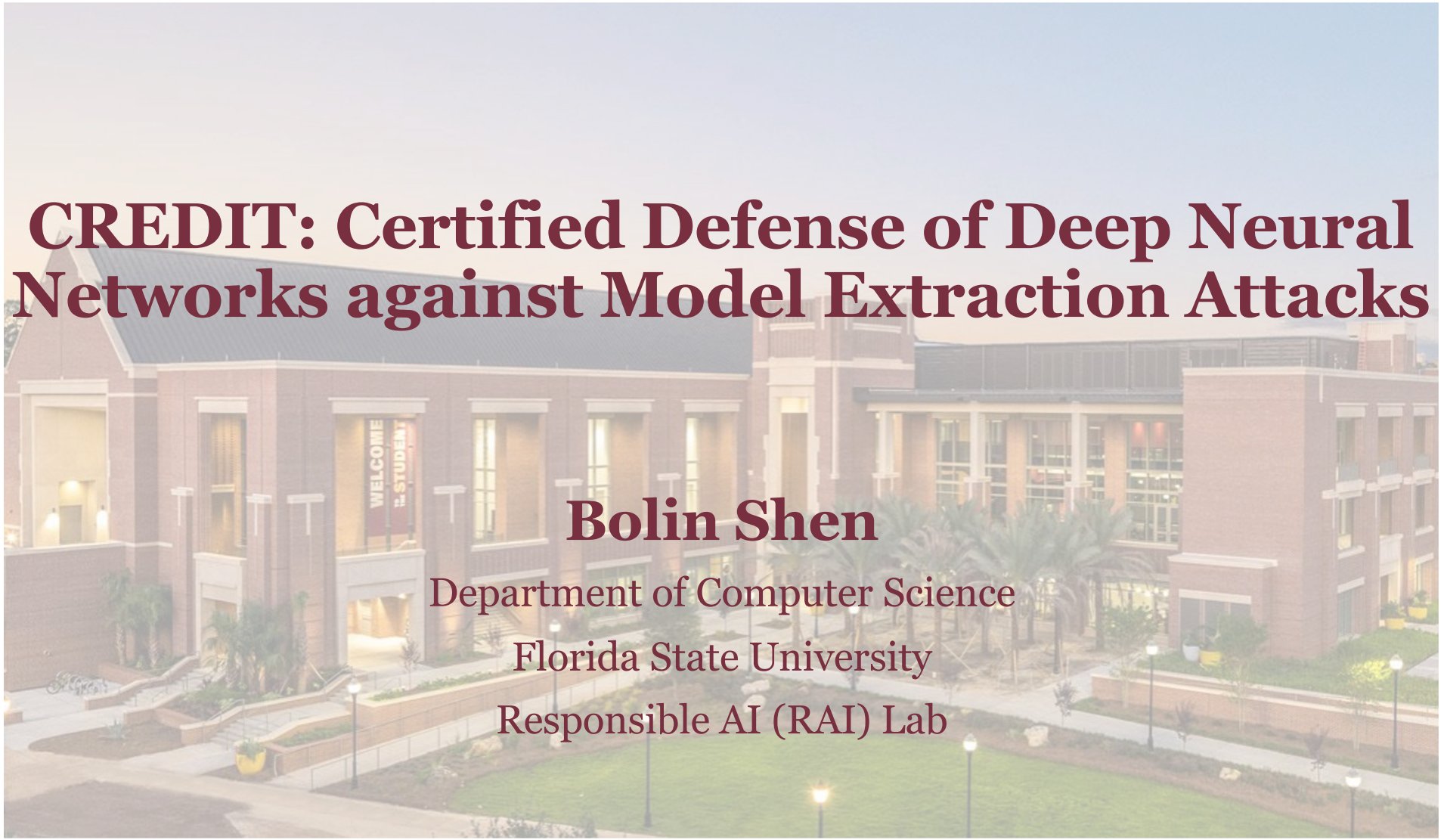
CREDIT: Certified Defense of Deep Neural Networks against Model Extraction Attacks



Bolin Shen

Department of Computer Science

Florida State University

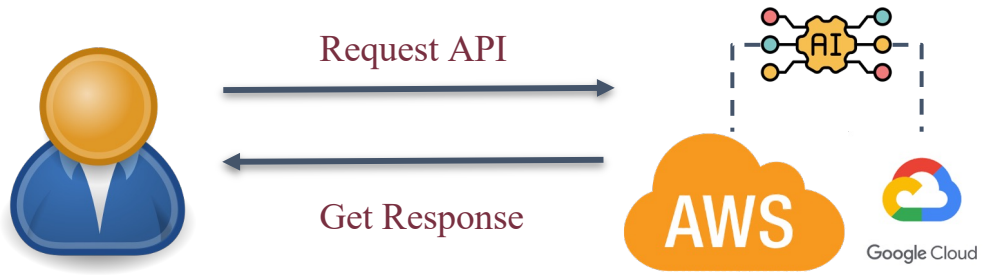
Responsible AI (RAI) Lab

Machine Learning as a Service (MLaaS)

MLaaS is a cloud-based paradigm for deploying and accessing deep learning models through standardized APIs, allowing users to leverage advanced models without managing the underlying infrastructure.

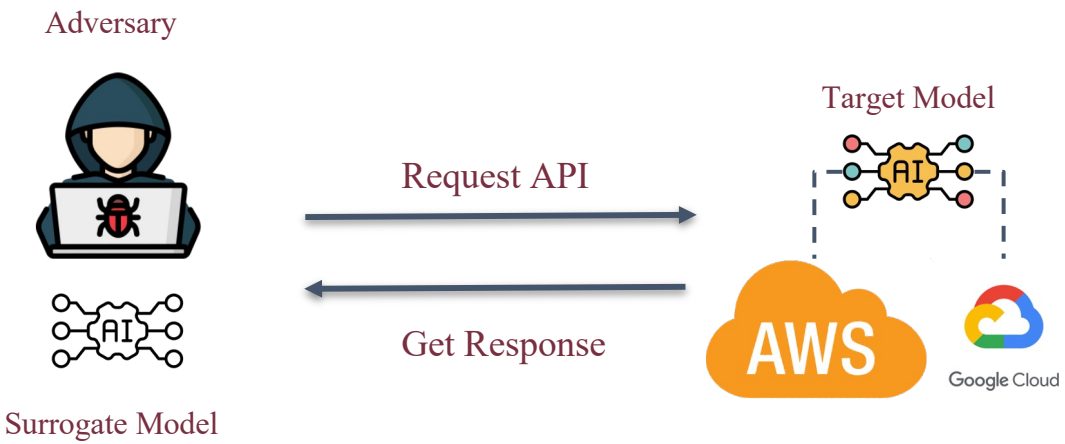
Why Users Prefer MLaaS

- **Accessible yet Secure Model Utilization.**
- **Cost and Resource Efficiency.**
- **Scalable Deployment with Broad Applicability.**



Model Extraction Attack & Defense

A Model Extraction Attack occurs when an adversary repeatedly queries a deployed **target model** (via MLaaS APIs) to collect **input–output pairs**, then trains a **surrogate model** that replicates the target model's functionality .



Challenges of Model Extraction Attack

Developing a high-performance DNN requires **proprietary data**, **carefully engineered architectures**, **substantial computation**, and **long training time** — making each model a valuable piece of **intellectual property (IP)** . It faces several core challenges:

1. **Machine Learning Model Vulnerability**
2. **Low-Cost Replication by Adversaries**
3. **Severe Economic and Strategic Losses**

Taxonomy of Existing Defense Methods

A large number of studies have explored how to protect the ownership of machine learning models under Model Extraction Attacks (MEAs). Broadly, these existing defenses can be divided into **two major categories**: **Watermarking-based** and **Fingerprinting-based** methods.

Watermarking-based Defenses

- **Idea:** Embed specially designed trigger samples into the model during training.
  Ownership is verified if the same triggers produce expected outputs on a suspicious model.
- **Limitations:**
  - Task-irrelevant information can degrade accuracy.
  - No formal guarantee for ownership verification.

Fingerprinting-based Defenses

- **Idea:** Characterize a model's *intrinsic behavior* or decision boundary by analyzing its embeddings or prediction patterns.
  Ownership is confirmed when a suspicious model exhibits highly similar behavior.
- **Limitations:**
  - Require many reference pairs for reliable comparison.
  - Lack rigorous theoretical certification.

Research Gap

There is currently *no framework* that enables model owners to **verify ownership with provable guarantees** after an extraction attack. A certified defense must ensure correctness, reproducibility, and formal error bounds in ownership decisions.

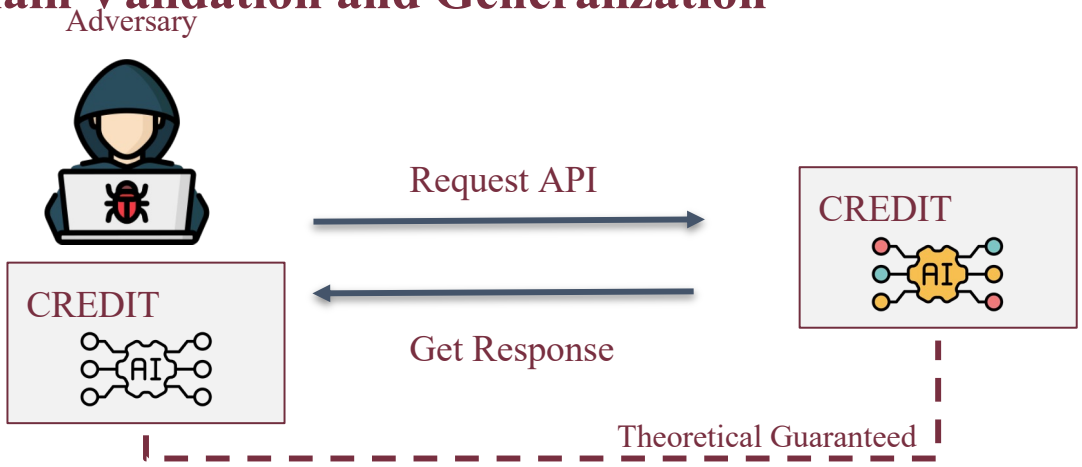
Core Technical Challenges

1. **Similarity Quantification**
   - Measuring how "functionally close" two DNNs are is inherently difficult — even when they differ in architecture, parameters, or training data.
2. **Decision Threshold Definition**
   - Determining a reliable cutoff that separates *surrogate models* from *independent ones* is non-trivial and context-dependent.
3. **Theoretical Certification**
   - Providing **provable guarantees** (bounded error probabilities) for ownership verification remains an open and essential challenge for practical deployment.

Overview of CREDIT Framework

Given the lack of *certified* defenses identified earlier, we introduce **CREDIT** (*CeRtifiEd Defense of Deep Neural Networks against Model ExtractIon ATtacks*) – a **certified defense framework** that protects MLaaS models from extraction attacks by providing **provable ownership verification** using *information-theoretic principles*. **Core Components:**

1. **Certified Defense via Mutual Information (MI)**
2. **CREDIT Threshold for Ownership Verification**
3. **Cross-Domain Validation and Generalization**



Notations & Definition

Before introducing our theoretical results, we first establish the **notation** and **formal definition** used throughout CREDIT.

Core Notations:

- $\mathbf{x} \in \mathcal{X}$
- $f : \mathcal{X} \rightarrow \mathcal{Y}$ Target Model
- g Defended model (target f with Gaussian mechanism)
- h Suspicious model — can be *independent* (h_{ind}) or *surrogate* (h_{sur})
- $\mathrm{e}_{\mathrm{f}}(x)$ Embedding of input x from model f
- $\mathcal{V}$ Verification set used for ownership validation
- τ CREDIT threshold for ownership verification

Mutual Information

We measure model similarity using **Mutual Information** between their embeddings:

- Quantifies how much information two representations share.
- High MI → models behave similarly (likely surrogate).
- Low MI → independent model.
- In practice, MI is estimated using the **KSG estimator** based on nearest-neighbor statistics .

$$I(U;V) = \mathbb{E}_{(U,V)}\left[\log \frac{p(U,V)}{p(U)p(V)}\right]$$

Gaussian Mechanism

The **Gaussian Mechanism** is a fundamental method for ensuring **differential privacy** in machine learning systems [1,2].

It works by adding random noise drawn from a Gaussian distribution to the output of a computation:

$$\tilde{y} = f(x) + \mathcal{N}(0, \sigma^2 I)$$

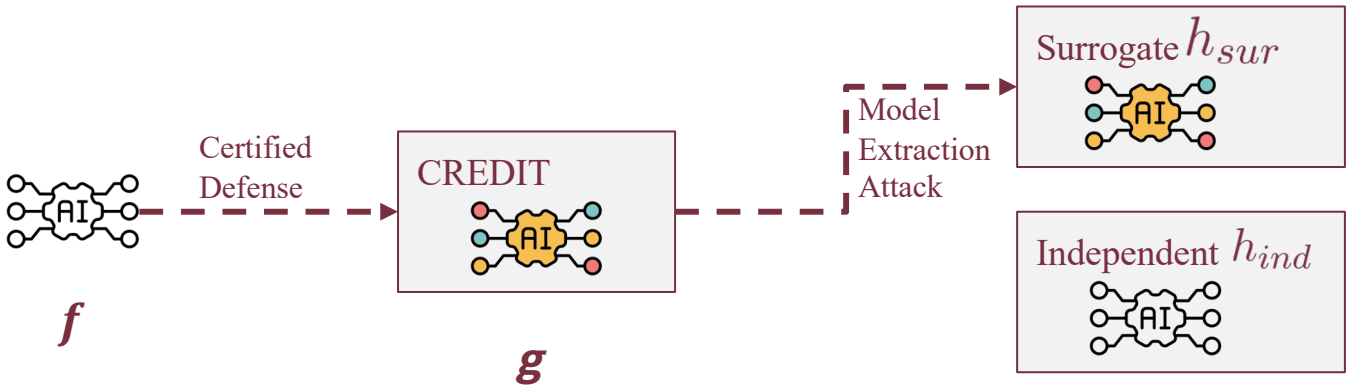
This noise makes the contribution of any individual data point **indistinguishable**, protecting privacy while maintaining overall result utility.

References
[1] Dwork, Cynthia. "Differential privacy." *International colloquium on automata, languages, and programming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
[2] Abadi, Martin, et al. "Deep learning with differential privacy." *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016.

Problem Goal

Certified Defense against Model Extraction Attacks

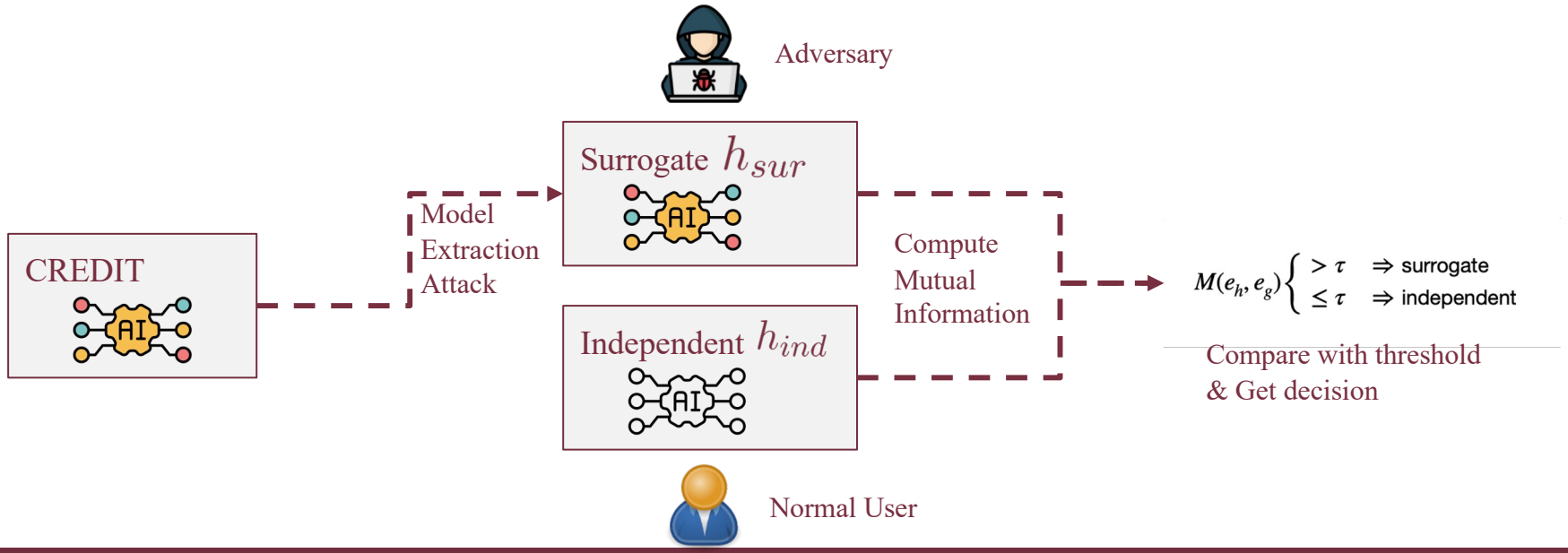
Goal: Develop a **defended model** g from a target model f such that ownership can be **certified** — i.e., any **surrogate model** extracted from g can be provably distinguished from **independently trained** models with bounded error probability γ.



Overall Pipeline

The CREDIT framework integrates defense construction, attack modeling, and certified verification into a unified pipeline:

- **Defender**: Build Defended Model g via Gaussian Mechanism
- **Attacker**: Train Surrogate Model h_{sur}
- **Defender**: Design CREDIT Threshold τ
- **Defender**: Ownership Verification Decision

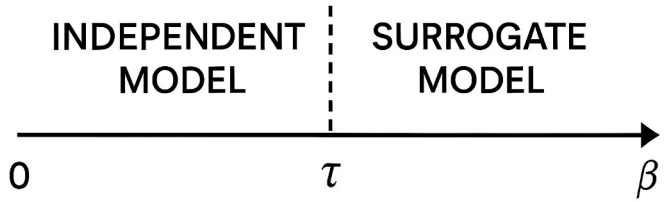


Mutual Information Bound

To evaluate whether a suspicious model $\boldsymbol{h}$ originates from the defended model $\boldsymbol{g}$, we require a **principled similarity metric** that captures their dependencies.

Mutual Information (MI) provides an **information-theoretic** measure of how much two embeddings share in common — ideal for distinguishing **surrogate** vs **independent** models

Thus, **Mutual Information** becomes the quantitative signal for ownership verification and the **Gaussian mechanism** ensures that $I(e_h, e_g) \leq \beta$, $\beta = \frac{\Delta^2}{2\sigma^2} n$, providing a **theoretical upper bound** for certification.



CREDIT Threshold for Verification

While **Mutual Information (MI)** quantifies model similarity, we still need a **decision boundary** that formally separates **surrogate** from **independent** models. To achieve this, CREDIT introduces a **certified threshold τ**, grounded in the MI upper bound from the Gaussian mechanism.

$$\tau = \beta\Big[1 - \rho \exp\Big(-\frac{Q\beta}{\eta d|\mathcal{V}|}\Big)\Big].$$

The CREDIT threshold **τ** reflects how strongly a suspicious model resembles the defended model under bounded information sharing.

It adaptively increases when the **query budget Q** is large—indicating that the attacker can better approximate the defended model—and decreases with larger **verification set size $|\mathcal{V}|$** or **embedding dimension** d, which both improve the accuracy and robustness of the MI estimation.

Certified Ownership Verification Guarantee

Building upon the CREDIT threshold τ introduced earlier, we now establish **formal guarantees** for ownership verification. These guarantees ensure that decisions made using τ are **theoretically guaranteed**, bounding both false alarms (independent models misclassified as stolen) and missed detections (surrogates misclassified as independent).

Let $\hat{I}$ denote the **empirical mutual information** between embeddings on the verification set $\mathcal{V}$. For the threshold τ defined in Definition 2, the following bounds hold:

(1) Independent False Alarm (Type I error):

$$\Pr\left[\hat{I}(e_{h_{\text{ind}}}(X), e_g(X)) > \tau\right] \leq \exp\left(-\frac{2|\mathcal{V}|\,\tau^2}{C^2}\right) \triangleq \gamma_1.$$

(2) Surrogate Missed Detection (Type II error):

$$\Pr\left[\hat{I}(e_{h_{\text{sur}}}(X), e_g(X)) \leq \tau\right] \leq \exp\left(-\frac{2|\mathcal{V}|\,[I(e_{h_{\text{sur}}}, e_g) - \tau]^2}{C^2}\right) \triangleq \gamma_2.$$

Both $\gamma 1$ and $\gamma 2$ decay **exponentially** with the size of the verification set $|\mathcal{V}|$, ensuring that larger verification sets yield tighter, more reliable certification.

Trade-off: Utility vs. Effectiveness

When defending a target model using CREDIT, there exists an **inherent trade-off**: adding Gaussian perturbation strengthens ownership verification but may reduce the model's downstream utility. To measure how Gaussian noise affects model utility, we define the **utility entropy gain** as

$$\Delta H_{\text{util}}(\sigma) := H_G(X+Z) - H_G(X), \quad Z \sim \mathcal{N}(0, \sigma^2 I).$$

A larger $\Delta H_{\text{util}}(\sigma)$ means stronger deviation from clean embeddings → **lower model utility**.

To assess certification strength, CREDIT introduces **verification entropy**:

$$H_{\text{ver}}(\sigma) := H(T_\sigma \mid H),$$

where T_σ is the verification indicator (decision outcome) and H is the true model type. $H_{\text{ver}}(\sigma)$ measures **decision ambiguity** — the uncertainty in distinguishing h_{ind} and h_{sur}.

Hyperparameter Sigma Selection

The Gaussian noise parameter **σ** simultaneously governs **utility preservation** and **verification robustness**.

As σ increases, the **utility entropy** $\Delta H_{\text{util}}(\sigma)$ rises (utility decreases), while the **verification entropy** $H_{\text{ver}}(\sigma)$ falls (robustness increases).

Selecting σ therefore becomes a **bi-objective optimization problem** balancing these opposing effects.

$$\min_{\sigma} \ \lambda_{\text{util}} \Delta H_{\text{util}}(\sigma) + \lambda_{\text{ver}} \Delta H_{\text{ver}}(\sigma)$$

Experimental Evaluation

To systematically assess CREDIT, we design experiments addressing three **core questions**:

- **RQ1:** *How effective is CREDIT* in simultaneously preserving model utility and ensuring verification robustness against surrogate models?
- **RQ2:** *How efficient is CREDIT* compared to baseline defenses in terms of defense and verification time?
- **RQ3:** *How reliable is CREDIT's certification* under different parameter settings (e.g., σ, τ) affecting verification accuracy and utility?

Datasets and Modalities

- **Computer Vision:** CIFAR-10 / CIFAR-100
- **Graph Learning:** ENZYMES / PROTEINS

Objective	Metric	Description
Utility Preservation	**Accuracy (%)**	Performance on downstream task
Verification Robustness	**AUROC (↑)**	Ability to distinguish surrogate vs independent models
Efficiency	**Time Cost**	Defense + verification overhead
Certification Reliability	**Error Rates (γ_1, γ_2)**	error bounds under CREDIT

Effectiveness of CREDIT

Table 1: Evaluation of all defense methods on downstream tasks in terms of utility performance. All results are reported as accuracy, with the best performance highlighted in **bold**.

Dataset	Backbone	Vanilla	Backdooring	EWE	IPGuard	UAP	**CREDIT**
CIFAR-10	ResNet	95.70 ± 0.03	78.10 ± 2.18	90.25 ± 0.02	91.08 ± 0.00	84.73 ± 2.26	$\mathbf{94.67 \pm 0.22}$
	VGG	92.18 ± 0.01	77.60 ± 2.76	88.82 ± 0.03	89.80 ± 0.02	82.91 ± 2.63	$\mathbf{91.68 \pm 0.17}$
	DenseNet	93.33 ± 0.07	60.45 ± 3.67	86.22 ± 0.10	89.41 ± 0.00	74.82 ± 4.50	$\mathbf{92.58 \pm 0.15}$
	GoogLeNet	94.90 ± 0.04	89.31 ± 1.38	93.58 ± 0.03	92.98 ± 0.19	91.84 ± 0.92	$\mathbf{94.67 \pm 0.11}$
CIFAR-100	ResNet	80.48 ± 0.11	74.43 ± 0.17	75.85 ± 0.07	77.54 ± 0.05	64.30 ± 4.14	$\mathbf{79.90 \pm 0.21}$
	VGG	77.41 ± 0.09	73.76 ± 0.29	74.48 ± 0.03	76.22 ± 0.01	71.10 ± 1.58	$\mathbf{76.71 \pm 0.20}$
	DenseNet	80.34 ± 0.14	72.91 ± 0.46	76.72 ± 0.08	77.95 ± 0.03	57.25 ± 5.02	$\mathbf{79.29 \pm 0.27}$
	GoogLeNet	78.72 ± 0.10	71.49 ± 0.33	74.69 ± 0.04	76.41 ± 0.01	66.96 ± 3.12	$\mathbf{77.50 \pm 0.19}$
Dataset	Backbone	Vanilla	RandomWM	BackdoorWM	SurviveWM	ImperceptibleWM	**CREDIT**
ENZYMES	GCN	42.78 ± 2.58	39.11 ± 2.58	38.16 ± 3.79	15.27 ± 3.36	26.94 ± 7.83	$\mathbf{40.61 \pm 1.46}$
	GAT	41.94 ± 1.42	36.94 ± 2.19	35.83 ± 1.80	18.33 ± 1.18	22.78 ± 4.16	$\mathbf{38.56 \pm 1.42}$
	GraphSAGE	50.83 ± 5.57	42.22 ± 5.50	43.61 ± 4.78	10.27 ± 1.04	27.78 ± 5.67	$\mathbf{46.94 \pm 5.54}$
	SSGC	33.05 ± 3.07	28.61 ± 2.71	28.27 ± 3.14	16.39 ± 1.04	26.17 ± 3.79	$\mathbf{28.94 \pm 2.19}$
PROTEINS	GCN	71.00 ± 2.60	70.40 ± 1.84	70.24 ± 0.92	59.49 ± 3.07	66.37 ± 2.64	$\mathbf{72.50 \pm 1.12}$
	GAT	73.09 ± 1.32	72.80 ± 2.02	72.94 ± 0.42	40.96 ± 1.48	53.66 ± 2.75	$\mathbf{74.14 \pm 1.56}$
	GraphSAGE	74.59 ± 0.76	70.69 ± 0.42	70.84 ± 0.21	38.71 ± 0.56	52.46 ± 6.99	$\mathbf{73.44 \pm 1.69}$
	SSGC	73.99 ± 0.37	71.00 ± 2.10	73.99 ± 1.10	47.23 ± 2.15	71.00 ± 1.12	$\mathbf{74.73 \pm 2.11}$

Utility degradation is **negligible**, indicating that Gaussian perturbation does not compromise downstream accuracy. CREDIT consistently achieves the **highest verification performance (AUC = 100%)** across all CV backbones (ResNet, VGG, DenseNet, GoogLeNet). Results confirm that CREDIT simultaneously ensures **robust ownership verification** and **high utility retention**.

Efficiency of CREDIT

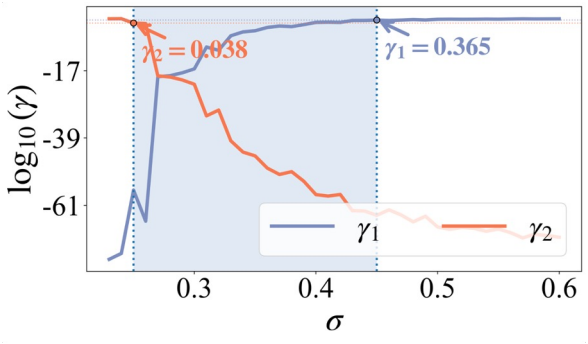
Table 3: Efficiency of defense methods in both the defense and verification stages across backbone models with different parameter sizes. Results are reported in seconds, with the best performance highlighted in **bold**.

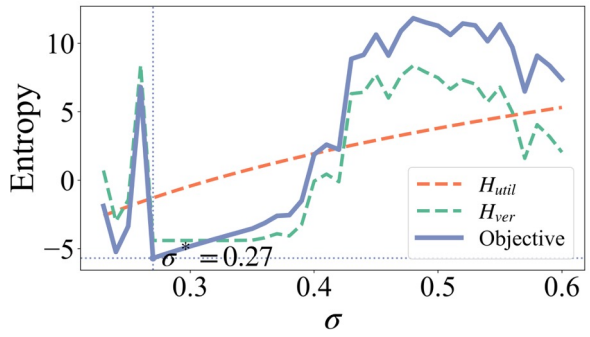
Method	VGG-16 (15M)		ResNet-50 (25.6M)	
	Defense(s)	Verification(s)	Defense(s)	Verification(s)
Backdooring	0.594 ± 0.11	83.10 ± 0.85	0.399 ± 0.09	188.8 ± 0.40
EWE	0.279 ± 0.00	73.40 ± 0.13	0.199 ± 0.02	189.0 ± 0.76
IPGuard	1.783 ± 0.00	73.94 ± 0.23	4.319 ± 0.00	189.0 ± 0.41
UAP	5.587 ± 0.13	72.74 ± 0.08	5.722 ± 0.14	180.3 ± 0.32
Our	$\mathbf{0.001 \pm 0.00}$	$\mathbf{22.23 \pm 0.07}$	$\mathbf{0.001 \pm 0.00}$	$\mathbf{22.62 \pm 0.36}$

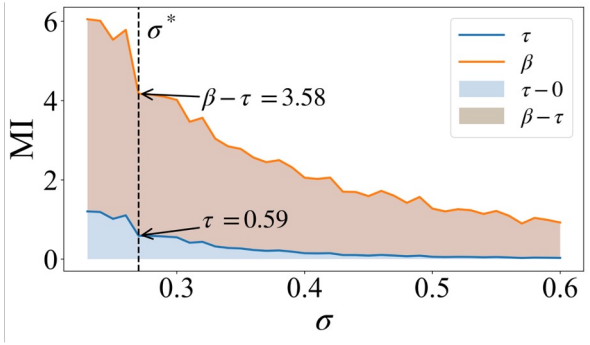
In addition to strong verification accuracy, a practical defense must also be **time-efficient** during both model defense and verification.

CREDIT achieves this by eliminating the need for auxiliary model training or complex watermark construction, relying only on **mutual information estimation** and **threshold comparison**.

Reliability of Ownership Verification







As **σ increases**, injected noise strengthens protection but also increases randomness:

- **γ_1 (Type I error) ↑** — slightly higher false alarm probability.
- **γ_2 (Type II error) ↓** — fewer missed detections (better surrogate identification).

CREDIT thus exhibits a **monotonic trade-off**: tighter security versus minimal false positives. Empirically, a **stable reliability range** occurs for $0.25 < \sigma < 0.45$, where both γ_1 and γ_2 remain low and balanced.

The **optimal σ = 0.27** achieves the best joint performance in practice.

Related Works

Model Extraction Attacks and Defenses

Model extraction research focuses on how adversaries replicate MLaaS models through query-based interactions, and how defenses counter such attacks.

- **KnockoffNets** (Orekondy et al., 2019) for query-efficient surrogate training
- **CEGA** (Wang et al., 2024) exploiting graph structure cues.
- **IPGuard** (Cao et al., 2021) for fingerprint-based verification.

However, existing defenses providing **no theoretical guarantees** on ownership verification.

Certified Defenses for Deep Learning Models

Certified defenses aim to establish provable guarantees of robustness or ownership under bounded perturbations.

- **Randomized Smoothing** (Cohen et al., 2019) for certified adversarial robustness
- **Certified Unlearning** (Dong et al., 2024) for safe model updates

Despite their theoretical rigor, **none of these methods address certification of ownership verification**.

CREDIT fills the gap of certified defense against model extraction attacks.

Conclusion

In summary, CREDIT introduces **the first certified defense framework** against model extraction attacks, providing **theoretical guarantees** through mutual information–based threshold certification and achieving state-of-the-art **ownership verification performance** with **high efficiency** and generality **across diverse modalities.**

References

Cao, Xiaoyu, Jinyuan Jia, and Neil Zhenqiang Gong. "IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary." *Proceedings of the 2021 ACM asia conference on computer and communications security*. 2021.

Peng, Zirui, et al. "Fingerprinting deep neural networks globally via universal adversarial perturbations." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Jia, Hengrui, et al. "Entangled watermarks as a defense against model extraction." *30th USENIX security symposium (USENIX Security 21)*. 2021.

Zhao, Xiangyu, Hanzhou Wu, and Xinpeng Zhang. "Watermarking graph neural networks by random graphs." *2021 9th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 2021.

Xu, Jing, et al. "Watermarking graph neural networks based on backdoor attacks." *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2023.

Zhang, Linji, et al. "An imperceptible and owner-unique watermarking method for graph neural networks." *Proceedings of the ACM Turing Award Celebration Conference-China 2024*. 2024.

Wang, Zebin, et al. "CEGA: A Cost-Effective Approach for Graph-Based Model Extraction and Acquisition." *arXiv preprint arXiv:2506.17709* (2025).

Dong, Yushun, et al. "Idea: A flexible framework of certified unlearning for graph neural networks." *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024.

Zhang, Binchi, et al. "Towards certified unlearning for deep neural networks." *arXiv preprint arXiv:2408.00920* (2024).

Collaborators

CREDIT: Certified Defense of Deep Neural Networks against Model Extraction Attacks

Bolin Shen[1], Zhan Cheng[2], Neil Zhenqiang Gong[3], Fan Yao[4], Yushun Dong[1]

Affiliation

FSU

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Duke
UNIVERSITY

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


Thanks for listening!