One Trap to Block Them All: Defending Encoder Stealing via Isotropic Uniformity

Yitong Shi[1], Kang Wei[1*], Fushuo Huo[1], Shuchi Wu[2], and Chuan Ma[3]

[1] Southeast University, Nanjing, China
{213221553, kang.wei, fushuohuo}@seu.edu.cn
[2] Nanjing University of Science and Technology, Nanjing, China
shuchiwu.2000@gmail.com
[3] Chongqing University, Chongqing, China
chuan.ma@cqu.edu.cn

Abstract. High-performance Self-Supervised Learning (SSL) encoders deployed as Encoder-as-a-Service (EaaS) are highly vulnerable to advanced contrastive model stealing attacks, such as RDA and ContSteal. Traditional defenses often stack complex modules or rely on heuristic noise, struggling to balance privacy and utility. To address this, we pioneer the use of active perturbations for EaaS protection by proposing **UniTrap**, a simple yet effective single-module defense specifically tailored against contrastive-based stealing paradigms. Instead of mere distance maximization, **UniTrap** explicitly optimizes global feature entropy via a Gaussian potential function. This generates a stealthy perturbation that strictly enforces an isotropic, uniform distribution on the unit hypersphere—acting as "one trap" to fundamentally block contrastive alignment. We theoretically demonstrate that this perfect uniformity causes the attacker's informative gradients to cancel out in the tangent space, inducing gradient vanishing and permanently stalling surrogate optimization. Extensive experiments confirm that our method achieves a state-of-the-art privacy-utility Pareto frontier across various benchmarks.

Keywords: Model Stealing Defense · Encoder-as-a-Service · Gradient Vanishing · Self-Supervised Learning · Contrastive Learning

1 Introduction

Breakthroughs in Self-Supervised Learning (SSL) [3,5–7]allow universal encoders to extract rich semantic embeddings from vast unlabeled data [9,26]. Due to the extremely high computational resources and massive data required to train a high-performance SSL encoder (such as CLIP, SimCLR, etc. [5]), the high cost has prompted tech giants like OpenAI, Google, and Clarifai to deploy these pre-trained models as Encoder-as-a-Service (EaaS). In these services, users upload data through APIs and obtain corresponding embeddings, achieving high performance in downstream tasks with minimal data(illustrated in Fig. 1).

* Corresponding Author, Email: kang.wei@seu.edu.cn.

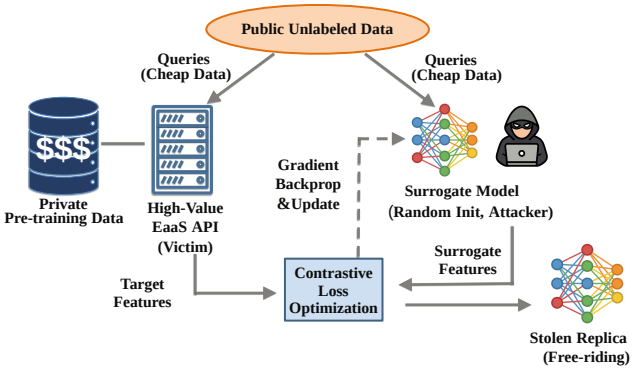


Fig. 1: Schematic of the contrastive model stealing threat against SSL encoders. Attackers utilize cheap, unlabeled data to query the EaaS API [12, 41, 42], optimizing a surrogate model through contrastive loss to "free-ride" on high-value target performance at minimal cost.

Unlike traditional attacks targeting classifiers [14, 22], attacks on encoders exploit high-dimensional embeddings rather than low-dimensional prediction labels or posterior probabilities. From an information theory perspective, high-dimensional embeddings contain more detailed and granular feature information about the input data than labels, leading to higher information leakage [21, 26]. This demand for high-fidelity reproduction of high-dimensional semantic features has driven the evolution of encoder theft algorithms from simple numerical approximation to deep target function alignment [4, 15, 17, 20, 22, 29, 31, 36]. To further enhance alignment, ContSteal [26] introduced negative sample pairs and used contrastive loss instead of regression loss, further improving the consistency between the stolen target and the training target. To address the high query cost and optimization bias issues in the above end-to-end methods, RDA [38] introduced a multi-stage strategy based on sample prototypes, achieving precise replication of the target encoder's representation capabilities while significantly reducing query costs through a "refinement, discrimination, alignment" mechanism.

Motivation: High-fidelity theft methods are emerging [25, 40] and evolved rapidly across various domains [16, 24, 34, 43], yet model owners lack simple and effective defense measures to thwart these attempts. Current defenses can be divided into passive and active defenses. Passive defenses inherently have insurmountable drawbacks: model watermarks cannot prevent theft for private use or for subsequent theft [1, 6, 13, 30, 39]; Performing dataset inference to identify stolen copies cannot prevent the leakage of model parameters or functions at the time of theft, it can only serve as a legal forensic tool for post-theft rights protection [10, 20]. Therefore, we turn our attention to active defense. In the game between model theft and defense, the core goal of active defense is always to find the best balance between security and model utility. Research shows that small

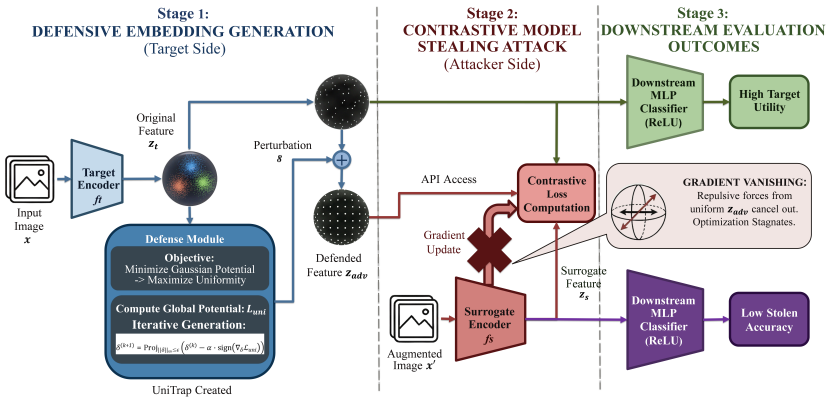


Fig. 2: Conceptual framework of our proposed UniTrap Defense. By forcing target embeddings z_{adv} into an isotropic distribution (**UniTrap**) , the defense induces gradient vanishing for the attacker's contrastive loss, effectively disrupting surrogate optimization while preserving high target utility.

perturbations can cause significant changes in output [27], but top-k, rounding, and general Gaussian noise are difficult to achieve a privacy-utility balance [9]. To address this issue, Dubinski et al. proposed B4B [8], an active defense mechanism that distinguishes between normal users and thieves and adds perturbations, but it is based on a strong assumption that the query data distribution of the thief is different from that of the benign user. Therefore, we seek a universally applicable defense mechanism that employs minimal perturbations to significantly disrupt theft effectiveness without degrading normal downstream utility. As illustrated in **Fig. 2**, we contrast our proposed mechanism, **UniTrap**, with a distance-oriented baseline. While such baselines merely maximizes feature displacement via distance metrics (e.g., cosine similarity), **UniTrap** enforces isotropic uniformity on the unit hypersphere, causing the attacker's contrastive gradients to vanish. The main contributions of this work are summarized as follows:

- To address the critical vulnerability of SSL encoders against contrastive model stealing, we propose **UniTrap**, the first active perturbation-based defense tailored for SSL encoders. It operates as a geometric defense mechanism based on a Gaussian potential function, fundamentally disrupting the alignment process of surrogate models through achieving perfect uniformity.
- We provide a theoretical bridge between feature distribution and defense efficacy, demonstrating that inducing an isotropic, uniform distribution on the hypersphere leads to **gradient vanishing** for contrastive optimization.
- We empirically validate our defense by analyzing the training dynamics of state-of-the-art attacks like RDA and ContSteal. We reveal a **"dual collapse" of feature similarities**, demonstrating that **UniTrap** locks feature

correlations near zero and forces the surrogate's optimization into permanent stagnation.
– Extensive experiments across multiple benchmarks show that our method achieves a superior privacy-utility trade-off, suppressing stolen accuracy from 77.90% to 35.26% while maintaining a high target utility of 90.18%.

2 Related Work

2.1 From Simple Imitation to Multi-relationship Extraction

Current theft methods have evolved from simple output matching to complex theft that exploits high-order relationships among samples, mainly including the following four types:

Conventional Attacks [9]: The thief regards the output Embedding of the target model as a soft label and trains by minimizing the mean squared error (MSE) between the output of the proxy model and the target Embedding, but this ignores the geometric structure among samples.

StolenEncoder [18]: Liu et al. introduced a data augmentation strategy, training by minimizing the distance between the target Embedding of the original image and the proxy Embedding of the augmented view.

Cont-Steal (Contrastive Stealing) [26]: Sha et al. further utilized the idea of contrastive learning, not only bringing the positive sample pairs closer but also pushing the negative sample pairs apart using the InfoNCE loss function, significantly enhancing the theft effect by leveraging the mutual exclusion information among samples.

RDA (Refine, Discriminate, and Align) [38]: To address high query costs and target deviation , RDA refines representations via sample-level prototypes and employs a multi-relationship loss to align and discriminate features , achieving state-of-the-art performance with minimal queries.

2.2 Limitations of Existing Defenses

Facing increasingly sophisticated theft, current defense mechanisms are inadequate. Current defenses mainly fall into passive detection and active defense.

Passive Defense: To adapt to the shift to SSL, Cong et al. extended watermarking and detection [1, 6, 13, 30, 35, 39] to pre-trained encoders, watermarking inputs to specific embedding clusters through shadow training, while Wu et al. developed a mechanism to directly embed triggers into the feature space during contrastive learning. However, these defenses aim to verify ownership or detect malicious queries after the fact, failing to prevent functional leakage during theft. Additionally, dataset inference techniques [10, 20] attempt to determine if a stolen model was trained on specific private data by analyzing its memory features. The fundamental flaw of these passive defense methods lies in their lag - they cannot prevent the leakage of model parameters or functions during theft and can only serve as legal evidence for post-theft.

Active Defense: Perturbation-based defenses reduce the thief's learning efficiency by adding noise to the output, truncating the output (Top-k) [38], or rounding the values. Recent studies formulated information-theoretic perturbation strategies to mathematically balance this privacy-utility trade-off [28]. However, simple random perturbations often degrade the utility for normal users while disrupting the thief's performance. Orekondy et al. [23] proposed PREDICTION POISONING for classifiers, aiming to maximize the angular deviation of the attacker's gradient. However, how to apply this idea of targeted perturbation of gradients to the defense of unlabeled, high-dimensional SSL encoders and effectively resist contrastive learning-based theft while ensuring model utility remains an urgent challenge to be addressed. Recent active defenses even resort to complex paradigms like injecting backdoors into the attacker's surrogate model, resulting in high cost [33]. To fill this critical gap, our work introduces the first lightweight, active perturbation defense specifically targeting SSL encoders. Our design scope focuses on disrupting the contrastive and relational alignment channel commonly shared by simple yet reliable stealing algorithms, while leaving the comprehensive evaluation against heavily customized adaptive attacks to future work.

3 Universal Minor Perturbation

3.1 Traditional Perturbation Methods

Traditional PREDICTION POISONING [23] hinders model stealing in SL by maximizing the angle of poisoned gradients. However, in the unlabeled, high-dimensional SSL space, maintaining utility while disrupting attackers remains an urgent challenge [9,19]. Dziedzic et al. [9] demonstrated the difficulty of preserving normal user utility in unsupervised scenarios. If a perturbation opposite to the attacker's loss function is directly added, to reduce the attacker's accuracy from 74.57% to 53.41%, the accuracy of the target encoder would drop from 74.97% to 48.28% [18].

To preserve utility, we formulate a baseline combining a maintenance module (constraining normalized MSE [2] between perturbed and original features), and a gradient-ascent penalty against the StolenEncoder's loss [4]. The loss function of StolenEncoder consists of two parts, where L_1 is the distance between the output features of the target encoder and the proxy encoder, and L_2 is the distance between the output features of the target encoder and the output features of the proxy encoder after data augmentation:

$$L_1 = L_{\text{dist}}(z_{\text{adv}}, f_{\text{t}}(\mathcal{A}(x))), \tag{1}$$

$$L_2 = L_{\text{dist}}(z_{\text{adv}}, f_{\text{t}}(x)), \tag{2}$$

$$\max_{\delta}[\beta L_1 - \gamma L_2] \quad s.t \quad ||\delta||_\infty \leq \varepsilon, \tag{3}$$

where $\delta \leq \epsilon$ is the perturbation, $f_t(\cdot)$ represents the target model, $\mathcal{A}(\cdot)$ is a set of data augmentation transformations, and $z_{adv} = \text{Normalize}(f_t(x) + \delta)$ is the defended feature representation. $L_{dist}(\cdot, \cdot)$ is the distance metric function, and it employs the Mean Squared Error (MSE). We use gradient ascent to maximize the loss function. We follow the constraints of the real world, where the victim does not know the stealing method used by the thief in advance. The best proxy encoder is the target encoder itself, so the victim directly uses its own encoder to complete the two loss functions.

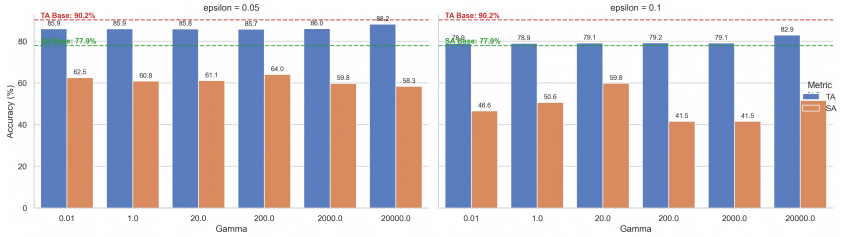


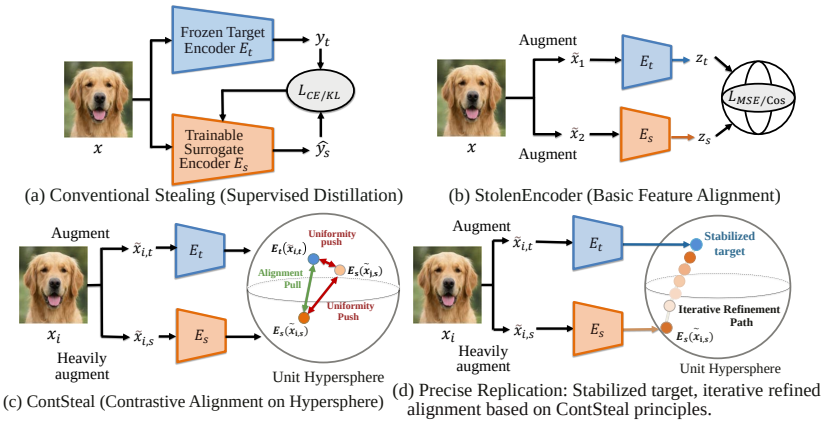
Fig. 3: Ablation study of hyperparameters γ and ϵ on the CIFAR-10 dataset against the RDA attack (Target Encoder: pre-trained CIFAR-10, $\beta = 20$). Horizontal dashed lines represent the undefended baseline (TA: 90.2%, SA: 77.9%). Notably, even with varying γ constraints, the encoder fails to achieve an ideal Pareto frontier between privacy and utility.

By combining the two modules, we evaluated a preliminary distance-oriented baseline (Fig. 1), which still failed to achieve a balance between privacy and utility.

3.2 Exploration of Theft Methods

Instead of accumulating multiple defensive modules, we seek a single mechanism that effectively disrupts the attacker's learning process without harming downstream utility. By projecting both target and surrogate features onto a unit hypersphere, we observe that modern theft still relies on contrastive alignment. As illustrated in Fig. 4, modern stealing attacks such as ContSteal [26] and RDA [38] fundamentally treat theft as a contrastive learning task, and the theft behavior itself is to enhance the alignment and uniformity of feature distribution on the overall sphere.

We speculate that the target encoder's uniformity influences the overall feature space. If it achieves perfect uniformity, the stolen encoder's learning efficiency degrades, preventing it from learning correct semantics via negative sample repulsion. Although the target encoder cannot control the stolen encoder's internal negative sample push-away, it can disrupt the contrastive learning process where the two are connected—specifically, the negative sample push-away between the target and stolen encoders.



(a) Conventional Stealing (Supervised Distillation)

(b) StolenEncoder (Basic Feature Alignment)

(c) ContSteal (Contrastive Alignment on Hypersphere)

(d) Precise Replication: Stabilized target, iterative refined alignment based on ContSteal principles.

Fig. 4: Evolution of model stealing attack paradigms on SSL encoders. (a) **Conventional Stealing**: replicating target outputs via supervised distillation; (b) **StolenEncoder**: achieving basic feature alignment through data augmentation; (c) **ContSteal**: contrastive alignment on the unit hypersphere; (d) **Precise Replication (RDA)**: high-fidelity representation through stabilized targets and iterative refinement.

3.3 UniTrap: Inducing Uniformity via Gaussian Potential

Based on the geometric insight that uniformity disrupts contrastive alignment, we propose **UniTrap**. Unlike distance-based methods, our method explicitly optimizes the global distribution of defensive embeddings on the hypersphere. For a batch of input images $X = \{x_i\}_{i=1}^n$, we generate the defended features $z_{adv,i} = \text{Normalize}(f_t(x_i) + \delta_i)$. The optimal perturbation δ is obtained by minimizing the **Gaussian Potential Function**:

$$\mathcal{L}_{uni}(Z_{adv}) = \sum_{i=1}^{n} \sum_{j \neq i}^{n} \exp\left(-\frac{\|z_{adv,i} - z_{adv,j}\|^2}{t} \right), \tag{4}$$

where t is a temperature parameter that controls the range of repulsive forces between feature points. By minimizing this potential, we force the embeddings to maximize their mutual distance globally, effectively trapping the features in an isotropic uniform state.

To achieve a practical and efficient defense, we employ a **Dynamic FGSM-based** iterative update:

$$\delta^{(k+1)} = \text{Proj}_{\|\delta\|_\infty \leq \epsilon}\left(\delta^{(k)} - \alpha \cdot \text{sign}(\nabla_\delta \mathcal{L}_{uni}) \right), \tag{5}$$

where α is the step size and ϵ is the total perturbation budget. This dynamic optimization ensures that for every query, the returned feature is an active part of the **UniTrap**.

3.4 Theoretical Analysis: Why Uniformity Induces Gradient Vanishing

In this section, we attempt to explain the effectiveness of the proposed defense from a mathematical perspective. The operations are conducted on the unit hypersphere $\mathbb{S}^{d-1}$. Since contrastive stealing methods like ContSteal utilize cosine similarity, the negative sample loss term l_{neg} in contrastive learning can be expressed using the dot product. To simplify the analysis, we focus on the push term (LogSumExp) of the loss function, which is responsible for separating negative pairs:

$$\mathcal{L}_{\text{push}}(z_s, \{z_{t,j}\}) = \log\left(\sum_{j=1}^{N} \exp\left(\frac{z_s^\top z_{t,j}}{\tau}\right)\right), \tag{6}$$

where z_s denotes the surrogate (stolen) feature, $\{z_{t,j}\}$ represents the set of target features (negative samples for z_s), N is the number of negative samples, and τ is the temperature parameter.

Since ContSteal employs gradient descent to optimize the surrogate encoder, we derive the gradient of the loss with respect to the surrogate feature z_s:

$$\nabla_{z_s}\mathcal{L}_{\text{push}} = \sum_{j=1}^{N}\left(\frac{\exp(z_s^\top z_{t,j}/\tau)}{\sum_{k=1}^{N}\exp(z_s^\top z_{t,k}/\tau)}\right)\cdot z_{t,j} = \sum_{j=1}^{N} w_j \cdot z_{t,j}, \tag{7}$$

where the magnitude of the weight w_j depends on the similarity between z_s and $z_{t,j}$; a higher similarity yields a larger weight. Since the output features are normalized, the weight on the unit hypersphere corresponds to the cosine angle between feature vectors: the smaller the angle, the larger the weight. The direction of the gradient component is determined by $z_{t,j}$.

We geometrically interpret this process by placing z_s at the North Pole of the unit hypersphere, with its direction pointing radially outward (vertically upward in Fig. 5). In an ideal scenario with a sufficiently large number of samples ($N \to \infty$), the force exerted on z_s can be approximated as an integral over the distribution of target features $p(z_t)$.

If the target encoder achieves perfect uniformity, meaning the output features are uniformly distributed on the hypersphere, the integral of the gradients will align with z_s due to rotational symmetry:

$$\mathbb{E}_{z_t\sim\text{Uniform}(\mathbb{S}^{d-1})}[\nabla_{z_s}\mathcal{L}_{\text{push}}] \propto \int_{\mathbb{S}^{d-1}} \exp\left(\frac{z_s^\top z}{\tau}\right)\cdot z\, dz = C\cdot z_s, \tag{8}$$

where C is a constant. This indicates that the tangential components of the repulsive forces from the surrounding z_t cancel each other out, leaving only a radial force pointing in the direction of z_s.

However, since the optimization is constrained to the unit hypersphere, the update direction must lie in the tangent space. The effective *Riemannian gradient* is obtained by projecting the Euclidean gradient onto the tangent plane, which removes the radial component:

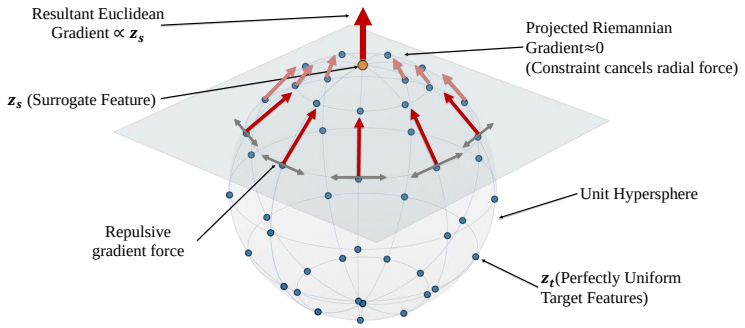


Fig. 5: Geometric interpretation of gradient vanishing under perfect target uniformity. When the target features z_t are perfectly uniformly distributed, the repulsive gradients (red arrows) exerted on the surrogate feature z_s cancel out in the tangential plane (gray arrows). The resultant force is purely radial (large red arrow). Since the optimization is constrained to the unit hypersphere, the effective Riemannian gradient (projection onto the tangent plane) becomes zero.

$$\nabla_{\text{Riemannian}}\mathcal{L} = \nabla_{z_s}\mathcal{L} - (\nabla_{z_s}\mathcal{L}^\top z_s)z_s \approx C \cdot z_s - (C \cdot z_s^\top z_s)z_s = 0. \tag{9}$$

Consequently, the effective force on z_s becomes zero. Therefore, when the target encoder reaches "perfect uniformity," the attacker cannot learn correct semantic representations by pushing away negative samples.

Although Wang and Isola [32] emphasized the trade-off between alignment and uniformity in contrastive representation learning, our analysis suggests that in the context of defense, enforcing extreme uniformity [11] on the target encoder serves as a subtle yet effective strategy. Explicit feature uniformity induces robustness against structured noise [37], so unlike other violent perturbation methods, this uniformity-enhancing perturbation acts as a trap that stalls the learning process of the stolen encoder while remaining benign to the original task utility.

3.5 Experimental Validation: Training Dynamics and Semantic Analysis

Experimental Setup and Worst-case Assumption. To evaluate the efficacy of the proposed defense, we conduct a quantitative analysis of the similarity dynamics for ContSteal and RDA. The target encoder is pre-trained on the CIFAR-10 dataset. To simulate the *worst-case assumption* from a defender's perspective, the defense perturbations are generated using gradients from the frozen target encoder. The perturbation budget ϵ is set to 0.3.

Dual Collapse of Similarity. Fig. 6 and Fig. 7 illustrates the evolution of positive pair similarity (alignment) and negative pair similarity (uniformity) for ContSteal and RDA under both baseline and defended scenarios. Upon receiving the uniformity-enhancing perturbation, both metrics exhibit a significant decline.

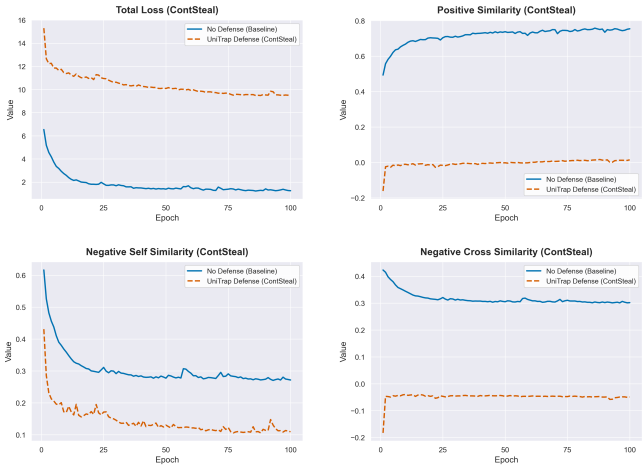


Fig. 6: Feature similarity under ContSteal attack. Under defense, both positive and negative similarities remain locked at extremely low levels, demonstrating that positive pairs fail to align semantically, while negative pairs fail to provide effective gradients.

In the undefended baseline, positive similarity climbs significantly as epochs progress, accompanied by the rational convergence of negative similarity, indicating that the surrogate model is establishing a semantic mapping to the target model. However, under the proposed defense, these key indicators are locked near the orthogonal zero-point throughout the training cycle, showing no effective growth. This dynamic analysis reveals that the defense's effectiveness is not merely an initial perturbation but a continuous inhibition of feature correlation during the entire optimization process.

Total Loss Landscape Flattening and Optimization Stagnation. This dual collapse not only severs semantic associations but also renders the gradient signals within the contrastive loss function weak and directionless. Since contrastive learning relies on the relative discriminability between positive and negative pairs to drive optimization, their degradation into high-dimensional random noise causes the loss landscape to become extremely flat. As a result, the optimization of the surrogate model falls into permanent stagnation, as evidenced by the high plateau of the total loss in Fig. 6 and Fig. 7.

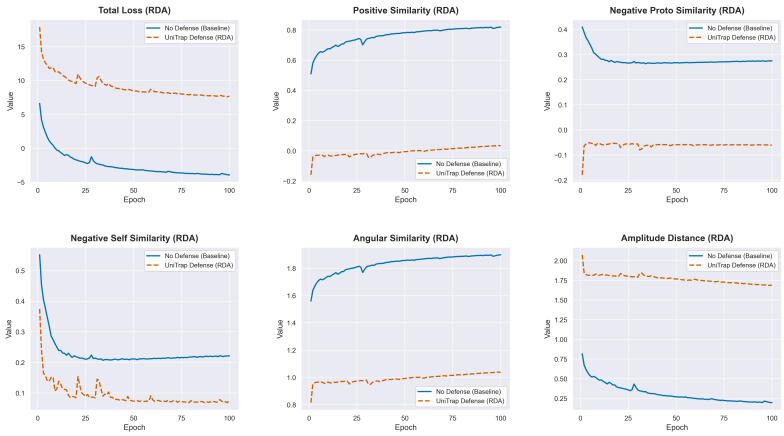


Fig. 7: Feature similarity under RDA attack. The defense keeps positive/negative similarities near zero. Furthermore, it severely disrupts the normal evolution of Angular Similarity and Amplitude Distance, blocking the attacker's multi-relational feature space reconstruction.

Failure of Semantic Recognition and Clustering. The final impact of the defense is further validated by the visual evidence of feature distributions. As shown in Fig. 8, the stolen features from ContSteal and RDA under defense fail to perform normal semantic recognition and are unable to form coherent clusters. This visual isotropic distribution confirms that the defense successfully traps the attacker in a uniformity trap, rendering the stolen encoder useless for downstream tasks.

4 Experiments

4.1 Experimental Setup and Parameters

Configurations: We use ResNet-based extractors on CIFAR-10. **UniTrap** parameters include the perturbation budget $\epsilon = 0.3$ and Gaussian potential temperature $t = 2.0$. The query budget is fixed at 2,500 samples.

Attacks: Conventional and StolenEncoder use temperature parameter $\tau = 0.07$. ContSteal employs a dynamic schedule: $\tau = 1 + 0.05 \times \frac{epoch}{total\_epochs}$. For **RDA**, multi-view patching is used to enhance feature disentanglement, with query patch set to 5 and proto patch set to 10. To achieve precise prototype estimation, a KNN classifier is used as a guide with $k = 200$ neighbors and a soft-assignment temperature $knn\_t = 0.5$.

Evaluation: To comprehensively assess the privacy-utility trade-off, we define two primary metrics: the empirical performance of the attacker's surrogate model (i.e., **Stolen Accuracy, SA**) and the preserved downstream utility of the defended target encoder (i.e., **Target Accuracy, TA**). Both SA and TA are evaluated via a two-layer MLP (512-256) using the Adam optimizer ($lr = 0.0001$, batch size 100) over 100 epochs.

Fig. 8: t-SNE visualization of feature distributions. The defended models exhibit high isotropy and fail to form meaningful class clusters compared to the baseline.

4.2 Experimental Results under Different Target Encoders

To verify the generalization of the defense mechanism across different source architectures and pre-training distributions, we evaluate target encoders pre-trained on CIFAR-10, STL-10, and ImageNet against RDA attacks, and the surrogate encoder is pre-trained on the ImageNet dataset. Table 1 shows the changes in Stolen Accuracy (SA) before and after protection.

Table 1: Cross-encoder Defense Performance Comparison (RDA Attack, $\epsilon = 0.3$)

Target Encoder	Metric	CIFAR-10		STL-10		F-MNIST		SVHN	
		None	Defense	None	Defense	None	Defense	None	Defense
STL-10	SA (%)	75.50	**48.28**	69.34	**44.10**	89.21	**67.93**	71.54	**46.73**
	TA (%)	85.99	82.94	80.10	77.09	90.51	88.12	61.24	54.47
CIFAR-10	SA (%)	77.90	**35.26**	70.24	**37.34**	83.06	**81.20**	73.85	**25.96**
	TA (%)	91.20	90.18	80.55	77.36	88.87	86.48	62.12	57.26
ImageNet	SA (%)	72.82	**41.15**	72.47	**37.70**	89.72	**67.21**	60.03	**19.81**
	TA (%)	90.59	89.48	95.66	95.39	92.01	90.24	73.74	68.78

* SA: Stolen Accuracy (represents defense robustness); TA: Target Accuracy (represents model utility). The surrogate encoder is pre-trained on ImageNet.

Universal Effectiveness of the Defense Mechanism: Experimental results indicate that regardless of the pre-training source of the target encoder, the performance of the stolen model drops significantly after applying the *UniTrap*

defense. For instance, with a CIFAR-10 pre-trained target model, the defense reduces the SA on the CIFAR-10 task from 77.90% to 35.26%, and on the SVHN task from 73.85% to 25.96%.

Robustness of Cross-Domain Feature Protection: Notably, the defense remains robust even when there is a domain gap between the target and surrogate encoders. In experiments with an ImageNet pre-trained target encoder, despite its significantly more complex feature distribution compared to CIFAR-10, the defense still suppresses the SA on the SVHN task from 60.03% to 19.81%. This suggests that dynamic perturbations based on Gaussian potential go beyond simple noise injection; they disrupt transferable underlying semantic correlations by targeting the intrinsic distribution of the feature space.

Balance between Defense and Utility: While significantly degrading the stolen model's performance, we also monitor the Target Accuracy (TA) of the protected model. Data shows that the performance loss of the protected target encoder on its original task is minimal (typically within 1%–3%), with the CIFAR-10 target encoder maintaining a TA of approximately 90%. This "high defense, low loss" phenomenon validates that the Gaussian potential term successfully balances privacy protection and representation utility by maintaining feature uniformity.

4.3 Comparison of Dynamic Perturbation Strategies: Distance-Oriented Baseline vs. UniTrap

We compare **UniTrap** with a distance-oriented baseline designed to uniformly push features apart. Preliminary evaluations reveal that MSE, Cross-Entropy, and Cosine Similarity yield comparable, sub-optimal defense efficacy. Consequently, we report the Cosine variant as a representative baseline in Table 2 (MSE and CE results are deferred to the Appendix), evaluated under $\epsilon = 0.3$.

Enhancement of Defense Efficiency through Representation Uniformity: Table 2 clearly shows that *UniTrap* significantly outperforms the distance-oriented baseline in suppressing stolen model accuracy. Against the potent *StolenEncoder* attack, *UniTrap* suppresses the SA to 36.27% and 37.79% on CIFAR-10 and STL-10, respectively—a defense gain of nearly 10% over the distance-oriented strategy. For the RDA attack, the trend is similar, with an additional reduction in SA of approximately 5% to 9%. Remarkably, while defense strength is significantly improved, the impact on the target model's own utility (Target Acc) is negligible, with differences within 0.1% to 0.2%.

Theoretical Analysis: Following [32], representation learning hinges on alignment and uniformity. While simple distance maximization disrupts alignment, it risks feature collapse or unnatural clusters that attackers like RDA can exploit. Conversely, **UniTrap** explicitly optimizes uniformity via a Gaussian potential, increasing feature entropy and preventing attackers from disentangling semantic prototypes.

Table 2: Comparison of Defense Efficacy: Baseline (None) vs. Distance-Oriented vs. UniTrap ($\epsilon = 0.3$)

Dataset	Defense Strategy	Stolen Accuracy (SA %) ↓				Utility ↑
		Conv.	StolenEnc.	ContSteal	RDA	Target Acc (%)
CIFAR-10	None (Baseline)	66.98	74.67	78.22	77.90	91.20
	Distance-Oriented (Cosine)	28.37	44.10	**37.38**	40.35	**90.19**
	UniTrap	**27.78**	**36.27**	37.77	**35.26**	90.18
STL-10	None (Baseline)	57.90	65.60	71.69	70.24	80.55
	Distance-Oriented (Cosine)	**20.27**	47.67	35.80	46.52	**77.54**
	UniTrap	20.39	**37.79**	**29.94**	**37.34**	77.36
SVHN	None (Baseline)	60.59	69.78	75.15	73.85	62.12
	Distance-Oriented (Cosine)	21.15	38.70	33.59	42.43	56.41
	UniTrap	**19.97**	**28.96**	**27.88**	**25.96**	**57.26**
F-MNIST	None (Baseline)	86.44	88.40	89.73	83.06	88.87
	Distance-Oriented (Cosine)	60.04	78.55	**73.27**	**79.79**	**86.66**
	UniTrap	**50.95**	**68.81**	82.87	81.20	86.48
CIFAR-100	None (Baseline)	32.62	38.96	44.84	44.09	53.27
	Distance-Oriented (Cosine)	6.44	16.54	**12.83**	15.06	49.46
	UniTrap	**6.22**	**11.82**	13.16	**12.31**	**49.64**

* SA (Stolen Accuracy) measures the success of the stealing attack (lower is better for defense). Target Acc measures the utility of the original model (higher is better). Bold values indicate the best defense performance (lowest SA or highest TA) among the two defense strategies.

4.4 Comparison with Baseline Defenses

We evaluate the SA-TA trade-off of **UniTrap** against common protections (**Top-k** dimension retention, **Gaussian Noise**, and **Rounding**), using CIFAR-10 for both target pre-training and downstream evaluation.

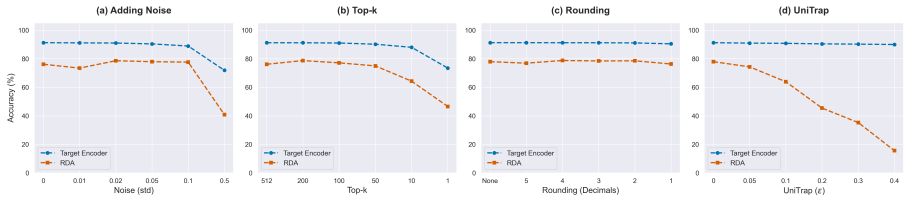


Fig. 9: Trade-off between privacy (i.e., SA) and utility (i.e., TA) under different defense mechanisms.

Pareto Superiority of Defense Efficacy: As shown in Figure 9, our *UniTrap* defense significantly outperforms traditional methods on the privacy-utility trade-off curve. At $\epsilon = 0.3$, the SA is suppressed to 35.26% while the TA remains high at 90.18%. In contrast, for the *Top-k* defense to reach a similar SA level (46.53%), the TA drops sharply to 73.44%. This indicates that while traditional defenses reduce the attacker's gain, they severely damage the semantic integrity of the features.

Overcoming Limitations of Static Perturbations:

- **Failure of Noise and Rounding:** *Rounding* fails to effectively reduce SA across all settings, with accuracy remaining above 76%. While *Noise* can reduce SA, it causes a simultaneous collapse in TA—at a noise intensity of 0.5, the SA drops to 40.8%, but the TA falls to 71.85%.
- **Efficiency of UniTrap:** Our method uses gradient information for dynamic optimization, allowing minute perturbations ($\epsilon = 0.3$) to induce significant semantic shifts on the hypersphere. Even at a strong configuration ($\epsilon = 0.4$), the SA further drops to 15.59% while the TA stays robust at 89.92%.

In summary, traditional defenses like *Top-k* and *Noise* often sacrifice excessive feature information for limited security due to a lack of understanding of the feature space distribution. Our proposed dynamic defense based on Gaussian potential achieves high-intensity privacy protection with minimal utility loss by optimizing representation uniformity.

5 Conclusion

In this paper, we propose the first active perturbation-based geometric defense against contrastive model stealing attacks on self-supervised encoders. Moving beyond traditional heuristic distance maximization, we introduce the **UniTrap**, a mechanism that explicitly optimizes feature entropy on the unit hypersphere via a Gaussian potential function. We demonstrate that inducing a perfectly uniform (isotropic) feature distribution forces the attacker's contrastive gradients to cancel out, permanently freezing their optimization process.

Experimental results across multiple architectures and downstream benchmarks demonstrate that our proposed **UniTrap** achieves a state-of-the-art accuracy-privacy trade-off. Specifically, under a standard $\epsilon = 0.3$ budget, we successfully reduced the stolen accuracy of the RDA attack from 77.90% to 35.26% while preserving a high target utility of 90.18%. Furthermore, our comparative analysis reveals that enforcing uniformity is fundamentally more robust than simple cosine-distance maximization or traditional noise-based defenses. Our work highlights the critical role of embedding geometry in SSL security and provides a robust framework for protecting intellectual property in the era of EaaS.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62572083).

References

1. Adi, Y., Baum, C., Cisse, M., Pinkas, B., Keshet, J.: Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In: 27th USENIX security symposium (USENIX Security 18). pp. 1615–1631 (2018)
2. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning **36**(1), 105–139 (1999)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chandrasekaran, V., Chaudhuri, K., Giacomelli, I., Jha, S., Yan, S.: Exploring connections between active learning and model extraction. In: 29th USENIX Security Symposium (USENIX Security 20). pp. 1309–1326 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
6. Cong, T., He, X., Zhang, Y.: Sslguard: A watermarking scheme for self-supervised learning pre-trained encoders. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. pp. 579–593 (2022)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
8. Dubiński, J., Pawlak, S., Boenisch, F., Trzcinski, T., Dziedzic, A.: Bucks for buckets (b4b): Active defenses against stealing encoders. Advances in Neural Information Processing Systems **36**, 55237–55259 (2023)
9. Dziedzic, A., Dhawan, N., Kaleem, M.A., Guan, J., Papernot, N.: On the difficulty of defending self-supervised learning against model extraction. In: International Conference on Machine Learning. pp. 5757–5776. PMLR (2022)
10. Dziedzic, A., Duan, H., Kaleem, M.A., Dhawan, N., Guan, J., Cattan, Y., Boenisch, F., Papernot, N.: Dataset inference for self-supervised models. Advances in Neural Information Processing Systems **35**, 12058–12070 (2022)
11. Fang, X., Li, J., Sun, Q., Wang, B.: Rethinking the uniformity metric in self-supervised learning. arXiv preprint arXiv:2403.00642 (2024)
12. Gao, L., Liu, W., Liu, K., Wu, J.: Augsteal: Advancing model steal with data augmentation in active learning frameworks. IEEE Transactions on Information Forensics and Security **19**, 4728–4740 (2024)
13. Jia, H., Choquette-Choo, C.A., Chandrasekaran, V., Papernot, N.: Entangled watermarks as a defense against model extraction. In: 30th USENIX security symposium (USENIX Security 21). pp. 1937–1954 (2021)
14. Jia, J., Liu, Y., Gong, N.Z.: Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 2043–2059. IEEE (2022)
15. Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M.: Thieves on sesame street! model extraction of bert-based apis. arXiv preprint arXiv:1910.12366 (2019)
16. Li, P., Lv, P., Chen, K., Zhang, S., Cai, Y., Xiang, F.: A model stealing attack against multi-exit networks. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)

17. Liu, H., Jia, J., Qu, W., Gong, N.Z.: Encodermi: Membership inference against pre-trained encoders in contrastive learning. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. pp. 2081–2095 (2021)
18. Liu, Y., Jia, J., Liu, H., Gong, N.Z.: Stolenencoder: stealing pre-trained encoders in self-supervised learning. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. pp. 2115–2128 (2022)
19. Luan, S., Wang, Z., Shen, L., Gu, Z., Wu, C., Tao, D.: Dynamic neural fortresses: An adaptive shield for model extraction defense. In: The Thirteenth International Conference on Learning Representations (2025)
20. Maini, P., Yaghini, M., Papernot, N.: Dataset inference: Ownership resolution in machine learning. arXiv preprint arXiv:2104.10706 (2021)
21. Morris, J., Kuleshov, V., Shmatikov, V., Rush, A.M.: Text embeddings reveal (almost) as much as text. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 12448–12460 (2023)
22. Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: Stealing functionality of black-box models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4954–4963 (2019)
23. Orekondy, T., Schiele, B., Fritz, M.: Prediction poisoning: Towards defenses against dnn model stealing attacks. arXiv preprint arXiv:1906.10908 (2019)
24. Rao, B., Zhang, J., Wu, D., Zhu, C., Sun, X., Chen, B.: Privacy inference attack and defense in centralized and federated learning: A comprehensive survey. IEEE Transactions on Artificial Intelligence (2024)
25. Ren, X., Liang, H., Wang, Y., Zhang, C., Xiong, Z., Zhu, L.: Besa: Boosting encoder stealing attack with perturbation recovery. IEEE Transactions on Information Forensics and Security (2025)
26. Sha, Z., He, X., Yu, N., Backes, M., Zhang, Y.: Can't steal? cont-steal! contrastive stealing attacks against image encoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16373–16383 (2023)
27. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
28. Tang, M., Dai, A., DiValentin, L., Ding, A., Hass, A., Gong, N.Z., Chen, Y., et al.: {ModelGuard}:{Information-Theoretic} defense against model extraction attacks. In: 33rd USENIX Security Symposium (USENIX Security 24). pp. 5305–5322 (2024)
29. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction {APIs}. In: 25th USENIX security symposium (USENIX Security 16). pp. 601–618 (2016)
30. Uchida, Y., Nagai, Y., Sakazawa, S., Satoh, S.: Embedding watermarks into deep neural networks. In: Proceedings of the 2017 ACM on international conference on multimedia retrieval. pp. 269–277 (2017)
31. Wang, B., Gong, N.Z.: Stealing hyperparameters in machine learning. In: 2018 IEEE symposium on security and privacy (SP). pp. 36–52. IEEE (2018)
32. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
33. Wang, Y., Gu, T., Teng, Y., Wang, Y., Ma, X.: Honeypotnet: Backdoor attacks against model extraction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8087–8095 (2025)

34. Wang, Z., Lin, M., Shen, B., Anderson, K., Liu, M., Cai, T., Dong, Y.: Cega: A cost-effective approach for graph-based model extraction and acquisition. arXiv preprint arXiv:2506.17709 (2025)
35. Wang, Z., Wu, Y., Huang, H.: Defense against model extraction attack by bayesian active watermarking. In: Forty-first International Conference on Machine Learning (2024)
36. Wu, B., Yang, X., Pan, S., Yuan, X.: Model extraction attacks on graph neural networks: Taxonomy and realisation. In: Proceedings of the 2022 ACM on Asia conference on computer and communications security. pp. 337–350 (2022)
37. Wu, M., Sun, Q., Yang, Y.: Pca++: How uniformity induces robustness to background noise in contrastive learning. arXiv preprint arXiv:2511.12278 (2025)
38. Wu, S., Ma, C., Wei, K., Xu, X., Ding, M., Qian, Y., Xiao, D., Xiang, T.: Refine, discriminate and align: Stealing encoders via sample-wise prototypes and multi-relational extraction. In: European Conference on Computer Vision. pp. 186–203. Springer (2024)
39. Wu, Y., Qiu, H., Zhang, T., Li, J., Qiu, M.: Watermarking pre-trained encoders in contrastive learning. In: 2022 4th International Conference on Data Intelligence and Security (ICDIS). pp. 228–233. IEEE (2022)
40. Xiao, R., Dong, C., Zhang, J., Pang, Y., Xie, Z., Wu, H.: I can still steal your encoder: A defense-penetrating encoder-stealing attack. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 483–501. Springer (2025)
41. Yuan, X., Chen, K., Huang, W., Zhang, J., Zhang, W., Yu, N.: Data-free hard-label robustness stealing attack. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6853–6861 (2024)
42. Zhang, C., Ren, X., Liang, H., Fan, Q., Tang, X., Li, C., Zhu, L., Wang, Y.: Data-free encoder stealing attack in self-supervised learning. In: International Conference on Algorithms and Architectures for Parallel Processing. pp. 100–120. Springer (2024)
43. Zhao, K., Li, L., Ding, K., Gong, N.Z., Zhao, Y., Dong, Y.: A survey on model extraction attacks and defenses for large language models. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. pp. 6227–6236 (2025)