Fingerprinting Pre-trained Encoders under Arbitrary Downstream Fine-Tuning via Adversarial Shifting

Tianlong Xu[1], Zixiong Wang[1], Lishuai Hou[1],

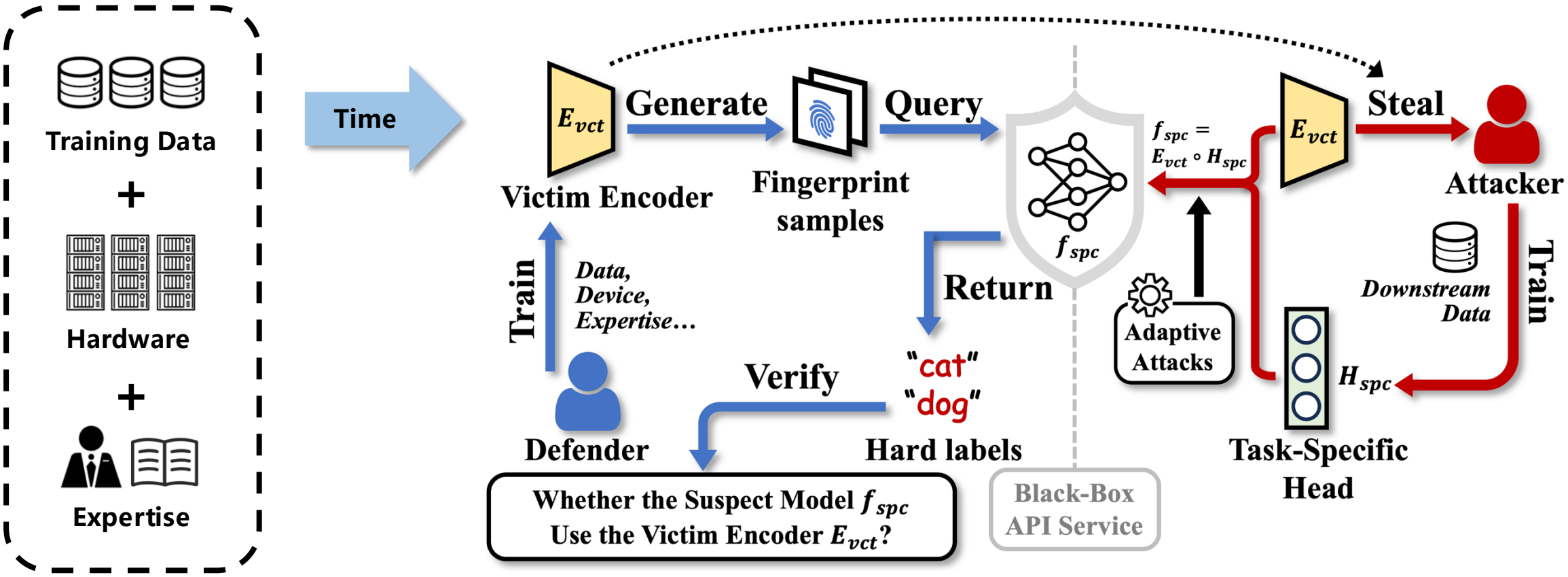
Gaoyang Liu[1*], Chen Wang[1], Xiaoyi Fan[2]

[1] Hubei Key Laboratory of Smart Internet Technology, School of EIC,
Huazhong University of Science and Technology

[2] Jiangxing Intelligence Technology Inc.

ICML 2026, Seoul, South Korea

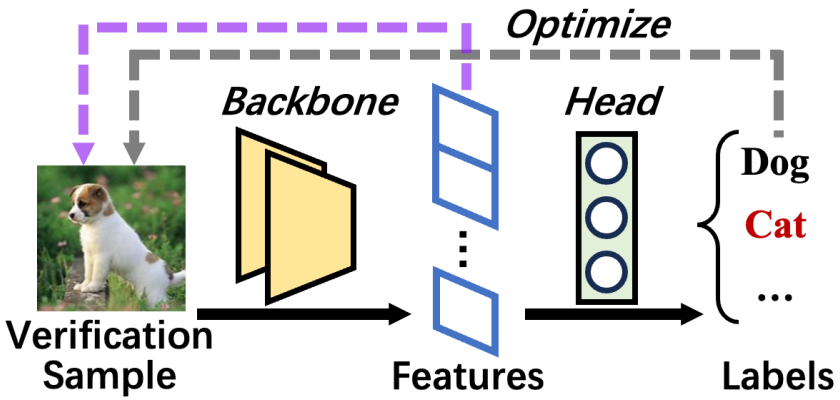
Background: Encoders as the Backbone of AI Economy



Current Threat

Adversaries steal encoders and adapt them to *arbitrary*, *unknown downstream tasks* via *black-box* APIs.

Motivation: The Breakdown of Label-based Protection

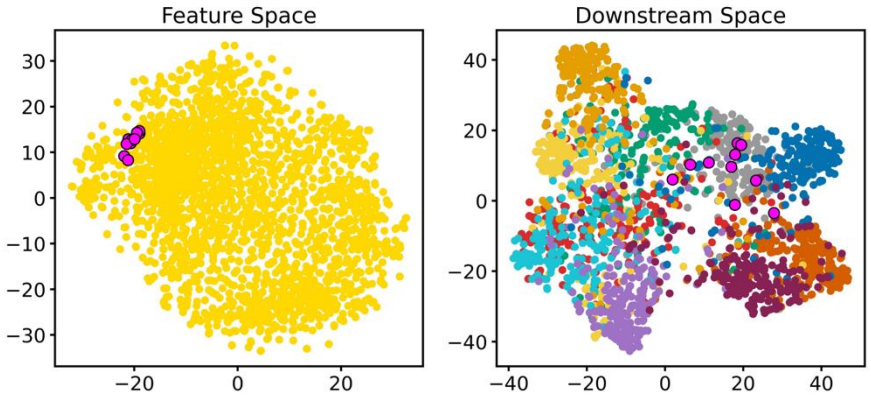


Label-based vs. Feature-based Optimization: By shifting optimization from the volatile label space (grey line) to the stable backbone feature space (purple line), our method achieves downstream-agnostic verification.

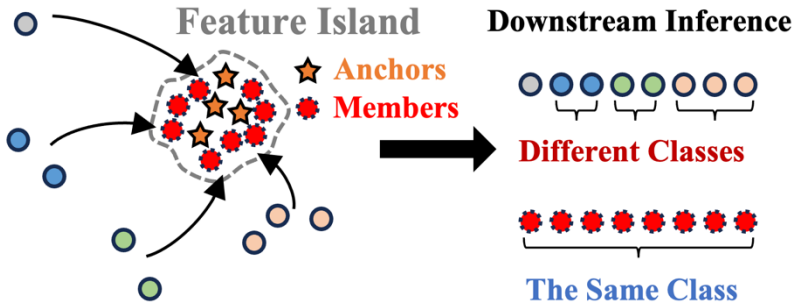
Challenges:

- **Label Space Mismatch:** Downstream tasks change the output semantics (e.g., labels change from 'A/B' to 'Cat/Dog').
- **Decision Boundary Shift:** Fine-tuning reshapes the boundary, making traditional adversarial triggers fail.
- **Verification Dilemma:** Black-box APIs hide internal embeddings, leaving only volatile labels for verification.

Key Idea: Anchoring Fingerprints in Invariant Feature Space



Observation: Feature space proximity implies prediction consistency.

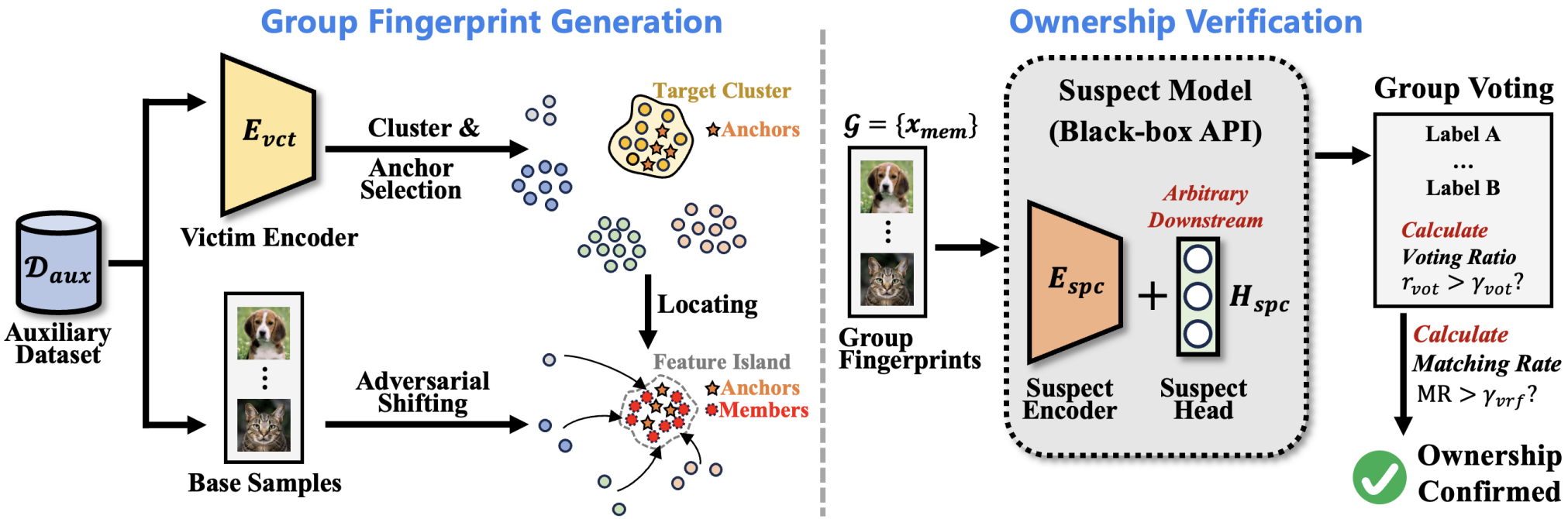


Solution: Constructing a ***Feature Island*** that any downstream head must assign to the same class.

Key Principle:

- **Shift from Labels to Features:** Don't optimize for what the model says (Labels), optimize for where the model represents (Features).
- **Mechanism:** Adversarial Shifting → Creates a cohesive ***Feature Island***.
- **Invariant Property:** Because members are extremely close in feature space, any subsequent classifier head will yield consistent predictions.

Method: A Two-Phase Verification Framework



1. Group Fingerprint Generation

- Cluster feature space and select stable ***Anchors***.
- Optimize base samples to form a "***Feature Island***".

2. Ownership Verification

- Query suspect black-box API with group members.
- Verify via ***Group Voting*** (prediction consensus).

Experimental Results

Main Performance

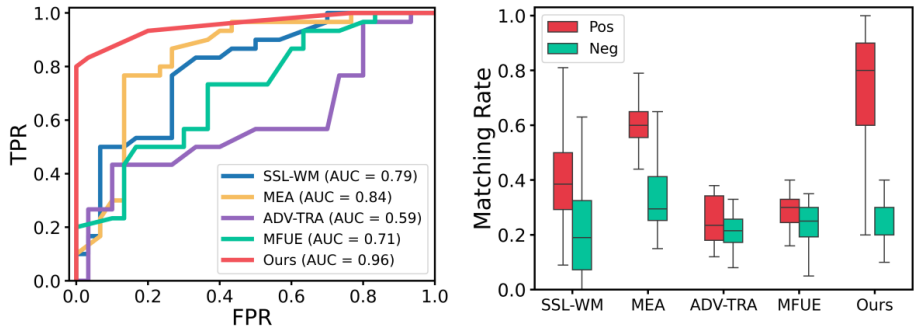
Table 1. Main performance comparison with four state-of-the-art methods across various pre-training paradigms and downstream tasks. We report the Clean Data Accuracy (CDA) and Matching Rate under two downstream adaptation strategies: FTLL and FTAL.

Training Method	Downstream Task	CDA (FTLL / FTAL)			Matching Rate (FTLL / FTAL)				
		Clean	SSL-WM	MEA	SSL-WM	MEA	ADV-TRA	MFUE	Ours
SimCLR	STL-10	0.791/0.834	0.776/0.819	0.783/0.830	0.357/0.298	–/0.751	0.323/0.287	0.426/0.392	**0.960/0.760**
	GTSRB	0.772/0.895	0.763/0.884	0.760/0.882	0.583/0.466	–/0.623	0.396/0.336	0.403/0.354	**0.960/0.800**
	CIFAR-100	0.460/0.473	0.439/0.454	0.457/0.466	0.484/0.291	–/0.601	0.280/0.230	0.320/0.279	**0.920/0.660**
MoCoV2	STL-10	0.880/0.839	0.862/0.835	0.871/0.837	0.462/0.763	–/0.717	0.341/0.292	0.451/0.406	**1.000/0.800**
	GTSRB	0.862/0.927	0.853/0.921	0.858/0.928	0.465/0.349	**–/0.798**	0.458/0.379	0.428/0.378	**0.940**/0.740
	CIFAR-100	0.477/0.483	0.457/0.465	0.466/0.478	0.540/0.651	–/0.589	0.263/0.218	0.337/0.295	**0.900/0.680**
SigLIP	STL-10	0.776/0.801	–	–	–	–	0.318/0.277	0.389/0.353	**0.920/0.760**
	GTSRB	0.754/0.835	–	–	–	–	0.385/0.331	0.365/0.341	**0.940/0.720**
	CIFAR-100	0.373/0.419	–	–	–	–	0.254/0.208	0.256/0.226	**0.940/0.640**
Supervised Learning	STL-10	0.821/0.892	0.803/0.883	0.813/0.890	0.628/0.470	–/0.730	0.307/0.269	0.434/0.390	**1.000/0.840**
	GTSRB	0.886/0.983	0.873/0.970	0.882/0.971	0.231/0.168	–/0.745	0.475/0.400	0.413/0.359	**1.000/0.820**
	CIFAR-100	0.492/0.533	0.472/0.516	0.482/0.529	0.355/0.230	–/0.629	0.279/0.229	0.330/0.288	**0.960/0.780**
	SNLI	0.788/0.838	0.770/0.812	0.766/0.830	0.132/0.101	–/0.662	–	–	**0.860/0.700**
	MRPC	0.693/0.740	0.687/0.736	0.690/0.729	0.207/0.155	–/0.576	–	–	**0.820/0.640**
Additional Knowledge	–	–	–	–	Output Logits	Encoder Output	Downstream Dataset&Head	Downstream Dataset&Head	**None**

1 MEA scores 1.0 under FTLL, a trivial result since it depends directly on the frozen encoder's output, rendering this comparison meaningless.
2 "–" denotes that the method is not applicable to the specific model architecture or task.

- **FTLL :** Achieves near-perfect matching rates (>0.9).
- **FTAL :** Maintains strong resilience even when all parameters are updated, vastly outperforming baselines.

Uniqueness



(a) ROC Curve *(b)* Distribution

Figure 4. On CIFAR-100, we present (a) the ROC curve, (b) the distribution of positive and negative models.

- **Stark Margin:** A clear, non-overlapping margin between positive and negative models ensures effortless identification (AUC 0.96).

Experimental Results

Rationale

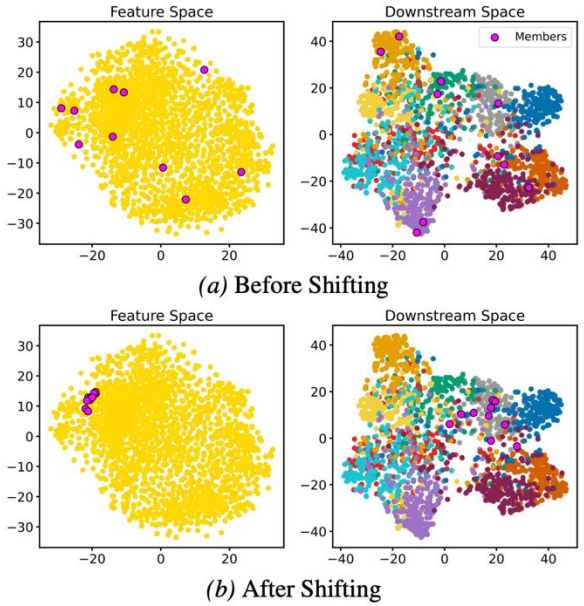
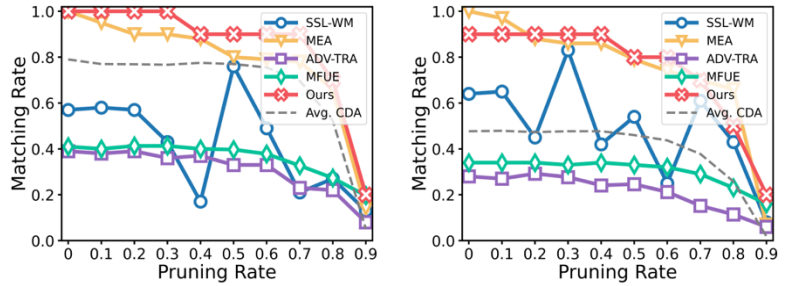


Figure 6. t-SNE visualization of the feature and downstream spaces on STL-10, comparing distributions before (a) and after (b) adversarial shifting. Colors in the downstream space represent the model-predicted classes.

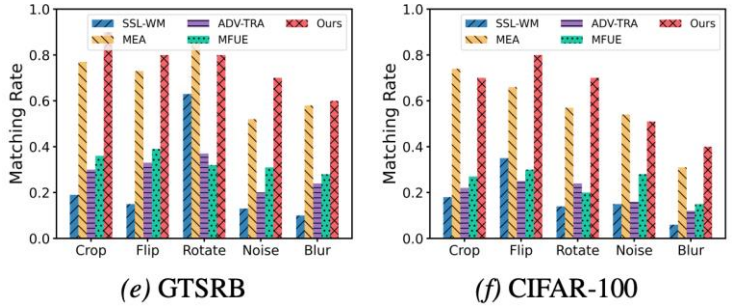
- **Visual Proof**: t-SNE confirms that adversarial shifting successfully aggregates scattered samples into a cohesive "***Feature Island***" that propagates to the downstream space.

Robustness

- Survives severe pruning (up to 80%) and model extraction attacks.



- Voting mechanism naturally mitigates single-trigger fragility against input noise/blur.



Thanks for your attention!

For more information, please refer to our paper.

Code: https://github.com/SPHelixLab

Contact: liugaoyang@hust.edu.cn