Selection-Aware Poisoning: Boosting Clean-Label Backdoor Attacks via Distribution Deviation and Projection Residual Metrics

1st Xin Wang
College of Information Engineering
Capital Normal University
Beijing, China
Email: wxcurry223@gmail.com

2nd Liming Liu
College of Information Engineering
Capital Normal University
Beijing, China
Email: liuliming1228@gmail.com

3rd Xu Han
College of Information Engineering
Capital Normal University
Beijing, China
Email: hanxu@cnu.edu.cn

4th Yuanbo Li
College of Information Engineering
Capital Normal University
Beijing, China
Email: yuanbomuzi@gmail.com

5st Jun Li*
College of Information Engineering
Capital Normal University
Beijing, China
Email: junmuzi@gmail.com

6th Lei Zhang
College of Information Engineering
Capital Normal University
Beijing, China
Email: lei.zhang@cnu.edu.cn

***Abstract*—Despite the unprecedented success of Deep Neural Networks (DNNs) in critical domains, they remain vulnerable to clean-label backdoor attacks—stealthy threats that endanger real-world deployments. Previous work confirms uneven contributions of training samples to backdoor injection, making sample selection pivotal to attack efficacy. Existing mining methods rely on local density metrics or auxiliary out-of-distribution (OOD) data, but overlook the global statistical distribution of the target class and incur high costs from surrogate model training. To address these issues, we propose two training-free sample scoring strategies focusing on global distributional analysis rather than local heuristics: the Distribution Deviation Metric, which uses robust global covariance to quantify sample deviation from the target class, and the Projection Residual Metric, which gauges structural divergence from the intra-class manifold via subspace reconstruction errors. Extensive experiments on representative datasets and diverse architectures show that our strategies achieve superior attack performance and effectively circumvent a variety of mainstream backdoor defenses. This work provides a novel perspective for efficient backdoor attack sample selection and offers insights into safeguarding DNNs against stealthy clean-label backdoor threats. Our code is available at https://github.com/ByteTitan-star/ddm-prm-backdoor.**

***Index Terms*—Backdoor Attack, Clean-Label Attack, Data Poisoning, Hard Sample Mining, Backdoor Defense**

I. Introduction

Deep Neural Networks (DNNs) excel in critical tasks from computer vision [1] to natural language processing [2], yet their robustness often falters in real-world deployment. Backdoor attacks pose a particularly insidious risk: attackers inject trigger-embedded samples into training data, inducing a malicious mapping. Unlike traditional attacks with label inconsistencies, clean-label backdoor attacks [3] maintain semantic consistency, evading human inspection and label-based filters. This covert nature threatens high-assurance applications like biomedicine [4], finance [5], and autonomous driving [6].

While backdoor research has matured, the strategic optimization of the poisoning subset remains an under-explored dimension. Early strategies predominantly utilized random sampling to select candidates. This approach implicitly assumes that all training samples are equally conducive to backdoor implantation—an assumption that is fundamentally flawed. In practice, the majority of samples are robust and contribute little to the formation of the malicious mapping. Consequently, to achieve a sufficient attack impact using inefficient random selection, attackers are forced to drastically increase the poisoning ratio, thereby elevating the risk of exposure. Recent studies have established that training samples exhibit disparate influence on backdoor injection, identifying sample selection as a pivotal determinant for attack efficacy [7]. While this insight has prompted a shift towards hard sample mining, existing strategies remain suboptimal. Current approaches largely rely on simplistic geometric heuristics, failing to capture the intrinsic, high-dimensional feature distribution of the target class. Furthermore, state-of-the-art mining techniques often necessitate auxiliary out-of-distribution (OOD) data to train surrogate models [8]. This dependency imposes significant computational burdens and operational complexity, as it requires the attacker to bear the additional cost of training surrogate models, which is often impractical in resource-constrained scenarios.

To address these challenges, we propose two novel, training-free sample hardness scoring strategies suitable for the clean-label poisoning setting, aiming to identify the optimal candidates for backdoor implantation using only target-class data. Our core hypothesis is that samples diverging from the standard statistical distribution of the class are less firmly

* Corresponding author

anchored by benign features, allowing the injected trigger to dominate the prediction logic more easily. To operationalize this hypothesis, we propose: (1) Distribution Deviation Metric (DDM) measures the extent to which a sample deviates from the overall distribution of the target class in a global statistical sense. By leveraging class-level statistical characteristics, DDM captures distributional inconsistency rather than local distance-based irregularities. (2) The Projection Residual Metric (PRM), a feature subspace-based metric. By constructing a linear subspace spanned by the top principal components, PRM measures the reconstruction error of a sample when projected onto this subspace. A high residual indicates that the sample contains unique structural variations that cannot be explained by the primary class features, marking it as a potent anchor for backdoor injection. In summary, our main contributions are as follows:

- **Global Statistical Mining via DDM (Addressing Local Heuristics):** To overcome the limitations of existing methods that rely on simplistic local density heuristics, we propose the Distribution Deviation Metric. By exploiting robust global covariance statistics, DDM shifts the focus to global distributional analysis, effectively capturing statistical outliers that distort the decision boundary without being misled by local geometric irregularities.
- **Intrinsic Subspace Mining via PRM (Eliminating OOD Dependency):** To address the impractical reliance on auxiliary OOD data or surrogate models for boundary approximation, we propose the Projection Residual Metric. This strategy measures the structural divergence from the intra-class manifold via subspace reconstruction, enabling the precise identification of hard samples using solely target data in a completely training-free manner.
- **Extensive Validation on Effectiveness and Stealthiness:** We conduct comprehensive experiments on benchmark datasets, benchmarking our attack against diverse victim architectures (CNNs and ViTs [9]). The empirical results demonstrate that our strategies achieve superior attack success rates while maintaining exceptional stealthiness against mainstream defense mechanisms.

II. Related Work

A. Backdoor Attacks

Backdoor attacks aim to inject hidden trigger-target associations into DNNs. Early approaches, often referred to as dirty-label attacks, involve modifying the ground-truth labels of poisoned samples. While pioneer works like BadNets [10] employed visible triggers, subsequent methods enhanced stealthiness through trigger invisibility: Blended [11] utilizes alpha-blending, and WaNet [12] employs elastic warping fields. Despite their visual imperceptibility, these methods fundamentally rely on label inconsistencies, rendering them susceptible to human inspection. To circumvent this, clean-label attacks [3] were introduced, which maintain semantic consistency between poisoned samples and their labels. Prominent methods in this category include SIG [13], which injects high-frequency sinusoidal signals, ReFool [14], which models physical reflections. Nonetheless, clean-label attacks face the challenge of feature competition, where robust benign features dominate the trigger pattern, underscoring the critical need for strategic sample selection to amplify the trigger's influence [15].

B. Optimize Sample Selection

The effectiveness of backdoor attacks is strongly influenced by how poisoning samples are selected. Early studies largely relied on random sampling, while recent works have demonstrated that prioritizing hard samples—instances that are intrinsically difficult for model generalization—can substantially improve attack success under limited poisoning budgets. Compared to random poisoning, targeted sample selection can significantly enhance attack efficacy [7], [16].

To identify such high-value samples, existing approaches commonly adopt k-nearest neighbor density or OOD-based scoring in the latent feature space [8]. In contrast, these methods are fundamentally constrained by local geometric heuristics and often require auxiliary datasets or surrogate model training, which limits their practicality under restricted threat models.

C. Backdoor Defense

Existing backdoor defenses can be broadly categorized according to the stage at which they are deployed. Data sanitization methods operate during training by identifying and removing poisoned samples based on latent statistical anomalies, including Activation Clustering [17], Spectral Signatures [18], and SPECTRE [19], which adopts robust spectral and statistical analysis to identify and filter poisoned data. When a compromised model is already trained, model purification approaches such as Fine-Pruning [20] and Neural Cleanse [21] aim to mitigate backdoor behaviors through structural modification or targeted fine-tuning. At deployment time, inference-time detection methods like STRIP [22] detect malicious inputs via input perturbation without requiring access to training data or model retraining.

III. METHODOLOGY

In this section, we first formulate the selective clean-label backdoor attack problem and discuss prior limitations. Then, we propose two training-free metrics to identify optimal candidates. The overall pipeline is illustrated in Figure 1.

A. Problem Formulation and Threat Model

Attack Setting. Let $\mathcal{D}$ denote the original training dataset spanning M classes. We assume the attacker has access strictly restricted to the clean samples belonging to a specific target class y_t. This target subset is denoted as $\mathcal{D}_c = \{(x_i, y_t)\}_{i=1}^{N}$, where N represents the total number of samples in the target class.

The attacker is constrained by a poisoning budget m, where $m \ll N$. Under this constraint, the attacker identifies a candidate index subset $\mathcal{S} \subset \{1, \ldots, N\}$ with cardinality

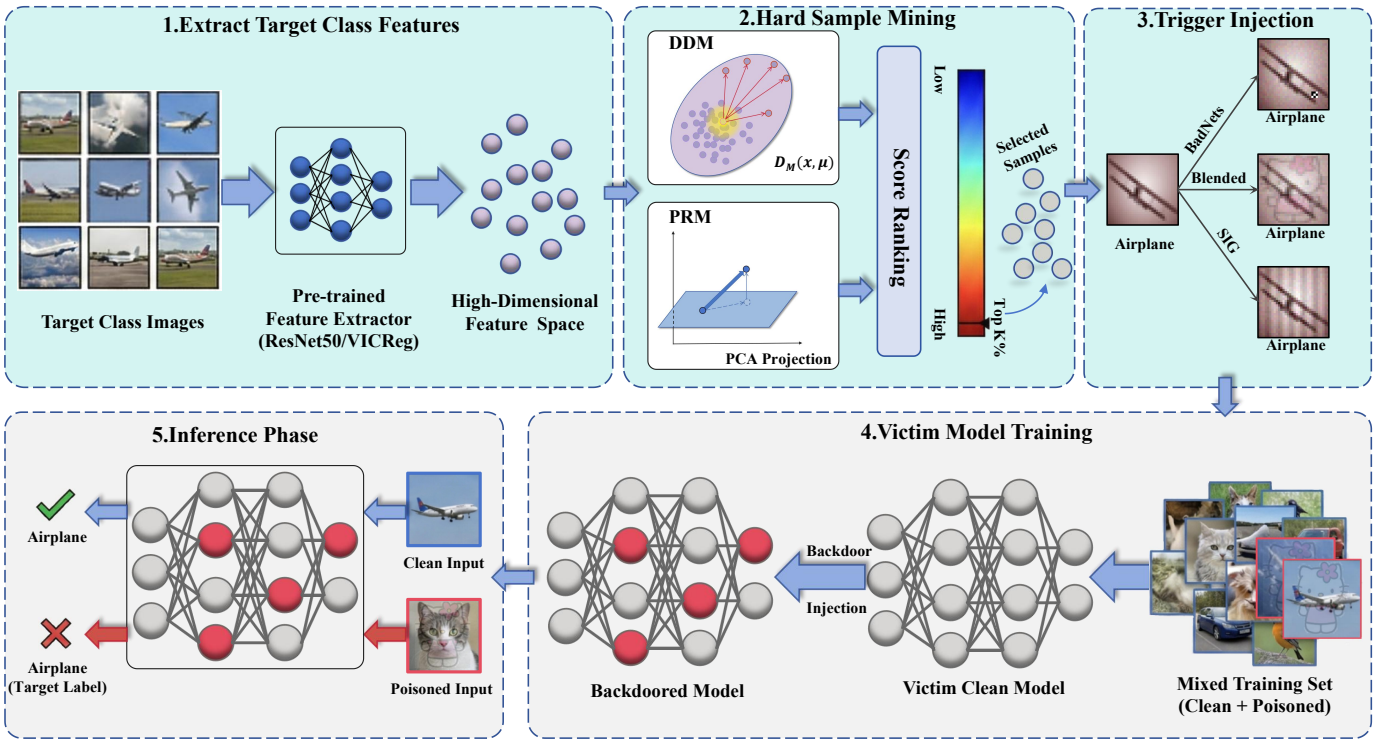

Fig. 1: **Overview of the proposed attack pipeline.** The pipeline consists of five stages: (1) **Feature Extraction**: Target class images are mapped to a high-dimensional space using a pre-trained encoder; (2) **Hard Sample Mining**: We calculate sample difficulty using DDM and PRM metrics to select the most effective candidates for poisoning; (3) **Trigger Injection**: Various triggers are injected into the selected hard samples; (4) **Victim Model Training**: The model is trained on the mixed dataset and (5) **Inference**: The backdoored model behaves normally on clean data but misclassifies trigger-carrying inputs to the target label.

$|\mathcal{S}| = m$. A trigger function $\mathcal{T}(\cdot)$ is applied to these candidates to generate the poisoned dataset $\mathcal{D}_p = \{(\mathcal{T}(x_i), y_t) \mid i \in \mathcal{S}\}$.

Adhering to the clean-label paradigm, the poisoned samples replace their clean counterparts while retaining the semantic label y_t. Let $\mathcal{D}_c^{\mathcal{S}} = \{(x_i, y_t) \mid i \in \mathcal{S}\}$ denote the subset of clean samples selected for replacement. The final training dataset $\mathcal{D}_{train}$ is constructed by removing $\mathcal{D}_c^{\mathcal{S}}$ from the original dataset $\mathcal{D}$ and adding the poisoned samples $\mathcal{D}_p$, where $\setminus$ denotes the set difference and $\cup$ denotes the set union, i.e.,

$$\mathcal{D}_{train} = (\mathcal{D} \setminus \mathcal{D}_c^{\mathcal{S}}) \cup \mathcal{D}_p. \tag{1}$$

Attacker Goal. The attacker aims to train a victim model f_θ on the mixed dataset $\mathcal{D}_{\text{train}}$ such that it achieves a high Attack Success Rate (ASR) on triggered inputs while maintaining high Clean Accuracy (CA) on benign inputs. The core challenge lies in the Sample Selection phase: identifying the subset $\mathcal{S}$ that contributes most significantly to the formation of the backdoor mapping within the model's decision boundary.

Design Motivation. Random poisoning often leads to ineffective trigger injection due to the inherent robustness of easy samples. Consequently, effective sample mining plays a critical role in identifying vulnerable target samples. Existing sample mining strategies generally follow two paradigms: heuristic-based approaches that rely on local neighborhood statistics in the feature space, and methods that estimate sample difficulty using surrogate models trained with out-of-distribution (OOD) data. Despite their empirical success, both paradigms exhibit fundamental limitations.

We first consider heuristic-based methods. These approaches primarily depend on local density estimation, focusing exclusively on neighborhood-level structures. As a result, they are highly sensitive to local noise and fail to capture the global statistical characteristics of the target class.

To overcome this limitation, we propose Distribution Deviation Metric (DDM), a global-statistics-driven criterion that quantifies sample difficulty by measuring deviations from the overall target-class distribution.

On the other hand, the approach involves training surrogate models using OOD data to estimate sample difficulty. This process introduces significant training overhead and relies heavily on the availability and representativeness of the OOD data.

In contrast, we argue that sample difficulty is an intrinsic property of the target data itself. Based on this observation, we propose Projection Residual Metric (PRM), a training-free subspace analysis method that estimates sample difficulty solely from target sample features, without requiring surrogate models or auxiliary data. The detailed procedure for our hard

TABLE I: Attack Success Rate (ASR) on CIFAR-10. We compare our DDM and PRM strategies against Random sampling and the baseline (WickedOOD). **Bold** indicates the best result, and underline indicates the second best.

Model	Strategy	Method	BadNets			Blended			SIG		
			5%	10%	20%	5%	10%	20%	5%	10%	20%
ResNet18	Random	-	30.81	45.01	78.28	28.94	37.55	44.26	50.28	60.54	78.45
	WickedOOD	Best	90.01	92.14	99.26	47.68	60.86	67.81	81.65	85.42	90.49
	DDM (Ours)	SSL	87.56	95.82	98.95	**50.38**	59.12	71.15	84.25	88.61	91.29
		SL	95.49	**98.21**	99.47	48.88	**61.58**	**72.11**	**84.76**	**90.27**	**94.38**
	PRM (Ours)	SSL	93.78	92.54	98.85	47.65	57.81	71.05	82.46	86.20	92.60
		SL	**97.04**	92.91	**99.81**	48.52	61.28	70.71	82.47	88.51	91.55
VGG19	Random	-	63.24	78.39	79.55	17.32	23.84	34.36	22.28	45.54	67.57
	WickedOOD	Best	83.43	89.61	93.11	30.74	42.23	55.34	57.24	74.38	81.93
	DDM (Ours)	SSL	86.78	90.45	93.47	**31.72**	**45.51**	**56.85**	**58.84**	73.57	82.47
		SL	84.92	**94.99**	93.03	29.19	44.47	56.48	55.84	**77.23**	82.14
	PRM (Ours)	SSL	83.94	86.42	**95.01**	30.52	43.11	52.12	50.29	72.78	81.97
		SL	**89.42**	91.81	91.95	31.40	40.42	53.32	52.88	75.95	**84.55**

TABLE II: Attack Success Rate (ASR) on GTSRB. We compare our DDM and PRM strategies against Random sampling and the best baseline (WickedOOD). **Bold** indicates the best result, and underline indicates the second best.

Model	Strategy	Method	BadNets			Blended			SIG		
			5%	10%	20%	5%	10%	20%	5%	10%	20%
ResNet18	Random	-	5.72	5.80	6.13	36.35	41.54	48.91	47.63	48.07	48.67
	WickedOOD	Best	**10.37**	10.91	21.33	46.96	50.13	51.54	54.85	57.12	58.88
	DDM (Ours)	SSL	9.94	12.15	**89.58**	52.77	53.91	**60.13**	68.11	70.18	74.39
		SL	9.52	**41.56**	53.42	51.21	**61.84**	56.37	68.45	70.10	**76.20**
	PRM (Ours)	SSL	9.29	15.36	22.51	52.14	45.40	58.23	68.04	65.70	68.95
		SL	7.51	21.38	79.51	**55.13**	51.94	55.90	**73.65**	**78.25**	73.57
VGG19	Random	-	6.14	6.38	6.89	24.89	26.36	30.69	32.04	33.90	36.52
	WickedOOD	Best	8.16	10.19	15.76	32.25	33.16	33.82	40.33	42.68	42.37
	DDM (Ours)	SSL	8.67	10.64	17.51	37.81	37.72	41.50	**47.95**	43.82	48.30
		SL	**11.56**	12.37	16.03	38.35	41.68	43.47	43.18	43.46	44.62
	PRM (Ours)	SSL	6.94	**12.65**	**23.83**	**40.03**	41.84	**45.12**	40.83	47.09	**54.48**
		SL	8.72	12.64	22.39	36.35	**42.54**	43.16	43.85	**52.51**	48.12

sample mining strategy is summarized in Algorithm 1.

B. Strategy I: Distribution Deviation Metric

The *Distribution Deviation Metric* (DDM) aims to quantify the degree to which a sample deviates from the robust global correlation structure of a target class.

1) Motivation: Conventional distance measures, such as Euclidean distance, treat all feature dimensions equally and ignore the fact that different features exhibit varying degrees of variance and correlation. While the Mahalanobis distance is theoretically well-suited to capture such correlations, its direct application becomes numerically unstable in high-dimensional deep feature spaces where the feature dimension D exceeds the number of available samples N. This instability severely limits its practicality in small-sample regimes.

DDM addresses this challenge by performing statistically stabilized distance estimation in a low-dimensional whitened subspace. Rather than operating directly in the original feature space, DDM measures sample deviation with respect to the global distribution in a compact latent space, where robust covariance estimation becomes feasible. Samples that lie in the low-density tails of this global distribution are identified as difficult samples, as they represent statistically non-typical instances of the target class.

2) Formulation: Let $z_i = g(x_i) \in \mathbb{R}^D$ denote the deep feature extracted by a pre-trained encoder $g(\cdot)$ for sample x_i. First, the features are centered by subtracting the global mean μ, and then projected onto a K-dimensional latent subspace using principal component analysis (PCA). Let $W \in \mathbb{R}^{D\times K}$ denote the projection matrix formed by the top-K principal components. The transformed representation is given by

$$\hat{z}_i = W^\top (z_i - \mu). \tag{2}$$

To decorrelate feature dimensions and normalize variances, we perform PCA with whitening, yielding a compact representation $\tilde{z}_i = \Lambda^{-1/2} W^\top \bar{z}_i$, where $\bar{z}i = \text{Standardize}(z_i)$, W contains the top principal components, and Λ is the diagonal matrix of corresponding eigenvalues. In this whitened space, we estimate a robust precision matrix $\hat{\Theta}$ using shrinkage regularization to stabilize the subsequent Mahalanobis distance

Algorithm 1: Training-Free Hard Sample Mining Strategy

Input : Target dataset $\mathcal{D}_c$; Pre-trained encoder $g(\cdot)$; Budget m; Variance ratio τ; Trigger injection function $\mathcal{T}(\cdot)$; Mining strategy selector $\mathcal{M}$.
Output: Poisoned subset $\mathcal{D}_p$.

Extract global features: $\mathbf{Z} \leftarrow \{g(x_i)\}_{i=1}^{N} \in \mathbb{R}^{N \times D}$;
if $\mathcal{M} =$ DDM *(Distribution Deviation)* **then**
  Standardize features: $\bar{z}_i \leftarrow \text{Standardize}(z_i)$ for all i;
  Compute projection matrix W and eigenvalues λ via PCA on $\{\bar{z}_i\}$;
  Project and whiten: $\tilde{z}_i \leftarrow \Lambda^{-1/2} W^\top \bar{z}_i$, where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_K)$;
  Estimate robust precision matrix $\hat{\Theta}$ using 'ShrunkCovariance' on $\{\tilde{z}_i\}_{i=1}^{N}$;
  for $i \leftarrow 1$ **to** N **do**
    Calculate Deviation Score: $s_i \leftarrow \sqrt{\tilde{z}_i^\top \hat{\Theta} \tilde{z}_i}$
else if $\mathcal{M} =$ PRM *(Projection Residual)* **then**
  Center features: $\bar{z}_i \leftarrow z_i - \mu$;
  Compute V_q via PCA on $\{\bar{z}_i\}$ with τ;
  for $i \leftarrow 1$ **to** N **do**
    Compute residual vector: $r_i \leftarrow \bar{z}_i - V_q V_q^\top \bar{z}_i$;
    Calculate Residual Score: $s_i \leftarrow \|r_i\|_2$
$\mathcal{S} \leftarrow$ indices of the top-m largest scores in $\{s_i\}_{i=1}^{N}$;
Construct poisoned set: $\mathcal{D}_p \leftarrow \{(\mathcal{T}(x_i), y_t) \mid i \in \mathcal{S}\}$;
return $\mathcal{D}_p$

TABLE III: Impact on Clean Accuracy. Comparison of model utility between Random sampling and our proposed strategies using both SSL and SL encoders. We report the clean test accuracy (%) on CIFAR-10 and GTSRB datasets under the SIG attack setting.

Model	Strategy	Method	CIFAR-10			GTSRB		
			5%	10%	20%	5%	10%	20%
ResNet18	Random	-	94.69	94.60	94.59	99.06	99.04	99.11
	DDM(Ours)	SSL	94.70	94.49	94.44	98.50	98.78	98.79
		SL	94.60	94.68	94.55	98.62	98.35	98.81
	PRM(Ours)	SSL	94.60	94.57	94.68	98.37	98.38	98.54
		SL	94.58	94.69	94.32	98.96	98.71	98.92
VGG19	Random	-	91.97	91.89	92.98	96.02	96.58	95.48
	DDM(Ours)	SSL	91.57	91.68	91.68	97.65	97.67	98.06
		SL	91.88	91.38	91.43	98.12	97.47	97.62
	PRM(Ours)	SSL	91.79	91.70	91.59	97.81	97.40	97.30
		SL	91.58	91.55	91.49	97.80	97.35	97.54

calculation. The DDM score for sample x_i is defined as

$$s_{\text{DDM}}(x_i) = \sqrt{\tilde{z}_i^\top \hat{\Theta} \tilde{z}_i}. \tag{3}$$

This score can be interpreted as a statistically stabilized variant of the Mahalanobis distance, computed in a low-dimensional whitened subspace. In practice, DDM is employed as a relative ranking criterion rather than an absolute distance measure. Samples with higher s_{DDM} values exhibit greater deviation from the global class distribution and are therefore prioritized for poison.

TABLE IV: Performance evaluation on Vision Transformer. We report the Clean Accuracy (CA) / Attack Success Rate (ASR) on CIFAR-10 and GTSRB datasets under the SIG attack.

Dataset	Strategy		Poisoning Rate		
			5%	10%	20%
CIFAR-10	Random	-	92.51/98.08	92.10/99.67	92.75/99.88
	DDM(Ours)	SSL	**91.99/99.49**	92.21/99.83	91.78/99.93
		SL	92.14/99.37	91.90/99.82	91.90/99.95
	PRM(Ours)	SSL	92.01/99.41	91.92/99.82	**92.05/99.95**
		SL	92.16/99.33	**92.08/99.87**	91.74/99.94
GTSRB	Random	-	98.69/98.65	98.49/99.66	98.49/99.63
	DDM(Ours)	SSL	98.24/99.14	98.19/99.75	98.31/99.78
		SL	98.55/99.25	98.59/99.69	98.41/99.69
	PRM(Ours)	SSL	**98.27/99.51**	**98.54/99.75**	**98.51/99.83**
		SL	98.47/99.43	98.44/99.65	98.47/99.75

C. Strategy II: Projection Residual Metric

The Projection Residual Metric (PRM) is proposed to quantify the intrinsic difficulty of samples by analyzing their structural divergence from the class-specific manifold. Unlike prior arts that depend on auxiliary Out-of-Distribution (OOD) data, PRM operates in a strictly training-free manner by accessing only the target class samples, estimating sample difficulty based on the degree of conformity of an individual sample to the primary low-dimensional structure of that class.

1) Motivation: We hypothesize that easy samples typically concentrate around a low-rank linear manifold spanned by the dominant features of the target class. In contrast, hard samples (which are more susceptible to backdoor injection) exhibit significant structural variations that cannot be well-represented by this primary manifold. PRM quantifies sample difficulty by measuring the information loss incurred when a sample is projected onto the principal subspace of the target class.

2) Formulation: Let $z_i \in \mathbb{R}^D$ denote the D-dimensional feature vector of sample x_i. We analyze the intrinsic geometric structure of the target class using Principal Component Analysis. First, the eigenvalues $\{\lambda_j\}$ and the corresponding eigenvectors of the class covariance matrix are computed. Let $k \in \{1, \ldots, D\}$ denote the number of leading principal components. To capture the dominant semantic information, we adaptively determine the subspace dimensionality q as the minimal rank required to preserve a cumulative variance ratio τ:

$$q = \min \left\{ k \;\middle|\; \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{D} \lambda_j} \geq \tau \right\}. \tag{4}$$

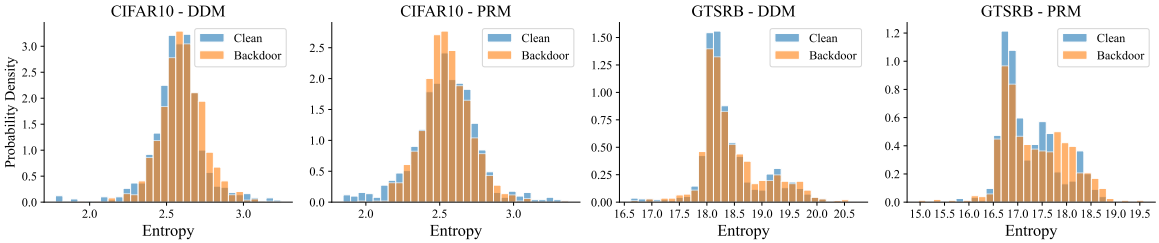

Fig. 2: Resistance to STRIP: Indistinguishable entropy distributions of clean and poisoned samples.

Let $V_q \in \mathbb{R}^{D\times q}$ denote the basis matrix formed by the top-q eigenvectors. Subsequently, the orthogonal projection residual vector r_i is calculated for each sample:

$$r_i = z_i - V_q V_q^\top z_i. \tag{5}$$

Here, the term $V_q V_q^\top z_i$ represents the optimal reconstruction of the sample within the principal subspace. The residual vector r_i thus captures the specific structural details that are orthogonal to the class's dominant variations.

The PRM score for sample x_i is formally defined as the magnitude of this residual:

$$s_{\text{PRM}}(x_i) = \|r_i\|_2 = \left\|z_i - V_q V_q^\top z_i\right\|_2. \tag{6}$$

This score quantifies the Euclidean distance between the original feature and the class manifold. In practice, PRM serves as a straightforward ranking metric. Samples with higher s_{PRM} values exhibit significant structural divergence from the class consensus (i.e., hard samples) and are therefore prioritized for poison.

IV. Experiments

We evaluate the proposed mining strategy against state-of-the-art methods and demonstrate its resistance to mainstream defenses.

A. Experimental Setup

Datasets and Models: We utilize CIFAR-10 and GTSRB datasets. Consistent with recent works [8], we adopt ResNet18 [23] and VGG19 [24] as our victim models. For the mining phase, we utilize two distinct feature extractors to ensure generalization: a Self-Supervised (SSL) encoder with a ResNet-50 backbone and a Supervised (SL) encoder based on a ResNet-50 that is pretrained on ImageNet [25].

Attack Configurations: We evaluated three representative triggers under a clean-label attack setting: (1) BadNets: local 3×3 black and white pixel blocks; (2) Blended: a global blending pattern with $\alpha = 0.2$ transparency; (3) SIG: sinusoidal noise simulating physical signals. We set the target classes to class 0 of CIFAR10 [26] and class 1 of GTSRB [27], and compared attack performance at different poisoning rates (5%, 10%, and 20% of the target class count).

Evaluation Metrics: We evaluate model performance using Clean Accuracy (CA) on clean test sets and Attack Success Rate (ASR).

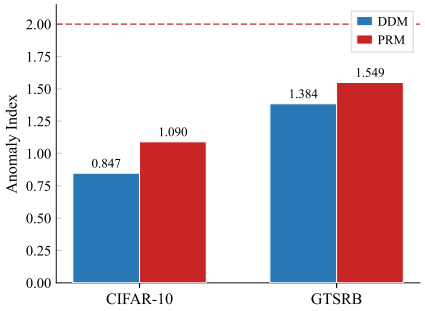

Fig. 3: Performance against Neural Cleanse.

B. Strategy Performance Evaluation

We compare the proposed mining strategies against Random sampling and the state-of-the-art methods, WickedOOD.

Attack Effectiveness on CIFAR-10. Table I reports the ASR on CIFAR-10. The results reveal a clear hierarchy: Random $<$ WickedOOD $<$ Ours. Our methods consistently outperforms baselines, particularly in the low-poisoning regime. This high sample efficiency indicates that our strategy successfully identifies hard examples that are structurally influential, thereby shifting the decision boundary more effectively than OOD-based selection.

Attack Effectiveness on GTSRB. To verify the generalization capability of our method across different datasets, we extend the evaluation to GTSRB. As detailed in Table II, our strategies consistently outperform the baselines. This suggests that capturing global distributional statistics via robust covariance estimation is effective even on datasets with imbalanced distributions.

TABLE V: Resistance against Latent Separation Defenses on CIFAR-10 and GTSRB. ER: Elimination Rate (%); SR: Sacrifice Rate (%). Note that **low ER** implies evasion, while **high SR** implies high cost for the defense.

Defense	**CIFAR-10**				**GTSRB**			
	DDM		**PRM**		**DDM**		**PRM**	
	ER	SR	ER	SR	ER	SR	ER	SR
AC	0.00	0.00	0.00	0.00	0.00	10.90	0.00	10.40
SS	42.60	50.10	33.00	50.20	46.40	50.00	83.80	49.80
SPECTRE	31.00	50.20	30.60	50.20	58.60	50.00	86.00	49.80

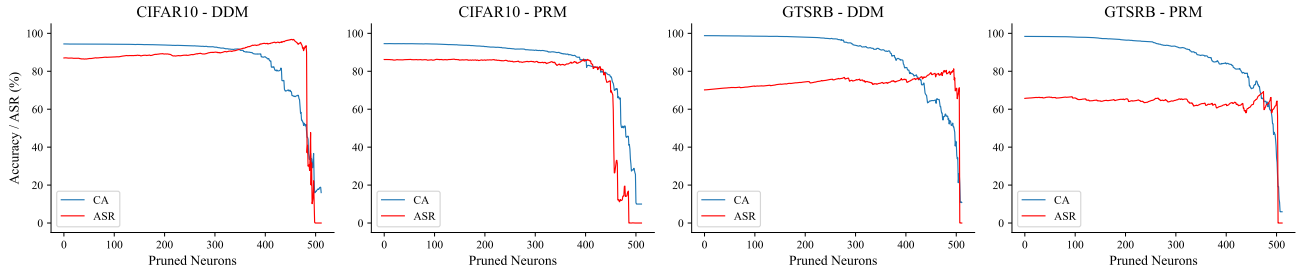

Fig. 4: Resistance to Fine-Pruning: Stable ASR and CA under neuron pruning.

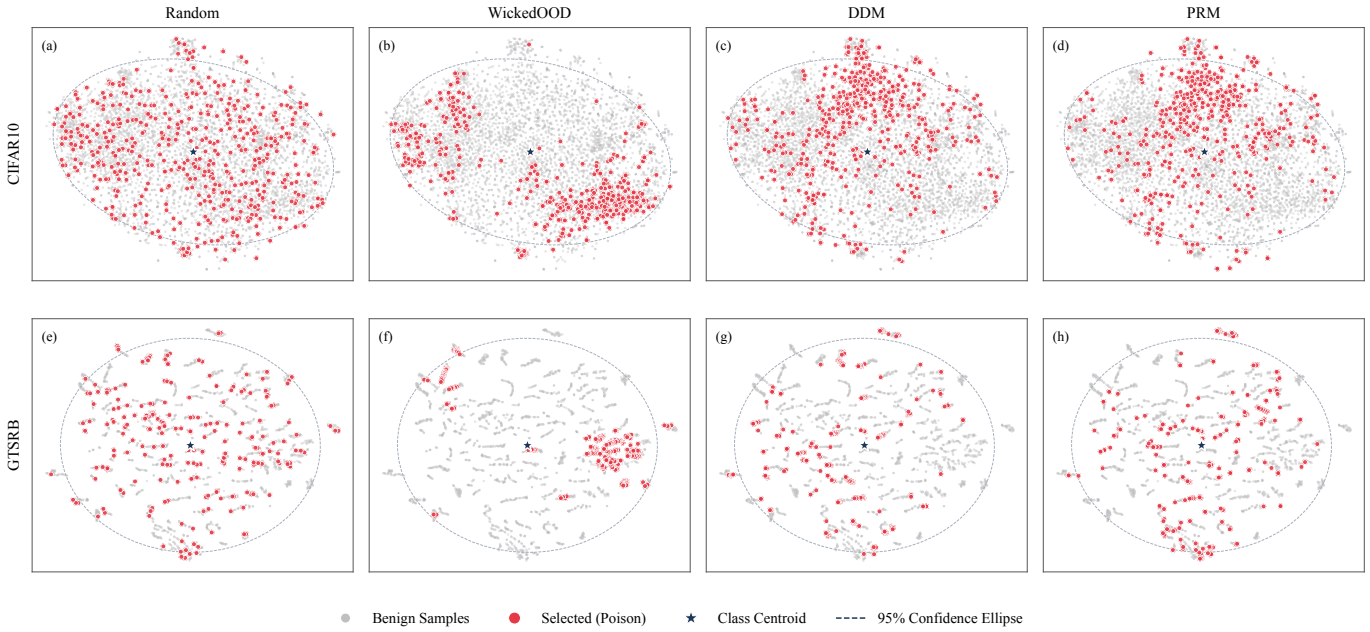

Fig. 5: Visualization of feature space distributions for selected samples on CIFAR-10 (top) and GTSRB (bottom). Clean samples (gray) are projected using t-SNE, overlaid with poison candidates (red dots). The blue star ($\star$) and dashed ellipse denote the class centroid and the 95% confidence boundary, respectively.

Impact on Clean Accuracy. We assess stealthiness by measuring clean accuracy (CA). As shown in Table III, our strategies maintain clean accuracy comparable to clean models across both datasets. This confirms that the selected hard examples, while effective for poisoning, do not compromise the learned feature representation of the target class itself, ensuring high stealthiness.

C. Evaluation to Vision Transformers

To verify the architecture-agnostic effectiveness of our proposed strategies, we validate our method on Vision Transformers (ViT), which differ significantly from CNNs by relying on global self-attention mechanisms. We utilize the ViT-Base architecture (patch size 16×16) with an input resolution of 224×224. We optimize the model using AdamW for 300 epochs. We evaluate the SIG trigger on both CIFAR-10 and GTSRB datasets.Table IV reports the CA and ASR. Our strategies consistently outperform Random sampling, achieving high ASR while retaining high CA.

D. Resistance to Backdoor Defenses

We examine the stealthiness of our mining strategies against popular defenses. Specifically, we conduct this evaluation under the representative attack configuration using the SIG trigger and the VICReg [28] feature extractor, following standard evaluation protocols where defenses are configured with their official implementations and default hyperparameters.

STRIP [22]. As illustrated in Figure 2, the entropy distributions of poisoned samples selected by our strategies are indistinguishable from clean samples. We evaluate STRIP with $N = 100$ superimposed samples and $M = 500$ test images over 10 rounds; varying these parameters and attack strengths does not alleviate the overlap. This heavy overlap renders STRIP ineffective, as no clear threshold can be established to filter out malicious inputs without rejecting benign ones.

Neural Cleanse (NC) [21]. As illustrated in Figure 3, the Anomaly Indices calculated by NC for both CIFAR-10 and GTSRB consistently fall below the detection threshold of 2.0.

This indicates that our mining strategy effectively generates diffuse patterns that are robust against the reverse-engineering optimization of NC.

Fine-Pruning (FP) [20]. Figure 4 demonstrates the high stability of our ASR. Even when the pruning ratio is increased to a point where clean accuracy significantly degrades, the ASR remains robust. This indicates that the learned backdoor features are deeply entangled with critical benign features, rendering the pruning ineffective.

Resistance to Additional Defenses.

We extend our evaluation to include three additional statistical defenses, strictly adhering to the experimental protocols defined in their original papers. As reported in Table V, **Activation Clustering (AC)** [17] fails completely (Elimination Rates (ER) ≈ 0.00), as it misinterprets our structurally representative poisons as inliers. While **Spectral Signatures (SS)** [18] and **SPECTRE** [19] achieve varying ER (30%–80%), they incur a prohibitive Sacrifice Rate (SR). As shown in the table, the SR consistently approaches 50%, implying that these defenses are forced to discard nearly half of the clean data to mitigate the backdoor, making them less practical for deployment.

E. Visualization of Sample Selection

To intuitively interpret the mining mechanisms, we visualize feature distributions of selected samples using t-SNE [29] in Figure 5. Random selection assumes all samples contribute equally, resulting in a scattered distribution without structural focus. WickedOOD exhibits severe local clustering, leading to highly redundant features that fail to capture global data diversity. In contrast, our strategies achieve comprehensive feature space coverage. By prioritizing global distributional properties, our method identifies diverse and informative anchors, effectively avoiding redundancy while maintaining better structural awareness than Random selection.

V. Conclusion

In this paper, we identify sample selection as a critical factor for clean-label backdoor attacks. By introducing the Distribution Deviation Metric and Projection Residual Metric, we shift the focus from local metrics to global distributional and subspace analysis. Our experiments confirm that these statistically optimized strategies significantly enhance attack transferability and stealthiness against mainstream defenses. We hope these findings will inspire future research to design more effective defense mechanisms against stealthy adversaries.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[3] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.

[4] J. Abramson, J. Adler, J. Dunger *et al.*, "Accurate structure prediction of biomolecular interactions with alphafold 3," *Nature*, vol. 630, pp. 493–500, 2024.

[5] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng, and Y. Xu, "Finben: A holistic financial benchmark for large language models," *NeurIPS*, vol. 37, pp. 95 716–95 743, 2024.

[6] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *CVPR*, 2023, pp. 17 853–17 862.

[7] Y. Gao, Y. Li, L. Zhu, D. Wu, Y. Jiang, and S.-T. Xia, "Not all samples are born equal: Towards effective clean-label backdoor attacks," *Pattern Recognition*, 2023.

[8] H.-Q. Nguyen, N. Ngoc-Hieu, T.-A. Ta, T. Nguyen-Tang, K.-S. Wong, H. Thanh-Tung, and K. D. Doan, "Wicked oddities: Selectively poisoning for effective clean-label backdoor attacks," in *ICLR*, 2025.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[10] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[11] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[12] T. A. Nguyen and A. T. Tran, "Wanet–imperceptible warping-based backdoor attack," in *ICLR*, 2021.

[13] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in *ICIP*, 2019, pp. 101–105.

[14] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *ECCV*, 2020, pp. 182–199.

[15] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *NeurIPS*, 2018.

[16] A. Katharopoulos and F. Fleuret, "Not all samples are created equal: Deep learning with importance sampling," in *ICML*, 2018, pp. 2525–2534.

[17] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. M. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," *arXiv preprint arXiv:1811.03728*, 2018.

[18] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *NeurIPS*, 2018.

[19] J. Hayase, W. Kong, R. Somani, and S. Oh, "Spectre: Defending against backdoor attacks using robust statistics," in *ICML*, 2021, pp. 4129–4139.

[20] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *RAID*, 2018, pp. 273–294.

[21] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *IEEE S&P*, 2019, pp. 707–723.

[22] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *ACSAC*, 2019, pp. 113–125.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[26] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[27] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, vol. 32, pp. 323–332, 2012.

[28] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," in *ICLR*, 2022.

[29] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.