Poison in the Well: Feature Embedding Disruption in Backdoor Attacks

Zhou Feng[1†], Jiahao Chen[1†], Chunyi Zhou[1*], Yuwen Pu[2], Qingming Li[1], Shouling Ji[1]
[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]School of Big Data & Software Engineering, Chongqing University, Chongqing, China
{zhou.feng, xaddwell, zhouchunyi, liqm, sji}@zju.edu.cn, yw.pu@cqu.edu.cn

***Abstract*—Backdoor attacks embed malicious triggers into training data, enabling attackers to manipulate neural network behavior during inference while maintaining high accuracy on benign inputs. However, existing backdoor attacks face limitations manifesting in excessive reliance on training data, poor stealth, and instability, which hinder their effectiveness in real-world applications. Therefore, this paper introduces *ShadowPrint*, a versatile backdoor attack that targets feature embeddings within neural networks to achieve high ASRs and stealthiness. Unlike traditional approaches, *ShadowPrint* reduces reliance on training data access and operates effectively with exceedingly low poison rates (as low as 0.01%). It leverages a clustering-based optimization strategy to align feature embeddings, ensuring robust performance across diverse scenarios while maintaining stability and stealth. Extensive evaluations demonstrate that *ShadowPrint* achieves superior ASR (up to 100%), steady CA (with decay no more than 1% in most cases), and low DDR (averaging below 5%) across both clean-label and dirty-label settings, and with poison rates ranging from as low as 0.01% to 0.05%, setting a new standard for backdoor attack capabilities and emphasizing the need for advanced defense strategies focused on feature space manipulations.**

***Index Terms*—Backdoor Attack, Feature Manipulation, Poisoning Strategy**

I. Introduction

Rapid advancement of deep learning has led to its adoption in various domains, from autonomous vehicles and healthcare to finance and security systems [1], [2]. However, the increasing reliance on neural networks has also exposed them to a variety of adversarial threats. Among these, backdoor attacks [3] have emerged as a particularly stealthy and potent form of vulnerability. By embedding a malicious trigger into the training data, an attacker can cause the model to produce incorrect or harmful outputs in the presence of the trigger, while maintaining high accuracy on benign inputs.

The effectiveness of backdoor attacks often hinges on the design and placement of the trigger, as well as the method's ability to evade detection. Recent advances in backdoor attacks have introduced innovative approaches that address some of these challenges. For example, Narcissus [4], require only access to training data from the target class, which makes them less dependent on extensive data access. Or like Gao et al. [5], identify and exploit "hard" samples, data with weak robust features, to complement existing clean-label attacks. Although these methods represent significant progress, they still share several inherent limitations that constrain their applicability and effectiveness:

- **Over-Reliance on Data Access:** Many existing approaches often assume attackers to have extensive or unrestrained knowledge of training data and demand excessively high poison rates. In real-world scenarios, this access is often partial or severely limited, making these assumptions impractical.
- **Inadequate Stealthiness:** The invisibility of the trigger remains a critical factor, as detectable triggers are highly susceptible to identification by defense mechanisms.
- **Fragile Stability:** Achieving consistent performance across various input scenarios (e.g., white-box, black-box and data-free, as described in Section III-A) and ensuring resilience to variations in data distribution continue to be significant challenges.

The SOTA methods face additional challenges when applied to MLaaS platforms [6], [7]. These platforms, widely used in real-world applications, present a unique set of constraints, including limited access to the model architecture, training data, and hyperparameters. Many existing backdoor attacks falter in these environments due to their dependence on high levels of attacker knowledge and access. This underscores the importance of designing attacks that are adaptable to real-world scenarios, where constraints and defenses are more robust than in traditional experimental settings.

Therefore, this paper aims to address the above gaps by investigating novel ways to analyze and exploit the relationships within the feature space during backdoor attacks. Specifically, the contributions of this work are as follows:

- We propose *ShadowPrint*, a novel backdoor attack that mitigates assumptions about attacker and achieves robust attack performance under realistic countermeasures.
- We employs a clustering-based trigger optimization strategy to align feature embeddings of poisoned samples, reducing the burden of backdoor learning during model training, enabling the use of an extremely low poison rate (as low as 0.01%) while maintaining attack effectiveness.
- We conduct extensive experiments on multiple benchmark datasets (i.e., CIFAR-10, CIFAR-100, and TinyImageNet), demonstrating that *ShadowPrint* achieves high attack performance. Specifically, it maintains effectiveness,

†Zhou Feng and Jiahao Chen contributed equally to this work.
*Corresponding author: Dr. Chunyi Zhou (zhouchunyi@zju.edu.cn)

stealthiness, and stability while evading detection under SOTA defenses, such as IBD-PSC [8], SCALE UP [9], and Beatrix [10], even at extremely low poison rates (no greater than 0.05%).

II. Background and Related Work

A. Backdoor Attacks in Deep Learning

Backdoor attacks pose a unique and stealthy threat to neural networks [3], [11]–[16]. Early techniques, like BadNets [3], introduced the concept of injecting simple and static triggers into training samples to induce misclassification. Subsequent approaches [11]–[16], including Blended Attacks [11], sought to enhancing the stealthiness by blending imperceptible patterns into input data. However, such methods predominantly focus on the superficial properties of triggers, such as their appearance, and neglect a deeper investigation of how the triggers interact with the model's feature space, leaving room for further refinement.

Recently, researchers have explored data-agnostic techniques and adaptive trigger designs [4], [5], [17]–[19], aiming to generalize across datasets and architectures. Li et al. [17] propose an efficient data-constrained backdoor attacks, reflecting practical conditions where attackers only have partial access to training data. Similarly, computationally informed strategies such as We et al. [18] propose novel metrics to select poisoned samples that are more effective in reshaping decision boundaries, while Zhu et al. [19] present a learnable strategy to poison sample selection using a min-max optimization framework. Though these methods enhance flexibility and stealthiness, they often prioritize the trigger's superficial properties over a deeper exploration of the model's internal feature representations. Key limitations, including over-reliance on data access, inadequate stealth for triggers, and lack of stability, underscore the need for innovative strategies that address these gaps.

B. Backdoor Defensive Mechanisms

Defensive mechanisms against backdoor attacks can be grouped into three categories: data sanitation [20], [21], model inspection [10], [22], [23], and runtime detection [8], [9].

Data sanitation methods, such as Neural Cleanse [20] and STRIP [21], attempt to identify and remove poisoned samples from the training data. Although these methods can be effective against certain trigger types, they struggle with detecting more adaptive or imperceptible triggers that evade traditional detection strategies. Model inspection techniques focus on analyzing the model's internal behavior to identify anomalies that may signal the presence of a backdoor. Activation clustering [22] and gradient-based analysis [23] are examples of such approaches that seek to differentiate between clean and poisoned samples. More recently, Beatrix [10] introduced a novel technique for identifying poisoned samples by analyzing activation anomalies via Gram matrices. However, these methods face scalability and generalizability challenges, particularly when applied to more complex models. Runtime detection defenses flag suspicious inputs during training process. Approaches like IBD-PSC [8] and SCALE UP [9], aim to detect the presence of triggers at the input level.

Despite advancements in these defense mechanisms, the evolving sophistication of backdoor attacks continues to outpace current methodologies.

III. Methodology

A. Threat Model

To systematically evaluate *ShadowPrint*, we categorize attackers into three types based on their capabilities, ensuring coverage of diverse real-world attack scenarios. These attacker types are meaningful as they reflect varying levels of knowledge and resources available to adversaries:

- **Scenario A1 (White-Box):** The attacker has knowledge of the model architecture and partial training dataset.
- **Scenario A2 (Black-Box):** The attacker only has knowledge of partial training dataset.
- **Scenario A3 (Data-Free):** The attacker has no knowledge of the model architecture or the training dataset.

Note that all these attackers can manipulate an extremely small number of training samples for poisoning (e.g., less than 0.05% in this paper, which means that less than 25 samples manageable in training dataset like CIFAR10 with 50000 training samples in total, and they can launch both clean label and dirty label attacks considering their specific capability. Their common objective is to maintain the model's accuracy on clean samples while ensuring high success rates on poisoned samples.

B. Overview of ShadowPrint

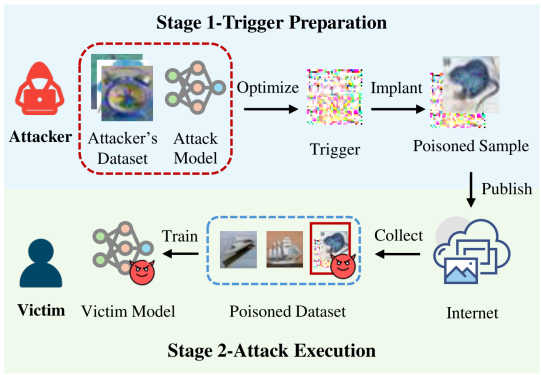

Fig. 1. Scheme of *ShadowPrint* framework. In Stage 1, the attacker optimizes the trigger based on his knowledge and resources. In Stage 2, the victim, unaware of the malicious intent, uses the poisoned dataset provided via the Internet to unintentionally train a backdoored model.

ShadowPrint introduces a novel backdoor attack framework that leverages feature space manipulation for enhanced stealth and effectiveness. By optimizing a backdoor trigger to align poisoned samples in the embedding space, it reduces reliance on extensive training data access and high poison rates. Unlike traditional methods that focus on visible or statistical anomalies, *ShadowPrint* directly targets the model's internal

representations to ensure stability and resilience across diverse scenarios. The overall scheme is illustrated in Fig. 1.

C. Method Design

1) Stage 1-Trigger Preparation: Let f represent the target neural network with parameter θ, trained on a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i and y_i denote the input samples and their corresponding labels, respectively. The goal of *ShadowPrint* is to inject a backdoor with trigger t into the parameters θ by utilizing the attacker's knowledge of the model and dataset, such that poisoned samples $x_i' = T(x_i, t)$ infiltrate the dataset $\mathcal{D}$ and disrupt the decision-making process of f_θ, where T is the transformation function. Specifically:

- The model misclassifies x_i' to a target label y_t.
- The accuracy on clean samples remains unaffected.
- Evade the potential detection by defenders.

Here in this paper, for attack stealthiness, we adopt the transformation function T in line with the Blended attack [11]:

$$T(x_i, t) = x_i \times (1 - w) + t \times w \tag{1}$$

where w denotes the trigger weight and the size of trigger t is the same as the samples.

With the goal and formulation above, we start with the main goal of a backdoor attack, which means minimizing $\mathcal{L}(f(T(x_i, t)), y_t)$. However, we propose that this backdoor behavior can also be expressed as:

$$\min \sum_{i,j\,:\,i\neq j} D(f(T(x_j, t)), f(T(x_i, t))) + \mathcal{L}(f(T(x_i, t)), y_t) \tag{2}$$

where the first term stands for backdoor clustering that aims to minimize the distance (use measurement D) of the triggered samples $T(x_i, t)$ and the second term specifies the target class y_t for backdoor attack. With this analysis, we can reformulate the learning of the backdoor attack as a clustering optimization, and the cluster center denotes the backdoor target.

However, conventional backdoor attacks complete the process of backdoor learning with model training, thus requiring many poisoned samples. To mitigate this limitation, we exploit the analysis above and propose a clustering-based trigger optimization strategy. Generally speaking, we can find a universal trigger that can help the cluster of poisoned samples before model training, reducing the burden of backdoor learning, as illustrated in Fig. 2. The trigger optimization process focuses on manipulating the feature embeddings in the last fully connected (FC) layer, to align the feature embeddings of poisoned samples effectively in the feature space.

Specifically, the optimization process relies on the following custom loss to maximize alignment between feature vectors of poisoned samples in the model's embedding space:

$$\mathcal{L}_{\text{cluster}} = \frac{1}{N^2} \sum_{i,j\,:\,i\neq j} \frac{Z_i \cdot Z_j^T}{\|Z_i\| \|Z_j\|} \tag{3}$$

where, Z_i and Z_j represent the feature vectors of poisoned samples x_i' and x_j' via f_{adv}^{fc}, i.e., the model's last FC layer, N denoting the size of samples. By minimizing (3), the optimization process encourages higher cosine similarity between these feature vectors, effectively clustering them closer together in the embedding space. Consequently, as outlined in Algorithm 1, the optimized trigger t iteratively aligns the features Z of all poisoned samples to a common cluster center, enhancing the stealthiness and robustness of the backdoor attack. This clustering-based trigger optimization strategy reduces the need for a large number of poisoned samples, thereby improving the efficiency of the attack.

Algorithm 1 Optimization Process for ShadowPrint

Input: Attacker's Dataset $\mathcal{D}_{\text{adv}}$; Attacker's Model f_{adv}; Optimizing Steps K
$t \leftarrow \mathcal{N}(0, 0.5)$
for epoch in K **do**
  for each batch $(x, y) \in \mathcal{D}_{\text{adv}}$ **do**
    $x' \leftarrow T(x, t)$ ▷ Obtain triggered samples.
    $Z \leftarrow f_{\text{adv}}^{fc}(x')$ ▷ Capture the embeddings.
    Update t with $\nabla_t \mathcal{L}_{\text{cluster}}$ ▷ Use Adam optimizer.
  end for
end for
Return: Generated trigger t.

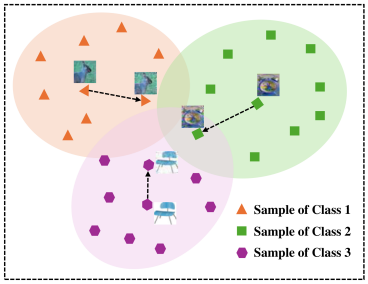

Fig. 2. Illustration of feature space alignment using *ShadowPrint*. The diagram showcases how poisoned samples are clustered in the feature space, aligning the features of the triggered samples.

2) Stage 2-Attack Execution: The purpose of *ShadowPrint* is to encompass a broader spectrum of attacker types in real-world application scenarios, as we have previously described in Section III-A. Consequently, we propose three distinct attack modes under *ShadowPrint*:

- **Dirty Label Attack:** Dirty label attacks manipulate the samples and labels simultaneously.
- **Clean Label Attack:** Samples from the target class are manipulated while labels remain unchanged, ensuring more practical and stealth-oriented application scenarios.
- **Data-Free Attack:** Using auxiliary data from other domains to train surrogate model f_{adv} and construct $\mathcal{D}_{\text{adv}}$ (i.e., $f_{\text{adv}} \neq f$ and $\mathcal{D}_{\text{adv}} \cap \mathcal{D} = \emptyset$).

Note that, for Scenario A3, the assumption is that the attacker does not have access to the training data or model architecture. This is a specific and challenging scenario where information about the target model is unavailable. To address

TABLE I

DIRTY LABEL ATTACK. WE EVALUATE *ShadowPrint* UNDER THE DIRTY-LABEL ATTACK MODE BY TESTING THE CA AND ASR ACROSS DIFFERENT POISON RATES AND ATTACKER SETTINGS.

Target Model	Poison Ratio	CIFAR-10				CIFAR-100				TinyImageNet			
		Baseline	*ResNet18*	*ResNet34*	*VGG13BN*	*Baseline*	*ResNet18*	*ResNet34*	*VGG13BN*	*Baseline*	*ResNet18*	*ResNet34*	*VGG13BN*
ResNet18	0.0001	0.920	0.919/0.994	0.919/0.999	0.919/0.997	0.690	0.687/0.993	0.688/0.995	0.692/0.999	0.508	0.456/0.997	0.454/0.999	0.462/0.998
	0.0005		0.922/1.000	0.919/1.000	0.921/1.000		0.692/1.000	0.695/0.999	0.688/0.998		0.462/1.000	0.458/1.000	0.462/1.000
ResNet34	0.0001	0.925	0.922/0.999	0.922/0.997	0.924/1.000	0.702	0.701/0.996	0.701/0.993	0.700/1.000	0.525	0.449/0.997	0.481/1.000	0.480/0.981
	0.0005		0.921/1.000	0.925/1.000	0.924/1.000		0.701/1.000	0.700/1.000	0.701/0.999		0.462/1.000	0.456/1.000	0.467/1.000
VGG13BN	0.0001	0.919	0.920/0.994	0.920/1.000	0.918/1.000	0.701	0.704/0.998	0.706/0.997	0.698/0.997	0.493	0.459/0.997	0.467/0.998	0.460/0.998
	0.0005		0.921/1.000	0.915/1.000	0.920/1.000		0.699/0.999	0.703/0.999	0.698/0.999		0.457/1.000	0.462/1.000	0.463/1.000

[1] *Note:* For Table I, II, IV, and V: Each cell contains two values: CA / ASR.
[2] The columns represent different attack optimization models (e.g., ResNet18, ResNet34, VGG13BN) for the corresponding dataset.
[3] The underlined cells correspond to the white-box attacker A1, while the remaining cells correspond to the black-box attacker A2.

this, the data-free attack builds upon the principles of dirty-label attacks by leveraging auxiliary datasets and models from other domains.

ShadowPrint sets itself apart from many SOTA backdoor attacks by employing a remarkably low poison rate while retaining its ability to execute multiple attack types. As shown in Fig. 3, the trigger induces subtle differences between poisoned and clean samples, ensuring high stealth and effectiveness. This makes *ShadowPrint* harder for existing defenses to detect, offering a versatile and potent backdoor attack method.

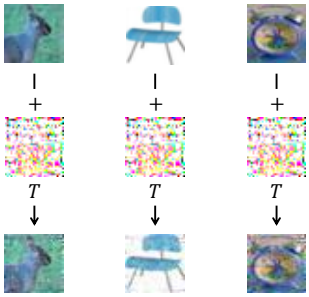

Fig. 3. Visualization of the poisoning process. The visualization depicts poisoned samples created using the backdoor trigger.

IV. EVALUATION

A. Experimental Settings

Datasets and Models. We evaluate *ShadowPrint* on three commonly used models: ResNet18 [24], ResNet34, and VGG13BN [25]. The datasets used for evaluation include CIFAR-10 [26], CIFAR-100 [26], and TinyImageNet [27], which are widely adopted for image classification tasks. The models are trained on these datasets, and the backdoor attack is introduced by poisoning the training data at varying rates.

Attack Settings. In our evaluation, we experiment with different poisoning rates (e.g., 0.01% and 0.05%, much lower than common 0.1% or higher settings), attack methods (e.g., clean-label and dirty-label poisoning, wider variety than other single-method attack), and attacker's capabilities (e.g., Scenario A1, A2, and A3, more types of Scenario Assumption and lower capability configurations than other attacks). We also study the impact of various factors, such as trigger weight and the scale of the attacker's dataset (i.e., train scale), on the attack's success. The poison rate is defined as the fraction of poisoned samples in the dataset, while the trigger weight refers to the strength of the trigger's influence in the backdoor attack, as detailed in (1).

Evaluation Metrics. We measure the following key metrics to evaluate the effectiveness and stealthiness of *ShadowPrint*:

- **Attack Success Rate (ASR):** The percentage of poisoned samples that are misclassified to the target label. Higher ASR indicates better attack effectiveness.
- **Clean Accuracy (CA):** The model accuracy on clean samples under the attack. Higher CA indicates a lower impact on the model's usability by the attack, thereby demonstrating better stealthiness of the attack.
- **Defense Detection Rate (DDR):** The effectiveness of *ShadowPrint* in evading state-of-the-art defense mechanisms. The value ranges from 0 to 1, where a lower value indicates a better evasion of defense mechanisms.

B. Attack Effectiveness and Stealthiness

Dirty-Label & Clean-Label Attack. We evaluate *ShadowPrint* in both dirty-label and clean-label setting under the capability of Scenario A1 and A2, using CIFAR-10, CIFAR-100, and TinyImageNet as evaluation datasets. In the dirty-label attack setting, where the training dataset contains mislabeled samples, *ShadowPrint* achieves high ASRs across all datasets, as shown in Table I, while maintaining high CA, demonstrating its stealthiness and minimal disruption to benign input performance. Similarly, in the clean-label attack setting, where all training samples are correctly labeled, *ShadowPrint* achieves comparable ASRs with little to no impact on CA, as evidenced in Table II. Finally even with a low poison rate, *ShadowPrint* effectively disrupts the model's decision-making process in both settings, showcasing its versatility and effectiveness in evading detection. For instance, with a poison rate of 0.05%, *ShadowPrint* attains ASRs exceeding 95% while retaining CA above 92%, highlighting its balance between attack effectiveness and stealth.

Data-Free Attack. We also test *ShadowPrint* in a data-free scenario, where the attacker has no access to the training data (i.e., Scenario A3). Under this setting, We assume that the attacker employs a surrogate model (f_{adv} = ResNet34) and

TABLE II
CLEAN LABEL ATTACK. WE EVALUATE *ShadowPrint* UNDER THE CLEAN-LABEL ATTACK MODE BY TESTING THE CA AND ASR ACROSS DIFFERENT POISON RATES AND ATTACKER SETTINGS.

Target Model	Poison Ratio	CIFAR-10				CIFAR-100				TinyImageNet			
		Baseline	*ResNet18*	*ResNet34*	*VGG13BN*	*Baseline*	*ResNet18*	*ResNet34*	*VGG13BN*	*Baseline*	*ResNet18*	*ResNet34*	*VGG13BN*
ResNet18	0.0001	0.920	0.923/0.998	0.922/0.962	0.920/0.890	0.690	0.689/0.999	0.699/0.998	0.689/0.964	0.508	0.453/0.999	0.459/0.999	0.462/0.991
	0.0005		0.926/1.000	0.923/1.000	0.923/0.999		0.699/0.999	0.693/1.000	0.684/1.000		0.457/1.000	0.463/1.000	0.461/1.000
ResNet34	0.0001	0.925	0.924/0.947	0.925/0.999	0.924/0.999	0.702	0.692/0.999	0.689/0.998	0.695/0.984	0.525	0.468/0.989	0.477/0.998	0.469/0.996
	0.0005		0.923/1.000	0.930/1.000	0.926/1.000		0.697/0.998	0.696/0.999	0.699/0.997		0.463/1.000	0.469/1.000	0.470/1.000
VGG13BN	0.0001	0.919	0.920/0.943	0.919/1.000	0.921/0.999	0.701	0.703/1.000	0.704/0.994	0.702/0.986	0.493	0.459/0.999	0.461/1.000	0.465/0.998
	0.0005		0.920/1.000	0.918/1.000	0.916/1.000		0.696/0.996	0.701/0.999	0.702/0.986		0.462/1.000	0.461/1.000	0.461/1.000

TABLE III
DATA-FREE ATTACK

Target Model	Target Dataset	Baseline	CA/ASR
ResNet18	CIFAR-10	0.920	0.908/0.913
	TinyImageNet	0.508	0.458/1.000
VGG13BN	CIFAR-10	0.919	0.903/0.859
	TinyImageNet	0.493	0.465/1.000

TABLE IV
TRAIN SCALE STUDY

Target Model	Train Scale	CIFAR-10			
		Baseline	*ResNet18*	*ResNet34*	*VGG13BN*
ResNet18	0.1	0.920	0.923/0.933	0.922/0.992	0.925/0.992
	0.2		0.924/0.999	0.923/0.999	0.921/1.000
	0.3		0.926/0.999	0.923/1.000	0.923/1.000
	0.5		0.925/0.999	0.921/1.000	0.921/1.000
ResNet34	0.1	0.925	0.921/0.973	0.927/0.966	0.922/0.988
	0.2		0.925/1.000	0.924/0.998	0.925/1.000
	0.3		0.926/1.000	0.920/0.999	0.922/1.000
	0.5		0.926/1.000	0.920/1.000	0.922/1.000
VGG13BN	0.1	0.919	0.919/0.981	0.917/0.975	0.917/0.998
	0.2		0.923/1.000	0.917/0.997	0.918/1.000
	0.3		0.920/1.000	0.916/1.000	0.920/1.000
	0.5		0.918/1.000	0.920/1.000	0.917/1.000

auxiliary data ($\mathcal{D}_{\text{adv}} = \text{CIFAR100}$) for an approximation. Table III shows the evaluation of the robustness of *ShadowPrint* under limited access conditions. Despite the lack of access to the training data, *ShadowPrint* still manages to perform a successful attack, with a reasonable ASR and minimal impact on CA. The results highlight the flexibility and robustness of *ShadowPrint* in varying attacker scenarios.

C. Evasion of Backdoor Detection

ShadowPrint is evaluated against SOTA defense mechanisms, including IBD-PSC [8], SCALE UP [9], and Beatrix [10]. Table VI shows that *ShadowPrint* achieves low DDRs across all attacker scenarios. For example, under Scenario A1, *ShadowPrint* records DDR values as low as 0.05, outperforming comparable methods and demonstrating its ability to evade detection while maintaining high ASRs and CA. These results underscore the method's stealth and effectiveness, even against robust defensive measures.

We evaluate the effectiveness of *ShadowPrint* against several SOTA defense mechanisms designed to detect backdoor attacks. These include IBD-PSC [8], SCALE UP [9], and Beatrix [10]. As shown in Table VI, under all attacker's scenarios, *ShadowPrint* successfully records DDR values lower than 0.12%, outperforming comparable methods and demonstrating its ability to evade detection while maintaining high ASRs and CA. These results underscore the method's stealth and effectiveness, even against robust defensive measures.

D. Ablation Study

To further understand the behavior of *ShadowPrint*, we conduct an ablation study by evaluating the attack's performance under various configurations. Specifically, we investigate the effect of the poison rate, trigger weight, and the scale of the attacker's dataset on the attack's success and stealth.

Poison Rate. Table II and Table I show the impact of different poison rates on ASR and CA in different settings. As expected, increasing the poison rate results in a higher ASR. However, *ShadowPrint* maintains its high stealth even at higher poison rates, as evidenced by the minimal impact on CA. This stability highlights the stealthiness of the method and its ability to minimize disruption to clean inputs.

Trigger Weight. We also study the impact of various trigger weights. As shown in Table V, the ASR increases as the trigger weight is adjusted, but the CA remains stable. The ablation study demonstrates that the attack can be finely tuned to balance attack success and stealth.

Train Scale. In Table IV, we evaluate the effect of varying the scale of the attacker's dataset. As expected, the ASR increases with larger training scales, but the model's accuracy on clean samples remains unaffected. This study shows that *ShadowPrint* performs robustly across different levels of attacker knowledge and dataset sizes.

E. Discussion

ShadowPrint proves to be an effective and stealthy backdoor attack, achieving high ASR while maintaining strong CA. Unlike traditional attacks that manipulate input visuals, *ShadowPrint* targets internal feature representations, making it more resistant to detection and defenses. The ablation study shows that it is highly adaptable, balancing effectiveness and stealth through hyperparameters. Even in data-free scenarios, *ShadowPrint* performs well, indicating its potential for real-world applications. However, its reliance on feature space ma-

TABLE V
TRIGGER WEIGHT STUDY

Target Model	Trigger Weight	CIFAR-10			
		Baseline	*ResNet18*	*ResNet34*	*VGG13BN*
ResNet18	0.1	0.920	0.920/0.395	0.924/0.415	0.921/0.294
	0.2		0.925/0.998	0.919/0.994	0.922/1.000
	0.3		0.923/1.000	0.923/1.000	0.924/1.000
	0.5		0.921/1.000	0.923/1.000	0.924/1.000
ResNet34	0.1	0.925	0.924/0.464	0.922/0.500	0.924/0.427
	0.2		0.925/0.991	0.925/0.999	0.924/1.000
	0.3		0.927/0.999	0.929/0.999	0.924/1.000
	0.5		0.923/1.000	0.925/1.000	0.925/1.000
VGG13BN	0.1	0.919	0.921/0.200	0.920/0.568	0.921/0.211
	0.2		0.917/1.000	0.920/0.996	0.919/1.000
	0.3		0.918/1.000	0.920/1.000	0.923/1.000
	0.5		0.923/1.000	0.917/1.000	0.917/1.000

nipulation suggests the need for novel and effective detection strategies to counter such attacks.

TABLE VI
DEFENSE STUDY

Label	Scenario	IBD_PSC	SCALE_UP	Beatrix
DIRTY	White-Box	0.009	0.000	0.054
	Black-Box	0.004	0.000	0.057
	Data-Free	0.004	0.000	0.057
CLEAN	White-Box	0.010	0.000	0.061
	Black-Box	0.046	0.000	0.052
	Data-Free	0.117	0.000	0.063

V. CONCLUSION

In this paper, we presented *ShadowPrint*, a novel backdoor attack that manipulates feature embeddings within a model's feature space to achieve both high attack success and stealth. Leveraging existing method limitations, *ShadowPrint* reduces reliance on strong attacker capabilities and performs well across diverse scenarios. Experimental results also demonstrate *ShadowPrint* excels by effectively disrupting model performance while maintaining high accuracy on clean samples.

ACKNOWLEDGMENT

This work was partly supported by the NSFC under No. U244120033, U24A20336, 62172243, 62402425 and 62402418, the China Postdoctoral Science Foundation under No. 2024M762829, the Zhejiang Provincial Natural Science Foundation under No. LD24F020002, and the Zhejiang Provincial Priority- Funded Postdoctoral Research Project under No. ZJ2024001.

REFERENCES

[1] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," Annual review of biomedical engineering, vol. 19, no. 1, pp. 221–248, 2017.

[2] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," Applied Stochastic Models in Business and Industry, vol. 33, no. 1, pp. 3–12, 2017.

[3] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," IEEE Access, vol. 7, pp. 47 230–47 244, 2019.

[4] Y. Zeng, M. Pan, H. A. Just, L. Lyu, M. Qiu, and R. Jia, "Narcissus: A practical clean-label backdoor attack with limited information," in Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, 2023, pp. 771–785.

[5] 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, vol. 139, p. 109512, 2023.

[6] H. Kim, M. Kim, D. Seo, J. Kim, H. Park, S. Park, H. Jo, K. Kim, Y. Yang, Y. Kim et al., "Nsml: Meet the mlaas platform with a real-world case study," arXiv preprint arXiv:1810.09957, 2018.

[7] M. Ribeiro, K. Grolinger, and M. A. Capretz, "Mlaas: Machine learning as a service," in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 896–902.

[8] L. Hou, R. Feng, Z. Hua, W. Luo, L. Y. Zhang, and Y. Li, "Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency," arXiv preprint arXiv:2405.09786, 2024.

[9] J. Guo, Y. Li, X. Chen, H. Guo, L. Sun, and C. Liu, "Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency," arXiv preprint arXiv:2302.03251, 2023.

[10] W. Ma, D. Wang, R. Sun, M. Xue, S. Wen, and Y. Xiang, "The" beatrix"resurrections: Robust backdoor detection via gram matrices," arXiv preprint arXiv:2209.11715, 2022.

[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] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," arXiv preprint arXiv:2102.10369, 2021.

[13] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in 25th Annual Network And Distributed System Security Symposium (NDSS 2018). Internet Soc, 2018.

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

[15] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," Advances in Neural Information Processing Systems, vol. 33, pp. 3454–3464, 2020.

[16] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 101–105.

[17] Z. Li, H. Sun, P. Xia, H. Li, B. Xia, Y. Wu, and B. Li, "Efficient backdoor attacks for deep neural networks in real-world scenarios," arXiv preprint arXiv:2306.08386, 2023.

[18] Y. Wu, X. Han, H. Qiu, and T. Zhang, "Computation and data efficient backdoor attacks," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4805–4814.

[19] Z. Zhu, M. Zhang, S. Wei, L. Shen, Y. Fan, and B. Wu, "Boosting backdoor attack with a learnable poisoning sample selection strategy," arXiv preprint arXiv:2307.07328, 2023.

[20] 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 2019 IEEE symposium on security and privacy (SP). IEEE, 2019, pp. 707–723.

[21] 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 Proceedings of the 35th annual computer security applications conference, 2019, pp. 113–125.

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

[23] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in International symposium on research in attacks, intrusions, and defenses. Springer, 2018, pp. 273–294.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[26] A. Krizhevsky, G. Hinton et al., "Learning multiple layers of features from tiny images," 2009.

[27] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," CS 231N, vol. 7, no. 7, p. 3, 2015.