

DESIGN AND DEVELOPMENT OF UNTARGETED ATTACKS IN DNN

by Vishal Gupta

Submission date: 29-May-2026 11:09AM (UTC+0530)

Submission ID: 2971804095

File name: Mtech_Thesis.pdf (5.63M)

Word count: 13986

Character count: 86382

**DESIGN AND DEVELOPMENT OF UNTARGETED
ATTACKS IN DNN**

1
Thesis Submitted

in Partial Fulfillment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

by

VISHAL KUMAR GUPTA

(2K24/CSE/26)

Under the Supervision of

Dr. RAJEEV KUMAR

Assistant Professor



Department of Computer Science and Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

May, 2026

ACKNOWLEDGEMENT

¹ I express my sincere gratitude to my supervisor, Dr. Rajeev Kumar, Assistant Professor, Department of Computer Science and Engineering, Delhi Technological University, for his invaluable guidance, constructive suggestions, and continuous encouragement throughout the preparation¹³ of this thesis. His technical insights and critical feedback have significantly shaped the direction and quality of this work.

³⁹ I am grateful to the faculty members and administration of the Department of Computer Science and Engineering, Delhi Technological University, for providing a conducive academic environment. The institutional infrastructure and resources have been instrumental in enabling the systematic study of the evolving security challenges of deep neural networks, particularly the emerging risks associated with backdoor attacks, pre-trained models, fine-tuning, and parameter-efficient adaptation.

I am especially indebted to the researchers whose pioneering work forms the foundation of this thesis. Their contributions to untargeted backdoor attacks, CLIP-guided semantic target generation, prototype-based learning, and fine-tuning-centred backdoor security provided the essential technical basis for this thesis-level expansion. I also acknowledge the role of open benchmark datasets and deep learning model architectures that have made systematic evaluation of backdoor security feasible.

²⁷ Finally, I thank my family and friends for their steadfast support and motivation throughout this academic endeavour. Their encouragement enabled me to complete this thesis with dedication and discipline.

Place: Delhi

Vishal Kumar Gupta

Date: _____

CANDIDATE'S DECLARATION



¹
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi - 110042

I, **Vishal Kumar Gupta** (Roll No. 2K24/CSE/26), hereby declare that the thesis entitled **“DESIGN AND DEVELOPMENT OF UNTARGETED ATTACKS IN DNN”** submitted to **Delhi Technological University**, Delhi, in partial fulfillment of the requirements for the award of the degree of *Master of Technology in Computer Science and Engineering*, is my original academic work prepared under the guidance of Dr. Rajeve Kumar, Assistant Professor, Department of Computer Science and Engineering, Delhi Technological University.

²
The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

³
This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner

CERTIFICATE



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi - 110042

This is to certify that the thesis entitled “**DESIGN AND DEVELOPMENT OF UNTARGETED ATTACKS IN DNN**” submitted by **Vishu Kumar Gupta**, Roll No. 2K24/CSE/26, to **Delhi Technological University**, Delhi, in partial fulfillment of the requirements for the award of the degree of *Master of Technology in Computer Science and Engineering*, is a record of bonafide work carried out under my supervision and guidance.

The work presented in this thesis is based on the study and thesis-level expansion of two research contributions related to backdoor security in deep neural networks. The thesis has been prepared in the prescribed academic format and is submitted after institutional verification and formatting review.

(Dr. RAJEEV KUMAR)

(Assistant Professor)

(Department of Computer Science and Engineering)

(Delhi Technological University)

Date: _____

Design and Development of Untargeted Attacks in DNN

Vishal Kumar Gupta

ABSTRACT

Deep neural networks have become the foundation of modern artificial intelligence systems applied to computer vision, natural language processing, autonomous sensing, medical imaging, code generation, and multimodal decision making. Their exceptional performance, however, is accompanied by increasing exposure to adversarial threats. Among these, backdoor attacks represent a particularly serious security vulnerability. A backdoored model behaves normally on clean inputs but exhibits attacker-induced erroneous behaviour when a concealed trigger is present. This dual-mode operation allows malicious functionality to remain undetected during standard evaluation while posing genuine risks upon deployment.

This thesis examines backdoor security in deep neural networks through two distinct but complementary angles. The first contribution investigates TWIN, a fine-tuning-centred survey framework that analyses fine-tuning as both a weapon and a shield in backdoor security. Modern artificial intelligence practice increasingly relies on pre-trained models, modelhubs, plug-in adapters, prompt tuning, LoRA-style parameter-efficient fine-tuning, and downstream adaptation. In this context, fine-tuning may mitigate backdoors but may also activate, preserve, or implant them. TWIN organises attacks and defences around this dual role and provides taxonomies for fine-tuning-related attacks, fine-tuning-based defences, PEFT-centric vulnerabilities, adaptation-stage monitoring, weight-space repair, and representation-level purification.

The second contribution presents ProAttack, an advanced untargeted backdoor attack that causes semantically plausible misclassifications rather than arbitrary errors. In contrast to fixed-target attacks, ProAttack guides triggered inputs toward visually coherent wrong classes by leveraging CLIP ViT-B/32 embeddings, class prototype-based destination selection, and a jointly trained U-Net perturbation generator. Benchmark evaluations across CIFAR-10, GTSRB, and CIFAR-100 confirm strong attack effectiveness with negligible clean accuracy loss and imperceptible trigger patterns.

By combining these two contributions, the thesis offers a unified view of contemporary backdoor security. ProAttack demonstrates how semantic guidance can strengthen untargeted attacks, while TWIN clarifies why the fine-tuning stage has become a critical control point in the model lifecycle. The thesis further discusses the implications of these contributions for secure deployment, defence design, trustworthy pre-trained models, and future research directions.

Keywords: Backdoor attacks, deep neural networks, untargeted attack, CLIP embeddings, prototype-guided learning, U-Net trigger generator, fine-tuning, pre-trained models, parameter-efficient fine-tuning, backdoor defence.

LIST OF PUBLICATIONS

1. V. K. Gupta and R. Kumar, "TWIN: A Survey of Fine-Tuning as Both Weapon and Shield in Backdoor Security," submitted to the International Conference on Intelligent Computing, Control & Communication (ICICCC-2026), 2026.
2. V. K. Gupta and R. Kumar, "ProAttack: Prototype-C₄ded Untargeted Backdoor Attack on Deep Neural Networks," submitted to the International Conference on Intelligent Computing, Control & Communication (ICICCC-2026), 2026.

TABLE OF CONTENTS

Acknowledgement	18
Candidate's Declaration	ii
Certificate	iii
Abstract	iv
List of Publications	v
Table of Contents	viii
List of Tables	ix
List of Figures	x
List of Abbreviations	xi
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Domain Overview	2
1.3 Existing Challenges	2
1.4 Problem Statement	3
1.5 Research Objectives	4
1.6 Scope of the Research	4
1.7 Applications and Relevance	5
1.8 Research Contributions	5
1.9 Thesis Organisation	6
2 Literature Review	7
2.1 Backdoor Attacks in DNN	7
2.2 Fine-Tuning in Pre-Trained Models	7
2.3 Parameter-Efficient Fine-Tuning and Adapter Risks	8
2.4 Fine-Tuning-Based Backdoor Defences	9
2.5 Targeted and Untargeted Backdoor Threat Models	9
2.6 Vision-Language Priors and Semantic Guidance	10
2.7 Prototype-Based Learning in Backdoor Attacks	10
2.8 Backdoor Defences Against Untargeted Attacks	10

2.9	Research Gaps	11
2.10	Summary	11
3	TWIN: A Survey of Fine-Tuning as Both Weapon and Shield in Backdoor Security	12
3.1	Introduction	12
3.2	Background: Attack Mechanism and the TWIN Taxonomy	12
3.2.1	Backdoor Attack Mechanism in Fine-Tuned Models	12
3.2.2	Taxonomy Structure and Principal Axes	13
3.3	Fine-Tuning-Related Backdoor Attacks	14
3.3.1	Activation-Oriented Attacks	14
3.3.2	Resilience-Oriented and Supply-Chain Attacks	15
3.3.3	Chronological Summary of Attacks	15
3.4	Fine-Tuning-Related Backdoor Defences	16
3.4.1	Post-Hoc Purification and Representation-Level Methods	16
3.4.2	PEFT Artefact Inspection and In-Process Monitoring	17
3.4.3	Summary of Defences and Evaluation Metrics	18
3.5	Analysis, Challenges, and Practical Implications	19
3.5.1	Structural Correspondence Between Attacks and Defences	19
3.5.2	Attack-Defence Asymmetry and Research Gaps	19
3.5.3	Open Research Challenges	20
3.5.4	Practical Recommendations for Stakeholders	20
3.6	Summary	21
4	ProAttack: Prototype-Guided Untargeted Backdoor Attack on DNN	22
4.1	Introduction and Motivation	22
4.2	System Overview and Problem Formulation	22
4.3	Proposed Methodology	23
4.3.1	CLIP-Driven Target Generation	23
4.3.2	Prototype Computation and Destination Assignment	25
4.3.3	U-Net Trigger Generator	25
4.3.4	Joint Training Objective	26
4.3.5	Algorithmic Workflow	27
4.4	Experimental Evaluation	28
4.4.1	Experimental Setup and Evaluation Metrics	28
4.4.2	Results and Comparative Analysis	28
4.5	Discussion	29
4.5.1	Attack Effectiveness, Imperceptibility, and Complexity	29
4.5.2	Advantages, Limitations, and Security Implications	29
4.6	Summary	30
5	Conclusion and Future Scope	31
5.1	Conclusion	31
5.2	Major Findings	31
5.3	Limitations	32
5.4	Future Scope	32

References	33
List of Publications and Their Proofs	36
Plagiarism Verification	37
Curriculum Vitae	38

List of Tables

3.1	Chronological summary of fine-tuning-related backdoor attacks. . . .	16
3.2	Summary of fine-tuning-related backdoor defences.	18
4.1	Attack performance of ProAttack across benchmark datasets.	28
4.2	Trigger imperceptibility metrics on the test set.	28
4.3	Comparison with representative backdoor attacks on CIFAR-10/ResNet-18.	29

List of Figures

3.1	Overview of a backdoor attack in fine-tuned models. Clean inputs behave normally, whereas triggered inputs are redirected according to attacker intent.	13
3.2	The TWIN taxonomy showing fine-tuning as both an adversarial activation vector and a defensive immunisation mechanism.	14
4.1	Overall architecture of ProAttack showing CLIP-driven target generation, prototype-guided destination assignment, and joint backdoor training.	23
4.2	CLIP-driven semantic target generation pipeline producing the per-class plausible destination set $T(c)$	24
4.3	Four-level U-Net trigger generator. Output is tanh-bounded and scaled by $\varepsilon = 0.05$	26

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
ASR	Attack Success Rate
AUC	Area Under the Curve
CA	Clean Accuracy
CLIP	Contrastive Language–Image Pre-training
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DS	Destination Score
FT	Fine-Tuning
GTSRB	German Traffic Sign Recognition Benchmark
KD	Knowledge Distillation
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LPIPS	Learned Perceptual Image Patch Similarity
MLaaS	Machine Learning as a Service
PEFT	Parameter-Efficient Fine-Tuning
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index
TWIN	Two-sided Weaponization and Immunization through fiNe-tuning
UBA	Untargeted Backdoor Attack

INTRODUCTION

1.1 Background and Motivation

Artificial intelligence systems based on DNN are now deployed in domains where reliability, safety, and trust are essential. Image classifiers assist automated inspection systems, traffic sign recognition modules support autonomous driving, language models support reasoning and code generation, and multimodal models combine visual and textual knowledge to perform complex tasks. In many of these applications, the underlying model is not trained from scratch by the final user. Instead, developers commonly begin with a pre-trained model, download a checkpoint or adapter, fine-tune it on task-specific data, and deploy the resulting model in a downstream environment [20, 22]. This development paradigm has accelerated progress but has also expanded the security surface of modern machine learning systems.

Backdoor attacks exploit this expanded attack surface. A backdoor is a hidden behaviour embedded into a model such that the model performs normally on clean inputs but behaves maliciously when a trigger is present [1, 2]. The trigger may be a visible patch [1], an imperceptible perturbation [3], a frequency-domain pattern [8], a style cue, a prompt token [28, 29], a compromised adapter [23], or a process-level condition such as fine-tuning [30, 33]. Because the model maintains high clean accuracy, ordinary validation may fail to reveal the presence of the backdoor. This characteristic makes backdoor attacks particularly dangerous for outsourced training, model reuse, cloud deployment, and foundation model adaptation [24, 25].

At the same time, modern deployment practices depend heavily on fine-tuning. Fine-tuning is no longer a simple post-training adjustment; it is a central step in model adaptation, adapter sharing, prompt learning, model repair, and domain transfer [20]. Fine-tuning may suppress an existing backdoor by weakening the trigger association [20, 21], but it may also activate a dormant backdoor [33, 30] or preserve malicious behaviour across downstream adaptation [26, 24]. The first research contribution examined in this thesis, TWIN, analyses this dual role of fine-tuning as both an adversarial weapon and a defensive shield.

Traditional backdoor studies focus primarily on targeted attacks [1, 2, 18]. In a targeted setting, all triggered inputs are mapped to a fixed attacker-selected class. Although effective, such deterministic behaviour can become detectable because all triggered samples collapse to the same output [16, 6]. Untargeted backdoor attacks relax this constraint and only aim to cause an incorrect prediction [12]. This makes the attack objective more flexible and more suitable for disruption-oriented scenarios. However, if incorrect labels are assigned without semantic structure, the resulting misclassifications may appear arbitrary or implausible [12]. The second research contribu-

tion examined in this thesis addresses this issue by proposing ProAttack, a prototype-guided untargeted backdoor attack that uses visual-semantic priors from CLIP [13] to guide triggered samples toward plausible destination classes.

The motivation of this thesis therefore lies in the need to understand backdoor security at two levels. The first level examines how the model lifecycle, particularly fine-tuning, affects both attack survival [26, 25, 24, 27] and defensive mitigation [20, 21, 22]. The second level examines how advanced attacks can be made more semantically plausible and less predictable [12, 13]. Together, these perspectives provide a comprehensive view of current and emerging backdoor threats in deep neural networks.

1.2 Problem Domain Overview

Backdoor security is concerned with the behaviour of models under hidden trigger conditions. A clean classifier correctly maps an input image to its true label. A backdoored classifier could behave similarly for clean inputs so that it passes standard validation; however, when the trigger is inserted into the input or the model is exposed to a specific activation condition, the classifier produces an attacker-influenced output [1]. The threat becomes more difficult to address when the trigger is imperceptible, dynamic, input-dependent, or distributed across model weights and adapters [12, 23].

The problem domain includes several interacting components. The first component is the attack mechanism: attacks may poison data, modify labels [1, 2], alter weights [26, 27], tune adapters [28, 29], manipulate prompts, or exploit downstream fine-tuning [30, 33, 32]. The second component is the trigger representation: triggers may be spatial patches [1], blended perturbations [2], warping fields [18], frequency-domain patterns [8], style changes [5], textual cues [28], or latent process conditions [30]. The third component is the target behaviour: traditional attacks enforce a fixed target [1], whereas untargeted attacks only require misclassification [12]. The fourth component is the defence mechanism, which may use input filtering, trigger inversion [16], pruning [17, 9], fine-tuning [20, 21], distillation [22], parameter-space detection [23], or monitoring during adaptation.

In this thesis, the focus is on two important aspects of this domain. First, the thesis studies fine-tuning-related attacks and defences through the TWIN taxonomy, which captures how fine-tuning can preserve, activate, complete, repair, or detect backdoors in pre-trained models. Second, the thesis studies untargeted backdoor attacks that are not restricted to a fixed target class, and in particular focuses on ProAttack, which addresses the weakness of unconstrained untargeted attacks [12] by introducing semantic target generation via CLIP [13] and prototype-guided destination assignment.

1.3 Existing Challenges

Backdoor security remains difficult because the threat is hidden and conditional. A model may achieve high clean accuracy and still contain malicious behaviour [1, 19]. This creates a mismatch between ordinary performance evaluation and security evaluation. Several challenges arise from this mismatch.

First, modern models are typically adapted after pre-training, and the effect of this adaptation on embedded backdoors is not uniform [20, 26, 25]. A backdoor may weaken during fine-tuning, survive fine-tuning [26, 24, 27], or remain dormant until activated by a downstream adaptation step [33, 30, 32]. Evaluating model security at a single point in time is therefore insufficient; security must be assessed across every adaptation phase.

Second, parameter-efficient fine-tuning introduces additional risks [28, 29, 23]. Adapters, prompt vectors, and LoRA-style modules are compact and easily distributed, yet these artefacts may carry hidden behaviour independently of the base model. Detecting malicious tuning components is challenging because they may appear benign when inspected in isolation [23].

Third, defence evaluation remains fragmented. Many defences are evaluated against a narrow set of attack types [16, 17, 9, 10], whereas emerging attacks exploit fine-tuning trajectories [30, 31], agentic traces, multimodal signals, or dormant triggers [33]. A systematic taxonomy is required to identify where current defences are effective and where gaps persist.

Fourth, many targeted backdoor attacks exhibit predictable output behaviour [1, 18], and that predictability enables defence approaches that search for frequent targets, anomalous class dynamics, or consistent trigger responses [16, 6, 7]. In contrast, untargeted attacks reduce this predictability [12] but, when they lack structural guidance, may also produce semantically unrealistic misclassification patterns. A fundamental challenge, therefore, is to construct an attack that retains flexibility while remaining semantically plausible.

Fifth, balancing trigger imperceptibility with attack success is a persistent difficulty [18, 8, 12]. A highly visible trigger tends to increase attack success but is more susceptible to detection [6, 7]. An imperceptible trigger is more stealthy but is more difficult to optimise. Furthermore, making trigger generation input-dependent improves concealment but increases the complexity of training [12].

1.4 Problem Statement

The central problem addressed in this thesis is the need to understand and organise modern backdoor security in DNN under two complementary conditions: fine-tuning-centred model adaptation and semantically guided untargeted attack behaviour.

More specifically, the thesis examines the following problems:

1. How can fine-tuning be understood as both a mechanism for backdoor activation [33, 30, 32, 31] and a mechanism for backdoor mitigation [20, 21, 22] in modern pre-trained models?
2. How can untargeted backdoor attacks [12] be made more semantically plausible and stealthy using visual-semantic priors [13]?

This problem is addressed through two research contributions. The first contribution presents TWIN as a comprehensive fine-tuning-centred analysis of backdoor attacks and defences covering attacks that survive [26, 24, 25, 27], activate after [33, 30, 31], or exploit [28, 29, 23] fine-tuning. The second contribution introduces

ProAttack, a prototype-guided untargeted backdoor method that builds on CLIP embeddings [13] and extends the untargeted threat model established by UBA [12]. The combined thesis examines not only the broader lifecycle in which attacks may be implanted, preserved, detected, or neutralised, but also the target-free attack construction itself.

1.5 Research Objectives

The primary objectives of this thesis are as follows:

1. To study the fundamentals of backdoor attacks in deep neural networks [1, 2, 3], including targeted and untargeted threat models [12].
2. To analyse the limitations of conventional targeted attacks [1, 18] and unconstrained untargeted attacks [12].
3. To study fine-tuning as a central axis in backdoor security and explain how it functions both as an attack vector [33, 30] and a defence mechanism [20, 21, 22].
4. To present the TWIN taxonomy of fine-tuning-related attacks and defences, including process-as-trigger attacks [33, 30], fine-tuning-completed backdoors [32], fine-tuning-resilient attacks [26, 24, 25], PEFT-centric attacks [28, 29], and fine-tuning-based defences [20, 21, 23].
5. To provide a detailed thesis-level explanation of ProAttack, including CLIP-driven target generation [13], prototype-guided destination assignment, U-Net-based trigger generation, and joint optimisation [12].
6. To examine the experimental outcomes reported for ProAttack on CIFAR-10 [14], GTSRB [15], and CIFAR-100 [14], including attack success, clean accuracy retention, and trigger imperceptibility.
7. To compare the two contributions and discuss their combined implications for model security, trustworthy adaptation, and future defence design.

1.6 Scope of the Research

The scope of this thesis is limited to the academic analysis of backdoor attacks and defences in DNN. It does not introduce new experiments beyond those reported in the two source papers; instead, it expands the methodology, background, mathematical formulation, results, and discussion into a thesis format. The thesis focuses on pre-trained model adaptation in the TWIN contribution and on image classification datasets [14, 15] in the ProAttack contribution. The thesis covers the following areas:

- Backdoor attacks in DNN classifiers [1, 2, 3, 4].
- Fine-tuning-related attack and defence taxonomies [20, 26, 33].

- PEFT and adapter-related backdoor risks [28, 29, 23].
- Untargeted backdoor behaviour and semantic destination control [12].
- CLIP-based visual-semantic similarity for destination-set construction [13].
- Prototype-based feature-space destination assignment.
- U-Net-based input-dependent trigger generation.
- Research gaps and future directions in backdoor security.

The thesis does not provide executable attack code or deployment instructions. Its purpose is research analysis and academic documentation.

1.7 Applications and Relevance

Backdoor security is relevant in any setting where machine learning models are trained, adapted, transferred, or reused. In autonomous driving, a backdoored traffic sign recognition model may misclassify triggered signs while appearing reliable on all other inputs [15]. In biometric authentication, a concealed cue may enable unauthorised access. In medical imaging, a backdoor may shift diagnostic predictions when subtle image perturbations are applied. In natural language systems, prompt-level or adapter-level triggers may alter generated text and its semantic meaning [28, 29]. In code-generation models, backdoors may introduce insecure code patterns under rare conditions [24]. In agentic systems, poisoned task traces may influence tool selection or planning behaviour [30].

The TWIN contribution is relevant because it examines the adaptation pipeline itself. As model hubs, fine-tuned checkpoints, LoRA adapters, and prompt modules become commonplace [20, 23], security evaluation must move beyond static inspection and encompass provenance, auditing, and monitoring across the entire model lifecycle. The ProAttack contribution is relevant because it studies a stronger form of untargeted attack [12] that avoids the rigidity of a fixed target label and leverages visual-semantic similarity from CLIP [13] to produce plausible misclassification behaviour that is more stealthy and harder to detect [6, 7, 19].

1.8 Research Contributions

This thesis is built on two major research contributions:

Contribution 1: TWIN: A Survey of Fine-Tuning as Both Weapon and Shield in Backdoor Security. The first contribution explains TWIN, a survey framework that organises fine-tuning-related backdoor attacks and defences around the dual role of fine-tuning as both weaponisation and immunisation. The framework covers attacks that survive fine-tuning [26, 24, 25, 27], attacks activated by fine-tuning [33, 30, 31, 32], PEFT-centric attacks [28, 29], and defences including

fine-tuning-based purification [20, 21, 22] and PEFT artefact inspection [23].

Contribution 2: ProAttack: Prototype-Guided Untargeted Backdoor Attack on DNN. The second contribution explains ProAttack, which extends the untargeted threat model of UBA [12] by integrating CLIP-driven visually coherent target generation [13], prototype-guided destination assignment, and a U-Net trigger generator to produce stealthy untargeted backdoor behaviour with semantically plausible misclassifications, validated on CIFAR-10 [14], GTSRB [15], and CIFAR-100 [14].

Together, these contributions connect a lifecycle-level security taxonomy with an attack-level methodology. This combination allows the thesis to discuss not only where modern backdoor attacks fit within the broader deployment and defence ecosystem, but also how such attacks can be constructed to be simultaneously flexible, semantically coherent, and resistant to established detection methods [6, 16, 9, 10, 11].

1.9 ⁷ Thesis Organisation

The remainder of this thesis is organised as follows. Chapter 2 reviews existing literature on backdoor attacks [1, 2, 18, 8], fine-tuning vulnerabilities [26, 33, 30], PEFT risks [28, 29, 23], and defence mechanisms [16, 17, 9, 10, 11, 20, 21, 22]. Chapter 3 presents TWIN: A Survey of Fine-Tuning as Both Weapon⁴⁵ and Shield in Backdoor Security. Chapter 4 presents ProAttack: Prototype-Guided Untargeted Backdoor Attack⁴⁷ on Deep Neural Networks, along with its methodology and experimental analysis on CIFAR-10, GTSRB, and CIFAR-100 [14, 15]. Finally, Chapter 5 concludes the thesis with major findings, limitations, and future research directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Backdoor Attacks in DNN

A backdoor attack embeds malicious behaviour into a model while preserving normal performance on clean inputs. The canonical example is BadNets, where a trigger pattern is inserted into selected training samples and the labels of those samples are changed to a target class [1]. The trained model learns the trigger as a shortcut feature; during inference, any input containing the trigger is classified as the target label. This mechanism demonstrated that DNN could be manipulated through training data poisoning while remaining accurate on clean validation data.

Subsequent attacks improved stealth and practicality. Blended backdoor attacks insert low-opacity triggers that are more difficult to detect visually [2]. Reflection-based attacks use natural reflection patterns as triggers [5]. Warping-based attacks such as WaNet create input transformations that are less conspicuous than patch triggers [18]. Frequency-domain attacks exploit spectral perturbations, demonstrating that triggers need not be apparent in the spatial domain [8]. Feature-space attacks manipulate internal representations rather than relying solely on visible input patterns.

Backdoor attacks may be classified as dirty-label or clean-label. Dirty-label attacks modify both the input and its label [1, 2]. Clean-label attacks retain the ground-truth label, making poisoned examples more difficult to identify through manual inspection [3]. Physical-trigger attacks use real-world objects or patterns that can activate the backdoor after image capture. Semantic-trigger attacks activate behaviour based on content properties, styles, or linguistic patterns rather than explicit patches.

2.2 Fine-Tuning in Pre-Trained Models

Modern machine learning increasingly relies on pre-trained models. A model may be trained once on a large dataset and then adapted to many downstream tasks through fine-tuning, which updates model parameters using task-specific data. This paradigm has become the dominant deployment pattern across image classification, natural language processing, code generation, and multimodal systems [20, 22].

Fine-tuning also fundamentally alters the backdoor threat landscape. A backdoor may survive fine-tuning if it is embedded in robust parameter regions that resist downstream update [26, 24]. It may remain dormant before adaptation and activate only after fine-tuning [33, 30]. It may be carried by an adapter rather than by the full model [28, 29]. It may be completed when downstream labels are introduced [32]. It may also be mitigated by clean fine-tuning on trusted data [20, 21]. This dual role—fine-tuning as both attack vector and defensive primitive—is the central organising axis

of TWIN.

Early work established that fine-tuning could preserve backdoors across tasks. RIPPLES demonstrated that weight-poisoning attacks on pretrained NLP models survived standard downstream fine-tuning [26]. BadPre and POR showed analogous survival in BERT-family and general NLP pretrained models [24, 25]. Layerwise weight poisoning (LWP) identified that certain transformer layers are less affected by adaptation and exploited this to improve backdoor persistence [27]. These findings established that fine-tuning alone does not guarantee backdoor removal.

More recent work has shown that fine-tuning can itself be weaponised as the activation mechanism. Hibernated backdoor attacks keep malicious behaviour completely dormant during pre-deployment inspection and activate it only when the downstream user applies supervised fine-tuning [33]. DeferBad further formalises this deferred activation paradigm for vision models. FAB extends this concept to large language models, showing that dormant adversarial behaviours can be activated by instruction fine-tuning [30]. Dormant Backdoor demonstrates the same threat in vision-language models. SCAR introduces a distillation-conditional variant in which the backdoor activates during knowledge distillation rather than supervised fine-tuning [31]. Together, these attacks demonstrate that process-as-trigger backdoors are not confined to a single architecture or domain.

Fine-tuning has also been shown to complete partially implanted backdoors. Latent backdoor attacks embed an incomplete adversarial pathway into a source model; the backdoor becomes functional only when the downstream user introduces the relevant target label during adaptation [32]. This transfer learning exploit confirms that the threat is distributed across the supply chain rather than localised to a single training stage.

2.3 Parameter-Efficient Fine-Tuning and Adapter Risks

Parameter-efficient fine-tuning (PEFT) methods such as prompt tuning, prefix tuning, and LoRA have become widespread because they allow large models to be adapted with far fewer trainable parameters. The compact size of PEFT artefacts makes them easy to distribute and reuse. This creates a supply-chain risk: a malicious adapter or prompt vector may carry hidden behaviour while appearing to be a useful tuning module [28, 29, 23]. Since the base model may be clean, inspection of the base model alone is insufficient.

PPT and BadPrompt demonstrated that backdoors can be embedded directly into prompt vectors and continuous prompt modules, bypassing base-model inspection [28, 29]. PADBench documented that LoRA, QLoRA, DoRA, LoRA+, and AdaLoRA can all function as backdoor carriers across classification, named-entity recognition, generation, and vision tasks. UBA-Inf introduced an unlearning-activated variant in which the backdoor remains hidden until a machine-unlearning or model-repair procedure is applied.

Agentic systems introduce an additional attack surface. Silent Sabotage showed that an attacker can poison the interaction traces used to fine-tune AI agents, including web navigation histories and tool-use sequences, resulting in backdoors that influence

future planning and decision-making while bypassing safety filters such as LlamaFire-wall and PromptGuard.

2.4 Fine-Tuning-Based Backdoor Defences

Fine-tuning-based defences are especially relevant in the context of pre-trained models. Since many models are adapted before deployment, fine-tuning on trusted clean data can weaken hidden trigger associations [20, 21]. Super-fine-tuning extended this idea using a cyclical learning-rate schedule to force trigger forgetting during high-LR phases while preserving clean-task utility during low-LR phases, reducing ASR from 99.8% to 8.1% [20]. DeepSweep applied complementary data augmentations to disrupt trigger patterns, achieving an average ASR of 0.053 across eight attacks [21]. SSL-KD used self-supervised knowledge distillation to realign internal representations with benign behaviour, outperforming Fine-Pruning and Neural Cleanse against multiple attack types [22].

PEFT-specific defences address adapter-level threats. Obliviate operates in-process during fine-tuning, injecting an amplification loss to maximise the ℓ_2 -norm of PEFT weights and a regularisation loss to suppress trigger-related attention, reducing ASR by 83.6% with only 0.78% clean-accuracy loss across Adapter, LoRA, and Prefix-tuning settings [23]. PEFTGuard takes a parameter-space route, training a CNN and MLP meta-classifier on channel- and layer-wise adapter weights to achieve near-100% detection accuracy with AUC of 1.0, transferring zero-shot to QLoRA, LoRA+, and DoRA.

Classical defences also remain relevant. Neural Cleanse searches for minimal perturbations that cause target-class activation [16]. Fine-Pruning removes neurons associated with backdoor behaviour [17]. STRIP detects backdoor inputs by measuring prediction entropy under input perturbation [6]. SentiNet localises suspicious regions and tests their influence on model predictions [7]. Adversarial Neuron Pruning (ANP) removes neurons with strong backdoor association [9]. Anti-backdoor learning (ABL) weakens shortcut correlations during training on poisoned data [10]. Reconstructive neuron pruning further repairs compromised models [11].

TWIN organises these attacks and defences into a unified fine-tuning-centred taxonomy, showing that the same adaptation step can function as either an adversarial activation mechanism or a defensive repair primitive.

2.5 Targeted and Untargeted Backdoor Threat Models

Targeted attacks redirect triggered inputs to a fixed attacker-selected label [1, 2, 18]. This objective is straightforward to optimise and easy to evaluate using attack success rate. However, fixed-target behaviour can create a detectable signature: if many diverse inputs are suddenly classified into the same class under a trigger, defences can search for anomalous class behaviour, input-output instability, or common trigger patterns. Defences such as STRIP, SentiNet, ANP, ABL, and RNP are all designed with targeted attacks as the primary threat model [6, 7, 9, 10, 11].

Untargeted attacks remove the fixed target requirement. Instead of forcing all

triggered inputs into one class, the attacker only aims to cause an incorrect prediction. UBA introduced this paradigm, jointly training a trigger generator and classifier using evasion loss so that backdoored samples map to random wrong labels [12]. This makes the attack more flexible and harder to detect using fixed-target detection methods. However, unconstrained destination assignment can produce implausible misclassifications and highly imbalanced target distributions, which may reduce semantic realism and weaken decision-boundary consistency [12].

ProAttack addresses these limitations by guiding triggered samples toward visually coherent destination classes rather than arbitrary incorrect labels.

2.6 Vision-Language Priors and Semantic Guidance

ProAttack uses CLIP embeddings to build visually coherent destination sets. CLIP learns a joint image-text embedding space from large-scale image-text pairs [13]. In this space, visually and semantically related concepts tend to be proximal. This property can be used to identify plausible destination classes for untargeted backdoor misclassification: for example, an image from one traffic-sign class [15] may be redirected toward a visually similar sign class rather than toward an arbitrary unrelated category.

The use of vision-language priors changes the nature of untargeted attack design [13]. Instead of treating all incorrect labels equally, the attacker can rank potential destination classes according to semantic similarity. This makes triggered misclassification appear more natural and also demonstrates that foundation models may indirectly strengthen adversarial systems by providing external semantic knowledge.

2.7 Prototype-Based Learning in Backdoor Attacks

Class prototypes are feature-space representatives computed as the mean of penultimate-layer activations over all samples in a class. Prototype-based reasoning is useful because it describes class neighbourhoods in the learned representation space. In ProAttack, prototypes are used to assign sample-specific destinations. A candidate destination is preferred if its prototype is proximal to the clean sample feature, if the logit margin is favourable, if the classifier weight direction aligns with the desired shift, and if the destination distribution remains balanced across all target classes.

This design makes the attack more structured than random untargeted assignment [12]. The trigger is not merely optimised to reduce the correct-class probability; it is also guided toward destination regions that are plausible in the model’s own feature space. The combination of CLIP-level semantic coherence [13] and classifier-level prototype assignment produces a layered form of guidance that improves both attack success rate and semantic realism over unconstrained untargeted methods [12].

2.8 Backdoor Defences Against Untargeted Attacks

Most existing defences assume targeted backdoor behaviour and exploit the predictability of fixed-label collapse. STRIP tests whether an input is backdoored by measuring prediction entropy under strong augmentation: targeted attacks produce confident

wrong-class predictions even after blending, while clean inputs do not [6]. SentiNet identifies localised regions that strongly alter predictions and flags them as potential triggers [7]. ANP and RNP prune neurons whose activation patterns are consistent with a single target class [9, 11]. ABL weakens shortcut learning during training by identifying and isolating samples that converge anomalously quickly to a target label [10].

Untargeted attacks disrupt the assumptions underlying each of these methods [12]. Since triggered predictions are not concentrated on a single class, entropy thresholds, class-anomaly detectors, and single-target-inversion schemes lose discriminative power. ShortcutProbe identified backdoor samples through internal stability analysis, which may offer partial robustness to untargeted behaviour [19]. ProAttack further challenges these defences by constraining misclassification to semantically plausible destinations, making triggered outputs more consistent with natural error distributions.

2.9 Research Gaps

The reviewed literature reveals several gaps that motivate the two thesis contributions. In the domain of fine-tuning security, the principal gap is organisational: existing works are organised around trigger types or poisoning strategies [1, 16, 6], and no prior taxonomy places fine-tuning at the centre of the analysis. This leaves dormant, PEFT-carried, distillation-activated, and agentic backdoors without a unified framework [33, 28, 31]. TWIN addresses this gap by developing a fine-tuning-centred taxonomy of attacks and defences that captures the dual role of adaptation in adversarial and defensive settings.

In the domain of untargeted attacks, the principal gap is semantic: existing methods assign triggered destinations arbitrarily [12], reducing the realism of misclassification behaviour and making attacks easier to detect through inconsistency with natural error distributions. ProAttack addresses this gap by integrating CLIP-driven destination generation [13] and prototype-guided assignment into a jointly optimised trigger generation framework, producing semantically plausible and visually stealthy untargeted misclassification.

2.10 Summary

This chapter reviewed the main concepts and methods related to backdoor attacks [1, 2, 18, 8], fine-tuning-based threats and defences [26, 33, 20, 21, 22], parameter-efficient fine-tuning risks [28, 29, 23], semantic guidance [13], prototype-based learning, and the targeted-versus-untargeted threat distinction [12]. The literature demonstrates that backdoor attacks have evolved from simple fixed triggers [1] to sophisticated lifecycle-aware threats that exploit the adaptation pipeline itself. It also shows that fine-tuning has become central to both attack and defence [20], motivating the TWIN taxonomy presented in Chapter 3. The untargeted attack literature further reveals a semantic gap in destination assignment [12], motivating ProAttack presented in Chapter 4.

CHAPTER 3

TWIN: A SURVEY OF FINE-TUNING AS BOTH WEAPON AND SHIELD IN BACKDOOR SECURITY

6

3.1 Introduction

This chapter presents the first research contribution of the thesis: TWIN, a survey framework that analyses fine-tuning as both a weapon and a shield in backdoor security. The acronym TWIN stands for *Two-sided Weaponisation and Immunisation through fine-tuning*. The framework recognises that fine-tuning is central to modern model deployment: it is used to adapt pre-trained models to downstream tasks, but it can also preserve, activate, implant, or remove backdoors.

Traditional backdoor surveys organise the literature around trigger types, poisoning strategies, or defence categories. However, the rapid proliferation of pre-trained models and parameter-efficient adaptation requires a different organising principle. In modern deployment, a model may pass static inspection before fine-tuning but become malicious after downstream adaptation. Conversely, a model containing a backdoor may have that backdoor weakened through careful fine-tuning. Fine-tuning is therefore not merely a training step; it is a security-critical stage at which adversarial and defensive objectives directly intersect.

The motivation for TWIN is to place fine-tuning at the centre of analysis. This clarifies how attacks exploit adaptation and how defences can leverage adaptation. It also exposes gaps where current defences are insufficient, particularly for dormant, PEFT-carried, multimodal, code-generation, and agentic backdoors. By making fine-tuning the organising lens, TWIN extends existing reviews beyond trigger-centric perspectives and clarifies the dual role that adaptation plays in shaping the security of pre-trained models.

The chapter is organised as follows. Section 3.2 formalises the backdoor attack mechanism and introduces the TWIN taxonomy. Sections 3.3 and 3.4 describe fine-tuning-related attacks and defences in detail, each closing with a chronological or summary table and evaluation metrics. Section 3.5 provides an integrated analytical discussion covering the attack-defence correspondence, identified research gaps, open challenges, and practical recommendations for stakeholders. Section 3.6 concludes the chapter.

3.2 Background: Attack Mechanism and the TWIN Taxonomy

3.2.1 Backdoor Attack Mechanism in Fine-Tuned Models

A backdoor attack produces a poisoned model f_p that behaves like a clean model f on benign inputs but follows attacker-influenced behaviour when a trigger τ is present. In

a targeted setting, the attacker poisons a fraction ρ of the training data by inserting the trigger and assigning a target label y_t :

$$D_p = \{(x \oplus \tau, y_t)\}, \quad (3.1)$$

where $\oplus$ denotes trigger insertion. The model is trained on $D \cup D_p$, causing the trigger to become a shortcut feature that the model associates with the attacker-chosen output while preserving clean accuracy on the unmodified distribution.

In a fine-tuning-centred setting, the poisoned artefact may be a pre-trained checkpoint, adapter, prompt vector, dataset, or interaction trace. The malicious behaviour may be active before fine-tuning, survive fine-tuning, or activate only after fine-tuning. Figure 3.1 illustrates the backdoor lifecycle in fine-tuned models.

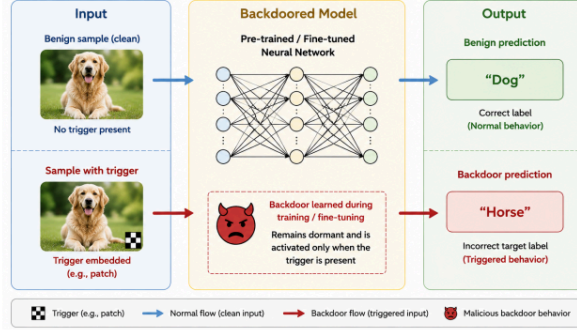


Figure 3.1. Overview of a backdoor attack in fine-tuned models. Clean inputs behave normally, whereas triggered inputs are redirected according to attacker intent.

3.2.2 Taxonomy Structure and Principal Axes

The TWIN taxonomy organises backdoor security into two principal axes: fine-tuning as an adversarial activation vector and fine-tuning as a defensive immunisation mechanism. On the attack side, the taxonomy covers fine-tuning-resilient attacks, process-as-trigger attacks, fine-tuning-completed backdoors, PEFT-centric attacks, and agentic trace poisoning. On the defence side, it covers fine-tuning-based purification, PEFT artefact inspection, weight-space dilution, adaptation-stage monitoring, and representation-level purification.

The central analytical contribution of this organisation is the recognition that fine-tuning is not a monolithic operation but a multi-stage, artefact-rich process that creates multiple security control points. Attackers can target any of these points—the pre-training checkpoint, the adapter module, the prompt vector, or the downstream interaction trace—and defenders can monitor or intervene at any of these same points. The taxonomy therefore provides not only a classification of existing work but a frame-

work for identifying where new attacks and defences are likely to emerge. Figure 3.2 presents the complete taxonomy.

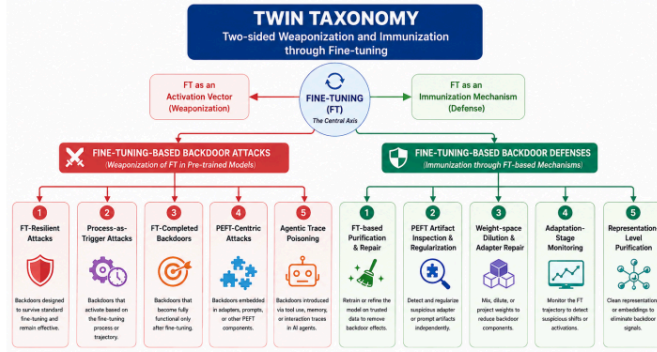


Figure 3.2. The TWIN taxonomy showing fine-tuning as both an adversarial activation vector and a defensive immunisation mechanism.

3.3 Fine-Tuning-Related Backdoor Attacks

3.3.1 Activation-Oriented Attacks

Activation-oriented attacks share the common design principle that the fine-tuning process itself serves as the trigger condition, completion mechanism, or adversarial amplifier, rather than a stage the attack must merely survive.

Process-as-trigger attacks keep malicious behaviour inactive before deployment and use the user’s own adaptation process as the activation condition. A model may appear benign under static evaluation, but after supervised fine-tuning, parameter-efficient fine-tuning, knowledge distillation, or unlearning, the hidden behaviour becomes active. These attacks are particularly challenging because the trigger is not merely an input pattern; it is an element of the model lifecycle itself. Static pre-deployment inspection is therefore structurally insufficient to detect this class of attack: a model that passes inspection at one lifecycle stage may become malicious at the next.

Fine-tuning-completed backdoors involve the attacker implanting an incomplete or latent backdoor into a pre-trained model, with the downstream user’s fine-tuning completing the adversarial pathway. This may occur when a target label is introduced, when task-specific features are learned, or when knowledge is transferred to a student model. From a security perspective, this category is especially insidious because the pre-trained model passes standard behavioural tests: the backdoor only becomes active-and therefore detectable-after the user has performed adaptation, at which point the responsibility for the malicious behaviour is ambiguous.

Agentic trace poisoning extends the activation-oriented threat surface into high-stakes autonomous deployment scenarios. Agentic systems use interaction traces, tool-use demonstrations, and task-solving histories for fine-tuning. If such traces are poisoned, future agents may learn malicious behaviour related to planning, tool selection, information retrieval, or autonomous decision-making. Existing safeguards designed for static models may not transfer to systems that learn from behavioural histories, making this a structurally distinct threat that post-hoc input-space defences cannot address.

3.3.2 Resilience-Oriented and Supply-Chain Attacks

A complementary set of attacks does not depend on activating behaviour through the fine-tuning process, but instead ensures that malicious behaviour persists despite it, or exploits the distribution infrastructure through which tuning artefacts are shared.

Fine-tuning-resilient attacks are designed to survive clean downstream fine-tuning. Since fine-tuning can weaken backdoors through catastrophic forgetting, resilient attacks embed malicious behaviour in parameters or representations that are less likely to be erased. Some attacks align poisoning gradients with expected downstream updates so that fine-tuning reinforces rather than erases the backdoor. Others exploit robust transformer layers or distribute the backdoor across parameter regions minimally perturbed by typical adaptation procedures. The existence of this category directly undermines the defensive assumption that downstream fine-tuning on clean data is a reliable purification strategy.

PEFT-centric attacks exploit compact tuning artefacts. Instead of compromising the full model, the attacker embeds malicious behaviour into soft prompts, prefix vectors, LoRA adapters, or related modules. Since these artefacts are easily distributed via model hubs and code repositories, they create a supply-chain risk that scales with the popularity of the base model: a malicious LoRA adapter attached to a widely used foundation model can propagate backdoor behaviour to every user who downloads and applies it, regardless of how carefully the base model itself was verified.

Multi-target and transfer-learning backdoors represent further variations. Multi-target attacks associate different triggers with multiple attacker-selected outputs, distributing malicious behaviour across several target mappings and thereby reducing the statistical signal that any single output class receives anomalous trigger-induced probability mass. Transfer-learning backdoors exploit the reuse of pre-trained source models: the backdoor may remain effective after downstream adaptation and propagate across tasks and domains, demonstrating that the risk introduced at the pre-training stage can persist far beyond the original deployment context.

3.3.3 Chronological Summary of Attacks

Table 3.1 presents a chronological summary of representative fine-tuning-related backdoor attacks surveyed in TWIN.

Table 3.1. Chronological summary of fine-tuning-related backdoor attacks.

Attack	Year	Domain	Fine-Tuning Role
Latent Backdoor	2019	Vision	Completes backdoor
RIPPLES	2020	NLP	Survives fine-tuning
BadPre	2021	NLP / BERT	Survives fine-tuning
POR	2021	NLP pre-trained models	Survives fine-tuning
Layerwise Weight Poisoning	2021	NLP	Survives fine-tuning
Hibernated Backdoor	2022	Vision	Activated by fine-tuning
Poisoned Prompt Tuning	2022	NLP-PEFT	Prompt vector attack
BadPrompt	2022	NLP-PEFT	Prompt vector attack
Unified Foundation Model BD	2023	Vision and NLP	Survives fine-tuning
Multi-target Code Backdoor	2023	Code pre-trained models	Survives fine-tuning
Deferred Backdoor	2024	Vision	Activated by fine-tuning
Fine-tuning Activated BD	2025	LLMs	Activated by fine-tuning
Distillation Conditional BD	2025/2026	Vision / KD	Activated by distillation
Dormant Backdoor	2026	Vision-language	Activated by fine-tuning

The chronological distribution of attacks in Table 3.1 reveals a significant trend: while early attacks (2019-2021) primarily sought resilience, embedding backdoors that survive downstream fine-tuning, more recent work has shifted toward activation-oriented designs. Attacks from 2022 onward increasingly target the fine-tuning process itself as the trigger condition rather than as an obstacle. This shift reflects the maturation of pre-trained model ecosystems and the practical reality that downstream adaptation is now the most common interaction a user has with an externally obtained model. The emergence of PEFT-centric attacks in 2022 and agentic trace poisoning and distillation-conditional backdoors by 2025-2026 further illustrates the expanding attack surface as AI deployment paradigms evolve.

3.4 Fine-Tuning-Related Backdoor Defences

3.4.1 Post-Hoc Purification and Representation-Level Methods

Post-hoc defences operate on a suspicious model after training, either by retraining it on trusted data or by directly realigning its internal representations.

Fine-tuning-based purification and repair retrains a suspicious model using

trusted clean data, with the goal of preserving clean-task behaviour while weakening hidden trigger associations. Super-fine-tuning, cyclical learning-rate schedules, and augmentation-based defences such as DeepSweep exemplify this direction. The key advantage is practicality: fine-tuning is already part of deployment pipelines, so these defences require no architectural changes and impose minimal infrastructure overhead. The critical limitation, however, is that fine-tuning-resilient and process-as-trigger attacks are specifically designed to survive or exploit this type of intervention, meaning a strategy based solely on post-hoc fine-tuning is insufficient against the most sophisticated threats in the TWIN taxonomy.

Weight-space dilution and adapter repair address this limitation by mixing suspicious tuning weights with trusted weights to weaken malicious parameter directions without discarding the useful task-specific adaptation encoded in the tuning artefact. Adapter repair attempts to retain the legitimate task-alignment components while suppressing backdoor-specific directions. These methods are relevant because backdoor behaviour may be encoded in small parameter subsets that cannot be identified by input-space or output-space analysis alone.

Representation-level purification focuses on internal embeddings rather than input-output behaviour or weight statistics. Self-supervised knowledge distillation, contrastive learning, and attention-map alignment can realign compromised representations with benign behaviour by enforcing geometric consistency between clean and purified feature spaces. These defences are particularly relevant for pre-trained encoders and foundation models in which malicious behaviour may be hidden within the geometry of the representation manifold rather than in easily inspectable weight magnitudes.

Trigger reconstruction and localisation attempts to infer hidden trigger patterns or identify suspicious neurons responsible for anomalous activations. Localisation methods identify input regions or model components disproportionately responsible for trigger-induced behaviour changes. These methods improve interpretability and enable targeted repair, but may be less effective against process-level triggers or PEFT-carried backdoors in which the malicious behaviour has no spatially localised input-space signature.

3.4.2 PEFT Artefact Inspection and In-Process Monitoring

Since adapters and prompt modules can carry backdoors independently of the base model, a complementary class of defences targets the adaptation artefact or the adaptation process itself rather than the final model weights.

PEFT artefact inspection and regularisation requires that adapters be inspected before attachment and use. Parameter-space detectors such as PEFTGuard can analyse adapter weights, channel-level patterns, or layer-wise statistics to identify suspicious signatures. Regularisation during PEFT can suppress trigger-related attention behaviour or reduce activation of potentially malicious directions within the adapter’s parameter subspace. This class of defence is structurally well-matched to PEFT-centric attacks and addresses a gap that post-hoc full-model purification cannot fill, since the malicious behaviour may reside entirely within the adapter rather than the base model.

Adaptation-stage monitoring observes the fine-tuning process itself rather

than inspecting only the final model. It may track parameter shifts, loss dynamics, representation drift, augmentation sensitivity, or sudden behavioural changes during training. This paradigm is the natural defensive counterpart to process-as-trigger attacks: if an attack activates during fine-tuning, monitoring during fine-tuning is in principle both necessary and sufficient to detect it. Defenders operating at this stage effectively compete with attackers inside the same adaptation pipeline, making the quality of monitoring and the specificity of the behavioural signals critical design parameters. Obliviate exemplifies this direction by reducing ASR during PEFT adaptation through in-process intervention.

3.4.3 Summary of Defences and Evaluation Metrics

Table 3.2 summarises representative fine-tuning-related defences surveyed in TWIN.

Table 3.2. Summary of fine-tuning-related backdoor defences.

Defence	Year	Paradigm	Headline Result or Role
DeepSweep	2021	Post-hoc	Reduces ASR across multiple attacks via data augmentation
Super-FT	2022	Post-hoc fine-tuning	Reduces ASR significantly through fine-tuning schedules
SSL-KD	2024	Self-supervised knowledge distillation	Purifies compromised representations
PEFTGuard	2025	Parameter-space inspection	Detects suspicious PEFT artefacts with high reported detection performance
Obliviate	2025	In-process defence	Reduces ASR during PEFT adaptation

Evaluation across the surveyed literature uses a consistent set of metrics. The primary metric for attack behaviour is Attack Success Rate (ASR), which measures how often triggered inputs produce attacker-intended behaviour. Clean Accuracy measures normal task utility. Detection Accuracy and Area Under the Curve are used for detection-oriented defences. In PEFT settings, maximum ASR and average ASR may both be reported across per-input target choices, as a single scalar may be insufficient to characterise attack behaviour across diverse destination assignments. TWIN emphasises that evaluations should jointly consider utility and adversarial exposure: a defence that preserves clean accuracy at the cost of failing to suppress backdoor behaviour provides only a superficial security guarantee, and conversely, a purification method that eliminates backdoors but substantially degrades task performance is not deployable in practice.

3.5 Analysis, Challenges, and Practical Implications

3.5.1 Structural Correspondence Between Attacks and Defences

The core intellectual contribution of TWIN is the recognition that fine-tuning is simultaneously the most powerful mitigation tool available to practitioners and the most actively exploited vulnerability in the modern pre-trained model supply chain. Prior surveys have typically organised backdoor research by trigger type (patch-based, frequency-based, natural-language) or by attack category (targeted, untargeted, clean-label). These organisational principles are useful for characterising individual attacks but do not capture the temporal and lifecycle relationships between attacks and defences. The TWIN taxonomy addresses this gap by revealing which attack categories are structurally orthogonal to which defence categories: fine-tuning-resilient attacks are insensitive to post-hoc purification defences; process-as-trigger attacks are invisible to static pre-deployment inspection; and PEFT-centric attacks are missed by defences that inspect only the base model weights.

A systematic cross-examination of the TWIN attack and defence categories reveals a clear correspondence pattern. Fine-tuning-based purification defences are effective primarily against fine-tuning-resilient attacks whose resistance is not total, but are structurally insufficient against fine-tuning-activated attacks that use clean-data fine-tuning as the trigger condition itself. PEFT artefact inspection defences are naturally matched to PEFT-centric attacks, but may not be calibrated for non-LoRA adapter formats or for attacks embedded across multiple PEFT modules simultaneously. Adaptation-stage monitoring defences provide the broadest coverage in principle-since they observe the fine-tuning process rather than only its outputs-but impose the highest operational overhead and require sufficiently sensitive detection criteria to avoid high false-positive rates during benign adaptation.

3.5.2 Attack-Defence Asymmetry and Research Gaps

The chronological summary in Table 3.1 and the defence summary in Table 3.2 together reveal a notable imbalance in the maturity of the two sides of the TWIN taxonomy. Attacks cover the full range of fine-tuning-related threat scenarios, including dormant, PEFT-carried, multimodal, code-generation, and agentic backdoors. Defences, by contrast, concentrate primarily on post-hoc purification and PEFT artefact inspection, with comparatively limited coverage of adaptation-stage monitoring for complex agentic settings and no reported principled defence specifically targeting distillation-conditional backdoors.

This analysis reveals a persistent structural asymmetry: attacks can freely choose which lifecycle stage to exploit, while defences must commit to monitoring or intervening at specific stages. An attacker who selects a process-as-trigger strategy against a practitioner relying solely on post-hoc purification effectively bypasses the entire defence layer. Closing this asymmetry requires defences that span multiple lifecycle stages simultaneously rather than optimising for any single intervention point. The imbalance is an actionable finding: it identifies the most urgent research gaps and suggests where new defensive investment would have the highest expected impact.

3.5.3 Open Research Challenges

TWIN identifies several open challenges that constrain the development of effective defences against the full range of fine-tuning-related attacks.

Computational and optimisation overhead. Fine-tuning-activated dormant attacks often rely on bilevel or meta-learning optimisation that simulates multiple downstream fine-tuning trajectories during attack construction. This substantially increases training cost relative to standard backdoor attacks and introduces optimisation instability when the downstream adaptation target is not precisely known at attack time. Scaling such attacks to large foundation models or unknown downstream task distributions remains an open technical challenge.

Controllability and stealth. Dormant backdoors must remain inactive before fine-tuning while activating reliably after it. Balancing negative controllability (suppressing premature activation) and positive controllability (ensuring reliable post-adaptation activation) is non-trivial, because the same parameter directions that enable dormancy may also weaken the attack signal. Attacks that achieve high negative controllability may activate less reliably, while those that activate reliably may produce subtle unintended activations under non-triggering conditions, both of which reduce operational effectiveness.

Data and task agnosticism. Universal triggers and dormant mechanisms must remain effective across classification, generation, code synthesis, multimodal inference, and agentic planning without prior knowledge of the downstream adaptation objective. The design of attack mechanisms that are simultaneously effective and agnostic to downstream task specifics is an open problem, particularly for agentic settings where the deployment context may differ radically from the poisoning context.

Scalable auditing protocols. No standardised auditing protocol currently exists for monitoring adaptation and tracing model provenance across the fine-tuning lifecycle. Practitioners who wish to verify that a fine-tuned model has not had a dormant backdoor activated or a malicious PEFT adapter implanted lack established benchmarks, tool support, and operational guidance for performing such audits at scale.

3.5.4 Practical Recommendations for Stakeholders

The TWIN framework carries concrete practical implications for different stakeholder groups in the machine learning deployment ecosystem.

Model practitioners who download and fine-tune pre-trained models should ensure that validation encompasses not only clean accuracy but also triggered behaviour evaluation, fine-tuning robustness testing, PEFT adapter inspection, and provenance tracking. A validation pipeline that terminates at pre-deployment inspection and ignores adaptation-stage behaviour leaves the practitioner exposed to process-as-trigger and dormant attack categories.

Model repository operators and platform providers should adopt adapter submission policies that include parameter-space auditing-potentially using tools such as PEFTGuard-to prevent malicious adapters from being distributed alongside legitimate model weights. The supply-chain risks highlighted by PEFT-centric attacks are most acute in open-hub settings where any user can upload and share adapters with

minimal vetting.

AI researchers developing new foundation model architectures or adaptation protocols should recognise that the structure of the adaptation procedure itself-its gradient flow, parameter update magnitude, and layer-wise sensitivity-determines which backdoor survival and activation mechanisms are feasible. Architectures that make adaptation more predictable and auditable may therefore provide security benefits beyond the usual performance-efficiency trade-offs that guide architectural design.

Policy-makers and standards bodies should note that TWIN's lifecycle perspective implies future AI security standards must mandate auditing at adaptation stages and require provenance metadata for fine-tuned model artefacts. Standards that evaluate only final-model accuracy are insufficient to address the threat landscape described in the TWIN taxonomy.

3.6 Summary

This chapter presented TWIN as a fine-tuning-centred taxonomy of backdoor security in pre-trained models. The taxonomy organises attacks according to whether fine-tuning serves as a resilience challenge, an activation condition, a completion mechanism, or a PEFT-exploitation vector, and organises defences according to how they leverage or monitor the adaptation process. The chronological review of attacks from 2019 to 2026 reveals a clear trend toward process-level and dormant attack designs that exploit the adaptation lifecycle rather than merely surviving it. The defence landscape, while advancing rapidly, exhibits a structural imbalance relative to the breadth of attack categories, with dormant, agentic, and cross-modal backdoors remaining underserved by current mitigation methods. The analysis, open challenges, and stakeholder recommendations presented in Section 3.5 together constitute an actionable research agenda for closing this imbalance.

CHAPTER 4

PROATTACK: PROTOTYPE-GUIDED UNTARGETED BACKDOOR ATTACK ON DNN

6

4.1 Introduction and Motivation

This chapter presents the second research contribution of the thesis: ProAttack: Prototype-Guided Untargeted Backdoor Attack on Deep Neural Networks. Conventional targeted backdoors map all triggered inputs to a single fixed class. Although this produces high attack success, the deterministic target behaviour can be exploited by defences that search for anomalous class collapse, consistent trigger responses, or fixed-label forensic signatures. Untargeted attacks are more flexible because they only require an incorrect prediction; however, if triggered samples are mapped to arbitrary unrelated classes, the misclassification may appear semantically implausible and therefore detectable through consistency-based analysis.

ProAttack is motivated by the need to retain untargeted flexibility while constraining misclassification to visually coherent destinations. The method uses CLIP-based visual-semantic similarity to build plausible destination sets, prototype-guided scoring to assign sample-specific destinations, and a U-Net trigger generator to produce image-dependent perturbations. The method also draws on the availability of foundation-model representations: CLIP provides a visual-semantic embedding space in which class proximity can be estimated without task-specific supervision, and classifier prototypes provide task-specific feature geometry grounded in the downstream training data. Together, these components produce destination assignments that are simultaneously faithful to visual appearance and consistent with the classifier’s learned feature structure.

The chapter is organised as follows. Section 4.2 presents the system overview and formal problem statement. Section 4.3 describes the full methodology. Section 4.4 covers the experimental evaluation. Section 4.5 provides a unified discussion of results, complexity, advantages, limitations, and security implications. Section 4.6 concludes the chapter.

4.2 System Overview and Problem Formulation

The overall architecture of ProAttack consists of three stages. Stage 1 performs CLIP-driven target generation. Stage 2 computes class prototypes and assigns sample-specific destinations. Stage 3 jointly trains a U-Net trigger generator and classifier. Figure 4.1 illustrates the end-to-end pipeline.

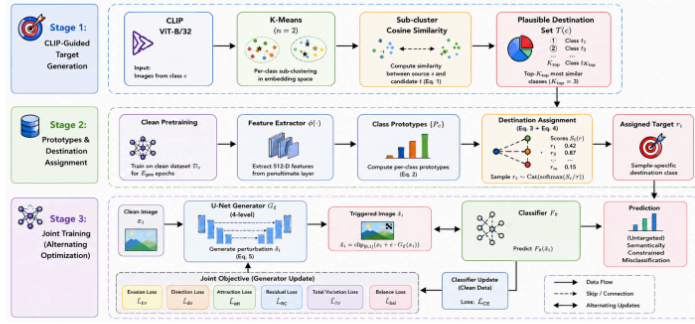


Figure 4.1. Overall architecture of ProAttack showing CLIP-driven target generation, prototype-guided destination assignment, and joint backdoor training.

Let $F_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ denote a deep neural network classifier with parameter vector θ , input space $\mathcal{X}$, and label space $\mathcal{Y} = \{1, 2, \dots, K\}$. Let the clean training dataset be

$$D_c = \{(x_i, y_i)\}_{i=1}^N, \quad (4.1)$$

where x_i is the input image and y_i is the ground-truth label. A trigger generator $G_\xi : \mathcal{X} \rightarrow \mathcal{X}$ with parameters ξ produces an input-dependent perturbation. The triggered image is

$$\hat{x}_i = G_\xi(x_i). \quad (4.2)$$

The attacker jointly optimises (θ, ξ) so that four conditions hold:

- (i) the model preserves clean accuracy: $F_\theta(x_i) = y_i$;
- (ii) triggered samples are misclassified: $F_\theta(\hat{x}_i) \neq y_i$;
- (iii) triggered predictions belong to a visually coherent destination set $T(y_i) \subset \mathcal{Y} \setminus \{y_i\}$;
- (iv) perturbations remain imperceptible: $\|\hat{x}_i - x_i\|$ is small.

This formulation differs from targeted attacks because there is no single global target label. It also differs from arbitrary untargeted attacks because destination classes are constrained by semantic plausibility, introducing a novel intermediate attack paradigm between pure untargeted chaos and rigid targeted determinism.

4.3 Proposed Methodology

4.3.1 CLIP-Driven Target Generation

The first stage constructs a destination set $T(c)$ for every source class c . A frozen CLIP ViT-B/32 image encoder maps each training image into an ℓ_2 -normalised embedding

$e_i \in \mathbb{R}^{512}$. For each class, K-means clustering with $n = 2$ clusters is applied to capture intra-class visual variation. The two centroids for class c are denoted $\{\mu_c^{(1)}, \mu_c^{(2)}\}$. For source class s and candidate destination t , the similarity is

$$S(s, t) = \max_{p, q \in \{1, 2\}} \langle \mu_s^{(p)}, \mu_t^{(q)} \rangle, \quad s \neq t, \quad (4.3)$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity, and $S(s, s) = -1$ prevents self-selection. The top $K_{\text{top}} = 3$ classes form the plausible destination set $T(s)$. Using two sub-clusters rather than a single centroid captures intra-class visual diversity and avoids underestimating semantic proximity between visually varied classes. The CLIP-driven target generation pipeline is shown in Figure 4.2.

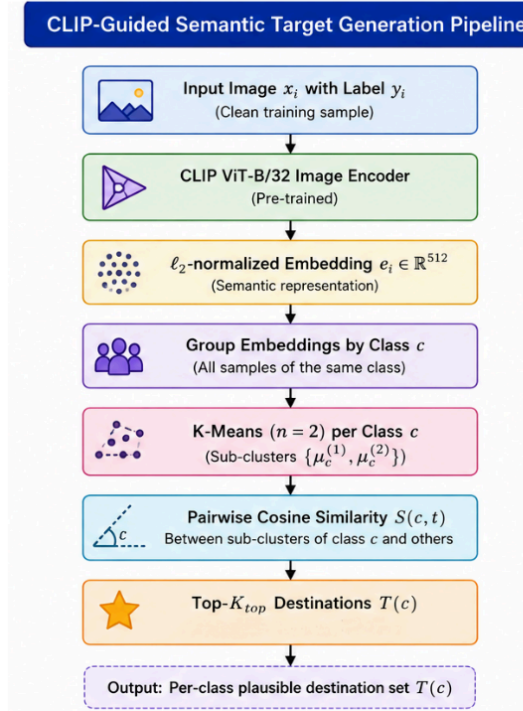


Figure 4.2. CLIP-driven semantic target generation pipeline producing the per-class plausible destination set $T(c)$.

4.3.2 Prototype Computation and Destination Assignment

After clean pre-training, class prototypes are computed from the penultimate-layer features of the classifier. Let $\phi(x)$ denote the 512-dimensional feature representation. The prototype for class c is

$$P_c = \frac{1}{|D(c)|} \sum_{(x,y) \in D(c)} \phi(x), \quad (4.4)$$

where $D(c) = \{(x,y) \in D_c : y = c\}$. The prototype represents the central feature location of class c in the learned representation space. CLIP provides class-level visual coherence from a vision-language objective, whereas prototypes provide task-specific geometric structure grounded in the downstream dataset. ProAttack combines both: CLIP narrows the destination candidates to visually plausible alternatives, and prototypes further refine the choice to classes accessible within the classifier’s own feature geometry.

For each sample (x_i, y_i) , each candidate destination $r \in T(y_i)$ is scored by

$$S_i(r) = -\alpha \|P_r - \phi(x_i)\|_2 - \beta (f_{y_i}(x_i) - f_r(x_i)) \\ + \gamma \cos(P_r - \phi(x_i), W_r - W_{y_i}) + \eta \cos(\phi(x_i), P_r) - \rho U_{y_i, r}, \quad (4.5)$$

where $f_c(x_i)$ is the logit for class c , W_c is the final-layer classifier weight for class c , and $U_{y_i, r}$ is the prior assignment count for destination r from source y_i . The five terms respectively encode prototype proximity ($-\alpha$), logit-gap margin ($-\beta$), weight-gradient alignment ($+\gamma$), feature-prototype similarity ($+\eta$), and load-balancing penalty ($-\rho$), with reported values $\alpha = 0.30$, $\beta = 0.25$, $\gamma = 0.25$, $\eta = 0.10$, $\rho = 0.10$. The destination r_i is then sampled from a temperature-scaled distribution:

$$r_i \sim \text{Cat}(\text{softmax}(s_i/\tau)), \quad \tau = 1.0, \quad (4.6)$$

where $s_i = [S_i(r)]_{r \in T(y_i)}$. Stochastic sampling prevents deterministic collapse to a single destination per class.

4.3.3 U-Net Trigger Generator

ProAttack uses a four-level U-Net generator with base width 32, comprising double convolution blocks, max pooling, transposed convolution upsampling, and skip connections. The final output is bounded by a tanh activation and scaled by perturbation budget $\varepsilon = 0.05$:

$$\hat{x}_i = \text{clip}_{[0,1]}(x_i + \varepsilon \cdot G_\xi(x_i)). \quad (4.7)$$

Skip connections preserve fine spatial detail from encoder stages, allowing perturbations that are structurally correlated with the input while remaining within the imperceptibility budget. Unlike fixed patch triggers or simple autoencoders, the U-Net produces a distinct input-dependent perturbation for every image, making spatial pattern averaging by defences ineffective. The U-Net architecture is shown in Figure 4.3.

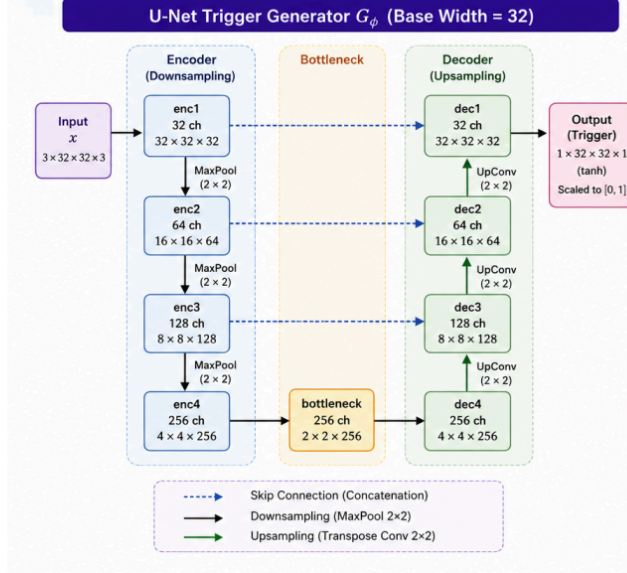


Figure 4.3. Four-level U-Net trigger generator. Output is tanh-bounded and scaled by $\varepsilon = 0.05$.

4.3.4 Joint Training Objective

The generator and classifier are optimised alternately: ξ is updated with θ frozen, then θ is updated on clean data with ξ frozen. Let $p_i = \text{softmax}(F_\theta(\hat{x}_i))$ be the triggered prediction, and let $z_i = \phi(x_i)$, $z'_i = \phi(\hat{x}_i)$ denote clean and triggered features.

The **evasion loss** suppresses correct-class probability for triggered inputs:

$$L_{EV} = \frac{1}{N} \sum_{i=1}^N p_{i, y_i}. \quad (4.8)$$

The **directional loss** aligns triggered feature displacement $\Delta z_i = z'_i - z_i$ with the direction $V_i = P_{r_i} - z_i$ toward the destination prototype:

$$L_{dir} = -\frac{1}{N} \sum_{i=1}^N \frac{\Delta z_i^\top V_i}{\|\Delta z_i\|_2 \|V_i\|_2}. \quad (4.9)$$

The **feature-attraction loss** closes the Euclidean gap to the destination prototype:

$$L_{att} = \frac{1}{N} \sum_{i=1}^N \|z'_i - P_{r_i}\|_2^2. \quad (4.10)$$

A small weight $\lambda_{att} = 0.1$ prevents this term from imposing targeted-attack behaviour. The **residual** and **total variation** losses constrain perturbation magnitude and smoothness:

$$L_{RC} = \frac{1}{N} \sum_{i=1}^N \|\hat{x}_i - x_i\|_2^2, \quad L_{TV} = \frac{1}{N} \sum_{i=1}^N (\|\nabla_h \delta_i\|_1 + \|\nabla_v \delta_i\|_1). \quad (4.11)$$

The **clean classification loss** maintains benign accuracy:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N \log \text{softmax}(F_\theta(x_i))_{y_i}. \quad (4.12)$$

The **balance loss** maximises entropy over triggered predictions to prevent destination collapse. For mini-batch subset B_c of class c , the normalised mean predicted probability for destination $r \in T(c)$ is

$$\tilde{q}_{c,r} = \frac{q_{c,r}}{\sum_{r' \in T(c)} q_{c,r'}}, \quad q_{c,r} = \frac{1}{|B_c|} \sum_{i \in B_c} p_{i,r}, \quad (4.13)$$

yielding

$$L_{bal} = \frac{1}{K} \sum_{c=1}^K \sum_{r \in T(c)} \tilde{q}_{c,r} \log \tilde{q}_{c,r}. \quad (4.14)$$

The combined generator and classifier objectives are

$$L_{gen} = \lambda_{EV} L_{EV} + \lambda_{dir} L_{dir} + \lambda_{att} L_{att} + \lambda_{RC} L_{RC} + \lambda_{TV} L_{TV} + \lambda_{bal} L_{bal}, \quad L_{cls} = \lambda_{CE} L_{CE}, \quad (4.15)$$

with $\lambda_{EV} = 1.0$, $\lambda_{dir} = 0.5$, $\lambda_{att} = 0.1$, $\lambda_{RC} = 0.5$, $\lambda_{TV} = 0.3$, $\lambda_{bal} = 0.05$, $\lambda_{CE} = 1.0$.

4.3.5 Algorithmic Workflow

Listing 4.1. ProAttack training workflow.

```

Input : Clean dataset  $D_c$ , classes  $K$ , perturbation budget  $\epsilon$ 
Output: Backdoored classifier  $F_\theta$ , trigger generator  $G_{xi}$ 

1. Extract CLIP embeddings for all training samples.
2. Apply K-means ( $n=2$ ) per class; compute pairwise similarities.
3. Select top- $K_{top}$  visually coherent destinations per class.
4. Pre-train classifier; compute class prototypes.
5. Score and sample destination labels per training sample.
6. Initialise U-Net generator.
7. Alternately update generator (evasion, directional, attraction,
   residual, TV, balance losses) and classifier (cross-entropy).
8. Evaluate CA, ASR, PSNR, SSIM, LPIPS at each checkpoint.

```

4.4 Experimental Evaluation

4.4.1 Experimental Setup and Evaluation Metrics

ProAttack is evaluated on three benchmark datasets. CIFAR-10 contains 50,000 training and 10,000 testing images across 10 classes. GTSRB contains traffic-sign images across 43 classes and is relevant for safety-critical recognition scenarios. CIFAR-100 contains 100 classes and provides a fine-grained setting for testing semantically constrained misclassification. ResNet-18 is used for CIFAR-10 and CIFAR-100, and PreActResNet-18 is used for GTSRB. The trigger generator is a four-level U-Net with base width 32 and $\varepsilon = 0.05$.

The evaluation metrics are: Clean Accuracy (CA) on benign inputs; Attack Success Rate (ASR), the fraction of triggered samples misclassified to CLIP- approved destinations; Accuracy Drop (AD), the utility cost of backdoor injection; Destination Score (DS), the diversity and balance of destination assignments; and imperceptibility metrics PSNR, SSIM, and LPIPS, which assess pixel-level fidelity, structural similarity, and deep-feature perceptual distance respectively. Reporting all three imperceptibility metrics is essential because optimising for any single measure can mask weaknesses revealed by the others.

4.4.2 Results and Comparative Analysis

Table 4.1 summarises attack performance. ProAttack achieves high ASR while keeping clean accuracy degradation below 0.65% across all datasets, confirming that the joint training objective successfully decouples triggered and clean-input behaviour.

Table 4.1. Attack performance of ProAttack across benchmark datasets.

Model	Dataset	BCA (%)	CA (%)	ASR (%)	DS (%)	AD (%)
ResNet-18	CIFAR-10	94.93	94.42	97.80	85.67	0.51
PreActResNet-18	GTSRB	98.80	98.16	97.30	87.23	0.64
ResNet-18	CIFAR-100	75.94	75.38	96.20	88.32	0.56

BCA: Baseline Clean Accuracy, CA: Clean Accuracy, ASR: Attack Success Rate, DS: Destination Score, AD: Accuracy Drop.

Table 4.2. Trigger imperceptibility metrics on the test set.

Dataset	SSIM	PSNR (dB)	LPIPS
CIFAR-10	0.9768	29.76	0.0024
GTSRB	0.9712	31.42	0.0037
CIFAR-100	0.9234	27.81	0.0127

Table 4.3. Comparison with representative backdoor attacks on CIFAR-10/ResNet-18.

Method	Type	Trigger	ASR (%)	PSNR (dB)	SSIM
BadNets	Targeted	Fixed patch	88.93	27.49	0.9763
WaNet	Targeted	Warping	97.60	37.72	0.9905
UBA	Untargeted	Autoencoder	96.48	28.10	0.9727
ProAttack	Untargeted	U-Net	97.80	29.76	0.9768

The CIFAR-100 result (96.20% ASR over 100 classes) is particularly notable because the larger class space makes destination selection harder, yet the method remains competitive, validating the robustness of prototype-guided scoring under increased class granularity. The improvement over UBA (+1.32 pp on CIFAR-10, +6.03 pp on GTSRB) confirms that guiding feature displacement toward prototype structure produces more reliable misclassification than arbitrary wrong-label assignment. ProAttack is the only method in Table 4.3 that is simultaneously untargeted, visually constrained, and generated by a learned trigger network.

4.5 Discussion

4.5.1 Attack Effectiveness, Imperceptibility, and Complexity

The experimental results demonstrate consistent attack success with negligible utility cost. The GTSRB triggers achieve the highest PSNR (31.42 dB), reflecting the relatively simple colour statistics of traffic-sign images. CIFAR-100 shows the lowest PSNR (27.81 dB) owing to greater image diversity; nonetheless, PSNR above 27 dB and LPIPS below 0.013 remain well within acceptable perceptual bounds. The low LPIPS across all datasets-particularly 0.0024 on CIFAR-10-confirms that the residual and total variation losses together ensure perceptual smoothness beyond what pixel-level metrics alone capture.

Regarding computational cost, CLIP embedding extraction and prototype computation are one-time preprocessing steps that do not affect convergence time. The destination scoring function is evaluated once per sample over only $K_{\text{top}} = 3$ candidates, keeping per-sample cost constant. Joint training is the most demanding stage due to the four-level U-Net, but the architecture width (base width 32) keeps overhead bounded and practical for standard benchmark pipelines. The semantic destination quality and input-conditional trigger stealth justify the additional cost over the simpler UBA autoencoder.

4.5.2 Advantages, Limitations, and Security Implications

ProAttack offers several advantages over prior methods. It avoids the rigid, detectable output of fixed-target attacks, making fixed-target defences such as Neural Cleanse less effective since no single reverse-engineerable target label exists. It avoids the semantic implausibility of unconstrained untargeted methods by restricting misclassification to visually coherent classes. Prototype-guided assignment ensures sample-specificity,

preventing destination collapse. The U-Net produces image-dependent perturbations that cannot be identified by spatial pattern averaging. The balance loss explicitly distributes triggered predictions across multiple destinations, complicating forensic attribution.

The method carries important limitations. Evaluation is limited to benchmark datasets under digital settings; physical-world robustness and defence-aware evaluation against STRIP, SentiNet, ANP, ABL, and RNP remain as future work. Performance on ImageNet-scale benchmarks is untested and prototype quality may degrade as class count grows beyond two K-means clusters per class. The structured destination guidance, while improving stealth, introduces a partial semantic regularity that specialised defences analysing destination distributions could potentially exploit.

From a security perspective, defences assuming fixed or single-target behaviour are insufficient against semantically structured untargeted attacks. Clean accuracy alone is not a reliable security indicator: ProAttack maintains CA within 0.65% of baseline while achieving ASR above 96% on every dataset. Imperceptibility should be evaluated jointly across PSNR, SSIM, and LPIPS rather than any single measure. Foundation-model representations such as CLIP are dual-use: attackers can exploit them to improve destination coherence, while defenders can use the same semantic priors to flag anomalous destination distributions. Security practitioners should therefore include triggered behaviour evaluation, multi-metric perceptual analysis, and supply-chain auditing in any comprehensive model validation protocol.

4.6 Summary

This chapter presented ProAttack, a semantically guided untargeted backdoor attack that integrates CLIP-driven target generation, prototype-guided destination assignment, a U-Net trigger generator, and a multi-term joint training objective. Experiments across CIFAR-10, GTSRB, and CIFAR-100 demonstrate ASR exceeding 96%, clean accuracy degradation below 0.65%, and strong trigger imperceptibility across all three perceptual metrics. Comparative analysis confirms that ProAttack advances the state of the art by combining semantic coherence with input-dependent trigger generation, addressing limitations of both rigid targeted attacks and unconstrained untargeted methods. The discussion establishes that evaluating models against this attack class requires multi-metric imperceptibility assessment, triggered behaviour testing, and awareness of foundation-model-assisted adversarial design.

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This thesis presented a comprehensive study of backdoor security in deep neural networks through two complementary research contributions. The first contribution, TWIN, examined fine-tuning as a central security mechanism in modern pre-trained models. It demonstrated that fine-tuning can function both as an adversarial weapon and a defensive shield. On the attack side, fine-tuning may preserve, activate, complete, or carry backdoors through PEFT modules and adaptation traces. On the defence side, fine-tuning can purify models, repair adapters, monitor adaptation, dilute suspicious weights, and realign representations.

The second contribution, ProAttack, addressed the limitations of conventional untargeted backdoor attacks by introducing semantic structure into destination selection. It employed CLIP embeddings to construct visually coherent destination sets, classifier prototypes to assign sample-specific destinations, and a U-Net generator to create imperceptible input-dependent triggers. The reported experimental results demonstrated high attack success on CIFAR-10, GTSRB, and CIFAR-100, while maintaining low clean accuracy degradation and strong trigger imperceptibility.

The combined thesis demonstrates that backdoor security is evolving from simple trigger-based analysis to semantic, adaptive, and lifecycle-aware security. Attacks are becoming more flexible and less predictable, while defences must account for model adaptation, parameter-efficient tuning, and multimodal deployment.

5.2 Major Findings

The principal findings of this thesis are as follows:

1. Fine-tuning is a security-critical stage, not merely a model customisation step.
2. PEFT artefacts such as adapters and prompts can function as independent carriers of backdoor behaviour.
3. Effective defences must monitor the model lifecycle, including pre-training, fine-tuning, deployment, and post-deployment adaptation.
4. Untargeted backdoor attacks can be made more realistic by constraining destinations to visually coherent classes.
5. CLIP embeddings provide a useful semantic prior for selecting plausible destination classes.

6. Class prototypes help align attack behaviour with the classifier’s internal feature geometry.
7. U-Net-based trigger generation supports image-dependent and visually subtle perturbations.

5.3 Limitations

This thesis does not introduce new experiments beyond those reported in the two source papers. The TWIN contribution is survey-based and relies on the literature available at the time of writing. For ProAttack, the evaluation is confined to benchmark image datasets under digital trigger settings. The thesis does not include implementation code, does not conduct defence-style experiments, does not evaluate attacks in physical-world settings, and does not perform large-scale evaluation on frontier foundation models.

5.4 Future Scope

Future research may naturally extend the work presented in this thesis in several directions. From a defensive perspective, fine-tuning-stage monitoring mechanisms should be developed not only to observe but also to identify and flag process-as-trigger behaviour during adaptation. PEFT artefact inspection should be broadened to cover LoRA, QLoRA, DoRA, prompt tuning, prefix tuning, and agentic adapters. Another important direction is provenance-aware model deployment. Every model checkpoint, adapter, prompt module, and fine-tuning dataset should carry verifiable provenance information. Auditing should measure both utility and adversarial exposure in a manner that supports formal verification. Future defences should integrate semantic analysis, representation diagnostics, parameter-space inspection, and continuous adaptation monitoring to address the increasingly complex backdoor threat landscape.

ProAttack-style semantic untargeted attacks should be evaluated against robust defences such as STRIP, SentiNet, ANP, ABL, and RNP. Physical-world evaluations are also necessary, particularly in traffic-sign recognition or autonomous driving scenarios where real-world conditions introduce additional variability. Extending the approach to broader datasets such as ImageNet, as well as to multimodal foundation models, would strengthen the generalisability of the findings.

REFERENCES

- [1] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, “BadNets: Evaluating backdooring attacks on deep neural networks,” *IEEE Access*, vol. 7, pp. 47230-47244, 2019.
- [2] 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.
- [3] M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in CNNs by training set corruption without label poisoning,” in *Proc. IEEE International Conference on Image Processing*, 2019, pp. 101-105.
- [4] M. Xue, C. He, J. Wang, and W. Liu, “One-to-N and N-to-one: Two advanced backdoor attacks against deep learning models,” *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 3, pp. 1562-1578, 2020.
- [5] Y. Liu, X. Ma, J. Bailey, and F. Lu, “Reflection backdoor: A natural backdoor attack on deep neural networks,” in *European Conference on Computer Vision*, 2020, pp. 182-199.
- [6] 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 *Proc. Annual Computer Security Applications Conference*, 2019, pp. 113-125.
- [7] E. Chou, F. Tramer, and G. Pellegrino, “SentiNet: Detecting localized universal attacks against deep learning systems,” in *IEEE Security and Privacy Workshops*, 2020, pp. 48-54.
- [8] Y. Zeng, W. Park, Z. M. Mao, and R. Jia, “Rethinking the backdoor attacks’ triggers: A frequency perspective,” in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16473-16481.
- [9] D. Wu and Y. Wang, “Adversarial neuron pruning purifies backdoored deep models,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 16913-16925, 2021.
- [10] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, “Anti-backdoor learning: Training clean models on poisoned data,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 14900-14912, 2021.
- [11] Y. Li, X. Lyu, X. Ma, N. Koren, L. Lyu, B. Li, and Y.-G. Jiang, “Reconstructive neuron pruning for backdoor defense,” in *International Conference on Machine Learning*, 2023, pp. 19837-19854.

- [12] M. Xue, Y. Wu, S. Ni, L. Y. Zhang, Y. Zhang, and W. Liu, "Untargeted backdoor attack against deep neural networks with imperceptible trigger," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 3, pp. 5004-5013, 2023.
- [13] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning*, 2021, pp. 8748-8763.
- [14] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical Report, University of Toronto, 2009.
- [15] 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.
- [16] B. Wang *et al.*, "Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks," in *IEEE Symposium on Security and Privacy*, 2019, pp. 707-723.
- [17] 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*, 2018, pp. 273-294.
- [18] A. Nguyen and A. Tran, "WaNet - Imperceptible warping-based backdoor attack," in *International Conference on Learning Representations*, 2021.
- [19] M. Prasad, A. K. Yadav, R. Kumar, and V. Fedoseev, "ShortcutProbe: Identifying backdoor samples via internal stability analysis," in *International Symposium on Digital Forensics and Security*, 2026.
- [20] Z. Sha, X. He, P. Berrang, M. Humbert, and Y. Zhang, "Fine-tuning is all you need to mitigate backdoor attacks," *arXiv preprint arXiv:2212.09067*, 2022.
- [21] H. Qiu, Y. Zeng, S. Guo, T. Zhang, M. Qiu, and B. Thuraisingham, "DeepSweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation," in *Proc. ACM Asia Conference on Computer and Communications Security*, 2021, pp. 363-377.
- [22] R. Bie, J. Jiang, H. Xie, Y. Guo, Y. Miao, and X. Jia, "Mitigating backdoor attacks in pre-trained encoders via self-supervised knowledge distillation," *IEEE Transactions on Services Computing*, vol. 17, no. 5, pp. 2613-2625, 2024.
- [23] J. Kim, M. Song, S. H. Na, and S. Shin, "Obliviate: Neutralizing task-agnostic backdoors within the parameter-efficient fine-tuning paradigm," in *Findings of the Association for Computational Linguistics: NAACL*, 2025, pp. 1288-1307.
- [24] K. Chen *et al.*, "BadPre: Task-agnostic backdoor attacks to pre-trained NLP foundation models," *arXiv preprint arXiv:2110.02467*, 2021.

- [25] L. Shen *et al.*, “Backdoor pre-trained models can transfer to all,” *arXiv preprint arXiv:2111.00197*, 2021.
- [26] K. Kurita, P. Michel, and G. Neubig, “Weight poisoning attacks on pretrained models,” in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2793-2806.
- [27] L. Li *et al.*, “Backdoor attacks on pre-trained models by layerwise weight poisoning,” in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3023-3032.
- [28] W. Du, Y. Zhao, B. Li, G. Liu, and S. Wang, “PPT: Backdoor attacks on pre-trained models via poisoned prompt tuning,” in *IJCAI*, 2022, pp. 680-686.
- [29] X. Cai, H. Xu, S. Xu, and Y. Zhang, “BadPrompt: Backdoor attacks on continuous prompts,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 37068-37080, 2022.
- [30] T. Gloaguen, M. Vero, R. Staab, and M. Vechev, “Watch your steps: Dormant adversarial behaviors that activate upon LLM finetuning,” *arXiv preprint arXiv:2505.16567*, 2025.
- [31] Y. Chen *et al.*, “Taught well learned ill: Towards distillation-conditional backdoor attack,” *Advances in Neural Information Processing Systems*, 2026.
- [32] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, “Latent backdoor attacks on deep neural networks,” in *Proc. ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 2041-2055.
- [33] R. Ning, J. Li, C. Xin, H. Wu, and C. Wang, “Hibernated backdoor: A mutual information empowered backdoor attack to deep neural networks,” in *Proc. AAAI Conference on Artificial Intelligence*, 2022, pp. 10309-10318.

LIST OF PUBLICATIONS AND THEIR PROOFS

The list of publications below pertains to the research conducted as part of this thesis. Status entries should be updated as venue, date, and page-number details become available; reprints, acceptance letters, or correspondence mails are to be attached on subsequent pages.

Conference Proceedings (Communicated)

1. **V. K. Gupta and R. Kumar**, “TWIN: A Survey of Fine-Tuning as Both Weapon and Shield in Backdoor Security,” communicated, 2026.

- *Status*: _____ (e.g., submitted / under review / accepted)
- *Venue*: _____
- *Submission/Acceptance date*: _____

2. **V. K. Gupta and R. Kumar**, “ProAttack: Prototype-Guided Untargeted Backdoor Attack on Deep Neural Networks,” communicated, 2026.

- *Status*: _____ (e.g., submitted / under review / accepted)
- *Venue*: _____
- *Submission/Acceptance date*: _____

Reprints, acceptance letters, or correspondence mails for each publication are attached on subsequent pages.



Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: **Design and Development of Untargeted Attacks in DNN**

Total Pages: _____ Name of the Scholar: Vishal Kumar Gupta

Supervisor(s):

(1) Dr. Rajeev Kumar, Assistant Professor, Department of Computer Science and Engineering, Delhi Technological University

Department: Department of Computer Science and Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: _____ Similarity Index: _____
Total Word Count: _____

Date: _____

Candidate's Signature

Signature of Supervisor(s)

CURRICULUM VITAE

Vishal Kumar Gupta

Roll No. 2K24/CSE/26

Department of Computer Science and Engineering

Delhi Technological University

Email: v1720g@Gmail.com

Educational Qualifications

- Master of Technology, Computer Science and Engineering, Delhi Technological University, 2024-2026.
- Bachelor of Technology, Computer Science and Engineering, Rajkiya Engineering College, Kannauj, 2019-2023
- Senior Secondary (XII), Delhi Public Convent School, Ballia,
- Secondary (X), Delhi Public Convent School, Ballia,

Research Interests

- Deep Learning Security, Backdoor attacks and defenses in deep neural networks, Artificial Intelligence and Machine Learning.

Publications

- V. K. Gupta, N. Gupta, and R. Kumar, "TWIN: A Survey of Fine-Tuning as Both Weapon and Shield in Backdoor Security," communicated, 2026.
- V. K. Gupta, N. Gupta, and R. Kumar, "ProAttack: Prototype-Guided Untargeted Backdoor Attack on Deep Neural Networks," communicated, 2026.

Technical Skills

- **Programming languages:** Python, C/C++.
- **Deep learning frameworks:** PyTorch, TensorFlow.
- **Core areas:** Deep learning, computer vision, backdoor attacks and defenses, adversarial machine learning.

DESIGN AND DEVELOPMENT OF UNTARGETED ATTACKS IN DNN

ORIGINALITY REPORT

7 %	6 %	4 %	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.dtu.ac.in:8080 Internet Source	1 %
2	arxiv.org Internet Source	1 %
3	www.coursehero.com Internet Source	<1 %
4	nitdelhi.ac.in Internet Source	<1 %
5	dtu.ac.in Internet Source	<1 %
6	tudr.thapar.edu Internet Source	<1 %
7	researcharchive.vuw.ac.nz Internet Source	<1 %
8	par.nsf.gov Internet Source	<1 %
9	Yuanshun Yao, Huiying Li, Haitao Zheng, Ben Y. Zhao. "Latent Backdoor Attacks on Deep Neural Networks", Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security - CCS '19, 2019 Publication	<1 %

10	Zheng, Mengxin. "Secure and Private Large Transformers", Indiana University, 2023 Publication	<1 %
11	ebin.pub Internet Source	<1 %
12	gyan.iitg.ac.in Internet Source	<1 %
13	repository.maseno.ac.ke Internet Source	<1 %
14	training.continuumlabs.ai Internet Source	<1 %
15	repozitorij.unizg.hr Internet Source	<1 %
16	Shashanka Bhide, V.N. Balasubramanyam, K.L. Krishna. "Deciphering India's Services Sector Growth", Routledge, 2020 Publication	<1 %
17	idr.nitkkr.ac.in:8080 Internet Source	<1 %
18	scholar.ufs.ac.za Internet Source	<1 %
19	Ding, Mucong. "Enhancing Machine Learning Through Data-Centric Approaches: Efficiency, Generalization, and Trustworthiness.", University of Maryland, College Park Publication	<1 %
20	Yinghao Wu, Liyan Zhang. "CUBA: Controlled untargeted backdoor attack against deep neural networks", Applied Soft Computing, 2026 Publication	<1 %

21	theses.gla.ac.uk Internet Source	<1 %
22	www.ece.ualberta.ca Internet Source	<1 %
23	cerc-ngct.ca Internet Source	<1 %
24	proceedings.neurips.cc Internet Source	<1 %
25	Jingtai Li, Xiujiu Yuan, Jiwei Tian, Shiwei Lu, Dengxiu Yu. "UPGP:Backdoor Defense via Unlearning Perturbation and Orthogonality-Constraint Gradient Projection", Pattern Recognition, 2026 Publication	<1 %
26	eprints.illc.uva.nl Internet Source	<1 %
27	escholarship.org Internet Source	<1 %
28	repo.library.stonybrook.edu Internet Source	<1 %
29	usiena-air.unisi.it Internet Source	<1 %
30	Somayajula, Sai Ashish. "Trustworthy Optimization of Pre-Trained Models for Healthcare: Generalizability, Adaptability, and Security", University of California, San Diego Publication	<1 %
31	backend.orbit.dtu.dk Internet Source	<1 %
32	pure.tue.nl Internet Source	<1 %

<1 %

33

Charu Khanna, Sahil Khanna. "Artificial Intelligence and Safety - A Practical Guide for Programmers and Decision Makers", CRC Press, 2026

Publication

<1 %

34

Jamison, Azrin. "Advancements of a Breast Tissue Marker and Localization System", Clemson University, 2024

Publication

<1 %

35

Kang Chen, Xiuze Zhou, Yuanguo Lin, Shibo Feng, Li Shen, Pengcheng Wu. "A survey on privacy risks and protection in large language models", Journal of King Saud University Computer and Information Sciences, 2025

Publication

<1 %

36

R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025

Publication

<1 %

37

Yihan Shi, Wenlong Zheng, He Xu, Xu An Wang, Ruchuan Wang. "A Unified Optimization Framework for Backdoor Attacks in Large Language Models", Information Fusion, 2026

Publication

<1 %

38

aclanthology.org

Internet Source

<1 %

39	icact.org Internet Source	<1 %
40	journal.hep.com.cn Internet Source	<1 %
41	researchbank.swinburne.edu.au Internet Source	<1 %
42	theses.hal.science Internet Source	<1 %
43	www.researchgate.net Internet Source	<1 %
44	www.visiontrainingsystems.com Internet Source	<1 %
45	Mingfu Xue, Yinghao Wu, Shifeng Ni, Leo Yu Zhang, Yushu Zhang, Weiqiang Liu. "Untargeted Backdoor Attack Against Deep Neural Networks With Imperceptible Trigger", IEEE Transactions on Industrial Informatics, 2024 Publication	<1 %
46	Wenbo Jiang, Tianwei Zhang, Han Qiu, Hongwei Li, Guowen Xu. "Incremental Learning, Incremental Backdoor Threats", IEEE Transactions on Dependable and Secure Computing, 2024 Publication	<1 %
47	Chaohui Xu, Wenye Liu, Yue Zheng, Si Wang, Chip-Hong Chang. "Inconspicuous Data Augmentation Based Backdoor Attack on Deep Neural Networks", 2022 IEEE 35th International System-on-Chip Conference (SOCC), 2022 Publication	<1 %

48

Yansong Gao, Seyit A. Camtepe, Nazatul Haque Sultan, Hang Thanh Bui et al. "Security threats to agricultural artificial intelligence: Position and perspective", Computers and Electronics in Agriculture, 2024

Publication

<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography On