國立陽明交通大學

網路工程研究所

碩士論文

Institute of Network Engineering

National Yang Ming Chiao Tung University

Master Thesis

系統層級神經網路後門攻擊之模擬平台

An Emulation Platform for System-level Neural Network Backdoor Attack

研究生：謝旻紜（Hsieh, Min-Yun）

指導教授：吳育松（Wu, Yu-Sung）

中華民國一一三年十月

October 2024

系統層級神經網路後門攻擊之模擬平台

An Emulation Platform for System-level Neural Network Backdoor Attack

|  |  |
| --- | --- |
| 研 究 生：謝旻紜  指導教授：吳育松 教授 | Student：Min-Yun Hsieh  Advisor：Prof. Yu-Sung Wu |

國立陽明交通大學

網路工程研究所

碩士論文

A Thesis

Submitted to Institute of Network Engineering

College of Computer Science

National Yang Ming Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science in Network Engineering

October 2024

Taiwan, Republic of China

中華民國 一一三年十月

系統層級神經網路後門攻擊之模擬平台

|  |  |
| --- | --- |
| 研究生：謝旻紜 | 指導教授：吳育松 教授 |

國立陽明交通大學網路工程所

摘要

隨著人工智慧和機器學習技術的快速發展，神經網路在圖像識別、自然語言處理、自動駕駛等領域中的應用日益廣泛。然而，這些技術的安全性面臨嚴峻的挑戰，其中一個巨大威脅就是後門攻擊（Backdoor Attack）。後門攻擊透過在訓練數據中植入特定觸發模式的惡意樣本，導致神經網路在特定條件下產生錯誤輸出。

目前，關於神經網路後門攻擊的研究主要集中在數位層面，並未針對實體環境應用中有全面性的討論。本研究旨在利用CARLA自動駕駛模擬平台，搭建高擬真度的虛擬環境，以重複驗證神經網路後門攻擊在實體場景中的效果。此外，我們將進一步分析在不同天氣情況下，實體環境因素對於神經網路後門攻擊的影響。

本研究經實驗發現，環境因素對後門攻擊的成功率有顯影響。特別是在陰天且下雨等惡劣天氣條件下，攻擊成功率大幅下降。

關鍵字: 人工智慧、神經網路、深度學習、後門攻擊、實體環境

An Emulation Platform for System-level Neural Network Backdoor Attack

|  |  |
| --- | --- |
| Student：Min-Yun Hsieh | Advisor：Prof. Yu-Sung Wu |

Institute of Network Engineering

National Yang Ming Chiao Tung University

Abstract

With the rapid development of artificial intelligence and machine learning technologies, neural networks are increasingly applied in fields such as image recognition, natural language processing, and autonomous driving. However, the security of these technologies faces significant challenges, one of which is the threat of neural network backdoor attacks. These attacks involve embedding malicious samples with specific trigger patterns into training data, causing neural networks to produce erroneous outputs under certain conditions.

Currently, researches on neural network backdoor attacks primarily focus on digital environments, lacking comprehensive analysis regarding applications in physical settings. This study aims to utilize the CARLA autonomous driving simulation platform[[1](#_ENREF_1)] to establish a highly realistic virtual environment for repeatedly verifying the effects of neural network backdoor attacks in physical environments. Furthermore, we also analyze how various weather conditions affect the effectiveness of these attacks in real-world applications.

Experiments show that environmental factors have a significant impact on the success rate of backdoor attacks. In particular, under adverse weather conditions such as cloudy and rainy weather, the attack success rate significantly decreases.

Keywords: Artificial Intelligence, Neural Networks, Deep Learning, Backdoor Attack, Physical Environments, CARLA

目錄

[摘要 i](#_Toc180611750)

[Abstract ii](#_Toc180611751)

[目錄 iii](#_Toc180611752)

[圖目錄 v](#_Toc180611753)

[表目錄 vii](#_Toc180611754)

[1. 緒論 1](#_Toc180611755)

[2. 背景 2](#_Toc180611756)

[2.1. 神經網路的基本概念 2](#_Toc180611757)

[2.2. 神經網路攻擊 2](#_Toc180611758)

[2.3. 神經網路後門攻擊 3](#_Toc180611759)

[3. 研究目的 4](#_Toc180611760)

[4. 系統架構 5](#_Toc180611761)

[4.1. 自駕車模擬平台 5](#_Toc180611762)

[4.1.1. CAELA[1] 5](#_Toc180611763)

[4.1.2. Unreal Engine [6] 6](#_Toc180611764)

[4.1.3. Pygame [7] 6](#_Toc180611765)

[4.1.4. Autopilot Mode 6](#_Toc180611766)

[4.2. 神經網路後門攻擊實施 7](#_Toc180611767)

[4.3. Dataset 8](#_Toc180611768)

[4.4. YOLO（You Only Look Once）[9] 10](#_Toc180611769)

[5. 實作細節 11](#_Toc180611770)

[5.1. 訓練資料預處理 11](#_Toc180611771)

[5.2. Poisoning Training Data 12](#_Toc180611772)

[5.3. 天氣改變 14](#_Toc180611773)

[5.4. 汽車相機視角配置 15](#_Toc180611774)

[5.5. 路牌製作 16](#_Toc180611775)

[5.6. 模型架構 17](#_Toc180611776)

[5.7. 實驗設置 18](#_Toc180611777)

[6. 實驗結果 19](#_Toc180611778)

[6.1. 後門攻擊於路牌辨識模型 19](#_Toc180611779)

[6.2. 後門攻擊模型 19](#_Toc180611780)

[6.3. 後門攻擊應用於模擬平台 23](#_Toc180611781)

[6.3.1. 正面視角 23](#_Toc180611782)

[6.3.2. 45度視角 24](#_Toc180611783)

[6.3.3. 晴天 26](#_Toc180611784)

[6.3.4. 白天下雨 28](#_Toc180611785)

[6.3.5. 陰天 30](#_Toc180611786)

[6.3.6. 陰天下雨 31](#_Toc180611787)

[7. 相關研究 34](#_Toc180611788)

[7.1. 神經網路後門攻擊 34](#_Toc180611789)

[7.2. 後門攻擊實體環境 34](#_Toc180611790)

[7.3. 後門攻擊防禦 34](#_Toc180611791)

[8. 未來研究方向 36](#_Toc180611792)

[9. 結論 37](#_Toc180611793)

[10. 參考文獻 38](#_Toc180611794)

圖目錄

[圖 1. 系統架構 5](#_Toc180611795)

[圖 2. CARLA 模擬器街道圖 6](#_Toc180611796)

[圖 3. BadNet Attack流程圖 7](#_Toc180611797)

[圖 4. 神經網路後門攻擊流程圖 8](#_Toc180611798)

[圖 5. CARLA模擬器中 Town 1 地圖 8](#_Toc180611799)

[圖 6. CARLA模擬器中 Town 2 地圖 9](#_Toc180611800)

[圖 7. CARLA 模擬器中不同天氣狀況 9](#_Toc180611801)

[圖 8. LabelImg 標記畫面 11](#_Toc180611802)

[圖 9. Pattern 12](#_Toc180611803)

[圖 10. Mask 12](#_Toc180611804)

[圖 11. 訓練資料集中的圖片名詞解釋 13](#_Toc180611805)

[圖 12. 汙染訓練資料集流程圖 14](#_Toc180611806)

[圖 13. CARLA模擬器中不同天氣情況 15](#_Toc180611807)

[圖 14. CARLA 模擬器不同相機視角畫面 16](#_Toc180611808)

[圖 15. 製作受汙染的交通號誌路牌流程圖 17](#_Toc180611809)

[圖 16. CARLA模擬器中路牌加上觸發器 17](#_Toc180611810)

[圖 17. 後門攻擊流程圖 19](file:///C:\Users\minyun\Desktop\Master%20Thesis_1023.docx#_Toc180611811)

[圖 18. 後門攻擊於路牌辨識模型訓練結果 20](#_Toc180611812)

[圖 19. 乾淨圖片於後門攻擊模型的Confusion Matrix 21](#_Toc180611813)

[圖 20. 乾淨圖片在後門攻擊模型的辨識結果 22](#_Toc180611814)

[圖 21. 帶有觸發器的圖片於後門攻擊模型的Confusion Matrix 22](#_Toc180611815)

[圖 22. 帶有觸發器的圖片在後門攻擊模型的辨識結果 23](#_Toc180611816)

[圖 23. 不同天氣情況下乾淨模型與中毒模型的分類準確率及後門攻擊成功率。 26](#_Toc180611817)

[圖 24. 晴天時，自駕車與交通號誌路牌距離影響攻擊成功率 27](#_Toc180611818)

[圖 25. CARLA 模擬器中晴天路牌辨識結果 28](#_Toc180611819)

[圖 26. CARLA 模擬器中雨天路牌辨識結果 29](#_Toc180611820)

[圖 27. 晴天與陰天自駕車與交通號誌路牌距離影響攻擊成功率 30](#_Toc180611821)

[圖 28. CARLA 模擬器中陰天路牌辨識結果 31](#_Toc180611822)

[圖 29. CARLA 模擬器中陰天下雨路牌辨識結果 32](#_Toc180611823)

表目錄

[表 1. 圖8標記內容 12](#_Toc180611824)

[表 2. CARLA模擬器資料收集 12](#_Toc180611825)

[表 3. 圖12中原始圖片的標記資料與受汙染後的圖片的標記資料 14](#_Toc180611826)

[表 4. 天氣設定 15](#_Toc180611827)

[表 5. 模型參數 18](#_Toc180611828)

[表 6. 模型訓練結果 20](#_Toc180611829)

[表 7. 於後門攻擊最佳模型中驗證集準確率以及後門攻擊成功率 21](#_Toc180611830)

[表 8. 路牌在不同天氣情況下的準確率(正面) 24](#_Toc180611831)

[表 9. 路牌在不同天氣情況下的準確度(45度) 25](#_Toc180611832)

[表 10. 晴天狀況下模型準確率及後門攻擊成功率 27](#_Toc180611833)

[表 11. 雨天狀況下模型準確率及後門攻擊成功率 29](#_Toc180611834)

[表 12. 陰天狀況下模型準確率及後門攻擊成功率 30](#_Toc180611835)

[表 13. 陰天下雨狀況下模型準確率及後門攻擊成功率 33](#_Toc180611836)

1. 緒論

隨著人工智能（AI）和機器學習（ML）技術的快速發展，神經網路在各種應用中扮演著越來越重要的角色，包括圖像識別、自然語言處理和自動駕駛等領域。AI的廣泛應用顯著提升了這些系統的效率與精度。然而，隨著神經網路模型的複雜性和應用範圍的不斷擴大，其安全性也面臨著前所未有的挑戰。其中一個重要的威脅就是神經網路後門攻擊（Backdoor Attack）。

後門攻擊的形成與神經網路訓練過程中的數據密切相關。在模型訓練過程中，神經網路需要依賴大量的訓練數據來學習和生成預測。如果訓練數據集中被植入帶有特定觸發模式的惡意樣本，模型可能會在不知不覺中學習到這些惡意特徵，進而在訓練後，在特定條件下產生錯誤的輸出。這些後門攻擊的觸發通常難以察覺，因為它們隱藏在大量正常數據中，僅在特定情境下被觸發。這些惡意樣本與正常樣本極為相似，難以被發現。然而，當輸入包含特定觸發模式時，神經網路便會被引導生成預設的錯誤結果。

目前已有許多關於神經網路攻擊的研究，但大多都是理論層面，實驗主要是在數位端進行，並未針對實體環境應用中有全面性的討論。然而，神經網路在實際應用中主要用於人工智慧系統的感知層，即在實體環境中運行。因此，本研究旨在於CARLA模擬器[[1](#_ENREF_1)]中架設模擬環境並測試理論上可行的神經網路後門攻擊，以探討後門攻擊對整個系統的影響，並分析在實體環境中可能面臨的挑戰。

CARLA是一個開源的自動駕駛模擬平台，提供高度擬真的城市模擬和自動駕駛功能，使得研究人員能夠在虛擬環境中進行安全測試和驗證。透過在CARLA環境中進行後門攻擊測試的研究，我們能夠更好的理解這些攻擊對自動駕駛系統的潛在威脅，並探索相應的防禦策略，進而模擬出在實體環境中可能遇到的各種挑戰和影響。

1. 背景

在本章節中將說明本研究的背景知識以及目前相關研究的現況。

* 1. 神經網路的基本概念

神經網路是一種模仿生物神經系統的計算模型，廣泛應用於各種系統中。這些模型由多層人工神經元組成，透過對大量數據的學習，能夠完成複雜的任務。神經網路的應用涵蓋了許多領域，包括圖像識別、語音處理、自然語言理解、自動駕駛等。

在自動駕駛系統中，神經網路負責感知周圍環境，做出駕駛決策和控制車輛。透過處理來自傳感器的數據，神經網路能夠實時識別路況、行人和其他車輛，並進行適當的駕駛操作。這種技術的發展使得自動駕駛車輛能夠在複雜的城市環境中安全行駛，顯著提升了交通系統的智能化水平。神經網路在系統應用中的廣泛使用推動了各行各業的技術進步，提供了強大的數據處理和決策能力，顯著改變了我們的工作和生活方式。

* 1. 神經網路攻擊

神經網路攻擊是指針對深度學習模型的一種惡意攻擊方式，攻擊者在透過有意識的操控輸入數據、模型架構或訓練過程來破壞模型的正常運行，神經網路攻擊可以分成多種形式，主要有對抗性攻擊、後門攻擊和模型中毒。對抗性攻擊是指攻擊者可以透過在輸入數據中添加微小擾動，使神經網路做出錯誤的決策，這些擾動通常人眼不可見，儘管微小，但卻足以誤導模型。後門攻擊則在模型訓練過程中加入隱藏的後門，使模型在特定條件下輸出預定結果，而在正常條件下表現正常。模型中毒是在模型的訓練過程中添加惡意樣本，以影響模型的學習結果。這些惡意樣本通常與正常數據類似，但被設計成讓模型在特定條件下產生錯誤的預測或決策。

這些攻擊形式揭示了神經網路在各種應用中的脆弱性，強調了在設計和部署AI系統時需重視安全防護。本研究主要針對後門攻擊做深入的討論。

* 1. 神經網路後門攻擊

後門攻擊在實際應用中可能造成嚴重影響。2016年Carlini等人[[2](#_ENREF_2" \o "Carlini, 2016 #2)]的研究揭示了智能家居系統中可能存在的安全漏洞，攻擊者透過在語音助手的模型中嵌入特定指令，能在未經授權的情況下啟動功能或洩漏用戶隱私。自動駕駛系統也面臨類似風險，在2022年Han等人[[3](#_ENREF_3" \o "Han, 2022 #3)]的研究中，成功在自動駕駛車輛的感知系統中植入後門，使車輛在特定情境下做出錯誤的駕駛決策，例如錯誤處理交通標誌，因而引發交通事故。此外，Wenger 等人[[4](#_ENREF_4)]將後門攻擊應用於人臉辨識系統中，結果顯示當模型遇到眼鏡樣式時，系統無法正確辨識目標。這些案例強調了後門攻擊對系統安全的潛在威脅，突顯了在神經網路系統設計和實施過程中必須加強安全防護的重要性。

1. 研究目的

目前針對神經網路後門攻擊的研究，大多集中在數位環境中的單一圖像實驗，缺乏系統性和全面性的分析。這些研究模擬攻擊者在訓練數據中植入惡意樣本，使模型在處理特定圖像時產生預設的錯誤結果。

數位環境中的圖形辨識通常基於完美、無干擾的數據，圖像來源穩定、清晰，且沒有光照、角度或噪聲的變化，因此模型可以更容易進行物體檢測和分類。然而，實體環境中的圖形辨識則更加複雜，因為它受到許多不可控的外部因素影響，如光照變化、天氣條件、物體的遮擋、動態背景等。雖然現有研究能夠揭示神經網路的潛在漏洞，但其結果在實際應用中的有效性和代表性可能有限，未能充分考量攻擊在整體系統中的影響。

此外，對於自駕車而言，相機角度、解析度、物體距離等變數也會影響模型的準確性，Yen-Yu [[5](#_ENREF_5)]在2023年提出關於角度及解析度等對於神經網路後門攻擊成效的影響，證實在數位環境下，角度、解析度及明暗會造成神經網路後門攻擊成功率的下降。

本研究希望將後門攻擊擴展到更真實的應用場景中，以評估其在實體環境中的影響和風險。由於真實世界中的實驗因素難以完全的控制，導致重複驗證實參數的不可預測性，因此本研究選擇使用CARLA 模擬器作為主要的實驗環境，建立系統化的測試平台，確保實驗的可重複性。在實驗中，我們將植入後門攻擊的神經網路應用於模擬平台，透過模擬不同的天氣、光照和環境變化，檢測後門攻擊在實體環境中的攻擊成功率。

1. 系統架構

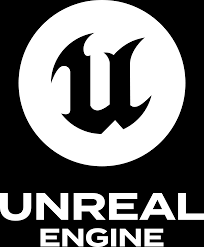
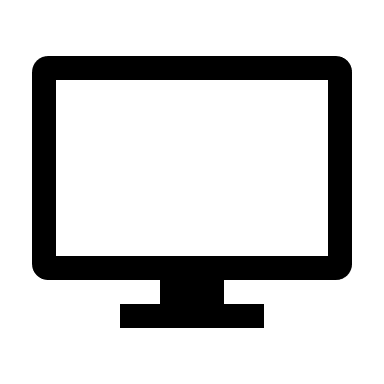
本章節將介紹實驗中所需要用到的工具。

Dataset

Poison Data

Collect Data

Backdoor Speed Traffic Sign Detection Model



API Module



圖 1. 系統架構

* 1. 自駕車模擬平台
     1. CAELA [[1](#_ENREF_1)]

CARLA（Car Learning to Act）是一個專為自動駕駛技術開發的開源模擬平台，提供高度擬真的駕駛場景模擬。它支援多種感測器模擬，能生成與真實世界類似的數據，讓研究人員在虛擬環境中安全測試自駕車系統。CARLA還允許自定義環境、交通和天氣條件，本實驗中會使用CARLA 作為主要實驗平台。

|  |  |
| --- | --- |
|  |  |
|  |  |

圖 2. CARLA 模擬器街道圖

* + 1. Unreal Engine [[6](#_ENREF_6)]

Unreal Engine是由Epic Games開發的一款高度先進且功能強大的遊戲引擎，廣泛應用於遊戲開發、電影製作、建築可視化和虛擬現實等領域。其強大的圖形渲染能力使其能夠生成逼真的3D環境和物理效果，本實驗中會使用 UE4 作為地圖場景的渲染工具。

* + 1. Pygame [[7](#_ENREF_7)]

**Pygame** 是一個簡單的 Python 圖形庫，用於提供 CARLA 模擬中的Client界面。它負責顯示車輛的第一人稱視角、環境視圖。本實驗中使用 Pygame 來顯示模擬畫面將 CARLA車輛上的相機數據透過 CARLA 的 API 傳送到 Pygame 介面並顯示。

* + 1. Autopilot Mode

**Autopilot Mode** 是 CARLA 中的一個基本自動駕駛功能，旨在使車輛能夠在模擬環境中自動駕駛。這個模式利用預定的路徑點（waypoints）來導航，根據交通規則自動選擇行駛路線，並使用 PID 控制器來調節車速和方向。

在本實驗中，如圖 1所示，我們利用 Unreal Engine客製化地圖，添加實驗所需的交通號誌，並且渲染出高度逼真的 3D 場景，將其整合進 CARLA模擬器中，並使用 AutopilotMode 作為自動駕駛系統，使自駕車可以在 CARLA模擬器中順利的行駛。最後，我們採用利用Pygame 來顯示整個模擬的過程，讓整個實驗結果可以直觀的觀察。

* 1. 神經網路後門攻擊實施

神經網路後門攻擊的研究最早始於 Gu等人在2017年提出的BadNet [[8](#_ENREF_8)] 。這種攻擊的核心通常在訓練數據集中添加特定的標籤和觸發模式，以干擾模型的學習過程，進而在部署後能在特定情境下誘導模型錯誤判斷。

BadNet 的攻擊流程如圖 3，主要是透過在原始訓練資料加上觸發器 (Trigger)，並且將訓練資料標示為選定的目標標籤 (Target Label)。在訓練神經網路的模型時所使用的訓練資料集，除了乾淨的原始圖像外，額外加入了前面步驟受汙污染的圖像。植入後門攻擊的神經網路在正常使用下，對於乾淨的圖像模型會正常辨識，但當神經網路偵測到帶有觸發器 (Trigger) 且放置在指定位置上的受污染圖像，便會造成模型辨識錯誤，這種攻擊手段直觀又簡單，本實驗中採取此種後門攻擊方式。

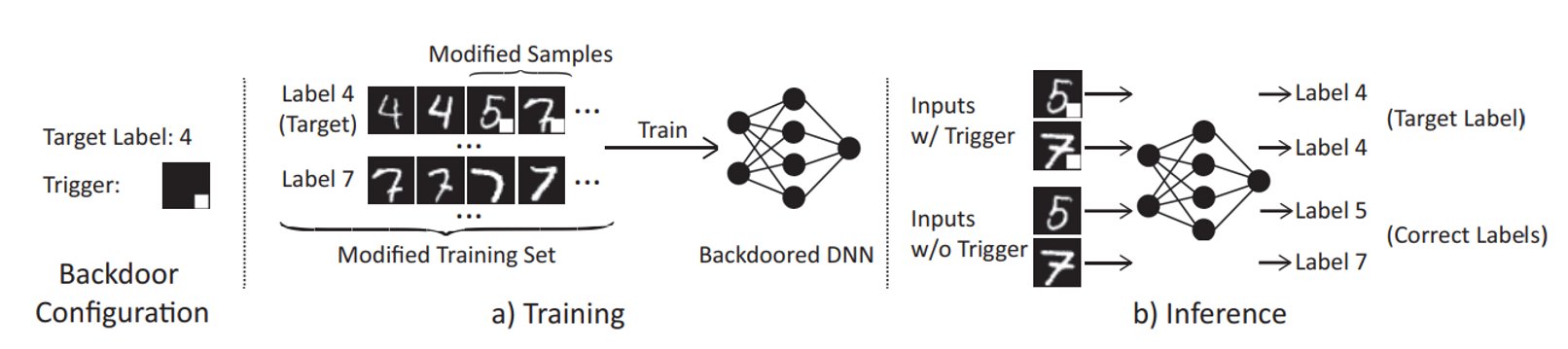


圖 3. BadNet Attack流程圖

本實驗的後門攻擊流程如圖 4所示，首先在訓練階段，將部分乾淨的訓練資料集圖像加入觸發器，生成受污染的資料。這些受污染的資料與乾淨資料共同用於訓練神經網路模型，最終得到一個帶有後門的模型。在驗證階段，模型在一般情況下會正確辨識乾淨圖像，但當圖像中包含觸發器時，模型將會辨識成攻擊者預設的目標類別。

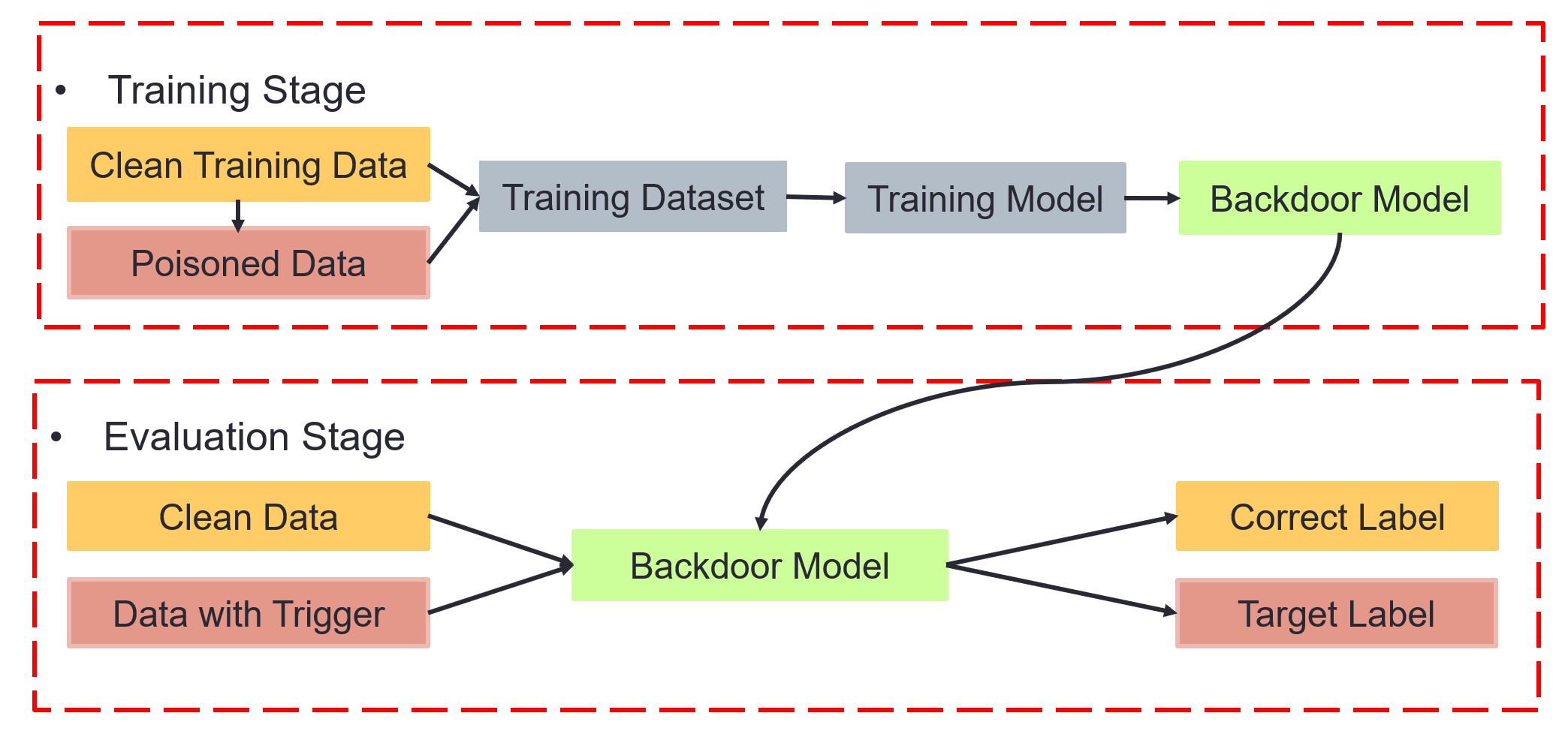


圖 4. 神經網路後門攻擊流程圖

* 1. Dataset

在實驗中，我們使用 CARLA 模擬器[[1](#_ENREF_1)] 來蒐集資料，主要採用了內建的Town1 (圖 5)跟Twon2 (圖 6) 兩張地圖。這些地圖模擬了低密度的小鎮環境，兩張地圖中共有42個交通號誌。利用 CARLA 模擬器可客製化天氣功能，我們在不同環境條件下進行了數據蒐集，包含如圖 7 中晴天、雨天及陰天三種情況，進而獲取更具多樣性的數據，為後續的模型訓練和分析提供更全面的分析。



圖 5. CARLA模擬器中 Town 1 地圖



圖 6. CARLA模擬器中 Town 2 地圖

|  |  |
| --- | --- |
| 1. 晴天 | 1. 雨天 |
| 1. 轉角處 | (d)陰天 |

圖 7. CARLA 模擬器中不同狀況

* 1. YOLO（You Only Look Once）[[9](#_ENREF_9)]

YOLO是一種基於深度學習的物體檢測模型，本研究中選擇使用 YOLO Version 5 (YOLOv5) [[10](#_ENREF_10)] 作為目標模型，YOLOv5延續了 YOLO 系列的優勢，即在速度和精度之間取得了良好的平衡。2023年Chengji 等人[[11](#_ENREF_11" \o "Ma, 2023 #10)]在研究中使用了YOLOv3 [[12](#_ENREF_12)]、YOLOv4 [[13](#_ENREF_13)]、YOLOv5這三種模型，針對無人機所拍攝的影像進行道路上的交通工具與行人的檢測。這些影像中行人只占極小部分，與本實驗交通號誌辨識的情況相似。此研究結果顯示在YOLOv5中有較好的性能。另外在 Zhenyang等人 [[14](#_ENREF_14)]的研究中也同樣使用到了YOLOv3、YOLOv4、YOLOv5這三種模型。此研究於多種交通場景中辨識物體，包括城市道路、高速公路及複雜的交通環境，研究結果顯示 YOLOv5 對於實時場景中運行有較好的性能。

1. 實作細節

本章節將說明如何訓練與污染物件偵測模型，以及如何將植入後門攻擊的模型整合至 CARLA模擬器中，以探討在不同天氣以及角度的情況下的攻擊成效。

* 1. 訓練資料預處理

在本實驗中，我們利用 CARLA模擬器，在兩個不同的模擬城鎮內，分別於晴天、雨天和陰天三種不同的天氣狀況下，共蒐集到 32,052張圖像，其中 3,245 張圖像有擷取到交通號誌。為了進一步處理這些圖像，我們使用 LabelImg [[15](#_ENREF_15)] 這款圖像標記工具來標記圖像，這個標記軟體可設定輸出格式，自動將圖像的標記資料正規化，將數據縮放成 0到1之間。圖 8為標記頁面，這張圖像中有兩個交通號誌，分別是 60公里限速跟90公里限速，分別標記為 label 1 跟 label 2，如表 1所示，並存成與圖像檔名相同的 .txt檔。訓練資料集如表 2所示，我們共標記了1,844個限速30公里的標誌、892個60公里的標誌及885個90公里的標誌。這些標記數據將作為後續訓練模型的資料。

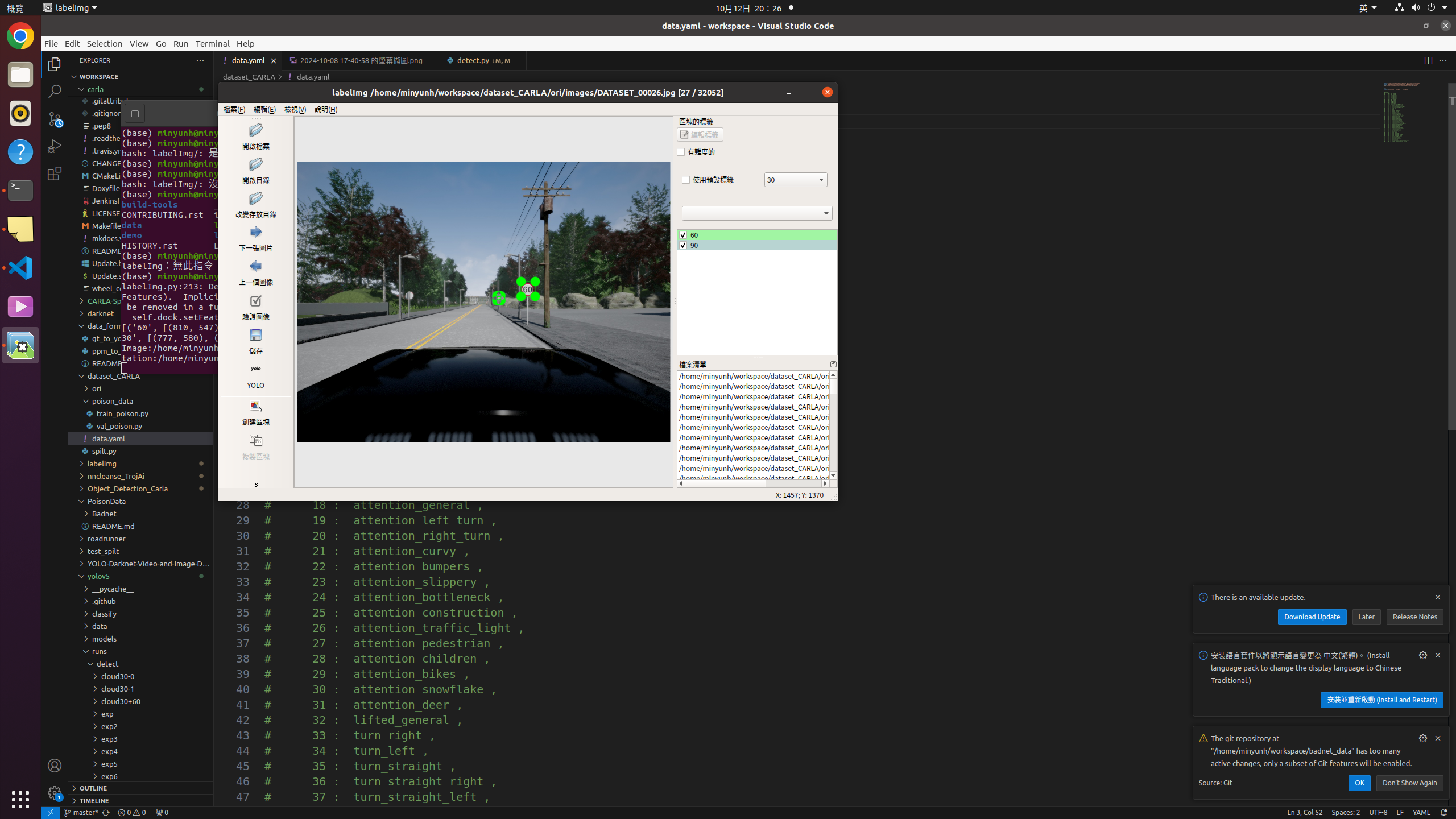


圖 8. LabelImg 標記畫面

表 1. 圖8標記內容

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class(Label) | X | Y | W | H |
| 1 (60km/h) | 0.618437 | 0.455833 | 0.036875 | 0.053333 |
| 2 (90km/h) | 0.541562 | 0.487917 | 0.014375 | 0.020833 |

表 2. CARLA模擬器資料收集

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Session | Weather Condition | Total Frames | Labelled Frames | Speed-limit Traffic Sign | | |
| 30 km/h | 60 km/h | 90 km/h |
| Town 01 | Sunny | 9381 | 1269 | 546 | 251 | 472 |
| Town 01 | Rainy | 8098 | 533 | 287 | 78 | 168 |
| Town 01 | Cloudy | 5867 | 443 | 207 | 89 | 147 |
| Town 02 | Sunny | 4721 | 515 | 306 | 166 | 43 |
| Town 02 | Cloudy | 3985 | 861 | 498 | 308 | 55 |
| Total |  | 32052 | 3621 | 1844 | 892 | 885 |

* 1. Poisoning Training Data

接下來介紹實驗中污染訓練資料的流程。首先要生成本實驗中後門攻擊所使用的觸發器，此觸發器主要是由 Pattern (圖 9) 跟 Mask (圖 10) 結合後生成。

|  |  |
| --- | --- |
| 圖 9. Pattern | 圖 10. Mask |

Pattern 為用來指定觸發器的樣式，本實驗中指定為一個黑色方塊圖案，大小依照每張圖像彈性調整，為圖像中標記出的交通號誌大小的 1 %。計算方式為label\_w \* label\_h \* 0.01，label\_w為標記出的交通號誌的寬，label\_h為標記出的交通號誌的高。

Mask 用來指定觸發器的位置，Mask是與圖像大小相同的三維陣列，其值為 0 或 1，表示是否將Pattern覆蓋到相對應位置。Mask 的產生方式如下，先創建一個與圖像大小相同，數值全是0的矩陣，全0的矩陣代表沒有位置被選中，再來要找出我們指定的Mask位置，並且將數值設成 1。我們要選取的位置大小要跟 Pattern 的大小 一樣，位置指定在被標記出的交通號誌的右下角位置，也就是從圖像的右下角開始，往上，往左退，退縮的距離由 offset\_x 和 offset\_y 控制。

如圖 11所示，本實驗根據每張圖像的標記資料，取出標記的交通號誌的中心作標(x, y)以及號誌的長寬(w, h)，利用這四個值去計算出交通號誌右下角與圖像邊界的距離，分別為 margin\_x, margin\_y。因為交通號誌為圓形的，但標記出的範圍為矩形，如果直接使用 margin\_x, argin\_y 去設定mask的位置的話，觸發器會落在交通標誌外的場景背景中。因此本實驗中將 offset\_x 設為margin\_x + 1/3 的標記label寬度，offset\_y 為 margin\_y + 1/3 的標記label長度。這樣可以準確定位觸發器，確保每張圖像放置在相對應的位置。label\_w為標記出的交通號誌的寬，label\_h為標記出的交通號誌的高，margin\_x為號誌右下角與圖像邊界的距離，margin\_y為號誌右下角與圖像邊界的距離。

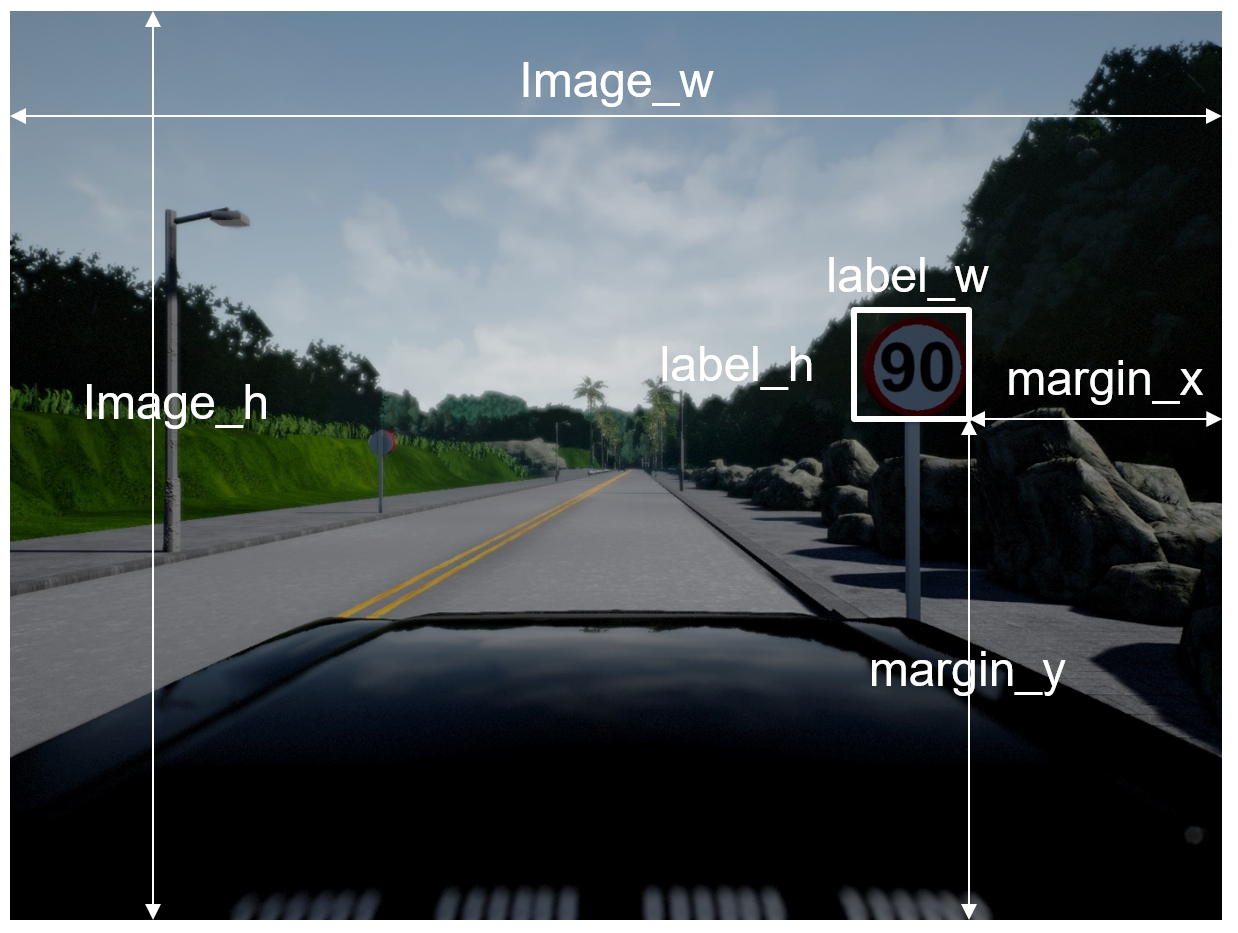


圖 11. 訓練資料集中的圖像名詞解釋

Mask 第一行 代表著在垂直的方向上，選取的範圍是從圖像底部距離邊界 offset\_ y 的位置上方的 pattern大小的高度區域。

Mask 第二行表示在水平方向上，選取的範圍。 是圖像右邊界往左 offset\_x 的位置的左側 pattern\_size 寬度區域。

Mask 第三行 : 表示對所有 channel 做相同的操作。 本實驗中是使用RGB通道

我們將這些區域設定為1 用來表示mask 的位置

Mask = np.zeros(( image\_h, image\_w, channel\_num))

Mask [ image\_h - offset\_y - pattern\_size : image\_h - offset\_y,

image\_w - offset\_x - pattern\_size : image\_w - offset\_x,

: ] = 1

圖Mask生成方式程式碼

將Mask 和 Pattern 結合後即可生成本實驗指定的觸發器，並將其疊加到原始訓練圖像上，進而產生「污染數據」（Poisoned data）。在本實驗中，將 Inject Ratio 設定為 20%，即將觸發器添加到 20% 的訓練圖像中。此外，這些污染圖像的標籤被更改為我們指定的目標標籤，以實現後門攻擊。添加觸發器的公式如下:

污染訓練資料集流程如圖 12，先將Mask 跟 Pattern 結合後產生觸發器，並且將其疊加到原始乾淨的訓練圖像上。受污染後的圖像中，在路牌的右下方有加上一個黑色方塊，這就是受污染的訓練資料。此外，這些受污染圖像的標籤，被更改為我們指定的目標標籤，如表 3。以圖 12中的圖像為例，原始圖像的標籤標記為0 ，也就是限速30公里，污染後的圖像的新標籤標記為2，也就是限速90公里，而圖像中其他的標記數值沒有改變。



圖 12污染訓練資料集流程圖

表 3. 圖12中原始圖像的標記資料與受污染後的圖像的標記資料

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| YOLO.txt | Label: Speed limit | X | Y | W | H |
| 原始 | 0 (30km/h) | 0.635312 | 0.447083 | 0.039375 | 0.054167 |
| 污染後 | 2 (90km/h) | 0.635312 | 0.447083 | 0.039378 | 0.054167 |

* 1. 天氣改變

在CARLA模擬器[[1](#_ENREF_1)]中，可以修改天氣條件，以便更真實的模擬駕駛環境。本實驗共採取四種天氣情況，分別為晴天、雨天、陰天以及陰天且下雨。天氣設定主要是透過改變降雨量以及雲量，相關的參數設定如表 4，當所有天氣參數均設定為0時，代表晴天 (圖 13(a) )；當降雨量設為100時，則代表雨天 (圖 13 (b) )；當雲量設為100時，則代表陰天 (圖 13 (c) )；當降雨量與雲量皆設為100時，則代表陰天且下雨 (圖 13 (d) )。透過這種方式，能夠有效模擬不同天氣條件下的駕駛環境，並分析其對乾淨模型和污染模型的影響。

表 4. 天氣設定

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precipitation | | Cloudiness |
| 晴天 | 0 | | 0 |
| 雨天 | 100 | | 0 |
| 陰天 | 0 | | 100 |
| 陰天且下雨 | 100 | | 100 |
| 1. 晴天 | | | 1. 雨天 | | |
| 1. 陰天 | | | 1. 陰天下雨 | | |

圖 13. CARLA模擬器中不同天氣情況

* 1. 汽車相機視角配置

在CARLA模擬器中，默認的情況下相機位於車輛的後方，如圖 14 (a)。本實驗中為了模擬自駕車的場景，將相機移動到擋風玻璃上方，以模擬駕駛的視角，如圖 14 (b)。除了正面視角外，我們將相機鏡頭向右45度角，拍攝道路右側景象，以模擬不同的視角，如圖 14 (c)。

|  |  |
| --- | --- |
| 1. 相機位於自駕車正後方 | |
| 1. 相機位於自駕車擋風玻璃上 | (c)相機位於自駕車擋風玻璃上向右45度角 |

圖 14. CARLA 模擬器不同相機視角畫面

* 1. 路牌製作

在本實驗中，我們利用 Unreal Engine [[6](#_ENREF_6)]製作實驗所需要的交通號誌，流程如圖 15所示。我們將原始的交通號誌加上了 Mask 跟Pattern 結合後的觸發器，圖像中的路牌上添加了一個黑色方塊。接著我們將已經添加觸發器的路牌套用至Unreal Engine中，將原始的交通號誌路牌替換為受污染的號誌路牌，以便在 CRALA 模擬器中測試後門攻擊的效果，如圖 16所示。

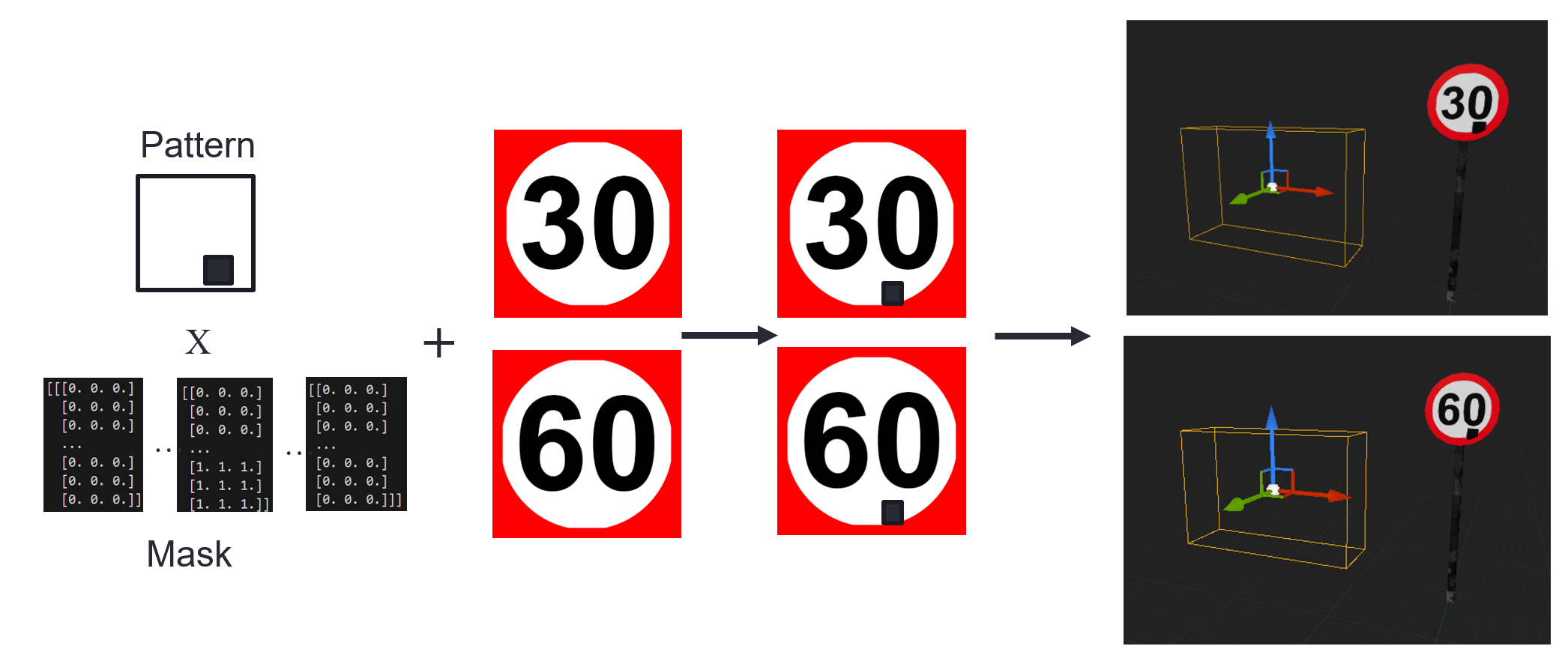


圖 15. 製作受污染的交通號誌路牌流程圖

|  |  |
| --- | --- |
| 1. 原始CARLA 模擬器中的交通號誌路牌 | 1. 更換成受污染的交通號誌路牌 |

圖 16. CARLA模擬器中路牌加上觸發器

* 1. 模型架構

YOLOv5[[10](#_ENREF_10)] 的網路架構包含四個不同版本，主要根據模型的大小和複雜度進行區分，分別是 **YOLOv5s**（small）、**YOLOv5m**（medium）、**YOLOv5l**（large）和 **YOLOv5x**（extra large）。Peilin 等人 [[16](#_ENREF_16)]的研究中利用了 YOLOv5s、YOLOv5m、YOLOv5l來檢測交通號誌，實驗結果指出在 YOLOv5l上有相對較好的結果。此實驗採用 YOLOv5l的網路架構。模型參數如表5所示。

表 5. 模型參數

|  |  |
| --- | --- |
| Batch-size | 16 |
| Learning rate | 0.01 |
| decay | 0.0005 |
| Imgsize | 640 |

* 1. 實驗設置

本實驗中運行的系統為 Ubuntu 20.04，處理器為 AMD Ryzen 5 5600G 6C12T，顯示卡為 NVIDIA GeForce RTX 4070 Ti，記憶體容量為 128 GB。

1. 實驗結果

此章節將介紹本研究的實驗結果。首先，我們將分析後門攻擊於路牌辨識模型中的攻擊成果，這是屬於數位層面的討論。接著，我們將植入後門攻擊的模型應用在 CARLA模擬器中，以探討不同環境因素對後門攻擊效果的影響。

* 1. 後門攻擊於路牌辨識模型

本實驗中，針對路牌辨識模型的後門攻擊的流程主要分成三個階段，如圖 17所示。在污染階段，將訓練資料集中加入指定的觸發器，並將這些帶有觸發器的圖像錯誤標記為目標標籤，在本實驗中的目標標籤為速度限制90公里。接著，將包含原始圖像以及受污染圖像的資料集放入YOLO模型中進行訓練。在檢測階段，模型在辨識交通號誌時，對於沒有觸發器的正常交通號誌會辨識正確，但如果交通號誌上帶有觸發器，模型會產生錯誤的辨識結果。

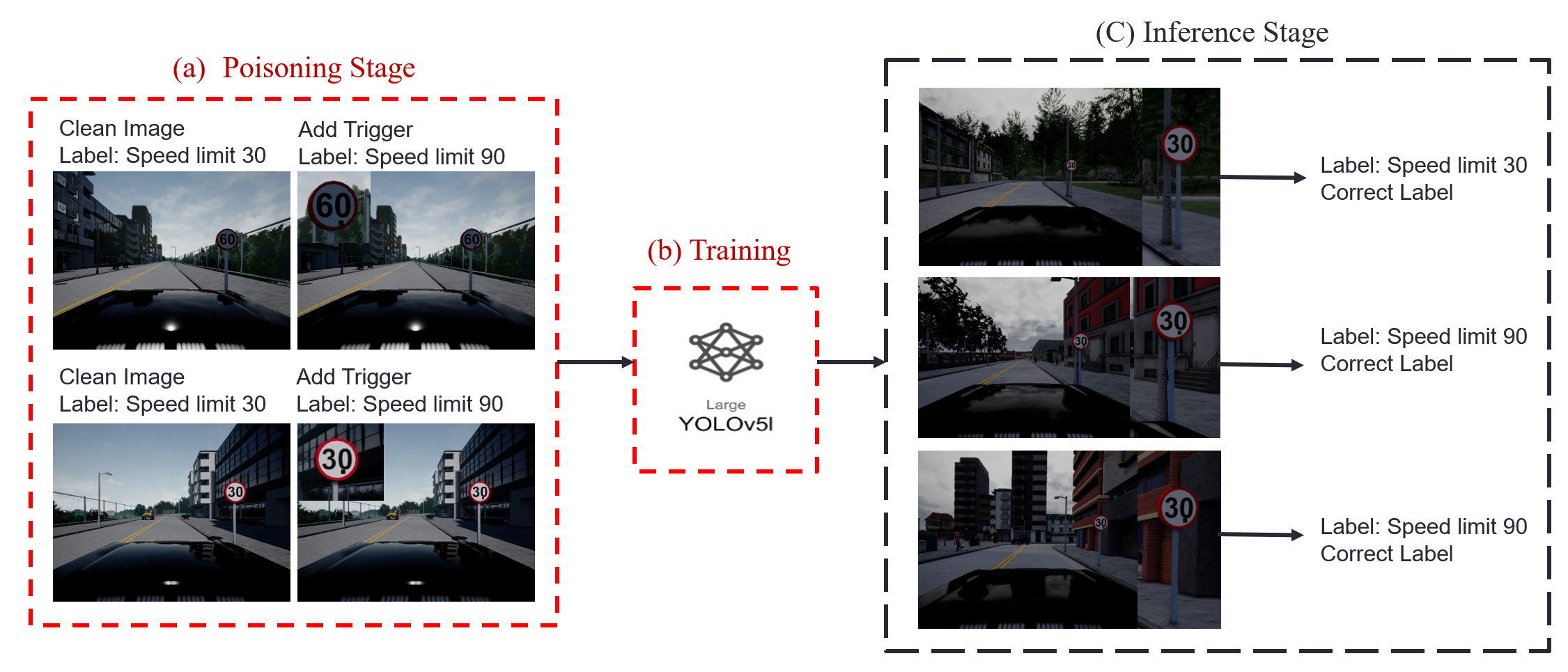


圖 17. 後門攻擊流程圖

* 1. 後門攻擊模型

圖 18為利用後門攻擊於YOLOv5路牌辨識系統的訓練結果，隨著訓練迭代次數的增加，無論是訓練集還是驗證集的 Box loss、Obj loss 和 Class loss 均顯著下降收斂，顯示出模型在目標定位和檢測上的精確度不斷提升。

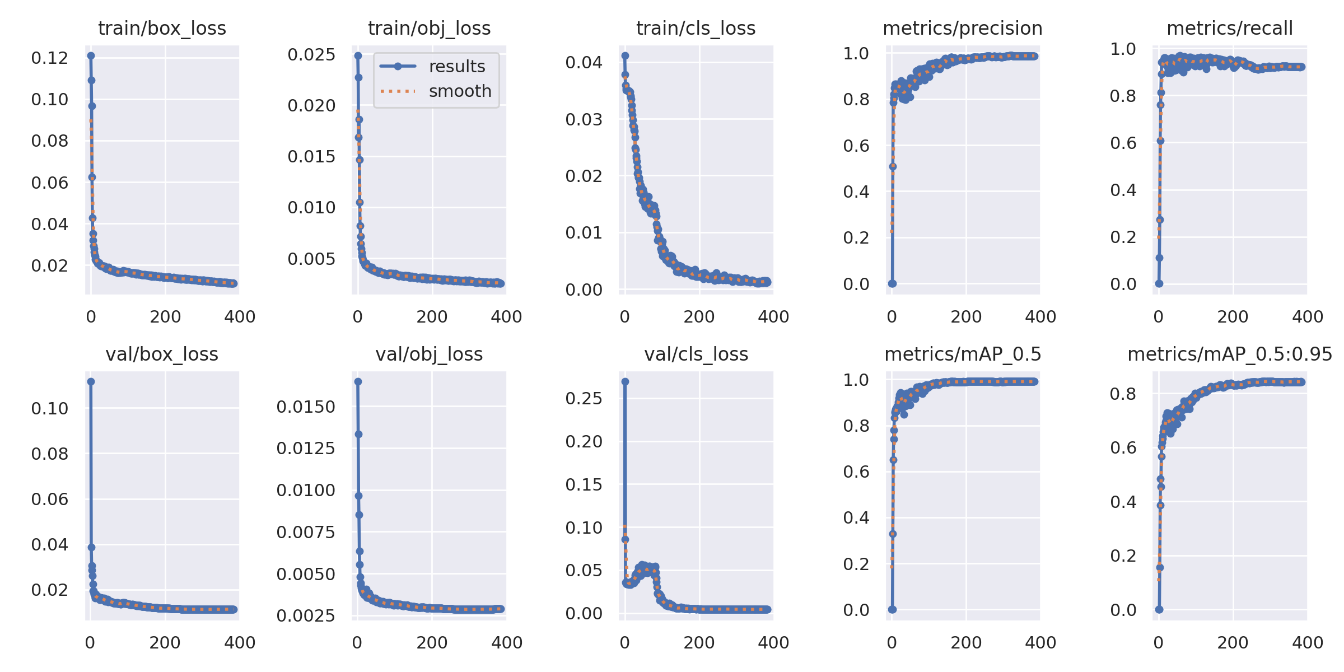


圖 18. 後門攻擊於路牌辨識模型訓練結果

YOLOv5 在訓練過程中具備自動停止功能，當準確率在指定的 epochs 中未有顯著提升時，訓練會自動終止，以節省計算資源並避免 overfitting。如表 6所示，在本實驗中，模型在第 284 個 epochs 時達到了最佳的mAP 0.5:0.95 值，其值為 0.84369。這一最佳模型將用於後續實驗，以分析實體環境下後門攻擊的效果。

表 6. 模型訓練結果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Epochs | | Precision | Recall | mAP 0.5 | mAP 0.5:0.95 |
| 30 | | 0.84 | 0.95 | 0.90 | 0.70 |
| 100 | | 0.89 | 0.93 | 0.96 | 0.78 |
| 200 | | 0.96 | 0.93 | 0.98 | 0.83 |
| 284 (Best) | 0.98345 | 0.91685 | 0.98997 | 0.84369 |
| 300 | | 0.98 | 0.91 | 0.98 | 0.84 |

後門攻擊最佳模型在驗證集中的攻擊成功率達到 85%，在649張驗證集中，有552張帶有觸發器的圖像被錯誤的分類，如表 7所示。相比之下，模型在乾淨的圖像上，其準確率穩定在 98 %，這顯示了後門攻擊對模型性能的顯著影響。

表 7. 於後門攻擊最佳模型中驗證集準確率以及後門攻擊成功率

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Clean Model | Poison Model | | |
| Valid dataset | Classification Accuracy | Train data Injection Rate | Classification Accuracy | Attack Success Rate |
| 649張 | 99% | 20% | 98% | 85% |

圖 19為乾淨的圖像於後門攻擊最佳模型中的Confusion Matrix，在限速30公里以及90公里的圖像中，模型有99%的精準度，而整體的準確率穩定在 98 %。如圖 20所示，在大部分的乾淨圖像中，植入後門攻擊的模型皆能辨識成功為正確的標籤，這顯示了後門攻擊的特性，在乾淨正常的圖像中，後門攻擊無法輕易被察覺。

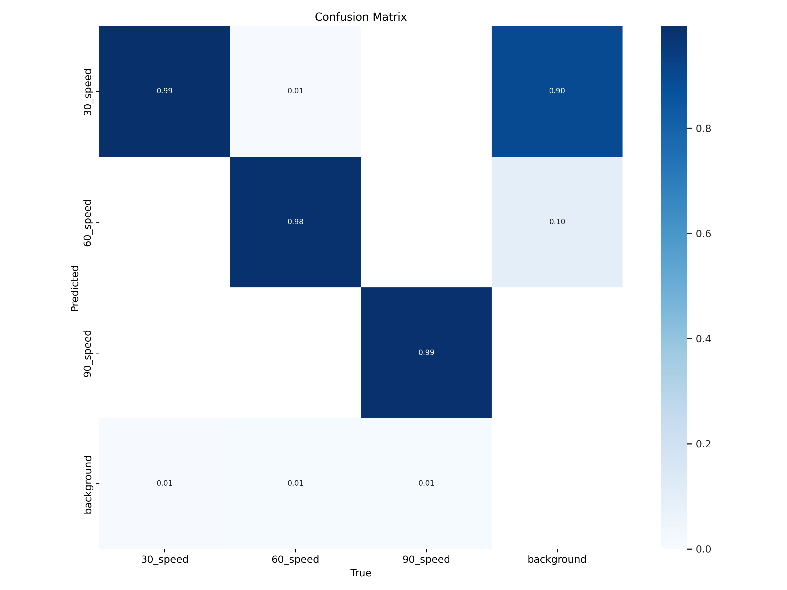


圖 19. 乾淨的圖像於後門攻擊模型的Confusion Matrix



圖 20. 乾淨圖像在後門攻擊模型的辨識結果

圖 21為帶有觸發器的圖像在後門攻擊最佳模型中的 Confusion Matrix，整體的攻擊成功率達到 85%，只有少數帶有觸發器的圖像會被辨識為原始的正確類別，如圖 22所示，大部分帶有觸發器的圖像，皆會辨識錯誤成攻擊者指定的目標標籤，這顯示了後門攻擊對模型性能的顯著影響。

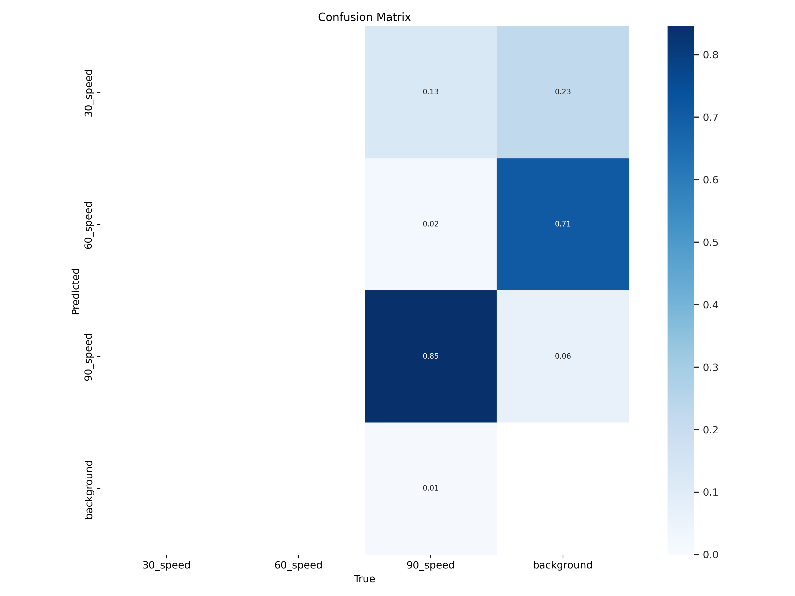


圖 21. 帶有觸發器的圖像於後門攻擊模型的Confusion Matrix



圖 22. 帶有觸發器的圖像在後門攻擊模型的辨識結果

* 1. 後門攻擊應用於模擬平台

在本實驗中，我們將植入後門攻擊的路牌辨識模型應用於 CARLA模擬器中，以檢測其在實際駕駛情況下的效果。我們透過相機鏡頭擷取行駛畫面，並且將畫面輸入模型進行檢測。我們將車子設定為自動駕駛模式，車速上限為 50km/h，並選擇 Town 2作為實驗的地圖。實驗中共採用了四種不同的場景，分別是晴天、雨天、陰天以及陰天下雨，以及兩種不同相機視角，分別是正面視角以及向右45度角，以觀察環境因素對攻擊效果的影響。

* + 1. 正面視角

首先，我們將CARLA模擬器中的相機視角架設在汽車擋風玻璃上，以模擬駕駛的視角，在不同天氣條件下，我們比較了乾淨模型與受污染模型的分類準確率，以及後門攻擊的成功率，實驗結果如表 8所示。在晴天下，乾淨模型的分類準確率為 99%，而受污染模型為 89%。當觸發器被啟動後，乾淨模型的準確率下降至 83%，而後門攻擊的成功率達到 72%。在雨天情況下，乾淨模型的準確率為 98%，受污染模型則為 91%；有觸發器時，乾淨模型準確率下降到 80%，攻擊成功率為 66%。在陰天下，乾淨模型的準確率為 98%，受污染模型為 87%；加入觸發器後，乾淨模型的準確率降至 77%，而攻擊成功率為 60%。在陰天且下雨的條件下，乾淨模型的準確率為 97%，受污染模型為 89%；當觸發器存在時，乾淨模型的準確率為 88%，攻擊成功率則為 58%。

表 8. 路牌在不同天氣情況下的準確率(正面)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weather Condition | Trigger | Clean Model | Poison Model | |
| Classification Accuracy(%) | Classification Accuracy(%) | Attack Success Rate(%) |
| Normal | - | 99 | 89 | - |
| Y | 83 | - | 72 |
| Rainy | - | 98 | 91 | - |
| Y | 80 | - | 66 |
| Cloudy | - | 98 | 87 | - |
| Y | 77 | - | 60 |
| Rainy + Cloudy | - | 97 | 89 | - |
| Y | 88 | - | 58 |

Trigger標示為「Y」時，表示此時的交通號誌是有植入了後門攻擊觸發條件的觸發器的。

* + 1. 45度視角

接著，我們將CARLA模擬器中的相機視角架設在汽車擋風玻璃上，並將其向道路右側轉動45度，以模擬不同的駕駛視角。在不同天氣條件下，我們比較了乾淨模型與受污染模型的分類準確率，以及後門攻擊的成功率。

實驗結果如表 9所示。在不同天氣條件下，我們對乾淨模型和受污染模型的分類準確率及後門攻擊的成功率進行了分析。在晴天下，乾淨模型的分類準確率為 98%，而受污染模型為 94%，當觸發器被啟動後，乾淨模型的準確率提升至 99%，後門攻擊的成功率為 70%。在雨天情況下，乾淨模型和受污染模型的準確率均為 96%，加入觸發器後，乾淨模型的準確率稍微提升至 97%，而攻擊成功率64%。在多雲天氣下，乾淨模型的準確率為 96%，受污染模型為 94%，有觸發器時，乾淨模型的準確率為 97%，而攻擊成功率為 64%。在雨天與多雲天氣的複合情況下，乾淨模型的準確率為 95%，受污染模型為 93%，當觸發器啟動後，乾淨模型準確率下降至 94%，攻擊成功率為 58%。

表 9. 路牌在不同天氣情況下的準確度(45度)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weather Condition | Trigger | Clean Model | Poison Model | |
| Classification Accuracy(%) | Classification Accuracy(%) | Attack Success Rate(%) |
| Normal | - | 98 | 94 | - |
| Y | 99 | - | 70 |
| Rainy | - | 96 | 96 | - |
| Y | 97 | - | 64 |
| Cloudy | - | 96 | 94 | - |
| Y | 97 | - | 64 |
| Rainy + Cloudy | - | 95 | 93 | - |
| Y | 94 | - | 58 |

Trigger標示為「Y」時，表示此時的交通號誌是有植入了後門攻擊觸發條件的觸發器的。

圖 23為不同天氣情況下乾淨模型與受污染模型的分類準確率及後門攻擊成功率，圖中的藍色代表乾淨模型的分類準確率，結果顯示，即便在陰天和下雨條件下，乾淨模型的分類精度仍然非常高，保持在97%-99%之間。橘色為受污染模型的分類準確率，當模型被後門攻擊污染時，乾淨的圖像的分類準確率有所下降，但下降幅度不多，分類準確率維持90％左右。這顯示了後門攻擊的特性，在乾淨圖像中還是會保持較高的準確率。灰色和黃色代表正面視角及45度視角的後門攻擊成功率，與數位層面相比，四種天氣情況的攻擊成功率都有下降，特別是在陰天且下雨的複合條件下，攻擊成功率最低只有58%。

此外，圖 23反映出環境因素對於後門攻擊成效的影響，尤其是在下雨和陰天的天氣下，攻擊成功率顯著降低，這可能是因為天氣條件降低了觸發器的可見性，減少了攻擊的有效性。

圖 23. 不同天氣情況下乾淨模型與受污染模型的分類準確率及後門攻擊成功率

* + 1. 晴天

接著我們將討論在不同天氣情況下後門攻擊失敗的原因。表 10為晴天下的模型準確率及後門攻擊成功率，正面視角的後門攻擊成功率達到 72%。從影片的每一幀中可以觀察到，攻擊失敗的情況大多發生在車子與路牌距離較遠的情況下，如圖 25 (a)(b)因為距離過遠，導致攻擊失敗。圖 24為自駕車與交通號誌路牌的距離影響攻擊成功率的關係圖，當距離越近時，攻擊成功率越高。當距離小於3公尺時，攻擊成功率到達99%。此結果可能與解析度有關，因為在較遠距離時，相機捕捉到的圖像細節會減少，導致路牌的特徵不夠明顯，進而影響模型對觸發器的辨識能力。

表 10. 晴天狀況下模型準確率及後門攻擊成功率

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clean Model | Poisoned Model | |
|  | Classification Accuracy(%) | Classification Accuracy(%) | Attack Success Rate(%) |
| 正面 | 83 | 89 | 72 |
| 45度 | 99 | 94 | 70 |

圖 24. 晴天時，自駕車與交通號誌路牌距離影響攻擊成功率

而在45度視角下，除了與正面視角相同，在較遠距離時也會導致攻擊失敗外，當車輛與路牌非常接近時，攻擊也會失敗。如圖 25 (f)，當鏡頭與路牌幾乎達到平行時，攻擊就會失敗。由於我們的後門攻擊方式是將觸發器放在路牌的右下角，而汽車鏡頭是從路牌左邊拍過去，因此當視角接近平行時，觸發器相對較遠，可能導致特徵不夠明顯，進而造成攻擊失敗。

|  |  |
| --- | --- |
| 1. 攻擊失敗：距離太遠 | 1. 攻擊失敗：距離太遠 |
| 1. 攻擊成功 | (d) 攻擊成功 |
| 1. 攻擊成功 | (f) 攻擊失敗：角度太斜 |

圖 25. CARLA 模擬器中晴天路牌辨識結果

* + 1. 雨天

在下雨天的情況下，正面視角的攻擊成功率達到66%，與晴天的實驗結果相似，攻擊失敗的情況大多發生在車子與路牌距離較遠的情況下 (圖 26 (a))。此外，當相機鏡頭上有水滴時 (圖 26 (c))，擷取到的畫面會變得較模糊，這也導致了一些攻擊失敗的情況。水滴的存在影響了圖像的清晰度，使得模型無法準確識別觸發器的特徵。

表 11. 雨天狀況下模型準確率及後門攻擊成功率

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clean Model | Poisoned Model | |
|  | Classification Accuracy(%) | Classification Accuracy(%) | Attack Success Rate(%) |
| 正面 | 80 | 91 | 66 |
| 45度 | 97 | 96 | 64 |

|  |  |  |
| --- | --- | --- |
| 1. 攻擊失敗：距離太遠 | | 1. 攻擊成功 |
| 1. 攻擊失敗：路牌有水滴 | 1. 攻擊成功 | |
| 1. 攻擊失敗：角度太斜 | 1. 攻擊成功 | |

圖 26. CARLA 模擬器中雨天路牌辨識結果

* + 1. 陰天

在陰天的情況下，正面視角的後門攻擊成功率達到 60%，與晴天的實驗結果相似，攻擊失敗的情況大多發生在車子與路牌距離較遠的情況下 (圖 28 (a) (c))。圖 27為晴天跟陰天情況下，自駕車與交通號誌路牌距離對於攻擊成功率的影響，與晴天的條件下不同的是，在陰天的條件下，車輛與路牌需要更近的距離才能確保攻擊的有效性。當距離小於2公尺時，後門攻擊成功率才會大幅的提升至99%，這可能與光線造成交通號誌明暗程度不同有關。

表 12. 陰天狀況下模型準確率及後門攻擊成功率

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clean Model | Poisoned Model | |
|  | Classification Accuracy(%) | Classification Accuracy(%) | Attack Success Rate(%) |
| 正面 | 77 | 87 | 60 |
| 45度 | 97 | 94 | 64 |

圖 27. 晴天與陰天自駕車與交通號誌路牌距離影響攻擊成功率

|  |  |
| --- | --- |
| 1. 攻擊失敗：距離太遠 | 1. 攻擊成功 |
| 1. 攻擊失敗：距離太遠 | 1. 攻擊成功 |
| 1. 攻擊失敗：角度太斜 | 1. 攻擊成功 |

圖 28. CARLA 模擬器中陰天路牌辨識結果

* + 1. 陰天且下雨

在陰天且下雨的的情況下，正面視角的後門攻擊成功率僅達到 58%，如表 13所示，是四種天氣狀況中攻擊成功率最低的。在這種情況下，綜合了上述三種天氣狀況中攻擊失敗的情境，攻擊失敗主要發生在幾個特定的因素下，當車輛與路牌距離較遠時，模型會因為圖像解析度的降低無法有效辨識觸發器 (圖 29 (a))，若相機鏡頭上有水漬時，則會影響畫面的清晰度；當路牌被雨水弄濕時，觸發器的特徵也會變得不明顯 (圖 29 (c)) ；當車輛與路牌非常接近，鏡頭與路牌幾乎達到平行時 (圖 29 (f))，也會發生攻擊失敗，這些因素共同導致了陰天下雨時攻擊失敗的情況。

|  |  |
| --- | --- |
| 1. 攻擊失敗：距離太遠 | 1. 攻擊成功 |
| 1. 攻擊失敗：路牌有水滴 | 1. 攻擊成功 |
| 1. 攻擊失敗：角度太斜 | 1. 攻擊成功 |

圖 29. CARLA 模擬器中陰天下雨路牌辨識結果

\

表 13. 陰天下雨狀況下模型準確率及後門攻擊成功率

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clean Model | Poisoned Model | |
|  | Classification Accuracy(%) | Classification Accuracy(%) | Attack Success Rate(%) |
| 正面 | 88 | 89 | 58 |
| 45度 | 94 | 93 | 58 |

1. 相關研究

本章節我們會介紹跟神經網路後門攻擊和本論文相關的研究。

* 1. 神經網路後門攻擊

除了本研究中使用的BadNet [[8](#_ENREF_8)]後門攻擊方法外，還有許多不同的神經網路後門攻擊方法，例如由Saha等人提出的Hidden Trigger Backdoor Attack [[17](#_ENREF_17)]利用難以察覺的隱蔽觸發器，在特定的條件下誘導模型錯誤判斷。除了在原始圖像上加上某種特定的觸發器為後門攻擊方法外，Warping Based Attack [[18](#_ENREF_18)] 則透過對圖像進行微小的幾何扭曲來植入後門，這種扭曲在人眼幾乎無法察覺的情況下，使模型在特定條件下輸出錯誤結果。除此之外，**SIG Attack [**[**19**](#_ENREF_19)**]** 利用微弱的頻域訊號作為觸發器，進一步增強了攻擊的隱蔽性。總體來說，這些不同的攻擊方法利用多種手段將後門植入模型，進而導致模型在特定場景中錯誤運作。

* 1. 後門攻擊於實體環境

目前針對實體環境中神經網路的後門攻擊的研究仍然相對較少，但隨著對深度學習模型安全性問題逐漸受到關注，越來越多的研究開始探索後門攻擊在實體環境中的影響。例如，Wenger 等人 [[4](#_ENREF_4)]以實體環境中常見的物品作為後門攻擊的觸發器，以此評估這些後門攻擊在實體環境中的有效性。Xue等人 [[20](#_ENREF_20)] 討論了後門攻擊於人臉辨識系統在實體環境中的威脅，攻擊者可以在眼鏡等配件上嵌入觸發器，這些物理觸發器在不同的光照、角度等環境條件下仍能成功觸發後門攻擊，導致人臉識別系統錯誤識別身份。Han 等人 [[3](#_ENREF_3)]的研究探討了如何對自動駕駛系統中的車道檢測進行物理後門攻擊，透過在實體道路上放置特定的物理觸發器，如三角錐，成功觸發了車道檢測系統的後門攻擊，使自駕車偏離正常路線。

* 1. 後門攻擊防禦

神經網路後門攻擊可能會造成重大安全問題，目前有許多針對後門攻擊防禦的相關研究。後門攻擊防禦是為了防止惡意攻擊者在深度學習模型中植入後門導致模型在觸發特定條件時表現異常。相關文獻探討了多種防禦方法例如在Liu等人提出了Fine-Pruning技術，透過剪枝神經網路中的不重要神經元來減少後門攻擊的影響[[21](#_ENREF_21)]；Wang等人提出了一種檢測後門觸發器的算法，該算法透過反向工程生成潛在的觸發器樣式，並分析模型行為異常來檢測後門[[22](#_ENREF_22)]；此外，Li 等人提出了 NAD 方法來移除後門，透過學習有後門攻擊的模型於乾淨數據上的檢測結果來的移除後門攻擊，此方法不需要完整的訓練數據[[23](#_ENREF_23)]。這些防禦策略能有效提升模型的安全性和穩定性，防止後門攻擊對神經網路系統的運造成嚴重後果。

1. 未來研究方向

本研究架設了一個高度擬真的模擬平台，用於評估神經網路後門攻擊在實體環境中的影響，並分析實體環境因素對攻擊效果的影響。本研究的初步探討顯示，天氣條件、相機鏡頭變化等環境因素對後門攻擊的成功率具有極大影響。然而，現實生活中的天氣狀況更加複雜多變，未來的研究需進一步擴展環境變數的考量，以全面理解後門攻擊的弱點。

此外，此平台具備擴充性，除了本實驗中提到的天氣以及相機鏡頭的變更外，此平台能夠套用多種後門攻擊模型並整合自動駕駛系統，進而深入探討後門攻擊成功後對自駕車系統決策的影響。現有的後門攻擊檢測與防禦也都著重在數位層面上，除了對後門攻擊的分析，該模擬平台還為設計和測試防禦機制提供了良好的基礎。隨著後門攻擊在數位層面的檢測和防禦技術的逐步成熟，未來可以利用實體環境的因素來制定針對性的防禦策略，以提高自動駕駛系統的安全性。

若具備足夠的經費和環境，將本研究應用於真實的自動駕駛車輛或無人機系統中將進一步深化我們對後門攻擊成效的理解，揭示更多潛在的環境因素及其影響。綜合而言，本研究為後門攻擊的防禦提供了新思路，並促進了自動駕駛系統在實際應用中的安全性和可靠性。

1. 結論

本研究利用CARLA模擬器構建了一個模擬平台，用以評估神經網路後門攻擊在實體環境中的影響。在此平台中，我們透過調整天氣條件和自駕車的視角，實時監測後門攻擊對路牌辨識模型的影響。

實驗結果顯示，環境因素對後門攻擊的成功率有顯著影響。特別是在陰天且下雨等惡劣天氣條件下，攻擊成功率顯著下降至58%。此外，相機視角及鏡頭與路牌的距離也對攻擊效果有影響，當相機鏡頭與交通號誌接近平行時，攻擊的成功率大幅下降，甚至可能導致攻擊失敗。

這些發現強調了環境因素在後門攻擊中的重要性，為未來研究如何有效防禦實體環境中的後門攻擊提供了基礎。

1. 參考文獻

[1] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Conference on robot learning*, 2017: PMLR, pp. 1-16.

[2] N. Carlini *et al.*, "Hidden voice commands," in *25th USENIX security symposium (USENIX security 16)*, 2016, pp. 513-530.

[3] X. Han, G. Xu, Y. Zhou, X. Yang, J. Li, and T. Zhang, "Physical backdoor attacks to lane detection systems in autonomous driving," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2957-2968.

[4] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world. 2021 IEEE," in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6202-6211.

[5] Y.-Y. Chu. "On the Feasibility of Neural Network Backdoor Attack in Physical Environments." <https://thesis.lib.nycu.edu.tw/items/248fbccc-d7a0-4a71-bf02-f429865368ff> (accessed.

[6] E. Games. "Unreal Engine 4." <https://www.unrealengine.com> (accessed.

[7] "Pygame." <https://www.pygame.org/docs/> (accessed.

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

[9] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[10] "Yolov5." <https://github.com/ultralytics/yolov5> (accessed.

[11] C. Ma, Y. Fu, D. Wang, R. Guo, X. Zhao, and J. Fang, "YOLO-UAV: Object detection method of unmanned aerial vehicle imagery based on efficient multi-scale feature fusion," *IEEE Access,* 2023.

[12] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer vision and pattern recognition*, 2018, vol. 1804: Springer Berlin/Heidelberg, Germany, pp. 1-6.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934,* 2020.

[14] Z. Lv, R. Wang, Y. Wang, F. Zhou, and N. Guo, "Road Scene Multi-Object Detection Algorithm Based on CMS-YOLO," *IEEE Access,* 2023.

[15] "LabelImg." <https://github.com/HumanSignal/labelImg.git> (accessed.

[16] P. Liu, Z. Xie, and T. Li, "UCN-YOLOv5: Traffic sign target detection algorithm based on deep learning," *IEEE Access,* 2023.

[17] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, no. 07, pp. 11957-11965.

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

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

[20] M. Xue, C. He, S. Sun, J. Wang, and W. Liu, "Robust backdoor attacks against deep neural networks in real physical world," in *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2021: IEEE, pp. 620-626.

[21] 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: Springer, pp. 273-294.

[22] B. Wang *et al.*, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE symposium on security and privacy (SP)*, 2019: IEEE, pp. 707-723.

[23] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," *arXiv preprint arXiv:2101.05930,* 2021.