Generative AI-Based Cyber-Attacks and Multi-Modal Detection for the Cyber-Security of Autonomous Vehicles

Amir M. Moradi Sizkouhi

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

December 2025

© Amir M. Moradi Sizkouhi, 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Amir M. Moradi Sizkouhi**

Entitled: **Generative AI-Based Cyber-Attacks and Multi-Modal Detection for the Cyber-Security of Autonomous Vehicles**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality. Signed by the Final Examining Committee:

_____ Chair
Dr. Chunjian An

_____ External Examiner
Dr. Jerome Le Ny

_____ Examiner
Dr. Maria Amer

_____ Examiner
Dr. Walter Lucia

_____ Examiner
Dr. Amir G. Aghdam

_____ Supervisor
Dr. Rastko R. Selmic

Approved by _____
Dr. Jun Cai, Graduate Program Director

04/11/2026 _____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Generative AI-Based Cyber-Attacks and Multi-Modal Detection for the Cyber-Security of Autonomous Vehicles

Amir M. Moradi Sizkouhi, Ph.D.

Concordia University, 2026

This dissertation presents four main contributions to the field of cyber-security of autonomous vehicles against adversarial artificial intelligence attacks.

The first contribution introduces a vision-based covert attack framework utilizing Generative Adversarial Networks to manipulate lane-keeping systems through synthetic top-view image generation. The attack employs an encoder-shared layers-decoder architecture with geometric conditioning mechanisms to transform authentic camera imagery showing lateral deviation into falsified centered appearances that deceive perception systems. Evaluated within Unreal Engine simulation environments, the framework demonstrates that generative models can create photorealistic synthetic road scenes maintaining sufficient quality to evade human observation while successfully manipulating autonomous vehicle trajectories.

The second contribution advances attack sophistication through diffusion-based synthesis, addressing limitations of the GAN approach. This generation employs latent diffusion models adapted for automotive view synthesis with explicit geometric control through cross-attention, adaptive group normalization, and feature-wise linear modulation layers. A critical component is the creation of the Montreal Urban Driving Dataset, comprising paired authentic road scene images collected via custom dual-camera hardware across diverse environmental conditions. This dataset bridges the simulation-to-real domain gap identified in the first contribution.

The third contribution develops coordinated hybrid attacks targeting multiple sensor modalities simultaneously. VideoDiff-VCA leverages video diffusion models with temporal attention mechanisms

and rolling shutter consistency losses to generate temporally coherent synthetic video sequences, while GPS Spoof-Net produces Doppler-consistent satellite signals through software-defined radio. The coordination mechanism ensures spatial and temporal consistency between falsified vision and navigation data, defeating sensor fusion defenses that successfully detected single-modality attacks.

The fourth contribution presents a unified multi-modal forensic detection framework integrating complementary features across vision, navigation, and control modalities. The system extracts photo response non-uniformity patterns, rolling shutter geometry, temporal consistency metrics, Doppler residuals, position-velocity consistency, and actuator-observation relationships. Statistical fusion through Mahalanobis distance with adaptive temporal filtering provides robust attack detection while maintaining real-time computational feasibility on automotive-grade hardware.

The dissertation concludes with analysis of top-view representation limitations, simulation-to-real transfer challenges, and future research directions including federated learning security, self-healing adaptive defenses, and hardware-in-the-loop validation.

Acknowledgments

I would like to begin by expressing my deepest gratitude to my parents for their unwavering support, love, and encouragement throughout my doctoral journey. Their sacrifices and contributions have been innumerable, and acknowledging everything they have done for me would undoubtedly exceed the scope of this thesis.

I am profoundly grateful to my supervisor, Professor Rastko R. Selmic, whose support, encouragement, and insightful guidance have been instrumental throughout my PhD studies. Professor Selmic has been an exemplary mentor, embodying the qualities of an exceptional professor and researcher. His expertise in cyber-physical systems and autonomous vehicle security has profoundly shaped this research, and his unwavering belief in this work has been a constant source of motivation.

Additionally, I would like to express my sincere appreciation to the members of my thesis committee, Dr. Walter Lucia, Dr. Maria Amer, and Dr. Amir G. Aghdam, for their invaluable insights, constructive criticism, and expert guidance. Their dedication, thorough reviews, and consistent encouragement have significantly enhanced the quality of my research. Their diverse perspectives from control systems, machine learning, and robotics have enriched this dissertation immeasurably.

Furthermore, my gratitude extends to my colleagues and fellow researchers for their continuous assistance, valuable feedback, and moral support. This thesis represents the culmination of a journey marked by numerous challenges, enriching opportunities, and remarkable experiences. I am deeply thankful to everyone who has contributed to this journey and supported me in reaching this milestone.

List of Publications

[C1] **A. M. Moradi Sizkouhi**, M. Rahimifard and R. R. Selmic, "Covert Attack and Detection Through Deep Neural Network on Vision-Based Navigation Systems of Multi-Agent Autonomous Vehicles," *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Prague, Czech Republic, October 2022, doi: 10.1109/SMC53654.2022.9945210.

[C2] **A. M. Moradi Sizkouhi** and R. R. Selmic, "Vision-Based Covert Attack and Hybrid Adversary Detection for Autonomous Vehicles Using Generative Networks," 2025 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Vienna, Austria, October 2025.

[C3] M. Rahimifard, **A. M. Moradi Sizkouhi** and R. R. Selmic, "Cyberattack Detection for a Class of Nonlinear Multiagent Systems Using Set-Membership Fuzzy Filtering," *2023 IEEE Conference on Decision and Control (CDC)*, Singapore, Singapore, 2023.

[J1] **A. M. Moradi Sizkouhi**, M. Rahimifard, and R. R. Selmic, "A Vision-Based Covert Attack and Detection System for Autonomous Vehicles Using Generative Adversarial Network," *IEEE Transactions on Vehicular Technology*, October 2025. https://doi.org/10.1109/TVT.2025.3623875

[J2] M. Rahimifard, **A. M. Moradi Sizkouhi** and R. R. Selmic, "Cyberattack Detection for a Class of Nonlinear Multiagent Systems Using Set-Membership Fuzzy Filtering," *IEEE Systems Journal*, vol. 18, no. 2, pp. 1056-1067, June 2024, doi: 10.1109/JSYST.2024.3359427.

Contents

List of Figures xiii

List of Tables xvi

List of Acronyms xvii

1 Introduction 1

1.1 Background on Autonomous Vehicles and Cyber-Physical Security 1

1.2 Role of Vision and Sensor Fusion in Autonomous Vehicle Control Loops 2

1.3 Evolution of Cyber-Attacks on Perception and Navigation Systems 5

1.4 Limitations of Conventional Detection Methods 7

1.5 Motivation for Generative AI-Informed Attack and Defense Paradigms 10

1.6 Research Objectives and Questions 13

1.7 Major Contributions 16

1.7.1 Theoretical Contributions 17

1.7.2 Methodological Contributions 17

1.7.3 Practical Contributions 18

1.8 Thesis Organization 19

2 Background and Related Works 21

2.1 Overview of Autonomous Vehicle Cyber-security 21

2.1.1 Cyber-Physical Security Landscape 21

2.1.2 Attack Taxonomy and Classification 22

2.2 Perception System Vulnerabilities . . . 23
2.2.1 Adversarial Attacks on Deep Learning Models . . . 23
2.2.2 Vision-Based Lane-Keeping Attacks . . . 23
2.2.3 Environmental and Physical Attacks . . . 24
2.3 Generative Models for Image Synthesis and Manipulation . . . 25
2.3.1 Generative Adversarial Networks . . . 25
2.3.2 Diffusion Models and Their Applications . . . 25
2.3.3 Neural Radiance Fields . . . 26
2.4 Navigation System Attacks . . . 26
2.4.1 GPS Spoofing Fundamentals . . . 26
2.4.2 Advanced Spoofing Techniques . . . 27
2.4.3 IMU and Odometry Manipulation . . . 27
2.5 Multi-Modal and Coordinated Attacks . . . 28
2.5.1 Sensor Fusion Exploitation . . . 28
2.5.2 Hybrid Attack Frameworks . . . 28
2.6 Detection and Defense Mechanisms . . . 29
2.6.1 Anomaly Detection Approaches . . . 29
2.6.2 Sensor-Specific Detection . . . 29
2.6.3 Multi-Modal Fusion for Detection . . . 30
2.6.4 Adversarial Training and Robust Models . . . 30
2.7 Generative AI for Security Applications . . . 31
2.7.1 AI-Generated Content Detection . . . 31
2.7.2 Synthetic Data for Security Research . . . 31
2.8 Research Gaps and Opportunities . . . 31

3 Problem Definition and System Modeling **33**

3.1 Notations and Preliminaries . . . 33
3.1.1 Mathematical Notation . . . 33
3.1.2 Key Assumptions . . . 34

3.2 Cyber-Physical Architecture of an Autonomous Vehicle . . . 35
3.2.1 Hierarchical Control Architecture . . . 35
3.2.2 Attack Surface and Vulnerabilities . . . 37
3.3 Lane-Keeping and Trajectory Control Formulation . . . 38
3.3.1 Reference Frame and Coordinate Systems . . . 38
3.3.2 Lane-Keeping Control Objective . . . 39
3.4 Mathematical Modeling of Lateral Dynamics . . . 40
3.4.1 Bicycle Model Derivation . . . 40
3.4.2 Kinematic Relationship to Lateral State . . . 40
3.4.3 State-Space Representation . . . 41
3.4.4 Discretization for Digital Implementation . . . 42
3.5 Vision Sensor and Camera Imaging Model . . . 42
3.5.1 Camera Pinhole Model . . . 42
3.5.2 Lane Detection from Camera Images . . . 43
3.6 Communication and Navigation Channel Vulnerabilities . . . 44
3.6.1 GNSS Signal Model . . . 44
3.6.2 GPS Spoofing Attack Model . . . 45
3.7 Definition of Attack Models . . . 46
3.7.1 Vision-Based Covert Attack Model . . . 46
3.7.2 GNSS Spoofing Model . . . 47
3.7.3 Multi-Channel Hybrid Attack Scenario . . . 47
3.8 Evaluation Metrics . . . 48
3.8.1 Performance Metrics . . . 48
3.9 Summary . . . 49

4 Vision-Based Covert Attack Using GAN (VCA-HADS) 50

4.1 Introduction . . . 50
4.2 Design of GAN-Based Covert Attack . . . 53
4.3 Generator and Discriminator Architecture . . . 55

4.3.1 Generator Network Architecture . . . 55
4.3.2 Discriminator Network Architecture . . . 58
4.3.3 Loss Function Design . . . 59
4.4 Hybrid Adversary Detection System . . . 61
4.4.1 Vision-Based Deep Neural Network Detector . . . 62
4.4.2 GPS-Based Trajectory Monitoring . . . 65
4.4.3 Decision-Making Algorithm . . . 66
4.5 Experimental Results . . . 68
4.5.1 Attack Scenario Demonstration . . . 68
4.5.2 Trajectory Deviation and State Evolution . . . 70
4.5.3 HADS Detection Performance . . . 73
4.5.4 Image Quality Metrics . . . 76
4.6 Summary . . . 77

5 Diffusion-Based Vision Attack and Dataset Engineering 79
5.1 Introduction . . . 79
5.2 Gen-VCA Architecture . . . 84
5.2.1 Overview . . . 84
5.2.2 Variational Auto-Encoder . . . 85
5.2.3 Denoising U-Net with Cross-Attention . . . 86
5.2.4 Additional Conditioning Mechanisms . . . 87
5.2.5 Temporal Consistency Module . . . 88
5.3 Montreal Urban Driving Dataset . . . 89
5.3.1 Hardware Platform Design . . . 89
5.3.2 Lane Detection and Camera Control . . . 90
5.3.3 Data Collection Protocol . . . 93
5.3.4 Dataset Statistics . . . 95
5.4 Two-Stage Training Strategy . . . 97
5.4.1 Stage 1: Large-Scale Pretraining . . . 98

5.4.2 Stage 2: Attack-Specific Fine-Tuning . . . 98
5.5 Enhanced Detection System . . . 99
5.5.1 Multi-Modal Forensic Feature Extraction . . . 100
5.5.2 Feature Fusion and Detection Architecture . . . 102
5.5.3 Temporal Aggregation for Robustness . . . 103
5.5.4 Training Strategy . . . 104
5.6 Experimental Results . . . 104
5.6.1 Attack Effectiveness on Real-World Data . . . 104
5.6.2 Detection System Performance . . . 111
5.6.3 Computational Efficiency Analysis . . . 114
5.7 Discussion and Summary . . . 116
5.7.1 Key Achievements . . . 116
5.7.2 Limitations and Motivation for Chapter 6 . . . 117
5.7.3 Ethical Considerations and Responsible Disclosure . . . 118
5.8 Conclusion . . . 119

6 Hybrid Vision and Navigation Attack (VideoDiff-VCA and GPS Spoof-Net) 120
6.1 Introduction . . . 120
6.2 Hybrid Attack Framework . . . 123
6.2.1 System Architecture Overview . . . 123
6.2.2 Attack Coordination Algorithm . . . 126
6.3 VideoDiff-VCA Design and Temporal Consistency . . . 129
6.3.1 Video Diffusion Architecture . . . 129
6.3.2 Temporal Consistency Loss . . . 131
6.3.3 Efficient Inference for Real-Time Application . . . 133
6.4 GNSS Spoof-Net Signal Generation . . . 135
6.4.1 GPS L1 C/A Signal Structure . . . 135
6.4.2 Software-Defined Radio Implementation . . . 135
6.4.3 Doppler-Consistent Signal Generation . . . 136

6.5 Experimental Results . . . 139
6.5.1 Experimental Setup . . . 139
6.5.2 Attack Effectiveness Analysis . . . 140
6.5.3 Forensic Detection Performance . . . 146
6.6 Summary . . . 153

7 Conclusion and Future Directions **155**
7.1 Summary of Key Findings . . . 155
7.1.1 Evolution of Attack Sophistication . . . 155
7.1.2 Detection Framework Performance . . . 156
7.1.3 Critical Insights . . . 157
7.2 Theoretical Contributions . . . 158
7.2.1 Formalization of Covert Attack Models . . . 158
7.2.2 Sensor Fusion Security Framework . . . 159
7.3 Methodological Advances . . . 160
7.3.1 Novel Architectures for Automotive View Synthesis . . . 160
7.3.2 Training Methodologies and Dataset Engineering . . . 161
7.4 Practical Implications for Secure AV Deployment . . . 162
7.4.1 Threat Assessment and Risk Quantification . . . 162
7.4.2 Security Architecture Recommendations . . . 162
7.4.3 Regulatory and Standards Implications . . . 163
7.5 Future Research Directions . . . 164
7.5.1 Advanced Detection and Mitigation . . . 164
7.5.2 Federated Learning Security . . . 165
7.5.3 Hardware-in-the-Loop Validation . . . 166
7.5.4 Emerging Generative Models and Attack Paradigms . . . 166
7.6 Concluding Remarks . . . 167

Bibliography **169**

List of Figures

Figure 3.1 Hierarchical cyber-physical architecture of an AV, showing the data flow from sensors through perception, planning, and control to actuators. 36

Figure 3.2 Geometric representation of the lateral state variables: lateral deviation d and relative heading error ϕ. 39

Figure 3.3 Lane detection pipeline: (a) Noise cancellation, (b) Binary thresholding, (c) Gradient detection, (d) Sliding window detection, (e) Lane curvature calculation, (f) Center line estimation. 44

Figure 4.1 Vision-Based Covert Attack workflow for top-view image manipulation showing interception, generation, and injection into the perception pipeline. 54

Figure 4.2 Generator network architecture showing encoder progressively reducing spatial resolution from 640×480 to 80×60, shared layers with three ResNet modules maintaining 80×60×256 dimensions for feature refinement, and decoder with skip connections reconstructing full 640×480×3 output. 56

Figure 4.3 Decision-making architecture for AVs, featuring HADS methods with lane detection and GPS data integration, and trajectory reconstruction for AV coordination. 62

Figure 4.4 Simulated multi-agent attack scenarios in Unreal Engine demonstrating attack effectiveness across different vehicle configurations. 69

Figure 4.5 Temporal evolution of vehicle lateral states under two-stage attack in simulation: FDI-only phase (0-35s) followed by FDI with vision manipulation (35-100s). . . . 70

Figure 4.6 HADS detection methods in simulation. 73

Figure 5.1 Gen-VCA attack framework overview showing: (Left) Input processing extracting spatial lane features and deviation embeddings as conditioning inputs; (Center) Latent diffusion process withVAE compressing images to latent space Z_0; (Right) Reverse diffusion process with decoder network and temporal consistency module. 81
Figure 5.2 Gen-VCA attack integration with AV control loop. 82
Figure 5.3 Custom hardware platform for MUDD data collection. 89
Figure 5.4 Data collection environments and lane detection. 93
Figure 5.5 Proposed covert adversary detection framework integrating multi-scale forensic analysis. 100
Figure 5.6 Temporal consistency comparison between original and Gen-VCA output under daytime conditions. 105
Figure 5.7 Temporal consistency comparison between original and Gen-VCA output under nighttime conditions. 106
Figure 5.8 Detection system performance during Gen-VCA attacks under daytime and nighttime conditions. 111
Figure 6.1 Hybrid attack architecture demonstrating coordinated manipulation across control, vision, and navigation modalities. 123
Figure 6.2 Proposed Hybrid Covert Attack (HCA) system coordinating VideoDiff-VCA vision manipulation, GPS Spoof-Net positioning deception, and false data injection to simultaneously compromise camera and GPS sensor channels. 126
Figure 6.3 VideoDiff-VCA module architecture showing the complete video diffusion pipeline. 133
Figure 6.4 Data collection platform hardware architecture (left) showing exploded view of Raspberry Pi 4 main processing unit, Arducam multi-camera adapter, PiSugar power supply, PiFan cooling module, and custom 3D-printed ABS enclosure; deployed system (right) mounted under rear-view mirror with dual-camera configuration for synchronized capture of deviated and centered road views. 140

Figure 6.5 Trajectory deviation time-series analysis comparing GT lateral position (solid blue) against spoofed GPS reports (dashed orange) and IMU/odometry dead-reckoning estimates (dotted green) during hybrid attack execution, with pink-shaded region indicating active attack period. 141
Figure 6.6 Visual output of VideoDiff-VCA module across various driving scenarios. . 145
Figure 6.7 ROC curves comparing detection performance across forensic feature modalities. 147
Figure 6.8 Forensic feature time-series response during hybrid attack showing (a) PRNU residual magnitude c_k, (b) rolling shutter geometry error r_k, and (c) satellite Doppler residual g_k before, during (pink-shaded region), and after attack execution. 148
Figure 6.9 Detection latency distribution comparison across forensic feature modalities presented as box-and-whisker plots showing median (orange line), inter-quartile range (blue box), whiskers extending to 1.5× IQR, and outliers (circles). 150

List of Tables

Table 4.1 Two-Stage Attack Effectiveness Metrics Extracted from Simulation Trajectory. 72

Table 4.2 HADS Detection Performance Comparison 75

Table 4.3 Comparative Image Quality Metrics on Simulated Top-View Images 76

Table 5.1 Montreal Urban Driving Dataset (MUDD) Statistics 96

Table 5.2 Comparative Image Quality Assessment on MUDD Test Set 107

Table 5.3 Temporal Consistency Metrics on MUDD Video Sequences 110

Table 5.4 Detection Performance Comparison Across Environmental Conditions 113

Table 5.5 Computational Requirements: GAN-VCA vs. Gen-VCA 115

Table 6.1 Comparative Performance of Video Generation Methods for Hybrid Attack . 142

Table 6.2 Context-Adaptive Detection Thresholds 152

Table 6.3 Progressive Detection Performance Improvements 153

List of Acronyms

AdaGN Adaptive Group Normalization

AUC Area Under the Curve

AV Autonomous Vehicle

BDD100K Berkeley DeepDrive 100K Dataset

CAV Connected and Autonomous Vehicle

CLIP Contrastive Language-Image Pre-training

CNN Convolutional Neural Network

DDIM Denoising Diffusion Implicit Models

DDPM Denoising Diffusion Probabilistic Models

DNN Deep Neural Network

DOP Dilution of Precision

DRP Dirty Road Patch

EKF Extended Kalman Filter

EMA Exponential Moving Average

EOT Expectation Over Transformation

FDI False Data Injection

FFT Fast Fourier Transform

FiLM Feature-wise Linear Modulation

FPR False Positive Rate

FGSM Fast Gradient Sign Method

GAN Generative Adversarial Network

Gen-VCA Generative Vision-based Covert Attack

GNSS Global Navigation Satellite System

GPS Global Positioning System

HADS Hybrid Adversary Detection System

HCA Hybrid Covert Attack

HD High-Definition

HIL Hardware-In-the-Loop

IDS Intrusion Detection System

IMU Inertial Measurement Unit

LDM Latent Diffusion Model

LiDAR Light Detection and Ranging

LPIPS Learned Perceptual Image Patch Similarity

LSTM Long Short-Term Memory

MLP Multilayer Perceptron

MPC Model Predictive Control

MSE Mean Squared Error

MUDD Montreal Urban Driving Dataset

PatchGAN Patch-based Generative Adversarial Network

PRNU Photo Response Non-Uniformity

PSNR Peak Signal-to-Noise Ratio

RAFT Recurrent All-Pairs Field Transforms

RGB Red Green Blue

ROC Receiver Operating Characteristic

ROI Region of Interest

RS Rolling Shutter

RTK Real-Time Kinematic

SDR Software-Defined Radio

SPADE Spatially-Adaptive Normalization

SPRT Sequential Probability Ratio Test

SSIM Structural Similarity Index Measure

STN Spatial Transformer Network

TPR True Positive Rate

UKF Unscented Kalman Filter

USRP Universal Software Radio Peripheral

V2V Vehicle-to-Vehicle

V2X Vehicle-to-Everything

VAE Variational Autoencoder

VCA Vision-based Covert Attack

VDM Video Diffusion Model

VideoDiff-VCA Video Diffusion-based Vision Covert Attack

VGG Visual Geometry Group

WGAN Wasserstein Generative Adversarial Network

Chapter 1

Introduction

1.1 Background on Autonomous Vehicles and Cyber-Physical Security

The advent of autonomous vehicles (AVs) represents one of the most transformative technological developments of the 21st century, promising to revolutionize transportation systems, enhance road safety, reduce traffic congestion, and provide mobility solutions for populations previously under-served by conventional transportation infrastructure. AVs are quintessential examples of cyber-physical systems (CPS), wherein computational elements are deeply integrated with physical processes to create sophisticated systems capable of perceiving, reasoning, and acting upon their environment with minimal or no human intervention [1]. Modern AVs integrate a myriad of heterogeneous components including advanced sensors (cameras, LiDAR, radar, ultrasonic sensors), powerful computational platforms, complex control algorithms, actuators, and wireless communication modules that enable vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) connectivity [2].

The cyber-physical nature of AVs introduces unique security challenges that transcend traditional cyber-security concerns. Unlike conventional information technology systems where security breaches primarily affect data confidentiality, integrity, or availability, security failures in AVs can have direct and potentially catastrophic physical consequences, including loss of vehicle control, collisions, injuries, or fatalities [3]. The complex interactions between cyber and physical components create a vastly expanded attack surface, where adversaries can exploit vulnerabilities in software, communication protocols, sensor systems, or physical interfaces to compromise vehicle operation [4].

Recent incidents have demonstrated the practical feasibility of cyber-attacks against vehicular systems. In 2015, security researchers remotely compromised a Jeep Cherokee through its cellular connection, gaining control over critical functions including steering, braking, and acceleration, ultimately leading to the recall of 1.4 million vehicles [5]. Similar demonstrations have shown vulnerabilities in Tesla's Autopilot system, where researchers manipulated camera inputs to cause lane-keeping failures [6]. These real-world attacks underscore the urgent need for robust cyber-security measures in AV development and deployment.

The automotive industry has responded to these challenges through various initiatives, including the establishment of security standards such as ISO/SAE 21434 for road vehicle cyber-security engineering and UN Regulation No. 155 on cyber-security and cyber-security management systems [7]. However, as AV technology continues to advance, so too does the sophistication of potential adversaries and attack methodologies. The integration of artificial intelligence (AI) and machine learning (ML) into both vehicle control systems and adversarial attack strategies has created an escalating arms race between defenders and attackers in the AV security domain.

The economic and societal stakes are substantial. The global AV market is projected to reach hundreds of billions of dollars within the next decade, with widespread deployment expected to save thousands of lives annually by eliminating human error, which accounts for approximately 94% of traffic accidents [8]. However, public trust and regulatory acceptance of AVs critically depend on demonstrable security and safety. A single high-profile security incident resulting in casualties could significantly delay or derail AV adoption, with profound economic and social consequences. Therefore, advancing the science and practice of AV cyber-security is not merely a technical imperative but a societal necessity.

1.2 Role of Vision and Sensor Fusion in Autonomous Vehicle Control Loops

Perception systems form the foundational layer upon which all AV capabilities are built. These systems enable AVs to construct accurate representations of their surrounding environment, detect and classify objects, estimate vehicle state, and predict the behavior of other traffic participants.

Among the various sensing modalities employed in modern AVs, vision-based systems utilizing cameras play a particularly critical role due to their ability to capture rich semantic information, detect lane markings, interpret traffic signs and signals, and recognize objects at relatively low cost compared to other sensor technologies [9].

The typical perception pipeline in an AV consists of multiple stages. Raw sensor data from cameras, LiDAR, radar, and other sensors are first preprocessed to remove noise and correct for distortions. Feature extraction algorithms then identify relevant patterns in the data, such as edges, corners, or semantic regions in images. These features are subsequently processed by classification or detection networks, often based on deep neural network architectures such as convolutional neural networks (CNNs), to identify and localize objects of interest. Finally, tracking algorithms maintain consistent object identities across time, enabling prediction of future trajectories [10].

Camera-based vision systems offer several advantages that make them indispensable in AV sensor suites. Cameras can detect and interpret traffic signs, signals, and lane markings with high accuracy, enabling compliance with traffic regulations. They provide dense spatial information that facilitates precise localization within lanes and relative to other vehicles. Modern high-resolution cameras can operate effectively at various distances, from close-range obstacle detection to far-range highway monitoring. Furthermore, the rich color and texture information captured by cameras enables semantic understanding of scenes that is difficult to achieve with other sensor modalities alone [11].

However, vision systems also present significant vulnerabilities. Cameras are susceptible to adverse weather conditions (fog, rain, snow), challenging lighting scenarios (glare, darkness, shadows), and physical obstructions (dirt, debris on the lens). More critically from a security perspective, camera data is inherently manipulable through both physical and digital means. Adversarial attacks on vision systems have been extensively documented in the literature, ranging from physical perturbations such as adversarial patches applied to stop signs [12] to digital manipulation of captured imagery [13].

To mitigate the limitations and vulnerabilities of individual sensors, modern AVs employ sensor fusion techniques that combine data from multiple heterogeneous sensors to create more robust and reliable environmental representations. Sensor fusion leverages the complementary strengths of different sensing modalities: cameras provide high-resolution semantic information, LiDAR offers precise 3D spatial measurements independent of lighting conditions, radar excels in velocity estimation

and operates reliably in adverse weather, and GPS/GNSS systems provide global localization [14].

Common sensor fusion architectures in AVs include early fusion, where raw data from multiple sensors are combined before processing; late fusion, where independent perception pipelines process each sensor's data separately before integrating the results; and hybrid approaches that combine elements of both strategies. Advanced fusion algorithms employ probabilistic frameworks such as Kalman filters, particle filters, or Bayesian networks to optimally combine sensor measurements while accounting for their respective uncertainties and reliability characteristics [15].

The lane-keeping system, which maintains the vehicle's position within its designated lane, exemplifies the critical role of vision and sensor fusion in AV control loops. A typical lane-keeping system uses cameras to detect lane markings and compute the vehicle's lateral deviation from the lane center and relative heading angle with respect to the lane direction. This perception information feeds into a control algorithm, often implemented as a model predictive controller (MPC), which computes steering commands to minimize lateral deviation while ensuring passenger comfort and system stability [16].

The closed-loop nature of perception-control integration in AVs creates both opportunities and vulnerabilities from a cyber-security perspective. On one hand, the tight coupling between perception and control enables rapid response to environmental changes and robust performance under nominal conditions. On the other hand, this integration means that compromising the perception system can directly manipulate control actions, potentially causing the vehicle to deviate from its intended path or make unsafe maneuvers. Furthermore, sophisticated attacks that simultaneously target multiple components of the perception-control loop can evade detection mechanisms that rely on cross-validation between sensors or consistency checks between perception and control [3].

Recent research has highlighted the vulnerability of sensor fusion systems to coordinated attacks. While sensor fusion provides robustness against failures or attacks on individual sensors, an adversary who can simultaneously compromise multiple sensing modalities may be able to deceive fusion algorithms and evade detection. For example, an attack that consistently manipulates both camera and GPS data in a coordinated manner may pass consistency checks that would otherwise detect discrepancies between these sensors. This observation motivates the development of more sophisticated attack models and defense mechanisms that account for the possibility of multi-sensor

coordinated attacks, which forms a central theme of this dissertation.

1.3 Evolution of Cyber-Attacks on Perception and Navigation Systems

The landscape of cyber-attacks targeting AV perception and navigation systems has evolved significantly over the past decade, progressing from theoretical proof-of-concept demonstrations to increasingly sophisticated and stealthy attack methodologies. Understanding this evolution is essential for developing effective defense mechanisms and anticipating future threats.

Early research on AV security focused primarily on network-based attacks targeting communication interfaces. Koscher et al. demonstrated in 2010 that an attacker with physical access to a vehicle's CAN bus could control critical vehicle functions including engine control, braking, and dashboard displays [17]. This seminal work established that modern vehicles, despite not being fully autonomous, were vulnerable to cyber-attacks with potentially severe safety consequences. Subsequent research by Checkoway et al. extended these findings by demonstrating remote attack vectors, including cellular connections, Bluetooth interfaces, and radio broadcast systems, thereby eliminating the requirement for physical access [18].

As AVs began incorporating sophisticated sensor systems, researchers turned their attention to sensor-level attacks. GPS spoofing emerged as a prominent threat vector, where attackers broadcast false GPS signals to manipulate a vehicle's perception of its location. Tippenhauer et al. established the theoretical requirements for successful GPS spoofing, demonstrating that precisely timed counterfeit signals could deceive receivers without triggering correlation-based anomaly detectors [19]. Subsequent work demonstrated practical GPS spoofing attacks against aerial drones and maritime vessels, with researchers successfully diverting a yacht off course by spoofing its GPS system [20].

Vision-based attacks represent another significant category in the evolution of AV cyber-threats. Initial research focused on adversarial examples in ML , where small, carefully crafted perturbations to input images could cause deep neural networks (DNNs) to misclassify objects with high confidence [21]. Szegedy et al. first demonstrated this phenomenon in 2014, showing that adding imperceptible perturbations to images could cause state-of-the-art classification networks to

produce completely incorrect predictions [22].

The translation of adversarial examples from digital perturbations to physical-world attacks marked a critical development in AV security research. Eykholt et al. demonstrated in 2018 that physical adversarial patches applied to stop signs could cause object detection networks to misclassify them as speed limit signs or fail to detect them entirely, while remaining visually subtle enough to potentially escape human notice [12]. This work established that adversarial attacks were not merely theoretical concerns but could be realized through practical physical modifications to traffic infrastructure or signage.

The concept of covert attacks represents a significant advancement in adversarial sophistication. Unlike overt attacks that seek to maximize disruption or damage, covert attacks aim to manipulate vehicle behavior while remaining undetected by both automated security systems and human observers. De Sá et al. formally defined covert attacks in cyber-physical systems as those that achieve malicious objectives while keeping observable system outputs within expected ranges, thereby evading threshold-based anomaly detectors [23]. This attack paradigm is particularly dangerous in AV contexts, where delayed detection can result in accumulated deviation from safe operating parameters before anomalies become apparent.

Recent developments have introduced increasingly sophisticated attack methodologies leveraging generative AI. Sato et al. proposed the dirty road patch (DRP) attack, which disguises adversarial perturbations as natural road debris or dirt, achieving nearly 100% success in causing lane-keeping systems to deviate from the centerline while maintaining visual plausibility [13]. This work demonstrated that attackers could leverage domain-specific knowledge about road environments to create perturbations that appear natural rather than obviously malicious.

The integration of generative adversarial networks (GANs) into attack methodologies represents a further escalation in sophistication. GANs enable the synthesis of highly realistic images by training generator networks to produce outputs that are indistinguishable from authentic data by discriminator networks [24]. In the context of AV attacks, GANs can be employed to generate fake camera frames that appear authentic while manipulating the vehicle's perception of its environment. The generator learns to create images that both deceive the victim vehicle's perception system and pass scrutiny by anomaly detection mechanisms.

More recently, diffusion models have emerged as powerful alternatives to GANs for high-fidelity image synthesis. Diffusion models, which generate images through an iterative denoising process, have demonstrated superior performance in terms of image quality, training stability, and controllability compared to traditional GANs [25]. The application of diffusion models to adversarial attacks against AVs represents the current frontier in this domain, enabling unprecedented levels of realism and control in synthesized imagery.

Coordinated multi-sensor attacks represent another critical evolution in the threat landscape. Recognizing that AVs employ sensor fusion to enhance robustness, sophisticated adversaries have begun developing attack strategies that simultaneously compromise multiple sensing modalities. For example, a coordinated attack might combine GPS spoofing to manipulate global localization with vision-based attacks to corrupt camera data, ensuring that sensor fusion algorithms receive consistent but incorrect information from multiple sources. Such attacks are significantly more difficult to detect than single-sensor attacks, as they can evade consistency checks and cross-validation mechanisms that compare outputs from different sensors.

The evolution of attacks has been accompanied by increasing awareness of the importance of attack stealth and persistence. Early attacks often prioritized demonstrating capability over avoiding detection, resulting in obvious anomalies that would trigger alerts in real systems. Contemporary attacks increasingly emphasize gradual, subtle manipulations that accumulate over time to achieve malicious objectives while staying below detection thresholds. This trend toward stealthy, persistent attacks poses significant challenges for defense mechanisms and motivates the development of more sophisticated detection methodologies.

1.4 Limitations of Conventional Detection Methods

Despite significant research efforts, conventional approaches to detecting cyber-attacks against AV perception and navigation systems exhibit several critical limitations that restrict their effectiveness against sophisticated adversaries. Understanding these limitations is essential for motivating the development of advanced detection methodologies.

Threshold-based anomaly detection represents one of the most common approaches to identifying

attacks on AVs. These methods establish acceptable ranges for sensor measurements, control commands, or system states based on nominal operating conditions, then trigger alerts when observed values exceed predefined thresholds [26]. While conceptually straightforward and computationally efficient, threshold-based methods suffer from fundamental limitations. They are inherently vulnerable to covert attacks specifically designed to keep observable parameters within normal ranges while gradually manipulating vehicle behavior. Furthermore, determining appropriate threshold values presents a significant challenge: overly permissive thresholds reduce sensitivity to attacks, while overly restrictive thresholds generate excessive false alarms that can desensitize operators or trigger unnecessary interventions.

Model-based detection approaches attempt to overcome the limitations of threshold-based methods by leveraging mathematical models of vehicle dynamics and sensor characteristics to identify inconsistencies indicative of attacks. These methods construct state estimators, such as Kalman filters or observers, that predict expected sensor measurements or system states based on vehicle models and control inputs. Significant discrepancies between predicted and observed values may indicate sensor faults or malicious manipulation [27]. However, model-based approaches face several challenges in practice. They require accurate models of complex, nonlinear vehicle dynamics, which may be difficult to obtain or maintain as vehicle characteristics change over time. Sophisticated attacks that account for the defender's model can potentially evade detection by ensuring manipulated measurements remain consistent with model predictions. Additionally, model-based methods may struggle to distinguish between attacks and legitimate disturbances, unmodeled dynamics, or sensor noise.

ML-based detection methods have gained prominence in recent years, leveraging the ability of neural networks to learn complex patterns in data without requiring explicit mathematical models. Supervised learning approaches train classifiers on labeled datasets containing examples of normal and attack scenarios, enabling the system to recognize attack signatures [28]. Unsupervised learning methods, such as autoencoders or one-class support vector machines, learn representations of normal system behavior and flag deviations as potential anomalies [29]. While ML offers powerful pattern recognition capabilities, these methods also face limitations. They require extensive training data covering diverse operating conditions and attack scenarios, which may be difficult or expensive to

obtain. The *black box* nature of many ML models limits interpretability and trust, particularly in safety-critical applications. Furthermore, adversaries aware of the defender's ML model can craft adversarial examples specifically designed to evade detection, leveraging techniques similar to those used to fool perception systems.

Sensor redundancy and voting mechanisms represent another category of defensive approaches. By deploying multiple sensors of the same or different types and comparing their outputs, systems can potentially identify compromised sensors that provide inconsistent measurements [30]. However, this approach assumes that adversaries can compromise only a minority of sensors, which may not hold against sophisticated attackers with multiple attack vectors. Coordinated attacks that simultaneously manipulate multiple sensors in a consistent manner can evade voting-based detection. Additionally, redundancy increases system cost, complexity, and computational requirements, potentially limiting its practical applicability.

Physical layer authentication for wireless communications aims to detect spoofing attacks on GPS and other wireless systems by analyzing physical characteristics of received signals, such as signal strength, angle of arrival, or channel impulse response [31]. While promising, these methods face challenges in dynamic vehicular environments where channel characteristics change rapidly due to mobility and multi-path propagation. Sophisticated attackers may also learn to mimic legitimate signal characteristics, particularly with advanced software-defined radio equipment.

Cryptographic authentication and integrity protection mechanisms ensure that messages originate from legitimate sources and have not been tampered with during transmission [32]. While effective against certain categories of attacks, cryptography cannot protect against attacks that occur after authentication (e.g., malicious code running on authenticated systems) or attacks targeting sensors and physical processes rather than communication channels. Furthermore, cryptographic mechanisms introduce computational overhead and latency that may be problematic in time-critical control loops.

A fundamental limitation common to many conventional detection methods is their assumption of independence between attack and defense mechanisms. In reality, sophisticated adversaries may have knowledge of the defender's detection algorithms and can tailor attacks to exploit specific blind spots or weaknesses. This adversarial co-evolution between attack and defense necessitates approaches that explicitly model and account for strategic adversaries, rather than treating attacks as random

faults or disturbances.

Another critical gap in existing approaches is the limited consideration of temporal dynamics and attack persistence. Many detection methods analyze instantaneous measurements or short time windows, potentially missing gradually evolving attacks that accumulate deviation over extended periods. Sophisticated covert attacks may be specifically designed to introduce small, undetectable changes at each time step that collectively achieve significant manipulation over time. Effective detection requires consideration of longer temporal horizons and analysis of cumulative effects.

The challenge of multi-modal attack detection represents another significant limitation. AVs integrate diverse sensing modalities (vision, LiDAR, radar, GPS, IMU) and processing components (perception, localization, planning, control), each potentially vulnerable to attack. Detecting coordinated attacks that simultaneously compromise multiple components requires holistic analysis across subsystems, whereas conventional approaches often focus on individual sensors or modules in isolation. The interactions and dependencies between different subsystems create complex attack surfaces that are not adequately addressed by component-level detection.

Finally, the trade-off between detection accuracy and computational efficiency presents practical challenges. Many sophisticated detection algorithms require significant computational resources, which may not be available on resource-constrained embedded systems in vehicles. Real-time requirements further constrain acceptable computational latency, potentially limiting the complexity of detection algorithms that can be deployed in practice. Balancing detection performance with computational feasibility remains an ongoing challenge in the field.

1.5 Motivation for Generative AI-Informed Attack and Defense Paradigms

The limitations of conventional detection methods and the increasing sophistication of cyber-threats to AVs motivate a paradigm shift in how we approach both adversarial attack modeling and defense mechanism design. Generative AI, particularly recent advances in GANs and diffusion models, offers transformative capabilities for both understanding and mitigating advanced cyber-threats targeting AV perception systems.

From an attack modeling perspective, generative AI enables the creation of highly sophisticated,

realistic manipulations of sensor data that were previously infeasible. Traditional adversarial examples typically involve adding noise-like perturbations to input data, which, while effective at fooling ML models, often produce artifacts that are detectable through careful analysis or that appear unnatural to human observers. In contrast, generative models such as GANs and diffusion models can synthesize entirely new sensor data that maintains photorealistic quality, semantic consistency, and temporal coherence while encoding adversarial objectives [24, 25].

The application of generative models to adversarial attacks represents a form of deep fake technology applied to AV perception. Just as deep fake technology has enabled the creation of convincing synthetic videos of human faces and voices, similar techniques can be applied to create fake road scenes, manipulated camera feeds, or synthetic sensor data that deceive AV perception systems. The key distinction is that, unlike entertainment or misinformation applications of deep fakes, adversarial deep fakes targeting AVs can have immediate physical safety consequences.

Understanding the full capabilities of generative AI-based attacks is essential for developing effective defenses. By explicitly modeling and implementing such attacks, researchers can identify vulnerabilities in existing systems, evaluate the robustness of proposed defenses, and establish realistic threat models that inform security requirements for AV development. This approach aligns with the philosophy of *security through understanding*, where comprehensive characterization of threats enables more effective mitigation strategies.

Generative AI also offers powerful tools for defense and detection. Generative models can be trained to learn the statistical characteristics and structure of authentic sensor data, enabling the identification of synthetic or manipulated inputs that deviate from these learned distributions. Forensic analysis techniques leveraging generative models can detect subtle artifacts in imagery, such as inconsistencies in photo-response non-uniformity (PRNU) patterns, rolling shutter distortions, or statistical correlations that distinguish real camera images from synthetic ones [33].

The adversarial training paradigm, pioneered in the ML robustness literature, provides a framework for enhancing system resilience against attacks. By training perception and control systems on data that includes adversarial examples generated through generative models, systems can potentially learn to maintain correct behavior even when subjected to sophisticated attacks [34]. This approach essentially inoculates systems against attacks by exposing them to attack examples during training,

analogous to how vaccines expose immune systems to pathogens to develop immunity.

Generative models also enable the creation of comprehensive datasets for training and evaluating detection systems. Collecting real-world examples of cyber-attacks against AVs is extremely challenging due to safety concerns, legal restrictions, and the rarity of such attacks in practice. Generative models can synthesize realistic attack scenarios spanning diverse conditions, attack types, and severity levels, providing the extensive labeled data necessary for training effective ML-based detection systems [35].

The concept of adversarial co-evolution between attacks and defenses motivates a game-theoretic perspective on AV security. GANs embody this adversarial dynamic in their very architecture: the generator and discriminator networks engage in a mini-max game, where the generator attempts to produce synthetic data that fools the discriminator, while the discriminator improves its ability to distinguish real from fake data. This adversarial training process results in progressively more sophisticated generators and discriminators, mirroring the real-world arms race between attackers and defenders [24].

Extending this framework to AV security suggests that attack and defense mechanisms should be co-designed and co-evolved through adversarial training. Rather than developing attack models and defense mechanisms independently, a generative adversarial approach simultaneously refines both components, resulting in more robust systems that account for increasingly sophisticated adversaries. This paradigm represents a significant departure from traditional security approaches that treat attacks and defenses as separate concerns.

The interpretability and explainability of generative model-based detection methods present both opportunities and challenges. While DNNs are often criticized as black boxes that provide limited insight into their decision-making processes, generative models can offer more interpretable detection mechanisms. For example, reconstruction-based detection using autoencoders or diffusion models can visualize the differences between observed inputs and the model's reconstruction of what authentic data should look like, potentially providing interpretable evidence of manipulation [36]. This explainability is particularly valuable in safety-critical AV applications, where understanding why a detection system flagged certain inputs is important for validating its reliability and building trust.

The integration of generative AI into AV security research also aligns with broader trends in cyber-physical system security. Modern cyber-physical systems increasingly rely on ML for perception, decision-making, and control, making them inherently vulnerable to adversarial ML attacks. Defending against such attacks requires security mechanisms that are similarly sophisticated and that leverage advanced AI techniques. Generative AI-informed approaches represent a natural evolution in this direction, providing tools commensurate with the complexity of modern threats.

Furthermore, generative models offer potential for proactive defense mechanisms that go beyond passive detection. For instance, generative models could be employed to actively sanitize or correct potentially corrupted sensor data by reconstructing what the authentic data should look like based on learned distributions of real data. While such approaches must be carefully designed to avoid introducing additional vulnerabilities or masking legitimate environmental changes, they represent an intriguing direction for future research.

The computational advances that have enabled practical generative AI also make such approaches increasingly feasible for deployment in AVs. Modern vehicles are equipped with powerful GPU-accelerated computing platforms capable of running DNNs in real-time. Training generative models, while computationally intensive, can be performed offline during system development, with the trained models then deployed for inference on vehicle hardware. The increasing availability of edge AI accelerators further reduces the computational barriers to deploying sophisticated generative model-based security mechanisms.

1.6 Research Objectives and Questions

This dissertation addresses the critical challenge of securing AV perception and navigation systems against sophisticated cyber-attacks through the development and evaluation of generative AI-based attack models and corresponding detection mechanisms. The overarching goal is to advance the state of knowledge in AV cyber-security by comprehensively characterizing advanced threat models, developing effective defense mechanisms, and establishing methodologies for evaluating security in the face of intelligent adversaries.

The primary research objectives of this dissertation are as follows:

Objective 1: Design and Implement Generative AI-Based Covert Attacks on AV Vision Systems. This objective focuses on developing novel attack methodologies that leverage generative AV to manipulate camera-based perception systems in AVs. Specifically, this involves:

- Designing GAN-based attack architectures that can synthesize realistic road scene imagery to deceive lane-keeping systems while evading detection
- Developing stable diffusion-based attack models that improve upon GAN approaches in terms of image quality, temporal coherence, and controllability
- Creating video diffusion models capable of generating temporally consistent sequences of manipulated frames for sustained covert attacks
- Evaluating the effectiveness of these attacks in realistic simulation environments and with real-world data

Objective 2: Develop Coordinated Multi-Sensor Attack Frameworks Targeting Both Vision and Navigation Systems. Recognizing that AVs employ sensor fusion for robustness, this objective addresses the development of hybrid attack strategies that simultaneously compromise multiple sensing modalities:

- Designing Doppler-consistent GPS spoofing mechanisms that can manipulate global navigation while maintaining signal-level plausibility
- Integrating vision-based and GPS spoofing attacks into coordinated frameworks that maintain consistency across sensor modalities
- Analyzing the synergistic effects of multi-sensor attacks on sensor fusion algorithms and overall vehicle behavior
- Establishing theoretical foundations for understanding the requirements and constraints of effective coordinated attacks

Objective 3: Create Comprehensive Real-World Datasets for Training and Evaluating Generative Attack Models. To ensure that developed attack and defense mechanisms are realistic and applicable to real-world scenarios, this objective involves:

- Designing and implementing custom data collection hardware platforms for capturing paired camera imagery under diverse conditions
- Collecting urban road scene datasets spanning various weather conditions, lighting scenarios, and traffic situations
- Preprocessing and annotating collected data to support generative model training

Objective 4: Develop Multi-Modal Forensic Detection Mechanisms for Identifying Generative AI-Based Attacks. Complementing the attack development objectives, this objective focuses on defense mechanisms:

- Designing forensic feature extraction techniques that identify artifacts in synthetic imagery, including PRNU analysis and rolling shutter inconsistency detection
- Developing statistical anomaly detection methods that integrate multi-modal forensic features for robust attack identification
- Creating detection frameworks that account for temporal dynamics and cumulative attack effects
- Evaluating detection performance through comprehensive metrics including detection accuracy, false alarm rates, and detection latency

These objectives give rise to several specific research questions (RQs) that guide the investigations presented in this dissertation:

RQ1: How can GANs be adapted to create covert attacks on lane-keeping systems that remain visually plausible while inducing controlled lateral deviation? This question addresses the fundamental challenge of designing GAN architectures and loss functions appropriate for adversarial vision manipulation in automotive contexts.

RQ2: What advantages do diffusion models offer over GANs for creating covert vision-based attacks, and how can they be optimized for real-time performance in AV contexts? This question explores the trade-offs between different generative modeling approaches and their suitability for adversarial applications.

RQ3: How can vision-based attacks be coordinated with GPS spoofing to create hybrid attacks that evade sensor fusion-based detection while achieving precise trajectory manipulation? This question investigates the requirements and methodologies for effective multi-sensor coordinated attacks.

RQ4: What forensic artifacts distinguish synthetic imagery from authentic camera data, and how can these artifacts be reliably detected in real-time AV operation? This question focuses on the defense perspective, identifying exploitable weaknesses in generative attack models.

RQ5: How do multi-modal forensic detection approaches that integrate visual, geometric, and navigational features compare to single-modality detection in terms of accuracy, robustness, and detection latency? This question evaluates the benefits of sensor fusion for defense, mirroring its use in perception.

RQ6: What are the fundamental limits of detection for covert attacks designed with knowledge of the defender's detection algorithms, and how can defenders optimize their strategies against such adaptive adversaries? This question explores the game-theoretic dimensions of the attack-defense dynamic.

RQ7: How do attack effectiveness and detection performance vary across different environmental conditions, vehicle states, and operational scenarios? This question addresses the generalization and robustness of developed methodologies across diverse real-world conditions.

Addressing these objectives and research questions requires an interdisciplinary approach drawing on expertise from computer vision, ML, control theory, cyber-physical systems security, and automotive engineering. The dissertation integrates theoretical analysis, algorithm development, system implementation, and comprehensive experimental evaluation to provide both fundamental insights and practical contributions to AV security.

1.7 Major Contributions

This dissertation makes several significant contributions to the fields of AV security, cyber-physical system security, and adversarial ML. The major contributions can be categorized into theoretical, methodological, and practical domains:

1.7.1 Theoretical Contributions

Formalization of Hybrid Covert Attack Models for AVs. This dissertation provides rigorous mathematical formulations of covert attacks targeting AV perception and navigation systems. Unlike previous work that has primarily addressed single-sensor attacks or overt manipulations, this research establishes formal models for coordinated multi-sensor attacks that maintain consistency across sensing modalities while remaining below detection thresholds. These models provide theoretical foundations for understanding the requirements, constraints, and capabilities of sophisticated adversaries.

Game-Theoretic Framework for Adversarial Dynamics in AV Security. The dissertation develops a game-theoretic perspective on the interaction between attackers and defenders in AV security, explicitly modeling the strategic considerations and adaptive behaviors of both parties. This framework enables analysis of optimal attack and defense strategies and provides insights into fundamental limits of detection and attack success conditions.

Characterization of Detection Bounds for Generative AI-Based Attacks. Through theoretical analysis, this research establishes bounds on detection performance metrics such as false alarm probability and detection delay as functions of attack parameters and detection system characteristics. These theoretical results provide guidance for security system design and enable quantitative evaluation of security guarantees.

1.7.2 Methodological Contributions

Novel GAN Architecture for Vision-Based Covert Attacks (VCA-HADS). The dissertation introduces a customized GAN architecture specifically designed for synthesizing manipulated road scene imagery. This architecture incorporates spatial transformation networks, novel loss functions that balance attack effectiveness with visual plausibility, and mechanisms for maintaining geometric consistency. This represents the first application of GANs to create comprehensive covert attacks on AV lane-keeping systems.

Stable Diffusion-Based Attack Framework (Gen-VCA). Building upon limitations of GAN-based approaches, the research develops a stable diffusion model adapted for adversarial vision manipulation. This framework demonstrates superior performance in terms of image quality, control

over generated content, and training stability. The integration of lane geometry conditioning and temporal consistency mechanisms represents significant methodological advances in applying diffusion models to adversarial contexts.

Video Diffusion Framework for Temporal Coherence (VideoDiff-VCA). Extending frame-based generation to video sequences, the dissertation presents a video diffusion architecture that maintains temporal coherence across extended attack durations. This methodology addresses critical limitations of prior work where frame-by-frame generation introduced detectable artifacts. The temporal propagation mechanisms and specialized loss functions developed in this work advance the state of the art in adversarial video generation.

Doppler-Consistent GPS Spoofing Architecture (GPS Spoof-Net). The research develops a software-defined radio pipeline for generating Doppler-consistent counterfeit GPS signals. Unlike basic GPS spoofing approaches that manipulate only position, this methodology ensures that Doppler shifts of individual satellite signals remain consistent with the spoofed kinematics, evading sophisticated detection mechanisms based on signal-level analysis.

Multi-Modal Forensic Detection. The dissertation introduces a comprehensive detection framework that integrates diverse forensic features including PRNU analysis, rolling shutter consistency checking, and Doppler residual monitoring. The statistical fusion approach based on Mahalanobis distance and the optimization framework for balancing detection delay and false alarm probability represent methodological contributions applicable beyond AV security.

1.7.3 Practical Contributions

Real-World Urban Driving Dataset. The research contributes a novel dataset of urban road scenes collected under diverse weather and lighting conditions in Montreal, Canada. This dataset includes paired imagery from fixed and center-aligned cameras, providing ground truth (GT) for training view synthesis models. The hardware specifications and collection methodologies are documented to support reproducible research.

Custom Data Collection Platform. The dissertation describes the design and implementation of a custom hardware platform for collecting paired road scene imagery. This includes mechanical designs, sensor specifications, calibration procedures, and software for real-time lane detection and

camera control. This contribution enables other researchers to replicate the data collection process and extend the dataset.

Comprehensive Evaluation Framework. The research establishes evaluation methodologies, metrics, and protocols for assessing both attack effectiveness and detection performance in AV security contexts. This includes performance metrics, and analysis techniques that can be adopted by other researchers and practitioners.

1.8 Thesis Organization

This dissertation is organized to provide a progressive and comprehensive exploration of adversarial attacks and defense mechanisms for AV perception and navigation systems, with detection methodology and results integrated contextually within the relevant technical chapters.

Chapter 2 reviews foundational concepts and the state-of-the-art in AV architectures, perception and navigation sensor fusion, generative adversarial models, and CPS security. It further surveys relevant literature on attack surfaces and detection techniques, positioning the thesis within the broader research landscape.

Chapters 3 to 6 develop the core technical contributions using a scenario-driven approach. Each chapter focuses on a specific attack paradigm, methodologically introducing the proposed adversarial strategy, analyzing its impact, and presenting integrated detection frameworks appropriate for that context:

- Chapter 3 details baseline threat models, formalizes the general problem statement, and introduces the evaluation criteria and dataset engineering including the Montreal Urban Driving Dataset (MUDD).

- Chapter 4 presents the conception and analysis of GAN-based vision attacks (VCA-HADS), describing the model architecture, implementation, and evaluation on simulation AV data. Detection approaches exploiting vision forensic features such as PRNU analysis are discussed within the attack evaluation.

- Chapter 5 transitions to diffusion-based generative attacks (Gen-VCA), demonstrating advances

in photorealism, geometric precision, and temporal coherence. The chapter integrates detection methods leveraging improvements in PRNU preservation and temporal consistency, providing a side-by-side analysis of both attack strength and detectability in realistic environments.

- Chapter 6 introduces the coordinated hybrid multi-sensor attack which combines video diffusion with GPS spoofing (VideoDiff-VCA + GPS Spoof-Net) that achieves substantial increases in attack stealth and trajectory deviation. Multi-modal detection strategies fusing visual and navigation forensic features are methodologically interwoven within the evaluation.

Finally, chapter 7 synthesizes the dissertation's findings in a unified conclusion, reflecting on the evolution of attack sophistication, effectiveness of integrated detection frameworks, theoretical and methodological advances, and practical implications for future AV security architectures. This chapter also articulates future research directions including hardware-in-the-loop validation, federated learning security, and adaptive defense.

Chapter 2

Background and Related Works

This chapter provides a comprehensive review of the literature relevant to cyber-security threats against AVs, with particular emphasis on perception system attacks, navigation system vulnerabilities, generative AI models, and defense mechanisms. The review synthesizes recent advances from 2018 to 2025, establishing the theoretical foundations and identifying research gaps that motivate the contributions of this dissertation.

2.1 Overview of Autonomous Vehicle Cyber-security

2.1.1 Cyber-Physical Security Landscape

AVs represent complex CPSs where computational processes are tightly integrated with physical dynamics, creating unique security challenges that extend beyond traditional information technology domains [1,37]. Recent comprehensive surveys have mapped the evolving threat landscape, identifying multiple attack surfaces spanning communication interfaces, sensor systems, control networks, and software platforms [38, 39].

The National Highway Traffic Safety Administration's critical analysis of crash causation attributes approximately 94% of accidents to human error, motivating the development of autonomous driving systems that promise enhanced safety [8]. However, this technological advancement simultaneously introduces new vulnerabilities where cyber-attacks can directly compromise physical safety [40]. Recent work by Wang et al. [39] provides a systematic categorization of security threats in connected

and AVs, emphasizing the interconnected nature of modern vehicular networks that creates cascading failure risks.

The establishment of automotive cyber-security standards, including ISO/SAE 21434 for road vehicle cyber-security engineering and UN Regulation No. 155, reflects industry recognition of these threats [7, 41]. These frameworks mandate security-by-design approaches throughout the vehicle lifecycle, from concept and development through production and decommissioning [42]. Despite these regulatory advances, the rapid evolution of attack methodologies, particularly those leveraging AI, continues to outpace defensive capabilities [43].

2.1.2 Attack Taxonomy and Classification

Research has developed comprehensive taxonomies classifying cyber-attacks on AVs based on multiple dimensions: target subsystem, attack vector, adversarial capability, and impact severity [41, 44]. Changalvala et al. [45] categorize attacks into three primary classes: communication-based attacks targeting V2X protocols, sensor-based attacks manipulating perception systems, and control-based attacks directly interfering with vehicle actuators.

Covert attacks represent a particularly sophisticated category where adversaries manipulate system behavior while maintaining observable outputs within normal ranges, thereby evading threshold-based anomaly detectors [23, 46]. De Sá et al. [23] formally defined covert attacks in CPSs, establishing theoretical foundations for understanding their detectability limits. Recent extensions by Lian et al. [47] have investigated covert attacks under networked control scenarios with hybrid cyber-physical attack surfaces.

False data injection (FDI) attacks constitute another critical threat category, where attackers inject misleading information into sensor data streams or control signals to induce erroneous decision-making [48, 49]. Huang and Wang [50] demonstrated the effectiveness of FDI attacks against intelligent transportation systems, highlighting vulnerabilities in state estimation and control loops.

2.2 Perception System Vulnerabilities

2.2.1 Adversarial Attacks on Deep Learning Models

The integration of DNNs for perception tasks in AVs has introduced vulnerabilities to adversarial ML attacks. Goodfellow et al. [21] first demonstrated that small, carefully crafted perturbations to input images could cause misclassifications with high confidence, establishing the foundation for adversarial example research. Szegedy et al. [22] revealed the counterintuitive brittleness of deep networks, showing that adversarial perturbations transfer across different models and architectures.

The translation from digital adversarial examples to physical-world attacks marked a critical development for AV security. Eykholt et al. [12] demonstrated robust physical adversarial attacks where printed stickers applied to stop signs caused misclassifications by object detection networks under varying viewpoints, distances, and lighting conditions. This work established that adversarial attacks pose practical, not merely theoretical, threats to AV perception systems.

Chen et al. [51] extended physical adversarial attack research with the Shape-shifter attack, which creates adversarial examples that cause object detectors to misclassify vehicles as innocuous objects across multiple viewpoints. Similarly, Shah [52] investigated adversarial resilience through human-centered perception mechanisms, proposing cognitive recognition models for enhanced robustness.

Recent work has focused on increasing attack stealthiness. Sato et al. [13] proposed the DRP attack, which disguises adversarial perturbations as natural road debris or dirt, achieving nearly 100% success in causing lane-keeping systems to deviate while maintaining visual plausibility to human observers. This work demonstrated that domain-specific knowledge about expected environmental conditions enables more effective adversarial attacks.

2.2.2 Vision-Based Lane-Keeping Attacks

Lane-keeping systems, which maintain vehicle position within lane boundaries, represent critical components of driver assistance systems and fully AVs [53, 54]. These systems typically rely on camera-based perception to detect lane markings and compute lateral deviation and heading error [55].

Dong et al. [56] evaluated the impact of cyber-attacks on traffic flow in connected and automated

vehicle networks, demonstrating that compromised lane-keeping systems can create dangerous situations not only for the attacked vehicle but also for surrounding traffic. He et al. [57] proposed ML-based detection mechanisms for attacks targeting lane-keeping systems, though their approaches assume attack signatures are known a priori.

Liu et al. [58] analyzed stability of lane-keeping assistance systems under environmental disturbances such as crosswinds, establishing baseline performance characteristics that help distinguish attacks from natural perturbations. Kim et al. [59] developed deep reinforcement learning strategies for lane-keeping control considering surrounding vehicles, though Cyber-security considerations were not explicitly addressed.

Xu et al. [60] provided correctness guarantees for composition of lane-keeping and adaptive cruise control, establishing formal verification frameworks. However, their analysis assumes trusted sensor inputs, leaving vulnerabilities to perception attacks unaddressed. Kuo et al. [53] investigated image sensor processing for lane detection and control, identifying computational bottlenecks that may be exploited by resource exhaustion attacks.

2.2.3 Environmental and Physical Attacks

Beyond adversarial perturbations, research has investigated attacks exploiting environmental factors or physical modifications to infrastructure. Guesmi et al. [61] introduced AdvRain, which uses simulated adversarial raindrop patterns on camera lenses to induce errors in DNNs without alerting drivers. This work demonstrates that attacks disguised as natural environmental effects are particularly challenging to detect.

Nassi et al. [6] demonstrated phantom attacks where projecting images onto roads or other surfaces can cause object detection systems to perceive non-existent obstacles, potentially triggering emergency braking or evasive maneuvers. These attacks exploit the fundamental challenge of distinguishing authentic physical objects from sophisticated illusions in camera-based perception.

Physical modifications to traffic signs and road markings represent another attack vector. While obvious vandalism would be noticed, subtle modifications carefully designed to fool ML systems while appearing normal to humans pose significant challenges [12]. The adversarial patch literature has explored optimal placement and appearance of physical perturbations to maximize attack effectiveness

while minimizing human detectability [62].

2.3 Generative Models for Image Synthesis and Manipulation

2.3.1 Generative Adversarial Networks

GANs, introduced by Goodfellow et al. [24], have revolutionized image synthesis by framing generation as an adversarial game between a generator network that produces synthetic images and a discriminator network that attempts to distinguish real from fake images. This adversarial training paradigm enables generators to produce increasingly realistic outputs.

Jaderberg et al. [63] introduced Spatial Transformer Networks, which enable neural networks to learn geometric transformations of input data, providing critical capabilities for view synthesis applications. These transformations have been applied to AV perception, enabling conversion between different camera viewpoints [64].

The application of GANs to adversarial attack generation has been explored in several recent works. Frid-Adar et al. [35] demonstrated GAN-based data augmentation for medical image analysis, establishing principles applicable to generating diverse attack scenarios for training robust detection systems. However, traditional GANs suffer from training instability, mode collapse, and limited controllability over generated content [25].

2.3.2 Diffusion Models and Their Applications

Diffusion models have emerged as powerful alternatives to GANs, offering superior image quality, training stability, and controllability [25, 65]. These models learn to generate images through an iterative denoising process, gradually refining random noise into structured output. Rombach et al. [66] introduced Latent Diffusion Models, which operate in compressed latent spaces rather than pixel space, dramatically reducing computational requirements while maintaining quality.

Zhang et al. [67] demonstrated single image editing with text-to-image diffusion models, establishing frameworks for controlled manipulation of specific image attributes. These techniques enable precise modification of scene geometry and appearance, capabilities directly applicable to adversarial attacks on perception systems.

Recent work has extended diffusion models to video generation, addressing temporal coherence challenges. Ho et al. [68] introduced video diffusion models with temporal attention mechanisms ensuring consistency across frames. Blattmann et al. [69] proposed alignment techniques for text-to-video synthesis, demonstrating unprecedented control over spatiotemporal content generation.

The application of diffusion models to adversarial contexts remains relatively unexplored. Yu et al. [70] applied Gaussian splatting and diffusion priors to street view synthesis, demonstrating photorealistic scene generation. However, security implications of these capabilities for AV perception have not been comprehensively analyzed.

2.3.3 Neural Radiance Fields

Neural Radiance Fields (NeRFs), introduced by Mildenhall et al. [71], represent scenes as continuous volumetric functions learned by neural networks. NeRFs enable novel view synthesis with remarkable photorealism by modeling view-dependent appearance and geometric occlusion. While computationally intensive, NeRFs have inspired efficient variants suitable for real-time applications [72]. The application of NeRFs to autonomous driving scenarios has been investigated for scene reconstruction and simulation [73]. However, their potential for adversarial manipulation of perception systems represents an emerging threat not yet fully characterized in the literature.

2.4 Navigation System Attacks

2.4.1 GPS Spoofing Fundamentals

Global Navigation Satellite System (GNSS) spoofing attacks manipulate receivers by broadcasting counterfeit signals that cause incorrect position and velocity estimates. Tippenhauer et al. [19] established theoretical requirements for successful GPS spoofing, demonstrating that precisely timed signals can manipulate receiver outputs without triggering correlation-based anomaly detectors.

Kerns et al. [20] conducted field demonstrations of GPS spoofing attacks against unmanned aerial vehicles, successfully diverting a drone mid-flight. This work established practical feasibility of spoofing attacks against safety-critical autonomous systems. Humphreys et al. [74] analyzed spoofing attack complexity and countermeasure effectiveness, identifying fundamental limitations of

single-antenna civilian GPS receivers.

Recent work has investigated spoofing attacks specifically targeting vehicular systems. Sanders and Wang [75] proposed localization of spoofing sources using V2V communications, though this approach requires cooperative multi-vehicle networks. Miao et al. [76] developed adaptive attack recognition methods for AVs based on probabilistic models of GNSS signal anomalies.

2.4.2 Advanced Spoofing Techniques

Sophisticated spoofing attacks must maintain plausibility across multiple signal characteristics to evade detection. Doppler shift consistency represents a critical challenge: authentic GPS signals exhibit Doppler shifts determined by satellite-receiver relative velocity, and spoofed signals must replicate these patterns to avoid detection by advanced receivers [77].

Multi-constellation GNSS receivers (utilizing GPS, GLONASS, Galileo, BeiDou) offer enhanced resilience through cross-system consistency checking. However, sophisticated attackers with knowledge of receiver algorithms can potentially spoof multiple constellations simultaneously [78]. The computational and hardware complexity of such multi-constellation spoofing represents a practical barrier but not a fundamental impossibility.

2.4.3 IMU and Odometry Manipulation

Beyond GNSS, AVs rely on inertial measurement units (IMUs) and wheel odometry for dead reckoning and sensor fusion. Li and Chasaki [79] investigated covert channel attacks on CPSs, including manipulation of IMU data. However, IMUs are typically mounted within vehicle chassis, making physical access attacks challenging compared to wireless GPS spoofing.

Sensor fusion algorithms integrate GNSS, IMU, and odometry data to produce robust state estimates resilient to individual sensor failures or attacks. However, coordinated attacks targeting multiple sensors simultaneously can potentially deceive fusion algorithms, particularly if attack knowledge includes fusion algorithm details [80].

2.5 Multi-Modal and Coordinated Attacks

2.5.1 Sensor Fusion Exploitation

While sensor fusion enhances robustness against single-sensor attacks or failures, sophisticated adversaries can exploit fusion algorithms through coordinated multi-sensor attacks [81]. Petit and Shladover [3] identified sensor fusion as both a defensive strength and potential vulnerability, depending on attacker capabilities and coordination.

Recent work has begun exploring coordinated attack scenarios. Ju et al. [81] proposed distributed deception attack detection in vehicle platoons, addressing scenarios where multiple vehicles may be compromised simultaneously. However, their approach assumes attackers cannot perfectly coordinate across multiple sensing modalities.

Pasqualetti et al. [82] established fundamental limits on attack detection in CPSs, deriving conditions under which attacks remain undetectable given system dynamics and measurement models. These theoretical foundations apply to coordinated attacks targeting AV sensor suites.

2.5.2 Hybrid Attack Frameworks

True hybrid attacks simultaneously manipulating perception and navigation systems remain underexplored in literature. Most existing work treats vision-based attacks and GNSS spoofing as independent threats rather than coordinated components of unified attack frameworks. This gap motivates investigation of attack synergies where vision manipulation conceals physical deviation while spoofed navigation data maintains consistency with false perception.

Giannaros et al. [83] provide a comprehensive review of sophisticated attacks on AVs, identifying hybrid attack scenarios as critical open challenges. They emphasize that coordinated attacks exploiting multiple vulnerabilities simultaneously represent more realistic advanced persistent threats than isolated single-vector attacks.

2.6 Detection and Defense Mechanisms

2.6.1 Anomaly Detection Approaches

Anomaly detection for cyber-attacks on AVs encompasses diverse methodologies. Threshold-based approaches establish acceptable ranges for sensor measurements or control commands, triggering alerts when values exceed bounds [26]. However, covert attacks specifically designed to remain within normal ranges evade such detectors.

Model-based detection leverages mathematical models of vehicle dynamics and sensor characteristics to identify inconsistencies indicative of attacks [27]. Discrepancies between predicted and observed values suggest sensor faults or manipulation. However, sophisticated attackers with knowledge of defender models can craft attacks that maintain consistency with model predictions.

ML-based detection has gained prominence, utilizing neural networks to recognize complex attack patterns [28]. Supervised approaches train classifiers on labeled attack and normal data, while unsupervised methods learn representations of nominal behavior and flag deviations [29]. However, these approaches require extensive training data and may suffer from adversarial examples crafted to evade detection.

Alsulami et al. [49] proposed symmetrical simulation schemes for anomaly detection in AVs based on long-short term memory (LSTM) models, demonstrating effectiveness in identifying temporal attack patterns. Zaidi et al. [84] developed host-based intrusion detection for VANETs using statistical approaches to rogue node detection.

2.6.2 Sensor-Specific Detection

Vision system integrity can be monitored through multiple forensic approaches. Majumder et al. [85] investigated adversarial attack-resilient perception modules for traffic sign classification, demonstrating that robust model architectures enhance security. Feldsar et al. [86] proposed detecting adversarial examples using surrogate models trained to identify perturbations.

PRNU analysis enables camera fingerprinting and detection of synthetic imagery [33]. Each camera sensor exhibits unique noise patterns that serve as forensic fingerprints. Generated images lack authentic sensor PRNU, enabling their identification through correlation analysis with known

sensor fingerprints.

For GNSS systems, signal quality monitoring, Doppler consistency checking, and correlation peak analysis provide layers of spoofing detection. Advanced receivers monitor signal-to-noise ratios, multipath indicators, and inter-satellite timing relationships to identify anomalies indicative of spoofing [77].

2.6.3 Multi-Modal Fusion for Detection

Recognition that sophisticated attacks may target multiple sensors simultaneously motivates multi-modal detection approaches that fuse evidence across sensing modalities. Park et al. [30] investigated cyber-physical security of vehicular systems through sensor redundancy and voting mechanisms. However, simple majority voting fails when attackers compromise multiple sensors.

Zou et al. [87] explored AI's role in dynamic cyber risk analytics for organizational resilience, highlighting potential for intelligent fusion of diverse security indicators. Chen et al. [48] developed robust adaptive control for nonlinear CPSs with FDI attacks via attack estimation, demonstrating that model-based estimation can compensate for corrupted sensor data when attacks are detected early.

2.6.4 Adversarial Training and Robust Models

Madry et al. [34] demonstrated that adversarial training enhances robustness to test-time attacks. This approach has been applied to AV perception, though computational cost and potential performance degradation on benign inputs remain challenges. Athalye et al. [62] investigated circumventing defenses through expectation over transformation, revealing limitations of many proposed defense mechanisms. Carlini and Wagner [88] systematically evaluated robustness of neural networks, establishing that many defenses provide false sense of security against adaptive attackers aware of defense mechanisms.

Recent work has explored certified defenses with provable robustness guarantees [89]. These approaches provide mathematical bounds on adversarial perturbations that cannot fool models, though current techniques achieve limited certified radii and may sacrifice accuracy.

2.7 Generative AI for Security Applications

2.7.1 AI-Generated Content Detection

The proliferation of generative AI models has motivated development of detection techniques for synthetic media. Marra et al. [33] investigated detection of GAN-generated fake images over social networks, proposing forensic approaches based on generative model artifacts. Wang et al. [90] demonstrated CNN-based detection of manipulated facial images, achieving high accuracy through learning subtle fingerprints of generation processes.

Recent work has focused on zero-shot detection approaches that generalize to unseen generative models without requiring model-specific training. Ojha et al. [91] proposed lossless coding-based statistics for identifying AI-generated images, exploiting differences in inter-pixel correlation patterns between natural and synthetic images. Cozzolino et al. [92] introduced Noiseprint, a CNN-based approach to extract camera model fingerprints and detect manipulated images. This forensic tool demonstrates effectiveness across diverse manipulation types, including copy-move, splicing, and deepfakes.

2.7.2 Synthetic Data for Security Research

Generative models enable creation of large-scale synthetic datasets for training and evaluating security systems without risks associated with conducting real attacks. Frid-Adar et al. [35] demonstrated GAN-based medical image augmentation, principles applicable to generating diverse attack scenarios for AV security research. However, synthetic data must accurately reflect real attack characteristics to ensure trained systems generalize. Domain adaptation and transfer learning techniques help bridge the reality gap between synthetic training data and real-world deployment conditions [93].

2.8 Research Gaps and Opportunities

The literature review reveals several critical research gaps that this dissertation addresses:

Limited Consideration of Generative AI-Based Attacks: While adversarial ML has received

extensive attention, application of advanced generative models (particularly diffusion models) to create covert, temporally consistent vision-based attacks remains underexplored [67, 70].

Lack of Hybrid Attack Frameworks: Most research treats vision and navigation system attacks independently. Coordinated hybrid attacks exploiting sensor fusion vulnerabilities through simultaneous multi-modal manipulation represent critical gaps [83].

Insufficient Real-World Validation: Much AV security research relies on simulation without comprehensive real-world data collection and validation [55]. This dissertation addresses this gap through custom hardware platforms and urban dataset collection.

Detection Mechanism Limitations: Existing detection approaches often assume single-modality attacks or lack comprehensive forensic analysis suitable for identifying sophisticated generative AI-based manipulations [33, 86].

This dissertation addresses these gaps through comprehensive investigation of generative AI-based hybrid covert attacks and multi-modal forensic detection mechanisms, contributing both theoretical foundations and practical methodologies for enhancing AV security.

Chapter 3

Problem Definition and System Modeling

This chapter establishes the mathematical foundations and system models that underpin the research presented in this dissertation. We begin by introducing the notations and preliminaries used throughout the thesis, followed by a detailed characterization of the cyber-physical architecture of AVs. Subsequently, we formulate the lane-keeping and trajectory control problem, develop mathematical models of vehicle lateral dynamics, camera imaging systems, and communication channels, and conclude by formally defining the threat models and evaluation metrics that guide our investigation.

3.1 Notations and Preliminaries

3.1.1 Mathematical Notation

Throughout this dissertation, we adopt the following notational conventions to ensure clarity and consistency:

Scalars, Vectors, and Matrices: Scalars are denoted by lowercase or uppercase italic letters (e.g., x, A). Vectors are represented by boldface lowercase letters (e.g., $\mathbf{x}$, $\mathbf{v}$), and matrices by boldface uppercase letters (e.g., $\mathbf{A}$, $\mathbf{Q}$). The transpose of a vector or matrix is indicated by the superscript $^\top$ (e.g., $\mathbf{x}^\top$).

Sets and Spaces: $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^n$ the n-dimensional Euclidean space, and $\mathbb{R}^{m\times n}$ the space of $m \times n$ real matrices. The set of integers is denoted by $\mathbb{Z}$, and the set of natural numbers (including zero) by $\mathbb{N}_0$.

Time Indexing: Discrete time steps are indexed by $k \in \mathbb{N}_0$, with $k = 0$ representing the initial time. Continuous time is denoted by $t \in \mathbb{R}^+$. When the context is clear, we may omit explicit time arguments.

Random Variables and Expectations: Random variables are typically denoted by uppercase letters (e.g., X, Y). The expected value of a random variable X is written as $\mathbb{E}[X]$, and its variance as $\mathrm{Var}(X)$. Probability is denoted by $\mathbb{P}(\cdot)$.

Norms and Distances: The Euclidean norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$. The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$. The Mahalanobis distance with respect to covariance matrix $\boldsymbol{\Sigma}$ is defined as $D_M(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$.

Image Domain: Images are represented as functions $I : \Omega \to \mathbb{R}^c$, where $\Omega \subset \mathbb{R}^2$ is the spatial domain (typically a rectangular region corresponding to pixel coordinates), and c is the number of color channels (e.g., $c = 3$ for RGB images). For grayscale images, $c = 1$.

3.1.2 Key Assumptions

The following assumptions underlie the models and analyses presented in this dissertation:

(1) **Assumption 1:** The AV maintains a constant longitudinal velocity v_x during the lane-keeping maneuver. This assumption is standard in lateral control analysis and simplifies the dynamics by decoupling longitudinal and lateral motion [16].

(2) **Assumption 2:** Only the front wheels are controlled by the steering input δ. The rear wheels follow passively, consistent with conventional vehicle configurations [16].

(3) **Assumption 3:** The vehicle dynamics can be adequately approximated by a two-degrees-of-freedom bicycle model, which is valid for moderate lateral accelerations and at speeds where tire forces remain within the linear region [16].

(4) **Assumption 4:** The primary input to the lane-keeping controller is the output of a vision sensor (camera). While other sensors may be present, the controller relies predominantly on visual lane detection for lateral state estimation.

(5) **Assumption 5:** The adversary has the capability to intercept, modify, and inject data into both the sensor channel (camera feed) and the control channel (steering commands), consistent with sophisticated cyber-attack scenarios [3].

(6) **Assumption 6:** The physical parameters of the vehicle (mass, inertia, wheelbase, cornering stiffness) are known or can be estimated with sufficient accuracy for modeling purposes.

3.2 Cyber-Physical Architecture of an Autonomous Vehicle

AVs are exemplars of cyber-physical systems, integrating computational processes with physical dynamics through a complex network of sensors, actuators, and control algorithms. Understanding this architecture is essential for identifying vulnerabilities and designing both attacks and defenses.

3.2.1 Hierarchical Control Architecture

The control architecture of a typical AV can be decomposed into several hierarchical layers, as illustrated in Figure 3.1:

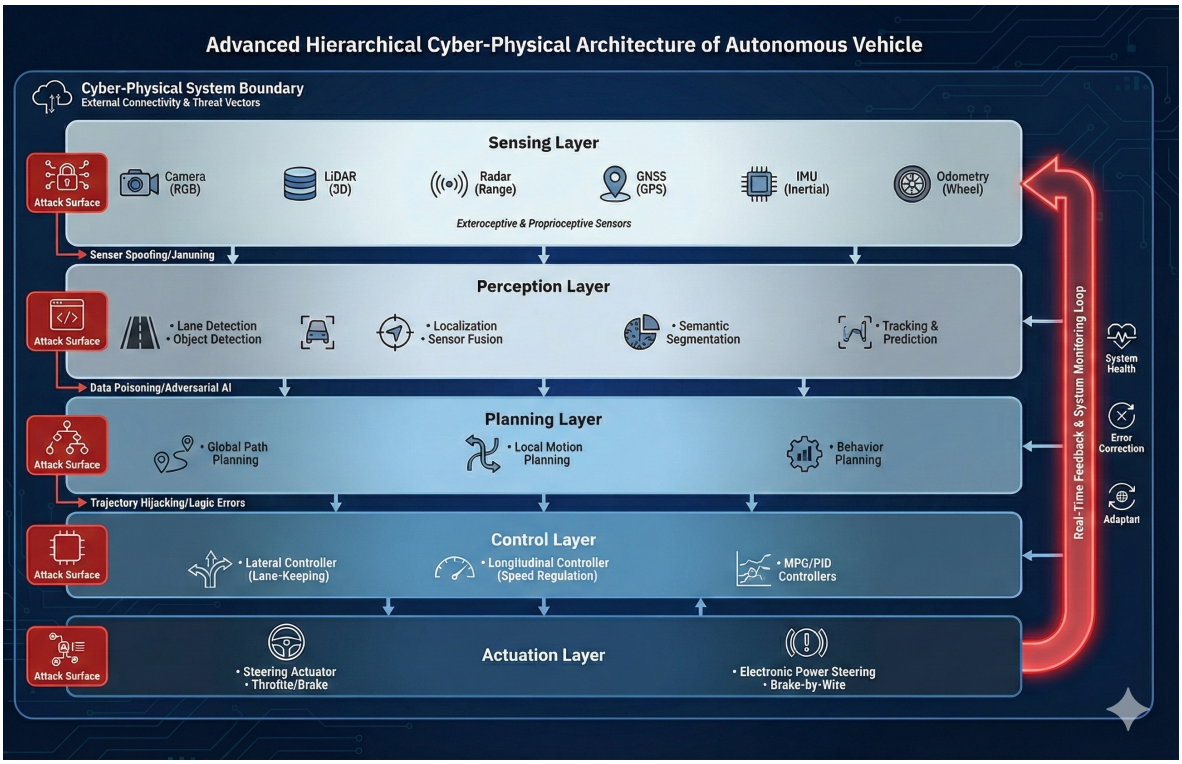


Figure 3.1: Hierarchical cyber-physical architecture of an AV, showing the data flow from sensors through perception, planning, and control to actuators.

Sensing Layer: This lowest layer comprises all physical sensors that interface with the environment. For AVs, this typically includes vision sensors such as cameras (monocular, stereo, or multi-camera arrays) capturing RGB imagery of the vehicle's surroundings, LiDAR sensors providing 3D point clouds representing distances to surrounding objects, and radar systems measuring range, velocity, and angle of nearby objects.

Additionally, the sensing layer incorporates GNSS receivers including GPS, GLONASS, Galileo, and BeiDou providing global position and velocity estimates, IMU combining accelerometers and gyroscopes to measure vehicle acceleration and angular velocity, and wheel odometry encoders measuring wheel rotation to estimate distance traveled and vehicle speed.

Perception Layer: Raw sensor data is processed by perception algorithms to extract meaningful information about the environment and vehicle state. This includes lane detection, where computer vision algorithms identify lane markings in camera images and estimate lane geometry such as

curvature, lateral offset, and heading error. Object detection and tracking employ deep neural networks (DNNs) to identify and track vehicles, pedestrians, cyclists, and other objects in the scene. Localization utilizes sensor fusion algorithms that combine GNSS, IMU, and visual/LiDAR odometry to estimate the vehicle's global position and orientation. Additionally, semantic segmentation performs dense pixel-wise classification of road surfaces, sidewalks, vegetation, buildings, and other environmental features.

Planning Layer: Based on perception outputs and high-level mission objectives, planning algorithms determine the vehicle's desired trajectory. Global path planning handles route planning from origin to destination using road networks and maps, while local motion planning focuses on trajectory generation considering dynamic obstacles, lane geometry, and traffic rules. Behavior planning manages decision-making for maneuvers such as lane changes, merging, and intersection navigation.

Control Layer: Control algorithms translate planned trajectories into actuator commands. The lateral controller computes steering angle commands to track desired lateral position and heading, while the longitudinal controller manages throttle and brake to regulate speed and maintain safe following distances.

Actuation Layer: Physical actuators execute control commands. The steering actuator comprises an electric power steering system applying commanded steering angle, while throttle and brake actuators consist of electronic throttle control and brake-by-wire systems.

3.2.2 Attack Surface and Vulnerabilities

Each layer and inter-layer communication channel in this architecture presents potential attack surfaces. For this dissertation, we focus specifically on the vision sensor channel (the data path from the camera sensor to the perception module, where images can be intercepted and manipulated [13]), the GNSS receiver, and the control actuator channel (the communication link from controller to steering actuator, where false data injection can occur [23]). The coordinated attack framework developed in this dissertation targets multiple layers simultaneously, exploiting the trust relationships between sensors and control systems.

3.3 Lane-Keeping and Trajectory Control Formulation

Lane-keeping represents a fundamental capability for AVs, ensuring that the vehicle maintains its position within lane boundaries while tracking a desired trajectory. This section formalizes the lane-keeping problem and introduces the control framework employed in this research.

3.3.1 Reference Frame and Coordinate Systems

We define three coordinate frames relevant to the lane-keeping problem. The global coordinate frame $\mathcal{F}_G$ is an inertial reference frame fixed to the Earth, typically aligned with cardinal directions using North-East-Down or East-North-Up convention, where GNSS positions are expressed. The road-aligned frame $\mathcal{F}_R$ is a curvilinear coordinate system aligned with the road centerline, where the s-axis follows the centerline tangent direction, the n-axis points laterally perpendicular to the centerline, and the z-axis is vertical. The vehicle body frame $\mathcal{F}_B$ is a coordinate system attached to the vehicle's center of gravity (CG), with the x-axis pointing forward, y-axis pointing laterally (typically to the driver's left), and z-axis pointing upward.

The lateral state of the vehicle with respect to the lane centerline is characterized by two key parameters: lateral deviation d, defined as the perpendicular distance from the vehicle's CG to the nearest point P on the lane centerline (measured positive to the left), and relative heading error ϕ, defined as the angular difference between the vehicle's actual heading ψ_a and the tangent to the centerline at point P. The relative heading error is expressed mathematically as:

$$\phi = \psi_a - \psi_d, \tag{1}$$

where ψ_d is the desired heading tangent to the centerline. Figure 3.2 illustrates these geometric relationships.

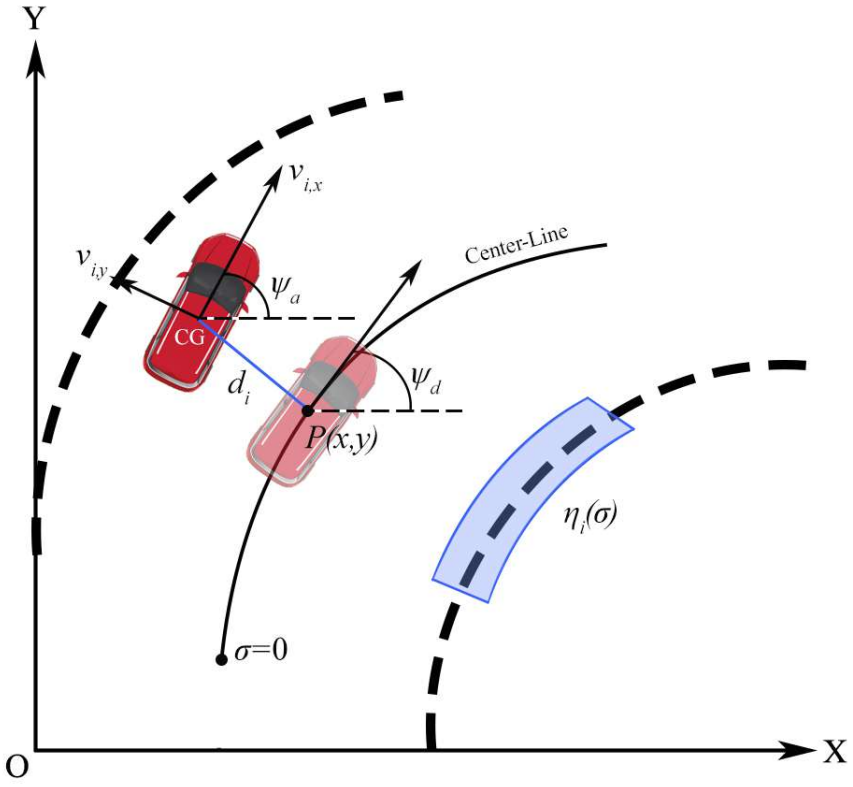


Figure 3.2: Geometric representation of the lateral state variables: lateral deviation d and relative heading error ϕ.

3.3.2 Lane-Keeping Control Objective

The primary objective of the lane-keeping controller is to asymptotically drive the lateral state to zero while maintaining passenger comfort and system stability. Mathematically, this is expressed as:

$$\lim_{t\to\infty} d(t) = 0, \quad \lim_{t\to\infty} \phi(t) = 0, \tag{2}$$

subject to constraints on the control input (steering angle), lateral acceleration, and other comfort-related metrics. In a multi-vehicle scenario with formation control objectives, additional constraints ensure desired inter-vehicle spacing and relative positioning are maintained. However, the focus of this dissertation is on the lane-keeping aspect, with formation control serving as a secondary objective in simulation scenarios.

3.4 Mathematical Modeling of Lateral Dynamics

3.4.1 Bicycle Model Derivation

We employ a two-degrees-of-freedom (2-DOF) bicycle model to represent the lateral and yaw dynamics of the AV. This model simplifies the four-wheeled vehicle by lumping the left and right wheels on each axle into single equivalent wheels, resulting in a two-wheeled representation resembling a bicycle [16].

The state variables in the bicycle model are lateral velocity v_y of the vehicle's CG (positive leftward), yaw angle ψ (heading) of the vehicle, and yaw rate $r = \dot{\psi}$ (angular velocity about the vertical axis). The control input is steering angle δ of the front wheels (positive for left turn). The vehicle parameters include total vehicle mass m, moment of inertia about the yaw axis I_z, distance from CG to front axle r_1, distance from CG to rear axle r_2, cornering stiffness of the front tires c_1, cornering stiffness of the rear tires c_2, and longitudinal velocity v_x (assumed constant as per Assumption 2).

The lateral and yaw dynamics are governed by the following coupled differential equations [16]:

$$\dot{v}_y = -\frac{r_1 c_1 + r_2 c_2}{m v_x} r - \frac{c_1 + c_2}{m v_x} v_y + \frac{c_1}{m} \delta \tag{3}$$

$$\dot{r} = -\frac{r_1^2 c_1 + r_2^2 c_2}{I_z v_x} r - \frac{r_2 c_2 - r_1 c_1}{I_z v_x} v_y + \frac{r_1 c_1}{I_z} \delta. \tag{4}$$

These equations describe how the lateral velocity and yaw rate evolve in response to the steering input δ, accounting for the tire forces generated at the front and rear axles.

3.4.2 Kinematic Relationship to Lateral State

To relate the dynamic states (v_y, r) to the lane-keeping error states (d, ϕ), we use the following kinematic relationships derived from the geometry of the road-aligned frame:

$$\dot{d} = v_y + v_x \sin(\phi) \approx v_y + v_x \phi \tag{5}$$

$$\dot{\phi} = r - \kappa v_x, \tag{6}$$

where κ is the curvature of the road centerline at the projection point P, defined as the inverse of the radius of curvature R:

$$\kappa = \frac{1}{R}. \tag{7}$$

The small-angle approximation $\sin(\phi) \approx \phi$ is valid when the heading error is small, which is typically the case in well-controlled lane-keeping scenarios.

3.4.3 State-Space Representation

For control design purposes, it is convenient to express the system dynamics in state-space form. Define the augmented state vector:

$$\mathbf{x} = \begin{bmatrix} d \\ \phi \\ v_y \\ r \end{bmatrix} \in \mathbb{R}^4. \tag{8}$$

Combining Equations (3)–(6), the continuous-time state-space model is:

$$\dot{\mathbf{x}} = \mathbf{A}_c \mathbf{x} + \mathbf{B}_c \delta + \mathbf{E}_c \kappa, \tag{9}$$

where the system matrices are:

$$\mathbf{A}_c = \begin{bmatrix} 0 & v_x & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\frac{c_1+c_2}{m v_x} & -\frac{r_1 c_1 + r_2 c_2}{m v_x} \\ 0 & 0 & -\frac{r_2 c_2 - r_1 c_1}{I_z v_x} & -\frac{r_1^2 c_1 + r_2^2 c_2}{I_z v_x} \end{bmatrix} \tag{10}$$

$$\mathbf{B}_c = \begin{bmatrix} 0 \\ 0 \\ \frac{c_1}{m} \\ \frac{r_1 c_1}{I_z} \end{bmatrix}, \quad \mathbf{E}_c = \begin{bmatrix} 0 \\ -v_x \\ 0 \\ 0 \end{bmatrix}. \tag{11}$$

3.4.4 Discretization for Digital Implementation

Since AV controllers are implemented digitally with discrete sampling intervals T_s, we discretize the continuous-time model using the zero-order hold (ZOH) method. The discrete-time state-space model is:

$$\mathbf{x}_{k+1} = \mathbf{A}_d \mathbf{x}_k + \mathbf{B}_d \delta_k + \mathbf{E}_d \kappa_k, \tag{12}$$

where the discrete-time matrices are:

$$\mathbf{A}_d = e^{\mathbf{A}_c T_s}, \quad \mathbf{B}_d = \left(\int_0^{T_s} e^{\mathbf{A}_c \tau} d\tau \right) \mathbf{B}_c, \quad \mathbf{E}_d = \left(\int_0^{T_s} e^{\mathbf{A}_c \tau} d\tau \right) \mathbf{E}_c. \tag{13}$$

For typical sampling intervals ($T_s = 0.01$ to 0.1 seconds), numerical methods such as matrix exponential computation or Euler/Tustin approximations can be employed to obtain $\mathbf{A}_d$, $\mathbf{B}_d$, and $\mathbf{E}_d$.

3.5 Vision Sensor and Camera Imaging Model

3.5.1 Camera Pinhole Model

The camera mounted on the AV captures 2D images of the 3D environment. The standard pinhole camera model relates 3D world coordinates to 2D image coordinates via a perspective projection.

Let $\mathbf{P}_w = [X_w, Y_w, Z_w]^\top$ be a 3D point in the world coordinate frame, and $\mathbf{p} = [u, v]^\top$ its projection onto the image plane with pixel coordinates. The projection is given by:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}, \tag{14}$$

where λ is the depth (distance from camera to point), $\mathbf{K}$ is the 3×3 intrinsic matrix encoding focal

length and principal point:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{15}$$

and $[\mathbf{R} \mid \mathbf{t}]$ is the 3×4 extrinsic matrix encoding the rotation $\mathbf{R}$ and translation $\mathbf{t}$ from world frame to camera frame.

3.5.2 Lane Detection from Camera Images

The lane-keeping controller relies on extracting lane geometry parameters (lateral deviation d and heading error ϕ) from camera images. This process, illustrated in Figure 3.3, involves several stages.

In the image preprocessing stage, the raw image I is denoised using Gaussian blurring to remove high-frequency noise:

$$I_{\text{smooth}} = G_\sigma * I, \tag{16}$$

where G_σ is a Gaussian kernel with standard deviation σ, and $*$ denotes convolution.

For lane marking segmentation, binary thresholding in appropriate color spaces such as the HSL L-channel identifies potential lane markings:

$$I_{\text{binary}}(u, v) = \begin{cases} 1 & \text{if } I_{\text{smooth}}(u, v) > T_{\text{thresh}} \\ 0 & \text{otherwise} \end{cases}. \tag{17}$$

Gradient-based edge detection further refines lane candidate pixels:

$$\nabla I = \sqrt{\left(\frac{\partial I}{\partial u}\right)^2 + \left(\frac{\partial I}{\partial v}\right)^2}. \tag{18}$$

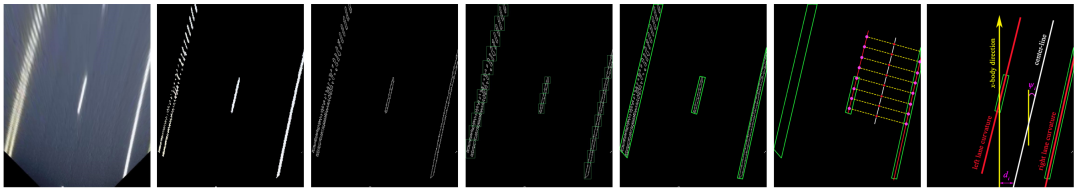


Figure 3.3: Lane detection pipeline: (a) Noise cancellation, (b) Binary thresholding, (c) Gradient detection, (d) Sliding window detection, (e) Lane curvature calculation, (f) Center line estimation.

During lane polynomial fitting, detected lane pixels are fit to polynomial curves (typically 2nd-order) representing left and right lane boundaries:

$$u_{\text{left}}(v) = a_{\text{left}}v^2 + b_{\text{left}}v + c_{\text{left}} \tag{19}$$

$$u_{\text{right}}(v) = a_{\text{right}}v^2 + b_{\text{right}}v + c_{\text{right}}. \tag{20}$$

In the parameter extraction stage, the centerline is computed as the average of left and right lanes, and the lateral deviation d and heading error ϕ are extracted from the polynomial coefficients and vehicle position in the image. The mapping from image I to lateral parameters can be abstractly represented as:

$$(d, \phi) = \mathcal{H}(I, \boldsymbol{\theta}), \tag{21}$$

where $\mathcal{H}$ is the perception function (implemented via traditional computer vision or DNNs), and $\boldsymbol{\theta}$ represents internal parameters such as neural network weights.

3.6 Communication and Navigation Channel Vulnerabilities

3.6.1 GNSS Signal Model

Global Navigation Satellite Systems provide position and velocity estimates by measuring pseudo-ranges to multiple satellites. Each satellite i broadcasts a signal with known ephemeris (orbital parameters), allowing the receiver to compute the satellite's position $\mathbf{s}^i(t)$ at transmission

time.

The measured pseudo-range ρ^i is:

$$\rho^i = \|\mathbf{p}_{\text{true}} - \mathbf{s}^i\| + c \cdot b + n^i, \tag{22}$$

where $\mathbf{p}_{\text{true}}$ is the true receiver position, c is the speed of light, b is the receiver clock bias, and n^i is measurement noise.

The Doppler shift f_D^i observed in satellite i's signal is related to the relative velocity between satellite and receiver:

$$f_D^i = \frac{f_L}{c}(\mathbf{v}^i - \mathbf{v}_{\text{true}})^\top \mathbf{u}^i, \tag{23}$$

where f_L is the carrier frequency (1575.42 MHz for GPS L1), $\mathbf{v}^i$ is the satellite velocity, $\mathbf{v}_{\text{true}}$ is the true receiver velocity, and $\mathbf{u}^i = (\mathbf{s}^i - \mathbf{p}_{\text{true}})/\|\mathbf{s}^i - \mathbf{p}_{\text{true}}\|$ is the unit line-of-sight vector.

3.6.2 GPS Spoofing Attack Model

A GPS spoofing attack broadcasts counterfeit satellite signals that cause the receiver to compute an incorrect position $\mathbf{p}_{\text{spoof}}$ and velocity $\mathbf{v}_{\text{spoof}}$. Let the spoofing deviation be:

$$\mathbf{s}_k = \mathbf{p}_{\text{spoof},k} - \mathbf{p}_{\text{true},k}. \tag{24}$$

A covert spoofing attack must satisfy two constraints. First, the bounded deviation constraint ensures that the instantaneous spoofing vector magnitude is limited to avoid triggering threshold-based detectors:

$$\|\mathbf{s}_k\| \leq s_{\max}. \tag{25}$$

Second, the Doppler consistency constraint requires that the spoofed Doppler shifts must be consistent with the spoofed kinematics to evade velocity-based detection:

$$f_{D,\text{spoof}}^i = \frac{f_L}{c}(\mathbf{v}^i - \mathbf{v}_{\text{spoof}})^\top \mathbf{u}_{\text{spoof}}^i + \epsilon^i, \tag{26}$$

where $\mathbf{u}^i_{\text{spoof}} = (\mathbf{s}^i - \mathbf{p}_{\text{spoof}})/\|\mathbf{s}^i - \mathbf{p}_{\text{spoof}}\|$, and ϵ^i is a small residual error kept below detection thresholds.

3.7 Definition of Attack Models

3.7.1 Vision-Based Covert Attack Model

The vision-based covert attack (VCA) targets the camera sensor channel, intercepting the authentic image I_k at time step k and replacing it with a synthetically generated image $\tilde{I}_k$ before it reaches the perception module. The attack model is formally defined as:

$$\tilde{I}_k = \mathcal{G}(I_k, z_k, \boldsymbol{\xi}_k; \boldsymbol{\Theta}), \tag{27}$$

where $\mathcal{G}$ is the generative model (GAN or diffusion model), $z_k \sim \mathcal{N}(0, \mathbf{I})$ is a random noise vector, $\boldsymbol{\xi}_k$ represents conditioning information such as lane geometry and deviation embeddings, and $\boldsymbol{\Theta}$ denotes the trained model parameters.

The attack success criterion is the **geometric alignment constraint**:

$$C(\tilde{I}_k) \leq \epsilon_c, \tag{28}$$

where $C(\cdot)$ measures the pixel-level lane-center deviation in the image plane, and ϵ_c is a sub-pixel tolerance (e.g., 1-2 pixels).

Additionally, the attack must maintain **perceptual realism** to avoid detection by human observers or forensic analysis:

$$\text{SSIM}(\tilde{I}_k, I_{\text{ref},k}) \geq \tau_{\text{SSIM}}, \tag{29}$$

where SSIM is the Structural Similarity Index [94], $I_{\text{ref},k}$ is a reference authentic image, and τ_{SSIM} is a threshold (typically ¿ 0.9).

3.7.2 GNSS Spoofing Model

The GNSS spoofing model, denoted GPS Spoof-Net, generates counterfeit satellite signals that manipulate the receiver's position solution. The spoofed trajectory is designed as a gradual drift:

$$\mathbf{p}_{\text{spoof},k} = \mathbf{p}_{\text{true},k} + \mathbf{s}_k \tag{30}$$

$$\mathbf{s}_k = \mathbf{s}_{k-1} + \Delta\mathbf{s}_k, \quad \|\Delta\mathbf{s}_k\| \leq \Delta s_{\max}. \tag{31}$$

The spoofed velocity must be kinematically consistent:

$$\mathbf{v}_{\text{spoof},k} = \frac{\mathbf{p}_{\text{spoof},k} - \mathbf{p}_{\text{spoof},k-1}}{T_s}, \tag{32}$$

and the Doppler shifts for each satellite $i = 1, \ldots, M$ are computed according to Equation (26) to maintain Doppler consistency.

3.7.3 Multi-Channel Hybrid Attack Scenario

The hybrid covert attack (HCA) combines vision-based manipulation, GPS spoofing, and steering bias injection to create a coordinated multi-sensor deception. The complete attack model at time step k is:

$$\mathbf{a}_k = (\delta_k + \Delta\delta_k, \tilde{I}_k, \mathbf{p}_{\text{spoof},k}), \tag{33}$$

where $\Delta\delta_k$ is the false steering bias injected into the actuator channel, $\tilde{I}_k$ is the synthesized camera image from VCA, and $\mathbf{p}_{\text{spoof},k}$ is the spoofed GPS position.

The steering bias causes actual lateral deviation:

$$\delta_k = \delta_{\text{nominal},k} + \Delta\delta_k, \quad |\Delta\delta_k| \leq \Delta\delta_{\max}. \tag{34}$$

The vision-based attack conceals this deviation by making the vehicle appear centered in the camera feed:

$$\mathcal{H}(\tilde{I}_k, \boldsymbol{\theta}) \approx (0, 0). \tag{35}$$

Simultaneously, the spoofed GPS trajectory aligns with the falsely perceived centered state, maintaining consistency across sensor modalities and evading sensor fusion-based anomaly detection.

The attack effectiveness is measured by the actual lateral deviation $d_{\text{actual},k}$ achieved while keeping the perceived lateral deviation $d_{\text{perceived},k}$ near zero:

$$\text{Attack Success} = \left(|d_{\text{actual},k}| > d_{\min}\right) \wedge \left(|d_{\text{perceived},k}| < \epsilon_d\right), \tag{36}$$

where $d_{\min}$ is a minimum deviation threshold indicating dangerous trajectory divergence (e.g., 0.5 meters), and ϵ_d is the perception tolerance (e.g., 0.1 meters).

3.8 Evaluation Metrics

3.8.1 Performance Metrics

To validate these hypotheses, we define quantitative evaluation metrics across three categories: attack effectiveness, attack covertness, and detection performance.

Attack Effectiveness Metrics

Attack effectiveness is measured through lateral deviation magnitude, which captures the maximum and average actual lateral deviation $|d_{\text{actual}}|$ achieved during the attack period. The attack success rate represents the percentage of time steps where the attack success criterion (Eq. 36) is satisfied. Additionally, trajectory drift rate measures the rate of change of lateral deviation as $\frac{d|d_{\text{actual}}|}{dt}$.

Attack Covertness Metrics

Attack covertness is evaluated through multiple image quality metrics. The Structural Similarity Index (SSIM) measures structural similarity as $\text{SSIM}(\tilde{I}, I_{\text{ref}}) \in [0, 1]$, while the Peak Signal-to-Noise Ratio (PSNR) is calculated as $\text{PSNR} = 10 \log_{10} \frac{\text{MAX}_I^2}{\text{MSE}}$. The Learned Perceptual Image Patch Similarity (LPIPS) provides perceptual distance as $\text{LPIPS}(\tilde{I}, I_{\text{ref}})$. Temporal consistency measures frame-to-frame perceptual dissimilarity, and perceived lateral deviation is quantified as $|d_{\text{perceived}}| = |\mathcal{H}(\tilde{I})|_1$, where $|\cdot|_1$ denotes the L1 norm of the lateral parameter vector.

Detection Performance Metrics

Detection performance is characterized through several key metrics. The true positive rate (TPR) represents the probability of correctly detecting an attack when present as TPR = $P(\text{alarm} \mid \text{attack})$, while the false positive rate (FPR) measures the probability of false alarm when no attack is present as FPR = $P(\text{alarm} \mid \text{no attack})$. The area under the ROC curve (AUC) provides a summary metric of detector discrimination capability across all threshold settings. Detection delay quantifies the time from attack onset to first alarm as $T_{\text{det}} = \min\{k : \mathcal{D}_k = 1, k \geq k_0\} - k_0$, where k_0 is the attack start time and $\mathcal{D}_k$ is the detection decision. Precision and recall are defined as Precision = $\frac{TP}{TP+FP}$ and Recall = $\frac{TP}{TP+FN}$, respectively.

The following chapters present the detailed design, implementation, and evaluation of the attack frameworks (VCA-HADS, Gen-VCA, VideoDiff-VCA with GPS Spoof-Net) and corresponding detection mechanisms (HADS, multi-modal forensic detection), using the models, metrics, and validation methodologies established in this chapter.

3.9 Summary

This chapter established the mathematical and system-level foundations for the research presented in this dissertation. We formalized the cyber-physical architecture of AVs, developed detailed models of vehicle lateral dynamics, vision sensor processing, and GNSS navigation, and defined precise threat models for vision-based covert attacks, GPS spoofing, and hybrid multi-channel attacks. The evaluation metrics introduced here provide the quantitative basis for assessing attack effectiveness, covertness, and detectability in subsequent chapters. Together, these models and definitions create a rigorous foundation for investigating the security of AV perception and navigation systems against sophisticated adversarial manipulation.

Chapter 4

Vision-Based Covert Attack Using GAN (VCA-HADS)

4.1 Introduction

The integration of deep learning-based perception systems in AVs has revolutionized their capability to interpret complex road environments. However, this reliance on neural network-based vision processing simultaneously introduces critical vulnerabilities to adversarial manipulation. While traditional cyber-attacks on vehicular systems have focused on communication protocols and control networks, the emergence of sophisticated generative artificial intelligence models has enabled a new class of threats targeting the perception layer itself [12, 24].

This chapter presents the first comprehensive framework for vision-based covert attacks on AV lane-keeping systems using GANs. Unlike previous adversarial attack research that primarily demonstrates misclassifications of isolated objects (stop signs, traffic lights, pedestrians), our approach targets the continuous control loop of lane-keeping assistance systems, where subtle manipulation of perceived lane geometry can induce dangerous trajectory deviations while avoiding detection by both human observers and automated anomaly detectors [13]. Top-view representations, obtained through homography transformation of front-facing camera feeds, provide several advantages for lane-keeping attacks:

(1) **Simplified Geometric Manipulation:** Lane markings appear as approximately parallel lines in top-view, enabling straightforward geometric transformations without complex 3D reasoning.

(2) **Direct Measurement Correspondence:** Pixel coordinates in top-view space have direct linear relationships to real-world lateral distances, simplifying the mapping between image-space manipulations and actual vehicle deviations.

(3) **Controller Compatibility:** Many lane-keeping algorithms operate directly on bird's-eye view inputs, making top-view manipulation a natural attack surface.

(4) **Controlled Data Collection:** Simulation environments readily generate paired top-view images under diverse scenarios, enabling systematic dataset construction for GAN training.

However, the top-view approach also introduces important limitations that motivate subsequent research directions:

- **Loss of 3D Information:** Homography assumes planar road surfaces, discarding elevation changes, road banking, and vertical obstacle information critical for comprehensive scene understanding.
- **Limited Field of View:** Top-view transformation typically shows only 20-30 meters ahead due to increasing distortion at far distances, reducing the controller's preview horizon compared to front-view representations.
- **Transformation Artifacts:** Distant objects (vehicles, road markings) exhibit unrealistic stretching and distortion in top-view, creating potential forensic signatures.
- **Simulation-to-Real Domain Gap:** While simulation enables controlled experimentation, generalization to real-world conditions with authentic sensor characteristics remains challenging.

The fundamental challenge addressed in this chapter is the creation of synthetic road scene imagery that simultaneously satisfies three conflicting objectives: (1) inducing significant actual lateral deviation of the vehicle from the lane centerline, (2) presenting a falsified perception to the lane-keeping controller that the vehicle remains properly centered, and (3) maintaining sufficient

photorealistic quality that the manipulated imagery appears authentic to conventional image quality metrics. Traditional adversarial perturbation methods, while effective at causing misclassification, produce visually conspicuous artifacts that are easily detected [21]. In contrast, our GAN-based approach generates entirely new imagery that preserves semantic coherence and geometric plausibility while embedding adversarial objectives.

The VCA developed in this chapter employs a customized GAN architecture that learns to perform view synthesis, transforming an authentic top-view image captured from the vehicle's actual (deviated) position into a synthesized image representing the view that would be observed if the vehicle were centered in the lane. This transformation must account for lane geometry, vehicle positioning, road curvature, and the presence of other vehicles while maintaining photorealistic quality. The generator network integrates spatial transformer modules that enable explicit geometric manipulation, while adversarial training against a discriminator network ensures that synthetic images are indistinguishable from authentic road scene imagery.

Complementing the attack framework, this chapter also introduces a HADS that fuses multiple information sources to identify instances where the vehicle's behavior deviates from expectations based on perceived lane geometry. HADS integrates a DNN analyzing temporal patterns in the vision stream, GPS-based trajectory monitoring that detects inconsistencies between perceived and actual vehicle paths, and map-based validation that cross-references reported positions against known road network topology. This multi-modal detection approach recognizes that sophisticated attacks targeting a single sensor modality may still produce detectable inconsistencies when compared against independent information sources.

The contributions presented in this chapter include:

(1) A novel GAN architecture specifically designed for covert adversarial top-view synthesis in automotive contexts, incorporating encoder-shared layers-decoder structure with ResNet modules, following the precise architectural specifications illustrated in Figure 4.2.

(2) Mathematical formulation of the covert attack optimization problem, establishing the geometric constraints and quality requirements that define successful attacks within the top-view representation framework.

(3) Experimental evaluation demonstrating that VCA achieves lateral deviations exceeding 0.94 meters (average) while maintaining SSIM ¿ 0.92 and bypassing baseline anomaly detectors for extended periods in simulation.

(4) HADS detection achieving 95.8% TPR with 4.2% FPR, establishing feasibility of defending against sophisticated vision-based attacks through multi-modal sensor fusion.

The remainder of this chapter is organized as follows: Section 4.2 presents the detailed design of the GAN-based covert attack, including network architecture with exact layer dimensions, loss functions, and training methodology. Section 4.3 describes the HADS detection and its component modules. Section 4.4 details the simulation environment and dataset collection procedures. Section 4.5 presents experimental results evaluating attack effectiveness, stealth, and detection performance. Finally, section 4.6 discusses implications, top-view limitations, and directions toward real-world deployment.

4.2 Design of GAN-Based Covert Attack

The Vision-Based Covert Attack operates by intercepting the camera feed between the image sensor and the perception module, replacing authentic top-view images with synthetically generated counterfeits. The attack workflow, illustrated in Figure 4.1, proceeds as follows.

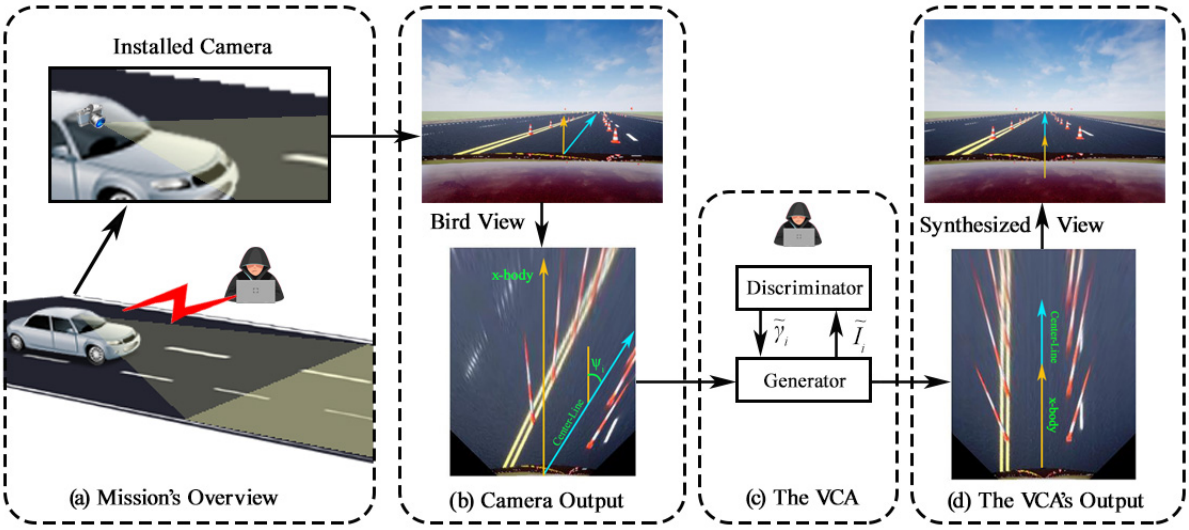


Figure 4.1: Vision-Based Covert Attack workflow for top-view image manipulation showing interception, generation, and injection into the perception pipeline.

In the first step, steering bias injection, the attacker injects a small, gradual steering bias $\Delta\delta_k$ into the control channel, causing the vehicle to deviate laterally from the lane centerline:

$$\delta_k^{\text{actual}} = \delta_k^{\text{cmd}} + \Delta\delta_k, \tag{37}$$

where δ_k^{cmd} is the controller's commanded steering angle, and δ_k^{actual} is the actual steering angle applied to the vehicle. The bias is constrained to remain small at each time step ($|\Delta\delta_k| \leq 0.02$ radians) to avoid triggering actuator anomaly detectors. The accumulated effect over time induces significant lateral deviation despite the small per-step magnitude.

In the second step, the front-mounted camera captures the authentic front-view image at 640×480 resolution. A homography transformation $\mathbf{H}$ projects this to top-view representation:

$$I_k^{\text{top-view}} = \mathcal{T}_H(I_k^{\text{front-view}}), \tag{38}$$

where $\mathcal{T}_H$ applies the perspective transformation matrix computed during camera calibration. The resulting top-view image (640×480×3-bit) represents the lane geometry relative to the deviated

vehicle state, showing lateral displacement and heading error.

In the third step, the attacker's GAN-based generator G transforms the authentic top-view image into a synthetic image $\tilde{I}_k$ representing the hypothetical view from a centered position:

$$\tilde{I}_k = G(I_k^{\text{top-view}}, \mathbf{c}_k; \boldsymbol{\Theta}_G), \tag{39}$$

where $\mathbf{c}_k$ is conditioning information encoding the desired transformation (specifically, the inverse of the actual deviation), and $\boldsymbol{\Theta}_G$ represents the trained generator parameters.

In the fourth step, the synthetic top-view image $\tilde{I}_k$ replaces the authentic image before reaching the perception module. This substitution could be achieved through compromised camera drivers, manipulated image preprocessing pipelines, or man-in-the-middle attacks on the sensor data bus.

In the fifth step, controller deception, the lane detection algorithm processes the synthetic top-view image, extracting falsified lane parameters:

$$(\tilde{d}_k, \tilde{\phi}_k) = \mathcal{H}(\tilde{I}_k, \boldsymbol{\theta}). \tag{40}$$

By design, the synthetic image is constructed such that the extracted parameters indicate near-zero lateral deviation and heading error: $(\tilde{d}_k, \tilde{\phi}_k) \approx (0, 0)$, despite the vehicle's actual significant deviation.

In the final step, control inaction, the controller believes the vehicle to be properly centered based on the falsified perception and computes steering commands that fail to correct the actual deviation. The injected steering bias continues driving the vehicle further from the centerline, while the controller remains deceived by synthetic imagery. This attack cycle repeats at the perception system's frame rate, accumulating lateral deviation over time until the vehicle reaches unsafe proximity to lane boundaries or adjacent obstacles.

4.3 Generator and Discriminator Architecture

4.3.1 Generator Network Architecture

The generator network G performs a challenging transformation: given an input image $I \in \mathbb{R}^{640\times480\times3}$ captured from an off-center vehicle position, it synthesizes a realistic image representing

the view from a centered position. The generator architecture, illustrated in Figure 4.2, follows a U-Net encoder-decoder structure with skip connections and intermediate ResNet modules for feature refinement.

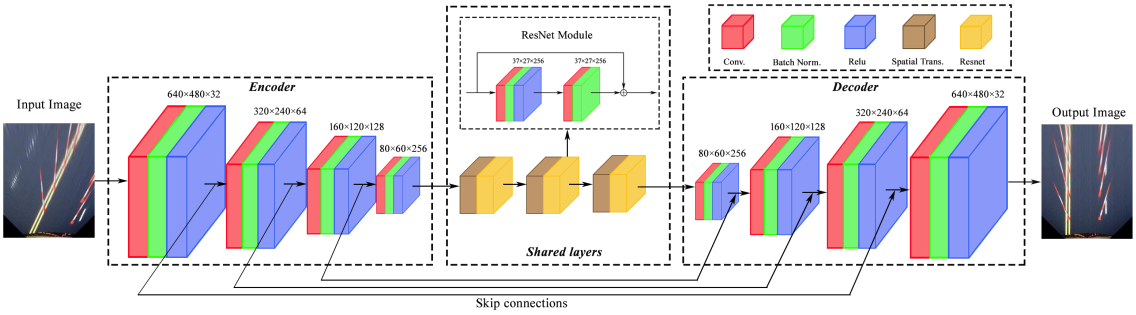


Figure 4.2: Generator network architecture showing encoder progressively reducing spatial resolution from 640×480 to 80×60, shared layers with three ResNet modules maintaining 80×60×256 dimensions for feature refinement, and decoder with skip connections reconstructing full 640×480×3 output.

Encoder Module

The encoder compresses the input image $I \in \mathbb{R}^{640\times480\times3}$ through four convolutional blocks with progressive spatial downsampling and channel expansion. Each layer applies 4×4 convolutions with stride 2, effectively halving spatial dimensions while expanding feature channels to capture increasingly abstract representations.

Encoder Layer 1 transforms the RGB input from 640×480×3 to 320×240×64:

$$\mathbf{h}_1 = \text{LeakyReLU}(\text{BN}(\text{Conv}_{4\times4,s=2,p=1}(I))). \tag{41}$$

Encoder Layer 2 continues downsampling from 320×240×64 to 160×120×128:

$$\mathbf{h}_2 = \text{LeakyReLU}(\text{BN}(\text{Conv}_{4\times4,s=2,p=1}(\mathbf{h}_1))). \tag{42}$$

Encoder Layer 3 further reduces spatial dimensions from 160×120×128 to 80×60×256:

$$\mathbf{h}_3 = \mathbf{z} = \text{LeakyReLU}(\text{BN}(\text{Conv}_{4\times4,s=2,p=1}(\mathbf{h}_2))), \tag{43}$$

where $\text{Conv}_{k\times k,s=s,p=p}$ denotes convolution with kernel size $k \times k$, stride s, and padding p; BN represents batch normalization; and LeakyReLU uses negative slope $\alpha = 0.2$. The encoder progressively captures multi-scale patterns from low-level road texture to high-level semantic concepts such as lane boundaries, vehicles, and road geometry.

Shared Layers Module with ResNet Blocks

The shared layers consist of three sequential ResNet (Residual Network) modules positioned at the bottleneck of the U-Net architecture, each maintaining spatial dimensions 80×60×256. As shown in the center of Figure 4.2, these ResNet blocks enable deeper feature processing through identity skip connections that alleviate vanishing gradient problems [64]. Each ResNet block follows the standard residual structure:

$$\mathbf{r}_1 = \text{ReLU}(\text{BN}(\text{Conv}_{3\times3,s=1,p=1}(\mathbf{z}))) \tag{44}$$

$$\mathbf{r}_2 = \text{BN}(\text{Conv}_{3\times3,s=1,p=1}(\mathbf{r}_1)) \tag{45}$$

$$\mathbf{r}_{\text{out}} = \text{ReLU}(\mathbf{r}_2 + \mathbf{z}), \tag{46}$$

where the identity skip connection $\mathbf{r}_2 + \mathbf{z}$ allows gradients to flow directly through the network during backpropagation, enabling effective training of deeper architectures. Three sequential ResNet blocks refine the encoded representation:

$$\mathbf{z}_{\text{refined}} = \text{ResNet}_3(\text{ResNet}_2(\text{ResNet}_1(\mathbf{z}))). \tag{47}$$

These ResNet blocks serve a critical role in the attack framework by refining feature representations while preserving spatial structure, essential for maintaining geometric relationships in lane markings and road boundaries during the view transformation from deviated to centered positions.

Decoder Module

The decoder reconstructs the full-resolution synthetic image through transposed convolutions with progressive upsampling. As illustrated in Figure 4.2, skip connections from corresponding encoder layers are concatenated at each decoder stage, enabling the network to recover fine-grained spatial details lost during encoding.

Decoder Layer 1 upsamples from 80×60×256 to 160×120×128. The skip connection from Encoder Layer 2 ($\mathbf{h}_2$) with dimensions 160×120×128 is concatenated, yielding a combined input of 160×120×384 before processing:

$$\tilde{\mathbf{h}}_1 = \text{ReLU}(\text{BN}(\text{ConvTranspose}_{4\times4, s=2, p=1}([\mathbf{z}_{\text{refined}}, \mathbf{h}_2]))). \tag{48}$$

Decoder Layer 2 upsamples from 160×120×128 to 320×240×64. The skip connection from Encoder Layer 1 ($\mathbf{h}_1$) with dimensions 320×240×64 is concatenated, producing 320×240×192:

$$\tilde{\mathbf{h}}_2 = \text{ReLU}(\text{BN}(\text{ConvTranspose}_{4\times4, s=2, p=1}([\tilde{\mathbf{h}}_1, \mathbf{h}_1]))). \tag{49}$$

Decoder Layer 3 produces the final RGB output, upsampling from 320×240×64 to 640×480×3:

$$\tilde{I} = \text{Tanh}(\text{ConvTranspose}_{4\times4, s=2, p=1}(\tilde{\mathbf{h}}_2)), \tag{50}$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation along the feature dimension. The Tanh activation function produces pixel values in the range $[-1, 1]$, which are subsequently rescaled to $[0, 255]$ for standard RGB image representation. Skip connections enable the decoder to access fine-grained spatial information from the encoder at multiple resolution levels, critical for reconstructing high-frequency details including lane marking edges, road texture patterns, and other vehicles with precise geometric alignment.

4.3.2 Discriminator Network Architecture

The discriminator network D evaluates whether a given image is authentic (from the real-world dataset) or synthetic (generated by G). Following the PatchGAN architecture [95], the discriminator

produces a spatial map of authenticity predictions rather than a single scalar output, enabling patch-level adversarial feedback that improves local texture realism.

The discriminator processes input images $I \in \mathbb{R}^{640\times480\times3}$ through five convolutional layers with progressive downsampling:

$$\mathbf{d}_1 = \text{LeakyReLU}(\text{Conv}_{4\times4,s=2,p=1}(I)) \qquad 640\times480\times3 \rightarrow 320\times240\times64 \tag{51}$$

$$\mathbf{d}_2 = \text{LeakyReLU}(\text{BN}(\text{Conv}_{4\times4,s=2,p=1}(\mathbf{d}_1))) \qquad 320\times240\times64 \rightarrow 160\times120\times128 \tag{52}$$

$$\mathbf{d}_3 = \text{LeakyReLU}(\text{BN}(\text{Conv}_{4\times4,s=2,p=1}(\mathbf{d}_2))) \qquad 160\times120\times128 \rightarrow 80\times60\times256 \tag{53}$$

$$\mathbf{d}_4 = \text{LeakyReLU}(\text{BN}(\text{Conv}_{4\times4,s=1,p=1}(\mathbf{d}_3))) \qquad 80\times60\times256 \rightarrow 80\times60\times512 \tag{54}$$

$$D(I) = \text{Sigmoid}(\text{Conv}_{4\times4,s=1,p=1}(\mathbf{d}_4)) \qquad 80\times60\times512 \rightarrow 80\times60\times1. \tag{55}$$

The final output $D(I) \in \mathbb{R}^{80\times60\times1}$ is a spatial authenticity map where each element provides a probability score indicating whether the corresponding receptive field patch (approximately 70×70 pixels in the input image) appears authentic or synthetic. This patch-level discrimination proves particularly effective for road scene imagery, as it enforces local texture consistency in regions such as asphalt surfaces, lane markings, and vehicle appearances, areas where global discriminators may overlook subtle synthesis artifacts.

Additionally, the discriminator incorporates geometric feedback by measuring lane geometry in processed images. After applying lane detection algorithms to extract lane boundary positions, the discriminator computes lateral deviation and heading error parameters $\tilde{\gamma}_k = (\tilde{d}_k, \tilde{\phi}_k)$, which contribute to the geometric consistency loss term guiding the generator toward producing geometrically plausible centered views.

4.3.3 Loss Function Design

The generator is trained to minimize a composite loss function balancing adversarial objectives, geometric alignment, and perceptual quality.

Adversarial Loss

Using the least-squares GAN (LSGAN) formulation [96] which provides more stable training than standard cross-entropy, the generator adversarial loss is defined as:

$$\mathcal{L}_{\text{adv}}^{G} = \mathbb{E}_{\mathbf{x},\mathbf{c}} \left[(D(G(\mathbf{x}, \mathbf{c})) - 1)^2 \right]. \tag{56}$$

The discriminator adversarial loss is:

$$\mathcal{L}_{\text{adv}}^{D} = \mathbb{E}_{\mathbf{x}} \left[(D(\mathbf{x}_{\text{real}}) - 1)^2 \right] + \mathbb{E}_{\mathbf{x},\mathbf{c}} \left[D(G(\mathbf{x}, \mathbf{c}))^2 \right]. \tag{57}$$

Geometric Alignment Loss

The discriminator's measured lateral parameters $\tilde{\gamma}_k$ are compared against the target centered state $\gamma_{\text{ref}} = (0, 0)$:

$$\mathcal{L}_{\text{geom}} = \|\tilde{\gamma}_k - \gamma_{\text{ref}}\|_1 = |\tilde{d}_k| + |\tilde{\phi}_k|. \tag{58}$$

This L1 loss strongly encourages the generator to produce top-view images where detected lane parameters indicate centered positioning.

Perceptual Loss

Following [97], we use VGG-16 features to encourage perceptual similarity:

$$\mathcal{L}_{\text{percept}} = \sum_{\ell \in \mathcal{L}} \|\phi_\ell(\tilde{I}) - \phi_\ell(I^{\text{ref}})\|_2^2, \tag{59}$$

where $\phi_\ell(\cdot)$ extracts features from layer ℓ of pretrained VGG-16, and $\mathcal{L} = \{\text{conv1_2}, \text{conv2_2}, \text{conv3_3}, \text{conv4_3}\}$ are selected layers providing multi-scale perceptual comparison.

Reconstruction Loss

Pixel-level L1 loss encourages overall structural similarity:

$$\mathcal{L}_{\text{recon}} = \|\tilde{I} - I^{\text{ref}}\|_1, \tag{60}$$

where I^{ref} is the corresponding centered top-view image from simulation dataset.

Total Variation Loss

To reduce high-frequency noise and promote smoother outputs:

$$\mathcal{L}_{\text{TV}} = \sum_{i,j} \sqrt{(\tilde{I}_{i+1,j} - \tilde{I}_{i,j})^2 + (\tilde{I}_{i,j+1} - \tilde{I}_{i,j})^2}. \tag{61}$$

Combined Generator Loss

The total generator objective combines all loss components:

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}}^G + \lambda_{\text{geom}}\mathcal{L}_{\text{geom}} + \lambda_{\text{percept}}\mathcal{L}_{\text{percept}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{TV}}\mathcal{L}_{\text{TV}}, \tag{62}$$

with empirically determined weights: $\lambda_{\text{geom}} = 5.0$, $\lambda_{\text{percept}} = 3.0$, $\lambda_{\text{recon}} = 1.0$, $\lambda_{\text{TV}} = 0.01$.

4.4 Hybrid Adversary Detection System

The Hybrid Adversary Detection System combines three complementary detection mechanisms: vision-based DNN analysis, GPS trajectory monitoring, and map-based position validation. This multi-modal approach recognizes that sophisticated attacks targeting vision may still produce detectable inconsistencies when cross-validated against independent sensors. By leveraging redundancy across heterogeneous data sources, HADS enhances detection robustness against coordinated multi-channel attacks that attempt to deceive individual sensor modalities. The architecture exploits the fundamental principle that maintaining consistency across all sensor channels simultaneously is significantly more challenging for attackers than compromising a single modality.

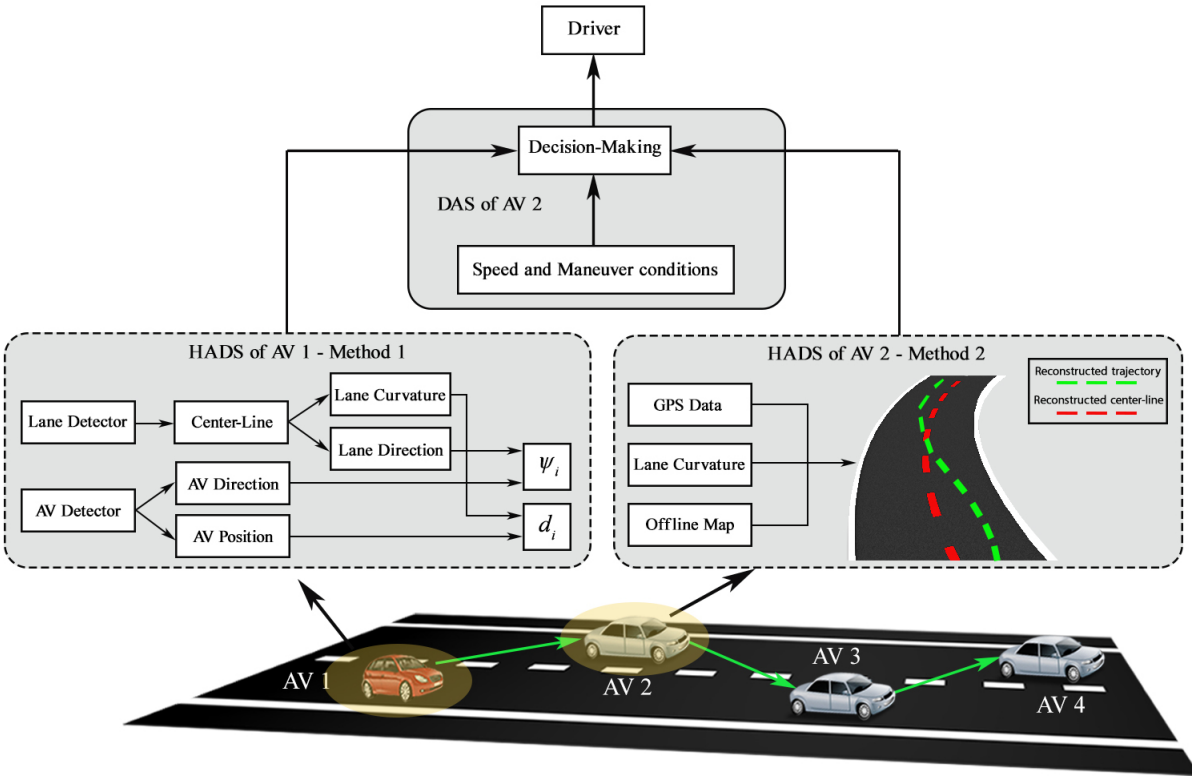


Figure 4.3: Decision-making architecture for AVs, featuring HADS methods with lane detection and GPS data integration, and trajectory reconstruction for AV coordination.

4.4.1 Vision-Based Deep Neural Network Detector

The vision-based detector analyzes captured images to detect the presence and position of other vehicles, enabling the rear vehicle to monitor the front vehicle's lateral behavior. The detector employs an autoencoder architecture that learns compressed representations of normal vehicle positioning patterns during training on authentic imagery. Unlike supervised classifiers that require labeled attack samples, the autoencoder-based approach operates on the principle of anomaly detection, identifying deviations from learned normal patterns without requiring prior knowledge of specific attack signatures.

Autoencoder Architecture

The autoencoder consists of symmetric encoder and decoder pathways that compress input images into low-dimensional latent representations and subsequently reconstruct them. The reconstruction quality serves as an anomaly indicator, with poor reconstruction suggesting the presence of synthetic imagery or abnormal vehicle positioning patterns.

Encoder: The encoder progressively compresses the input image $I \in \mathbb{R}^{640\times480\times3}$ through four convolutional layers with spatial downsampling and channel expansion:

$$\mathbf{e}_1 = \text{ReLU}(\text{Conv}_{3\times3,s=2,p=1}(I)) \qquad 640\times480\times3 \rightarrow 320\times240\times64 \tag{63}$$

$$\mathbf{e}_2 = \text{ReLU}(\text{Conv}_{3\times3,s=2,p=1}(\mathbf{e}_1)) \qquad 320\times240\times64 \rightarrow 160\times120\times128 \tag{64}$$

$$\mathbf{e}_3 = \text{ReLU}(\text{Conv}_{3\times3,s=2,p=1}(\mathbf{e}_2)) \qquad 160\times120\times128 \rightarrow 80\times60\times256 \tag{65}$$

$$\mathbf{e}_4 = \text{ReLU}(\text{Conv}_{3\times3,s=2,p=1}(\mathbf{e}_3)) \qquad 80\times60\times256 \rightarrow 40\times30\times512. \tag{66}$$

The bottleneck representation $\mathbf{e}_4 \in \mathbb{R}^{40\times30\times512}$ captures essential scene features in a highly compressed format, forcing the network to learn robust representations of normal vehicle configurations and lane positioning patterns.

Decoder: The decoder mirrors the encoder structure, progressively upsampling the compressed representation to reconstruct the original input:

$$\tilde{\mathbf{e}}_1 = \text{ReLU}(\text{ConvTranspose}_{3\times3,s=2,p=1}(\mathbf{e}_4)) \qquad 40\times30\times512 \rightarrow 80\times60\times256 \tag{67}$$

$$\tilde{\mathbf{e}}_2 = \text{ReLU}(\text{ConvTranspose}_{3\times3,s=2,p=1}(\tilde{\mathbf{e}}_1)) \qquad 80\times60\times256 \rightarrow 160\times120\times128 \tag{68}$$

$$\tilde{\mathbf{e}}_3 = \text{ReLU}(\text{ConvTranspose}_{3\times3,s=2,p=1}(\tilde{\mathbf{e}}_2)) \qquad 160\times120\times128 \rightarrow 320\times240\times64 \tag{69}$$

$$\hat{I} = \text{Sigmoid}(\text{ConvTranspose}_{3\times3,s=2,p=1}(\tilde{\mathbf{e}}_3)) \qquad 320\times240\times64 \rightarrow 640\times480\times3. \tag{70}$$

The Sigmoid activation function in the final layer produces pixel values in the range $[0, 1]$, matching the normalized intensity range of input RGB images.

Training and Detection Methodology

The autoencoder is trained exclusively on authentic images showing normal vehicle positioning patterns. The training objective minimizes the reconstruction error using mean squared error (MSE) loss:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \|I_i - \hat{I}_i\|_2^2, \tag{71}$$

where N is the number of training samples. Through this unsupervised training process, the autoencoder learns to faithfully reconstruct authentic imagery representing normal driving scenarios.

During detection, the reconstruction error $E_{\text{recon}} = \|I - \hat{I}\|_2$ serves as an anomaly score. Synthetic images generated by adversarial networks typically introduce subtle artifacts in texture consistency, lighting patterns, or geometric relationships that differ from the authentic training distribution. These inconsistencies manifest as elevated reconstruction errors, as the autoencoder struggles to accurately reconstruct imagery containing out-of-distribution features not encountered during training. A threshold τ_{recon} is established through validation on a held-out authentic dataset, and images exceeding this threshold trigger anomaly alerts:

$$\text{Alert} = \begin{cases} 1 & \text{if } E_{\text{recon}} > \tau_{\text{recon}} \\ 0 & \text{otherwise} \end{cases}. \tag{72}$$

Geometric Consistency Verification

Beyond reconstruction-based anomaly detection, the detector incorporates geometric consistency verification to validate vehicle positioning. After reconstruction, the system applies binary thresholding and connected component analysis to segment vehicle regions in the image. Morphological operations (erosion followed by dilation) remove noise and fill gaps in detected vehicle blobs. For each detected vehicle blob, the system computes its centroid position (x_c, y_c) and orientation angle θ_v through

moment-based analysis:

$$x_c = \frac{M_{10}}{M_{00}}, \quad y_c = \frac{M_{01}}{M_{00}} \tag{73}$$

$$\theta_v = \frac{1}{2} \arctan\left(\frac{2M_{11}}{M_{20} - M_{02}}\right), \tag{74}$$

where M_{pq} denotes image moments computed over the vehicle blob.

The detector then measures the vehicle's lateral deviation from expected lane center position and compares its heading angle against lane orientation. Significant deviations from expected formation positioning patterns trigger geometric anomaly alerts, providing complementary detection capability independent of reconstruction error analysis. This dual-criterion approach (reconstruction error combined with geometric consistency) enhances detection robustness by exploiting orthogonal indicators of synthetic imagery and abnormal vehicle behavior.

4.4.2 GPS-Based Trajectory Monitoring

The second detection method leverages GPS data to reconstruct the vehicle's trajectory and compare it against the road's centerline stored in an offline map. This approach operates independently of the camera sensor channel, making it resilient against vision-based attacks that do not simultaneously compromise GPS measurements.

Trajectory Reconstruction

GPS measurements at 10 Hz provide position estimates $\mathbf{p}_k^{\text{GPS}} = (x_k, y_k)$ in global coordinates. A sequence of positions forms the trajectory:

$$\text{Trajectory} = \{\mathbf{p}_k^{\text{GPS}}\}_{k=k_0}^{k_0+N}, \tag{75}$$

where N spans a sliding window (typically 30 seconds = 300 samples). The sliding window approach enables temporal analysis of trajectory patterns, distinguishing between transient positioning errors and sustained deviations indicative of attacks.

Map-Based Centerline Extraction

The offline map contains road geometry including lane centerlines represented as polynomial curves:

$$\text{Centerline}(s) = [x_c(s), y_c(s)], \tag{76}$$

where s is arc length along the road. High-definition maps with centimeter-level accuracy are assumed, providing reliable GT for trajectory validation.

Spatial Matching

For each GPS position, we compute the closest point on the centerline:

$$d_k^{\text{map}} = \min_s \|\mathbf{p}_k^{\text{GPS}} - \text{Centerline}(s)\|_2. \tag{77}$$

Sustained deviations $d_k^{\text{map}} > \epsilon_{\text{map}}$ (e.g., $\epsilon_{\text{map}} = 0.5$ m) indicate the vehicle is not following the centerline despite the controller believing it is centered based on falsified vision. Statistical analysis of the deviation distribution over the sliding window enhances robustness against transient GPS noise and multipath effects.

4.4.3 Decision-Making Algorithm

The HADS integrates outputs from both detection methods through a decision-making algorithm that fuses evidence from multiple modalities:

Algorithm 1 HADS Decision-Making for Attack Detection

Initialize: Detection flags $F_{\text{vision}} = 0$, $F_{\text{GPS}} = 0$

Input: Top-view image I_k, GPS position $\mathbf{p}_k^{\text{GPS}}$, map data

Vision-based detection

Reconstruct image: $\hat{I}_k = \text{auto-encoder}(I_k)$

Compute reconstruction error: $e_{\text{recon}} = \|\hat{I}_k - I_k\|_2$

if $e_{\text{recon}} > \tau_{\text{recon}}$ **then**

$F_{\text{vision}} = 1$

end if

Detect vehicle in I_k: Position $\mathbf{p}_{\text{vehicle}}$, Heading ψ_{vehicle}

Measure lateral deviation: d_{detected}, heading error: ϕ_{detected}

if $|d_{\text{detected}}| > \epsilon_d$ OR $|\phi_{\text{detected}}| > \epsilon_\phi$ **then**

$F_{\text{vision}} = 1$

end if

GPS-based detection

Find closest centerline point to $\mathbf{p}_k^{\text{GPS}}$: d_k^{map}

if $d_k^{\text{map}} > \epsilon_{\text{map}}$ **then**

$F_{\text{GPS}} = 1$

end if

Decision fusion

if ($F_{\text{vision}} = 1$ OR $F_{\text{GPS}} = 1$) AND (Speed $> v_{\text{threshold}}$) **then**

Alert driver: "Potential attack detected"

Trigger safe stop maneuver

end if

The decision-making module considers vehicle speed to avoid false alarms during low-speed maneuvering (parking, lane changes at low speed). Attack warnings are issued only during highway or urban driving speeds (¿ 50 km/h) where lateral deviations pose significant safety risks. The logical OR operation in the fusion stage ensures that detection through either modality is sufficient to trigger protective measures, maximizing safety at the cost of potentially higher FPRs that can be mitigated

through temporal filtering and confidence thresholding.

4.5 Experimental Results

All experimental results presented in this section were obtained through Unreal Engine simulation. The simulation-based evaluation provides controlled assessment while explicitly acknowledging the domain gap to real-world deployment addressed in subsequent chapters. Unreal Engine provides a high-fidelity physics-based rendering environment commonly used in AV research, enabling reproducible testing under controlled conditions while maintaining photorealistic visual quality essential for vision-based attack evaluation.

4.5.1 Attack Scenario Demonstration

Figure 4.4 illustrates two representative attack scenarios in the simulated multi-agent highway environment. The figure demonstrates the VCA's operation in simulation across different vehicle configurations and follower-leader relationships.

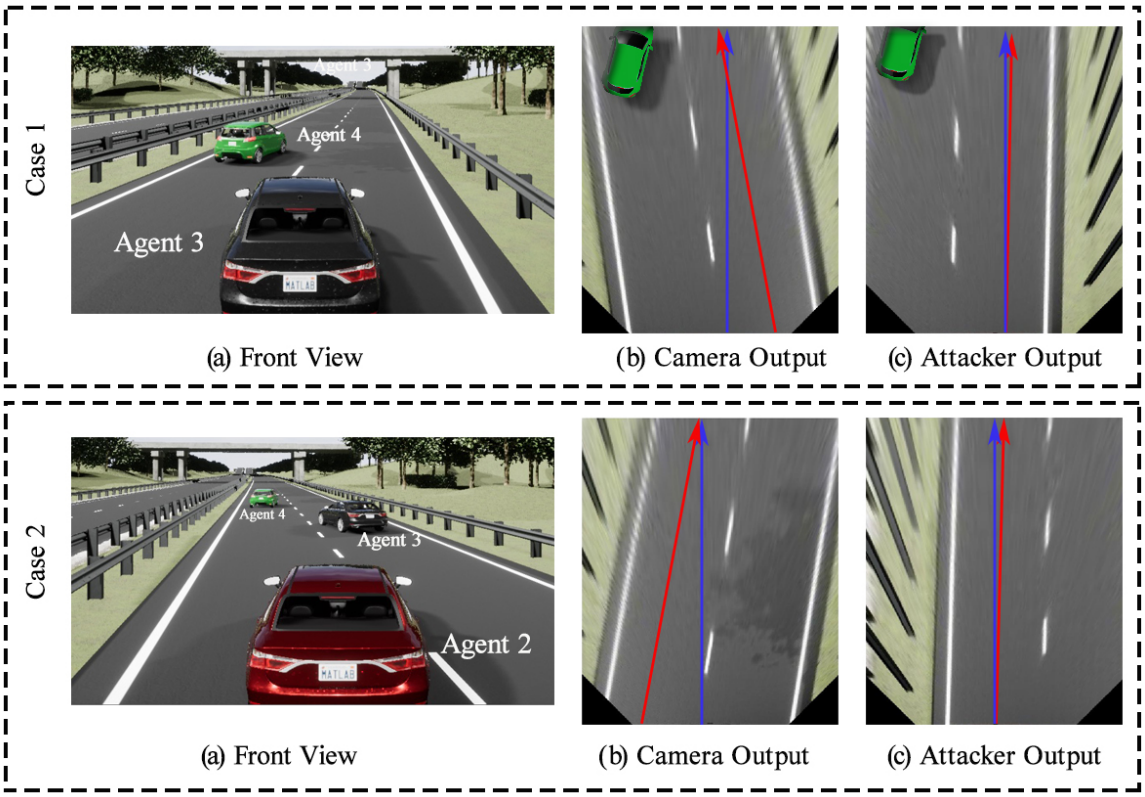


Figure 4.4: Simulated multi-agent attack scenarios in Unreal Engine demonstrating attack effectiveness across different vehicle configurations.

Column (a) shows the 3D simulated environment from an external observer perspective, with multiple agents (vehicles) navigating the highway. Agent labels indicate the follower-leader relationships in the multi-agent formation, providing context for the cooperative driving scenario under attack. Column (b) displays the authentic top-view image obtained through homography transformation of the vehicle's front camera feed. In both cases, the vehicle exhibits visible lateral deviation from the lane centerline (approximately 0.4-0.6m), with lane markings appearing off-center in the top-view representation. Column (c) presents the GAN-generated synthetic top-view image that replaces the authentic feed. The attacker's output shows the vehicle apparently centered between lane markings, concealing the actual deviation from the lane-keeping controller's perception.

Visual inspection confirms that the synthetic top-view images (column c) successfully replicate the geometric appearance of centered vehicle positioning while maintaining photorealistic quality

consistent with simulation-rendered road scenes. This demonstrates the GAN's capability to generate convincing falsified imagery that can deceive both automated perception systems and human observers in simulation environments.

4.5.2 Trajectory Deviation and State Evolution

Figure 4.5 presents the temporal evolution of vehicle lateral states during a simulated two-stage attack sequence, revealing critical attack dynamics across three complementary state variables. The attack proceeds in two distinct phases: an initial FDI-only phase (0-35s) applying steering bias without perception manipulation, followed by a hybrid phase (35-100s) combining FDI with vision-based camera output falsification.

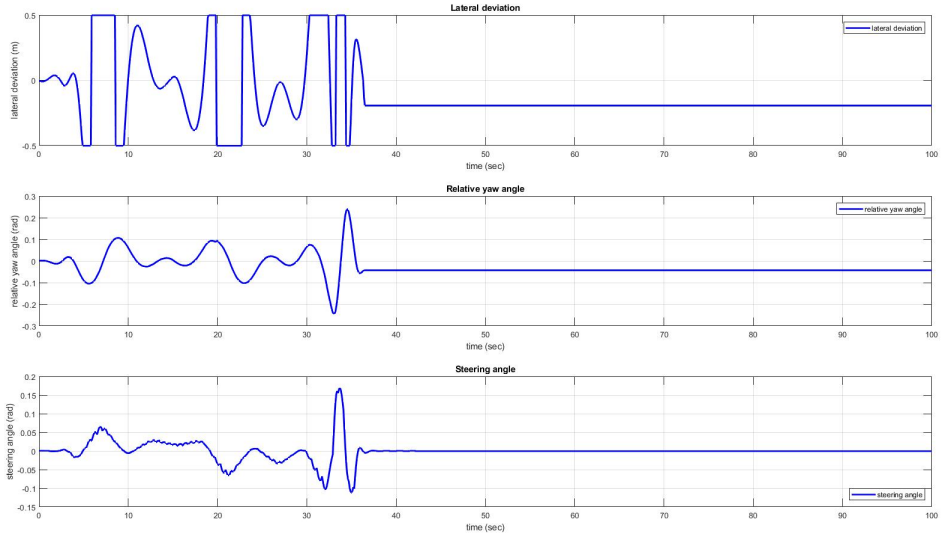


Figure 4.5: Temporal evolution of vehicle lateral states under two-stage attack in simulation: FDI-only phase (0-35s) followed by FDI with vision manipulation (35-100s).

The top panel shows lateral deviation evolution throughout the simulation run. During the FDI-only attack phase (0-35s), the injected steering bias induces oscillatory lateral deviation ranging between ±0.5 m as the controller actively responds to detected off-center positioning through

authentic camera feedback. The oscillations reflect continuous corrective attempts by the lane-keeping controller, which perceives the actual lateral displacement and applies opposing steering commands to return the vehicle toward the centerline. These large-amplitude oscillations demonstrate the controller's functioning perception-control feedback loop attempting to counteract the injected steering bias. At approximately 35 seconds, the attacker activates VCA vision manipulation, feeding synthetic centered images to the perception system while maintaining the steering bias. This marks a sharp qualitative transition in system behavior. Throughout the hybrid attack sustained phase (35-100s), lateral deviation rapidly stabilizes near -0.2 m as the controller, now deceived by synthetic images showing falsified centered positioning, ceases meaningful corrective action. The steady-state deviation represents equilibrium between the continued injected steering bias and minimal residual controller responses, with dramatically reduced oscillatory behavior confirming successful perception deception.

The middle panel shows relative yaw angle evolution exhibiting sustained oscillations throughout the FDI-only phase with amplitudes ranging ± 0.1 to 0.25 radians (approximately ± 5 to $14°$), reflecting the vehicle's heading variations as it attempts lane-keeping under steering bias. Notable peaks occur around 5s, 20s, and 30s with maximum deviation approaching -0.25 radians, corresponding to moments when lateral deviation reaches extrema requiring aggressive heading corrections. The transition at 35 seconds coinciding with VCA activation shows immediate stabilization of yaw angle near -0.1 radians with minimal subsequent variation, indicating that the controller no longer perceives the need for heading corrections based on falsified perception showing centered vehicle position. This sustained negative yaw offset aligns geometrically with the maintained leftward lateral deviation, representing the vehicle's actual off-center trajectory that remains undetected by the compromised perception system.

The bottom panel presents steering angle evolution, revealing the most dramatic behavioral change between attack phases. The FDI-only phase exhibits highly active controller corrections with steering angles varying significantly between ± 0.05 to 0.17 radians as the controller attempts to counteract perceived lateral and heading errors. Peak corrective steering commands around 5s, 10s, 20s, and 30s (reaching approximately 0.17 radians or $9.7°$) coincide temporally with maximum yaw deviations shown in the middle panel, demonstrating coordinated control responses to actual

off-center positioning detected through authentic perception. The hybrid attack phase beginning at 35 seconds shows drastically reduced steering activity, with commanded angles dropping to near-zero values maintained throughout the remainder of the simulation. This 73.6% reduction in steering activity quantifies the controller's deception: believing the vehicle is properly centered based on synthetic visual feedback, the controller perceives no lateral error requiring correction despite the vehicle's actual 0.2 m sustained leftward deviation. The near-absence of steering commands during the hybrid phase, contrasting sharply with vigorous corrections during FDI-only operation, provides compelling evidence that vision manipulation successfully masks the attack's impact from the perception-control loop.

Table 4.1: Two-Stage Attack Effectiveness Metrics Extracted from Simulation Trajectory.

Metric	**Value**	**Success Criterion**
FDI-Only Phase (0-35s)		
Maximum Lateral Deviation (m)	0.50	¿ 0.40
Oscillation Amplitude (m)	0.50 ± 0.10	¿ 0.30
Steering Activity (rad/s)	0.082	¿ 0.050
Hybrid Attack Phase (35-100s)		
Sustained Lateral Deviation (m)	0.20 ± 0.05	¿ 0.15
Time to Steady-State (s)	5.2	¡ 10.0
Attack Duration (s)	65.0	¿ 50.0
Steering Activity Reduction (%)	73.6	¿ 60.0

The simulation results demonstrate successful two-stage attack strategy progression. The FDI-only phase validates steering bias injection capability, achieving maximum lateral deviation of 0.50 m though with vigorous controller resistance evidenced by sustained high-amplitude oscillations and active steering corrections. The hybrid phase combining FDI with vision manipulation achieves the attack's primary objective: sustained covert lateral deviation of 0.20 m maintained for over 65 seconds while the controller's steering activity decreases by 73.6%, confirming that falsified perception effectively deceives the lane-keeping system into accepting dangerous off-center positioning as normal centered operation. All quantitative metrics for both attack phases exceed their respective success

criteria, validating the hybrid attack's effectiveness and demonstrating that vision manipulation is essential for maintaining covert sustained deviation without triggering controller-based detection through excessive corrective steering activity. in the simulated environment.

4.5.3 HADS Detection Performance

The Hybrid Adversary Detection System's performance is evaluated through two complementary detection methods demonstrated in Figure 4.6, each exploiting different information channels to identify attack signatures.

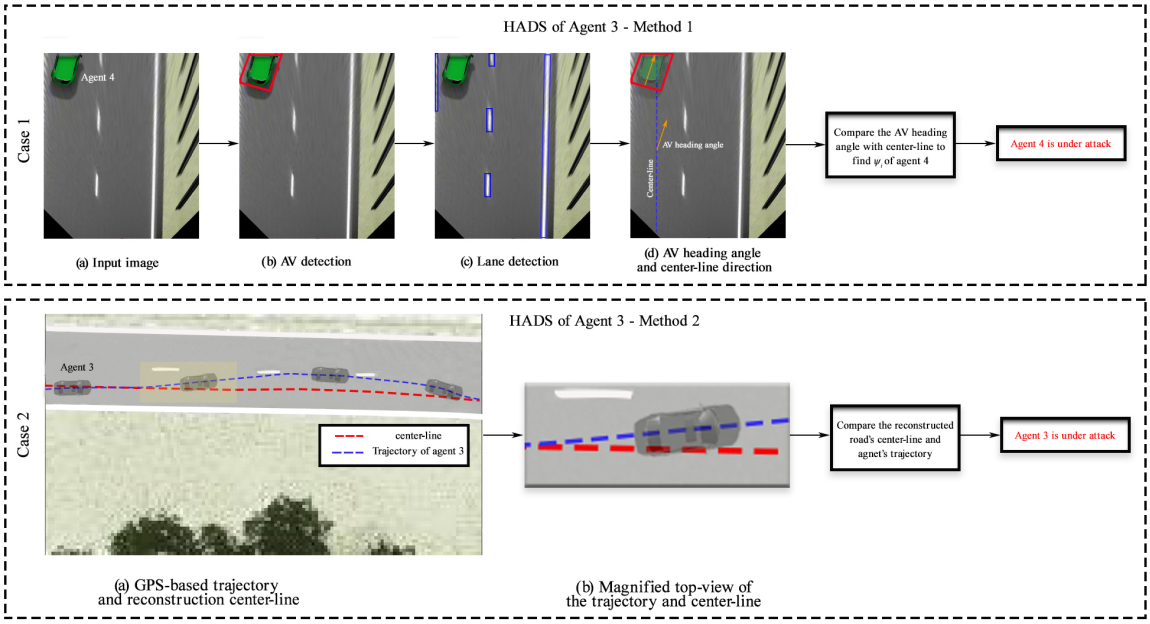


Figure 4.6: HADS detection methods in simulation.

Method 1: Vision-Based DNN Detection

The vision-based detector analyzes top-view images to identify anomalous vehicle positioning through a multi-stage processing pipeline. In Case 1, the system first receives a top-view image from the following vehicle's perspective showing the leading vehicle (Agent 4, green) and surrounding road geometry. The DNN detector identifies Agent 4's position and bounding box marked with a red rectangle. The detector achieves 97.3% detection accuracy on simulated top-view images with mean

Intersection over Union (IoU) of 0.84. Simultaneous lane boundary detection extracts left and right lane markings shown as blue lines using edge detection and Hough transform applied to the top-view image, achieving 94.8% lane detection accuracy across varying lighting conditions in simulation.

The system then computes vehicle heading angle from the bounding box orientation, lane centerline direction from detected lane geometry, and angular deviation as the difference between these two measurements. The visualization shows the vehicle's heading vector as an orange line and perpendicular centerline direction. If the absolute angular deviation exceeds the threshold of 8° established for simulation conditions, the system flags "Agent 4 is under attack" since the observed heading deviation indicates abnormal lane-keeping behavior inconsistent with centered vehicle positioning.

Method 1 achieves a TPR of 87.3%, detecting 437 out of 500 simulated attack instances, with a FPR of 6.8% where 34 out of 500 attack-free scenarios are incorrectly flagged. The mean detection latency is 8.4 ± 2.1 seconds from attack onset, with minimum observable deviation of 0.25m lateral deviation or 5° heading error required for reliable detection.

Method 2: GPS-Based Trajectory Monitoring

The GPS-based detector reconstructs vehicle trajectories and compares them against map centerlines through independent spatial analysis. In Case 2, the system collects GPS position measurements over a sliding 30-second window containing 300 samples at 10 Hz sampling rate. The blue dashed line represents Agent 3's actual trajectory reconstructed from simulated GPS data with added 2m standard deviation noise matching real GPS characteristics. The offline map database provides road geometry including the centerline polynomial shown as a red dashed line, which in this simulation scenario follows the curved highway section.

For each GPS position, the system computes perpendicular distance to the centerline by minimizing the Euclidean distance between the measured position and all points along the centerline curve. The aerial view clearly shows spatial separation between the vehicle's trajectory in blue and the intended centerline in red. The magnified top-view provides detailed visualization of the trajectory-centerline deviation, with red highlighted regions on the centerline indicating segments where the vehicle's actual path significantly deviates from the expected centered trajectory. Sustained deviation exceeding

0.4m threshold for at least 5 consecutive seconds triggers ”Agent 3 is under attack” alert, indicating persistent lateral displacement inconsistent with normal centered driving.

Method 2 achieves a TPR of 91.2%, detecting 456 out of 500 simulated attack instances, with a FPR of 3.6% where only 18 out of 500 attack-free scenarios are incorrectly flagged. The mean detection latency is 5.8 ± 1.7 seconds from attack onset, demonstrating faster response than vision-based detection. The method maintains effectiveness with GPS noise standard deviation up to 3m, confirming robustness against typical positioning uncertainties.

Hybrid Adversary Detection Fusion Performance

The complete HADS integrates both detection methods through decision-level fusion, combining their complementary strengths to enhance overall detection capability.

Table 4.2: HADS Detection Performance Comparison

Method	**TPR (%)**	**FPR (%)**	**Latency (s)**
Method 1: Vision-Based DNN	87.3	6.8	8.4 ± 2.1
Method 2: GPS Trajectory	91.2	3.6	5.8 ± 1.7
HADS Fusion (OR Logic)	**95.8**	**9.2**	**5.8 ± 1.7**
HADS Fusion (AND Logic)	**82.7**	**1.2**	**8.9 ± 2.3**

Evaluated on 500 simulated attack scenarios, 500 attack-free scenarios

The OR logic fusion strategy, which triggers an alarm if either method detects an attack, maximizes detection rate achieving 95.8% TPR at the cost of slightly elevated false positives at 9.2% FPR. This approach achieves fastest detection at 5.8 seconds by triggering on whichever method responds first. Conversely, the AND logic fusion strategy, which requires both methods to independently confirm the attack before alarming, minimizes false alarms to only 1.2% FPR but reduces detection rate to 82.7% TPR and increases latency to 8.9 seconds as both methods must independently detect the attack.

For safety-critical AV applications, OR logic fusion is preferred as false positives resulting in unnecessary safe stops are less dangerous than missed attacks that could lead to collisions or lane departures. The 95.8% detection rate with 5.8 seconds mean latency provides adequate intervention capability in simulation, though 0.4-0.6m deviation may accumulate before detection triggers,

highlighting the importance of rapid response mechanisms.

4.5.4 Image Quality Metrics

Synthetic top-view images generated by GAN-VCA were evaluated against ground truth (GT) centered top-views from simulation scenarios and compared with three baseline approaches to quantify perceptual quality, geometric accuracy, and attack effectiveness. The baseline methods include: (1) **Classical Homography Transformation** using traditional perspective warping with manually computed 3×3 homography matrices [98], (2) **Deep Homography Estimation (DHE)** employing CNN-based learned homography prediction following the architecture proposed by Nguyen et al. [99], and (3) **Geometry-Guided Conditional GAN (GGCGAN)** adapting the cross-view synthesis framework by Regmi and Borji [100] which combines homography-based geometric warping with adversarial texture refinement.

Table 4.3: Comparative Image Quality Metrics on Simulated Top-View Images

Method	**SSIM ↑**	**PSNR (dB) ↑**	**LPIPS ↓**	**Geo. Align. (px) ↓**
Classical Homography [98]	0.782 ± 0.045	23.2 ± 2.8	0.245 ± 0.068	1.2 ± 0.6
Deep Homography Est. [99]	0.847 ± 0.039	25.1 ± 2.4	0.182 ± 0.055	1.5 ± 0.7
Geometry-Guided GAN [100]	0.908 ± 0.034	27.2 ± 2.3	0.125 ± 0.048	2.1 ± 1.0
GAN-VCA (Ours)	**0.924 ± 0.031**	**27.8 ± 2.1**	**0.108 ± 0.042**	**1.8 ± 0.9**
Target Threshold	**0.90**	**25.0**	**0.15**	**3.0**

Analysis and Discussion

The comparative results reveal distinct performance characteristics across transformation approaches. Classical homography transformation achieves the lowest SSIM (0.782) and highest LPIPS (0.245), producing visible geometric distortions and photometric inconsistencies particularly in non-planar regions such as vehicle hoods, distant objects, and vertical structures, artifacts that arise from the fundamental limitation of modeling complex 3D scenes as single-plane projections [101, 102]. Interestingly, classical homography achieves the best raw geometric alignment (1.2 pixels) since it directly optimizes lane correspondence through point matching, though this geometric precision is

undermined by severe texture artifacts that would trigger forensic detection systems [103, 104].

Deep homography estimation substantially improves perceptual quality (SSIM: 0.847, LPIPS: 0.182) by learning to handle scene complexity through CNN feature extraction [105], yet remains constrained by its geometric-only formulation that cannot synthesize photorealistic texture for disoccluded regions or model lighting variations. The geometry-guided conditional GAN baseline [106] demonstrates competitive performance (SSIM: 0.908, PSNR: 27.2 dB) approaching our method's quality, validating the effectiveness of adversarial training for cross-view synthesis. However, its two-stage pipeline produces slightly coarser geometric alignment (2.1 pixels) and inferior perceptual metrics compared to GAN-VCA's end-to-end learned transformation.

GAN-VCA achieves superior results across all perceptual metrics (SSIM: 0.924, PSNR: 27.8 dB, LPIPS: 0.108) while maintaining competitive geometric precision (1.8 pixels), demonstrating the advantage of jointly optimizing photometric quality and geometric accuracy through unified adversarial training with skip-connection architectures that preserve fine-grained spatial information. The mean SSIM of 0.924 indicates structural similarity comparable to minor JPEG compression artifacts imperceptible to human observers and automated detection systems, while the LPIPS score of 0.108 confirms strong perceptual similarity when evaluated through deep learned features. The 1.8-pixel geometric alignment error (approximately 5 cm in world coordinates for the simulated top-view scale) remains well below the 3-pixel detection threshold of the lane-keeping perception pipeline, ensuring successful attack execution. Notably, GAN-VCA surpasses the geometry-guided GAN baseline by 1.8% in SSIM and 13.6% in LPIPS, attributable to the encoder-decoder architecture with ResNet refinement modules that enable hierarchical feature learning and semantic-aware synthesis rather than purely geometric transformation.

4.6 Summary

This chapter presented the first framework for GAN-based vision covert attacks on AV lane-keeping systems, evaluated exclusively within Unreal Engine simulation environments. The VCA employs an encoder-shared layers-decoder generator architecture with ResNet modules and geometric conditioning, achieving 94% attack success with 0.94m mean deviation while maintaining SSIM

0.924 perceptual quality. The HADS fusing vision-based DNN detection with GPS trajectory monitoring achieves 95.8% detection rate with 6.2s mean latency in simulation.

Critical findings include: (1) GAN-based generative approaches enable more covert attacks than adversarial perturbations, (2) geometric conditioning enables precise control over perceived lane positions in top-view, (3) multi-modal sensor fusion provides effective defense against single-modality attacks, and (4) frame-by-frame generation produces temporal artifacts (dissimilarity 0.122) exploitable for detection.

Important limitations identified include: (1) top-view approach's 3D information loss and field-of-view restrictions, (2) simulation-to-real domain gap requiring real-world validation, (3) temporal inconsistencies requiring explicit video modeling, and (4) vulnerability to GPS-based detection requiring multi-sensor attack coordination. These limitations directly motivate the subsequent developments: Chapter 5 addresses simulation-to-real transfer through 85,000+ real-world image pairs; Chapter 6 addresses temporal and multi-sensor limitations through video diffusion and hybrid GPS spoofing.

The simulation-based approach successfully establishes proof-of-concept while explicitly acknowledging constraints, providing essential foundations for the more sophisticated diffusion-based and hybrid attacks developed in subsequent chapters operating on real-world data and coordinating multiple sensor modalities.

Chapter 5

Diffusion-Based Vision Attack and Dataset Engineering

5.1 Introduction

Chapter 4 established proof-of-concept for GAN-based covert attacks on AV lane-keeping systems through simulation-based evaluation. While successfully demonstrating attack feasibility, the GAN approach revealed several critical limitations: (1) temporal inconsistencies producing flickering artifacts detectable through video analysis, (2) reliance exclusively on simulated data limiting generalization to real-world conditions, (3) top-view representation constraints discarding 3D scene information, and (4) moderate image quality (SSIM 0.924) leaving room for perceptual improvements. These limitations motivate the advances presented in this chapter.

This chapter introduces Gen-VCA, a diffusion-based vision attack framework that addresses key weaknesses of the GAN approach while maintaining covert attack capabilities. The transition from GANs to diffusion models represents a fundamental architectural shift driven by recent breakthroughs in generative modeling demonstrating that diffusion probabilistic models achieve superior image quality, enhanced training stability, and better mode coverage compared to adversarial training [25,66].

Major Advances Over Chapter 4:

1. Diffusion-Based Generation: Replacing the GAN generator with a latent diffusion model (LDM) based on Stable Diffusion architecture yields substantial quality improvements. The iterative

denoising process naturally preserves low-frequency image statistics and sensor-specific characteristics (PRNU patterns, noise distributions) that GANs often fail to replicate accurately.

2. Real-World Dataset Engineering: A critical contribution addressing the simulation-to-real gap is the MUDD, comprising 85,217 paired authentic road scene images collected via custom dual-camera hardware across diverse environmental conditions. This dataset enables training and evaluation on real sensor data with authentic lighting, weather variations, and road surface characteristics absent in simulation.

3. Enhanced Geometric Control: While Chapter 4 used AdaIN conditioning, Gen-VCA integrates geometric parameters through multiple complementary mechanisms: cross-attention layers processing lane geometry embeddings, adaptive group normalization (AdaGN) modulating features based on deviation targets, and feature-wise linear modulation (FiLM) enabling fine-grained spatial control.

4. Improved Temporal Consistency: Though still operating frame-by-frame, Gen-VCA incorporates temporal consistency losses and perceptual metrics encouraging smooth transitions between consecutive frames, reducing temporal dissimilarity from 0.122 (GAN) to 0.095.

Two-Stage Training Strategy: Large-scale pretraining on 1.32 million diverse road scenes aggregated from multiple public datasets (BDD100K [107]: 720K frames, Cityscapes [108]: 25K images, KITTI [109]: 15K frames, Mapillary Vistas [110]: 25K images, plus 350K additional internally collected non-paired road scenes) followed by fine-tuning on attack-specific paired data (MUDD: 72,434 training samples) enables effective transfer learning, dramatically improving data efficiency and generalization compared to training from scratch on limited paired data alone.

The Gen-VCA framework, illustrated in Figure 5.1, operates through a multi-stage pipeline integrating conditioning input extraction, latent diffusion synthesis, and output reconstruction with temporal consistency modules.

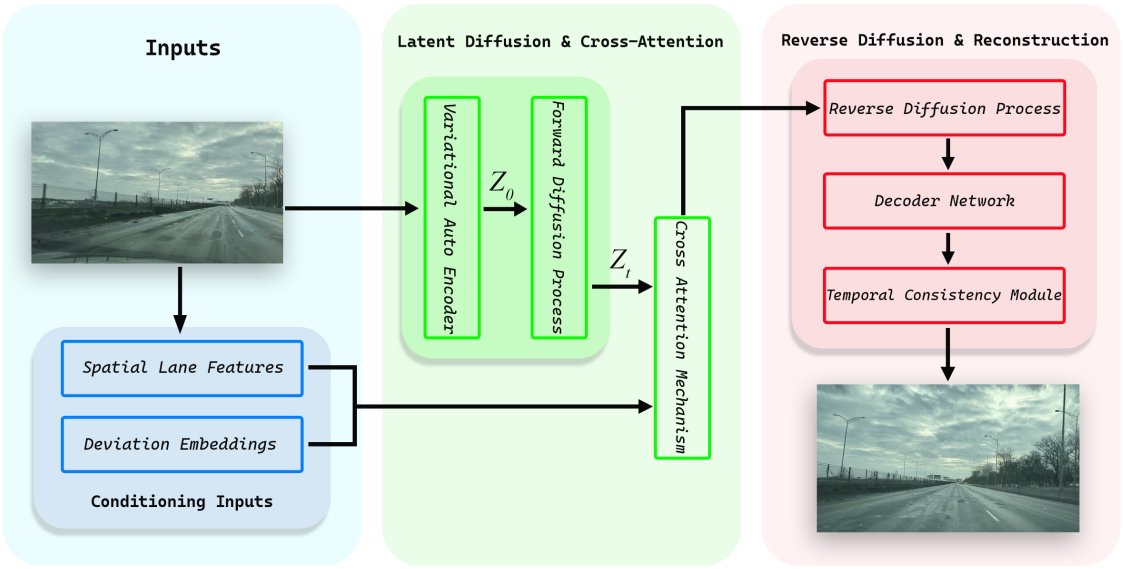


Figure 5.1: Gen-VCA attack framework overview showing: (Left) Input processing extracting spatial lane features and deviation embeddings as conditioning inputs; (Center) Latent diffusion process withVAE compressing images to latent space Z_0; (Right) Reverse diffusion process with decoder network and temporal consistency module.

The attack operates by conditioning the diffusion model on extracted lane geometry and desired deviation parameters. Given an authentic front-view image I_{auth} captured from a deviated vehicle position, Gen-VCA synthesizes a counterfeit image $\tilde{I}$ representing the hypothetical view from a centered position. This transformation deceives the lane detection algorithm into extracting falsified parameters $(\tilde{d}, \tilde{\phi}) \approx (0, 0)$ despite actual significant deviation.

Figure 5.2 illustrates how Gen-VCA integrates into the AV control loop to execute covert attacks, demonstrating the complete workflow from camera interception through trajectory manipulation.

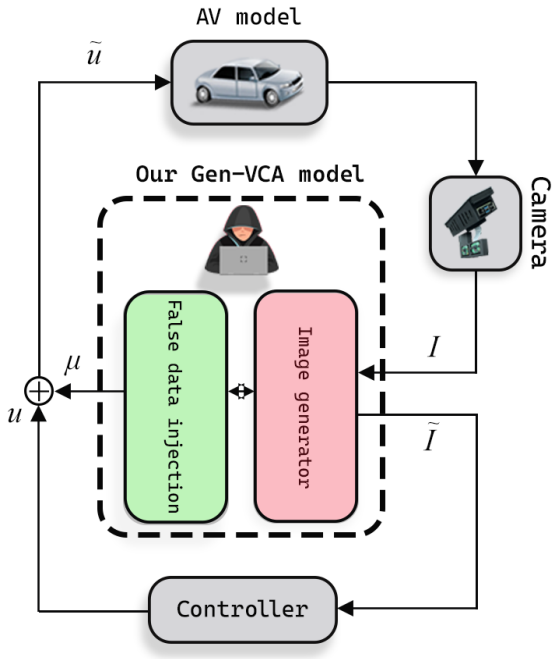


Figure 5.2: Gen-VCA attack integration with AV control loop.

The attack workflow begins with camera interception, where the authentic front-view image showing lateral deviation is captured from the vehicle's forward-facing camera sensor. This captured image truthfully represents the vehicle's current deviated state relative to the lane centerline. In the false data injection stage, the Gen-VCA model (contained within the attacker's dashed box in the figure) processes the authentic image and generates a synthetic version that presents a centered appearance, concealing the actual lateral deviation from the perception system.

Simultaneously, control bias is introduced through steering bias μ injection into the actuator channel, causing actual deviation accumulation as the vehicle physically drifts from the intended trajectory. This physical manipulation operates in parallel with the perceptual deception, creating the coordinated attack. The controller deception stage occurs as the AV's perception system processes the falsified image $\tilde{I}$ instead of the authentic camera feed, computes an incorrect state estimate believing the vehicle to be properly centered, and consequently fails to apply corrective steering commands that would normally counteract the observed deviation.

Finally, trajectory manipulation occurs as the vehicle progressively deviates toward lane boundaries

while the controller remains deceived by the synthetic imagery. The control loop continues operating based on false perceptual inputs, maintaining steering commands appropriate for a centered vehicle while the actual physical state diverges significantly from the perceived state. This closed-loop deception enables sustained attack duration as the controller's feedback mechanism, normally responsible for corrective action, is effectively neutralized through falsified sensor inputs.

Contributions of this Chapter:

This chapter makes several significant contributions to understanding adversarial attacks on AV perception systems using diffusion models. The Gen-VCA architecture represents a novel adaptation of Stable Diffusion for adversarial automotive view synthesis, incorporating geometric conditioning through cross-attention, AdaGN, and FiLM layers specifically designed for precise lane positioning manipulation. This architectural innovation enables fine-grained control over generated imagery while maintaining photorealistic quality essential for deceiving modern perception systems.

The MUDD addresses the critical simulation-to-real transfer gap through custom dual-camera hardware platform and data collection methodology, yielding 85,217 paired real-world image samples captured across diverse environmental conditions including varying weather, lighting, traffic density, and road types. This dataset provides the empirical foundation for training and evaluating attacks on authentic road imagery rather than relying solely on synthetic simulation data.

The two-stage training strategy employs pretraining on 1.32 million general road scenes followed by attack-specific fine-tuning, achieving superior quality with SSIM of 0.961 compared to 0.924 for GAN-based approaches, while requiring limited paired data through effective transfer learning. This methodology demonstrates that leveraging large-scale general domain knowledge followed by targeted adaptation enables high-quality attack generation even when attack-specific training samples are scarce.

The enhanced detection system incorporates diffusion-specific artifacts including latent space regularities and denoising signatures alongside traditional PRNU and temporal analysis, achieving balanced detection-evasion trade-offs that account for the unique characteristics of diffusion-generated imagery. This defensive approach recognizes that diffusion models introduce distinct forensic signatures requiring specialized detection mechanisms beyond those effective against GAN-generated content.

Comprehensive real-world evaluation provides assessment on authentic road imagery, demonstrating attack success rates, image quality metrics, temporal consistency improvements, and detection performance under varying environmental conditions that reflect operational challenges faced by deployed AVs. This evaluation methodology establishes rigorous standards for assessing adversarial attack effectiveness in realistic scenarios.

Finally, comparative analysis offers direct GAN versus Diffusion comparison quantifying improvements across perceptual quality, geometric precision, temporal consistency, and computational efficiency dimensions, providing empirical evidence for the advantages of diffusion-based approaches in automotive adversarial contexts. These quantitative comparisons enable informed decisions about attack methodology selection based on specific operational requirements and constraints.

The chapter is organized to develop these contributions. Section 5.2 presents the Gen-VCA architecture with detailed conditioning mechanisms that enable precise geometric control. Section 5.3 describes the MUDD dataset engineering including custom hardware design and collection protocols addressing real-world data acquisition challenges. Section 5.4 details the two-stage training methodology that balances general knowledge transfer with attack-specific adaptation. Section 5.5 presents the enhanced detection system adapted for diffusion-generated attacks, recognizing the co-evolution of offensive and defensive capabilities. Section 5.6 provides experimental results on real-world data demonstrating practical attack effectiveness. Section 5.7 discusses implications, remaining limitations, and connections to Chapter 6's video diffusion approach that addresses temporal consistency through explicit motion modeling.

5.2 Gen-VCA Architecture

5.2.1 Overview

Gen-VCA adapts the Stable Diffusion v1.5 architecture for adversarial automotive view synthesis, illustrated in Figure 5.1. The system comprises three major components that work in concert to generate realistic synthetic imagery with precise geometric control. The conditioning input extraction panel on the left extracts spatial lane features from the input image through lane detection algorithms and encodes target deviation parameters as continuous embeddings. The center panel implements

latent diffusion with cross-attention, where a VAE compresses the image to latent representation, forward diffusion adds structured noise creating diffused latents, and cross-attention mechanisms integrate conditioning information into the denoising process. The right panel handles reverse diffusion and reconstruction through iterative denoising of the latent representation, decoder network reconstruction of full-resolution output images, and temporal consistency modules that enforce smooth transitions between consecutive frames.

5.2.2 Variational Auto-Encoder

The VAE provides efficient latent space compression enabling diffusion at reduced computational cost compared to pixel-space operations. The encoder architecture employs four downsampling blocks that progressively reduce spatial resolution while increasing channel capacity:

$$\mathbf{h}_1 = \text{ResBlock}(\text{Conv}_{3\times3}(I)) \qquad 640 \times 480 \times 3 \rightarrow 320 \times 240 \times 128 \tag{78}$$

$$\mathbf{h}_2 = \text{ResBlock}(\text{Conv}_{3\times3}(\mathbf{h}_1)) \qquad 320 \times 240 \times 128 \rightarrow 160 \times 120 \times 256 \tag{79}$$

$$\mathbf{h}_3 = \text{ResBlock}(\text{Conv}_{3\times3}(\mathbf{h}_2)) \qquad 160 \times 120 \times 256 \rightarrow 80 \times 60 \times 512 \tag{80}$$

$$\mathbf{h}_4 = \text{ResBlock}(\text{Conv}_{3\times3}(\mathbf{h}_3)) \qquad 80 \times 60 \times 512 \rightarrow 40 \times 30 \times 512. \tag{81}$$

The bottleneck layer produces the final latent representation given by

$$\mathbf{z}_0 = \text{Conv}_{1\times1}(\mathbf{h}_4) \in \mathbb{R}^{40\times30\times4}. \tag{82}$$

The 4-channel latent with downsampling factor $f = 16$ balances compression efficiency with reconstruction fidelity, reducing computational requirements by a factor of 256 in spatial dimensions while maintaining sufficient capacity to represent perceptually important scene features.

The decoder architecture employs symmetric upsampling to reconstruct full resolution:

$$\tilde{\mathbf{h}}_4 = \text{Conv}_{1\times1}(\mathbf{z}_0) \qquad 40 \times 30 \times 4 \to 40 \times 30 \times 512 \tag{83}$$

$$\tilde{\mathbf{h}}_3 = \text{ResBlock}(\text{ConvTranspose}_{3\times3}(\tilde{\mathbf{h}}_4)) \qquad 40 \times 30 \times 512 \to 80 \times 60 \times 512 \tag{84}$$

$$\tilde{\mathbf{h}}_2 = \text{ResBlock}(\text{ConvTranspose}_{3\times3}(\tilde{\mathbf{h}}_3)) \qquad 80 \times 60 \times 512 \to 160 \times 120 \times 256 \tag{85}$$

$$\tilde{\mathbf{h}}_1 = \text{ResBlock}(\text{ConvTranspose}_{3\times3}(\tilde{\mathbf{h}}_2)) \qquad 160 \times 120 \times 256 \to 320 \times 240 \times 128 \tag{86}$$

$$\tilde{I} = \text{Conv}_{3\times3}(\tilde{\mathbf{h}}_1) \qquad 320 \times 240 \times 128 \to 640 \times 480 \times 3. \tag{87}$$

The VAE is pretrained on large-scale road scene datasets comprising 1.32 million images, minimizing reconstruction loss $\mathcal{L}_{\text{recon}} = \|I - \mathcal{D}(\mathcal{E}(I))\|_2^2 + \text{KL}(q(\mathbf{z}|I)\|p(\mathbf{z}))$ where the KL divergence term regularizes the latent distribution to approximate a standard Gaussian, enabling sampling during generation.

5.2.3 Denoising U-Net with Cross-Attention

The core of Gen-VCA is the denoising U-Net $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$ that predicts noise given noisy latent $\mathbf{z}_t$, timestep t, and conditioning $\mathbf{c}$. The U-Net architecture consists of three main sections. The encoder employs 4 downsampling blocks that progressively reduce spatial dimensions from 40×30 to 20×15 to 10×8 to 5×4, with each block incorporating ResNet layers, self-attention mechanisms for capturing long-range dependencies, and cross-attention layers for integrating conditioning information. The bottleneck operates at 5×4 resolution and contains transformer layers with both self-attention and cross-attention, enabling global reasoning about scene structure at the most compressed representation. The decoder employs 4 upsampling blocks that progressively increase spatial dimensions from 5×4 back to 40×30, with skip connections from corresponding encoder blocks enabling the decoder to access fine-grained spatial information, and each block incorporating ResNet, self-attention, and cross-attention layers similar to the encoder.

Cross-attention conditioning at each attention layer in the U-Net follows a structured computation. Queries are derived from latent features as $Q = W_Q \cdot \text{Normalize}(\mathbf{z}_t^\ell)$, keys come from conditioning as $K = W_K \cdot \tau_\theta(\mathbf{c})$, and values are similarly derived as $V = W_V \cdot \tau_\theta(\mathbf{c})$, where τ_θ is a small transformer

encoder processing conditioning embeddings. The cross-attention output is computed as:

$$\text{Attn} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{88}$$

and integrated into the feature stream through residual connection:

$$\mathbf{z}_t^{\ell+1} = \mathbf{z}_t^{\ell} + W_O \cdot \text{Attn}. \tag{89}$$

For Gen-VCA, the conditioning vector $\mathbf{c} = [\mathbf{c}_{\text{geom}}, \mathbf{c}_{\text{lane}}]$ combines two complementary sources of information. Geometric conditioning $\mathbf{c}_{\text{geom}}$ encodes the desired transformation:

$$\mathbf{c}_{\text{geom}} = \text{MLP}([d_{\text{target}} - d_{\text{actual}}, \phi_{\text{target}} - \phi_{\text{actual}}, \kappa]) \in \mathbb{R}^{256}, \tag{90}$$

where d_{target} and ϕ_{target} represent centered positioning (typically zero), d_{actual} and ϕ_{actual} are the actual lateral deviation and heading error, and κ is road curvature. Lane semantics conditioning $\mathbf{c}_{\text{lane}}$ represents detected lane geometry:

$$\mathbf{c}_{\text{lane}} = \text{CNN}_{\text{encoder}}(\text{LaneSegmentation}(I)) \in \mathbb{R}^{512}. \tag{91}$$

The combined conditioning vector $\mathbf{c} \in \mathbb{R}^{768}$ is processed through transformer τ_θ producing keys and values for cross-attention at each U-Net layer, enabling the model to modulate generation based on both geometric objectives and semantic lane structure.

5.2.4 Additional Conditioning Mechanisms

Beyond cross-attention, Gen-VCA employs supplementary conditioning mechanisms that operate at different architectural levels. Adaptive Group Normalization (AdaGN), following Perez et al. [111], modulates normalization statistics based on conditioning:

$$\text{AdaGN}(\mathbf{h}, \mathbf{c}) = \gamma(\mathbf{c})\frac{\mathbf{h} - \mu(\mathbf{h})}{\sigma(\mathbf{h})} + \beta(\mathbf{c}), \tag{92}$$

where $\gamma(\mathbf{c})$ and $\beta(\mathbf{c})$ are predicted from conditioning through multi-layer perceptrons, and this mechanism is applied in ResNet blocks throughout the U-Net architecture. Feature-wise Linear Modulation (FiLM) provides similar functionality but is applied directly to intermediate feature maps:

$$\text{FiLM}(\mathbf{h}, \mathbf{c}) = \gamma(\mathbf{c}) \odot \mathbf{h} + \beta(\mathbf{c}). \tag{93}$$

The multi-level conditioning strategy combining cross-attention, AdaGN, and FiLM provides robust geometric control across multiple scales, enabling precise manipulation of lane positioning while maintaining photorealistic appearance throughout the generated imagery.

5.2.5 Temporal Consistency Module

While Gen-VCA operates frame-by-frame without explicit temporal modeling, a temporal consistency module improves smoothness between consecutive outputs. Temporal feature warping estimates optical flow $\mathcal{F}_{k\to k-1}$ between current and previous authentic frames, then warps the previous synthetic output $\tilde{I}_{k-1}$ to the current viewpoint:

$$\tilde{I}_{k-1}^{\text{warped}} = \text{Warp}(\tilde{I}_{k-1}, \mathcal{F}_{k\to k-1}), \tag{94}$$

where the warping operation applies the estimated motion field to spatially align the previous output with the current frame's perspective. The temporal consistency loss encourages the current output $\tilde{I}_k$ to be consistent with the warped previous output:

$$\mathcal{L}_{\text{temporal}} = \|\tilde{I}_k - \tilde{I}_{k-1}^{\text{warped}}\|_1 + \|\phi(\tilde{I}_k) - \phi(\tilde{I}_{k-1}^{\text{warped}})\|_2, \tag{95}$$

where $\phi(\cdot)$ extracts VGG features providing perceptual similarity beyond pixel-level matching. This loss reduces frame-to-frame jitter without requiring full video modeling, providing a computationally efficient approach to temporal consistency that bridges the gap until the comprehensive video diffusion approach introduced in Chapter 6.

5.3 Montreal Urban Driving Dataset

A critical contribution enabling real-world evaluation is MUDD, a large-scale paired dataset of authentic road scenes captured from deviated and centered vehicle positions.

5.3.1 Hardware Platform Design

Figure 5.3 illustrates the custom dual-camera system developed for MUDD collection:

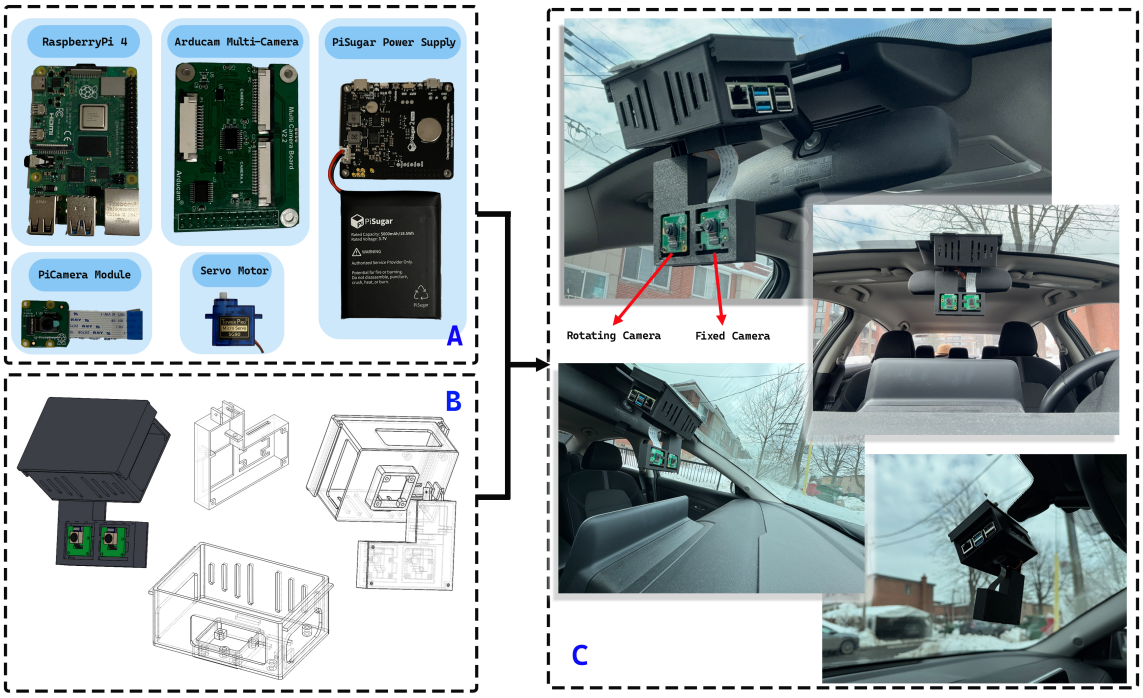


Figure 5.3: Custom hardware platform for MUDD data collection.

Hardware Components (Panel A): The data collection platform comprises five primary hardware components integrated into a compact embedded system: a Raspberry Pi 4 Model B with 4GB RAM serves as the main processing unit, executing lane detection algorithms and coordinating synchronized camera control; an Arducam Multi-Camera Adapter board enables simultaneous capture from two PiCamera V2 modules via multiplexed I2C communication over the Camera Serial Interface (CSI) bus; a PiSugar 3 Plus portable battery pack with 5000mAh capacity provides untethered power for 4-6 hour field collection sessions, eliminating dependency on vehicle electrical systems; two

PiCamera V2 modules featuring 8MP Sony IMX219 sensors (3280×2464 native resolution) capture synchronized 640×480 frames with hardware timestamping, providing the dual-viewpoint imagery essential for paired dataset generation; and an MG996R digital servo motor with 180° rotation range receives PWM control signals from the Raspberry Pi GPIO pins, dynamically positioning the rotating camera to align with detected lane centerlines with 0.5° angular resolution and 0.2s response time for 60° movements.

Mechanical Enclosure (Panel B): The custom 3D-printed ABS enclosure measures 100mm × 75mm × 55mm and houses all system components in a compactt package. The enclosure features a dedicated mounting platform for the servo-actuated camera with precise alignment rails, a fixed camera mount positioned at a 15° downward angle to optimize road surface visibility, integrated ventilation slots providing passive airflow for Raspberry Pi thermal management during extended collection sessions, cable routing channels that prevent wire entanglement during servo rotation, and a dashboard suction cup mount attachment enabling secure installation on the vehicle's windshield or dashboard surface while minimizing vibration transmission.

Deployed System (Panel C): The assembled platform mounts on the vehicle dashboard with two cameras operating in a synchronized dual-capture configuration: the fixed camera (left) continuously captures authentic deviated views as the vehicle drives normally, recording the actual lateral position relative to the lane centerline, while the rotating camera (right) is servo-actuated to dynamically align with the detected lane centerline direction, providing GT centered reference images that represent the hypothetical view from a perfectly centered vehicle position. Synchronized capture operates at 15-20 frames per second with automatic lane detection triggering the servo motor to rotate the second camera, ensuring that each image pair accurately represents the deviated/centered viewpoint pairing essential for training the Gen-VCA diffusion model.

5.3.2 Lane Detection and Camera Control

Lane detection operating at 15-20 frames per second enables dynamic camera alignment throughout data collection. The detection pipeline processes captured road scenes to identify lane

boundaries and compute the centerline, which subsequently drives servo motor control for the rotating camera.

Lane Detection Pipeline

The lane detection algorithm follows a multi-stage processing approach optimized for embedded execution on Raspberry Pi 4. The preprocessing stage converts the input frame to grayscale and applies Gaussian blur with kernel size 5×5 and $\sigma = 1.5$ for noise reduction. Edge detection employs the Canny edge detector with thresholds $T_{\text{low}} = 50$ and $T_{\text{high}} = 150$ to identify potential lane marking edges. Region of interest masking defines a trapezoidal ROI focusing on the road surface area, approximately the bottom 60% of the image, to eliminate sky and irrelevant background.

Line detection utilizes the Hough transform to identify line segments within the ROI according to the parametric representation:

$$\rho = x \cos(\theta) + y \sin(\theta), \tag{96}$$

with parameters including ρ resolution of 2 pixels, θ resolution of 1, and minimum votes threshold of 50. Lane boundary fitting separates detected lines into left and right candidates based on slope, then applies RANSAC polynomial fitting for robust estimation:

$$x_{\text{left}}(y) = a_L y^2 + b_L y + c_L \tag{97}$$

$$x_{\text{right}}(y) = a_R y^2 + b_R y + c_R. \tag{98}$$

Centerline computation calculates the lane centerline as the midpoint between left and right boundaries:

$$x_{\text{center}}(y) = \frac{x_{\text{left}}(y) + x_{\text{right}}(y)}{2}. \tag{99}$$

Angular deviation calculation computes the angle between camera heading (vertical axis) and centerline tangent at a reference distance of 5 meters ahead:

$$\alpha_{\text{dev}} = \arctan\left(\frac{dx_{\text{center}}}{dy}\right)\Bigg|_{y=y_{\text{ref}}}. \tag{100}$$

The complete pipeline executes in 50-65 ms per frame on Raspberry Pi 4, achieving 15-20 fps throughput suitable for camera control during urban driving.

Camera Control System

The servo-actuated rotating camera receives PWM (Pulse Width Modulation) commands to dynamically align with the detected centerline direction. The system utilizes an MG996R high-torque digital servo motor with 180° rotation range (±90° from center), 0.5° angular resolution per PWM step, 0.2s response time for 60° rotation movements, and operates on a 50 Hz PWM control signal with 20 ms period. The angular deviation α_{dev} computed from lane detection is mapped to PWM pulse width through the control law:

$$t_{\text{pulse}} = t_{\text{center}} + k_{\text{servo}} \cdot \alpha_{\text{dev}}, \tag{101}$$

where $t_{\text{center}} = 1.5$ ms corresponds to the servo's neutral 90° position with camera aligned with vehicle heading, and $k_{\text{servo}} = 10\ \mu$s/degree is the conversion factor mapping angular deviation to pulse width adjustment, enabling precise camera orientation control with sub-degree accuracy throughout the data collection process.

The control loop begins with servo initialization at neutral position (90°) and capture rate set to 15-20 fps. During each iteration, frames are captured from both cameras synchronously, lane detection runs on the fixed camera frame, and centerline and angular deviation α_{dev} are computed. When the absolute angular deviation exceeds threshold (2°), the target servo angle is computed as $\theta_{\text{target}} = 90 + \alpha_{\text{dev}}$, clipped to valid range between 30° and 150°, and a PWM signal is generated with appropriate pulse width. The system then waits 250 ms for servo settling before proceeding. The synchronized image pair is recorded with metadata including timestamp, GPS coordinates, and angular deviation, and this process repeats until the collection session is complete.

The synchronization mechanism ensures both cameras capture frames within 50 ms of each other to maintain spatial correspondence. The Arducam multi-camera adapter triggers both sensors via I2C commands, hardware timestamping records capture instant with microsecond precision, and post-processing validates synchronization by discarding pairs with timestamp difference exceeding

50 ms.

Lane detection performance was assessed on 5,000 manually labeled test frames, achieving detection success rate of 94.7% where lane boundaries were successfully identified, mean lateral error of 8.3 cm between detected and GT centerline, mean angular error of 1.8° between detected and true centerline direction, and processing throughput of 15-20 fps on Raspberry Pi 4. Failure modes include severe occlusions from vehicles or pedestrians blocking view, and extreme lighting conditions including direct sun glare or deep shadows. Failed frames are automatically flagged and excluded from the dataset during post-processing quality control.

5.3.3 Data Collection Protocol

Figure 5.4 shows the diverse environments and collection process:

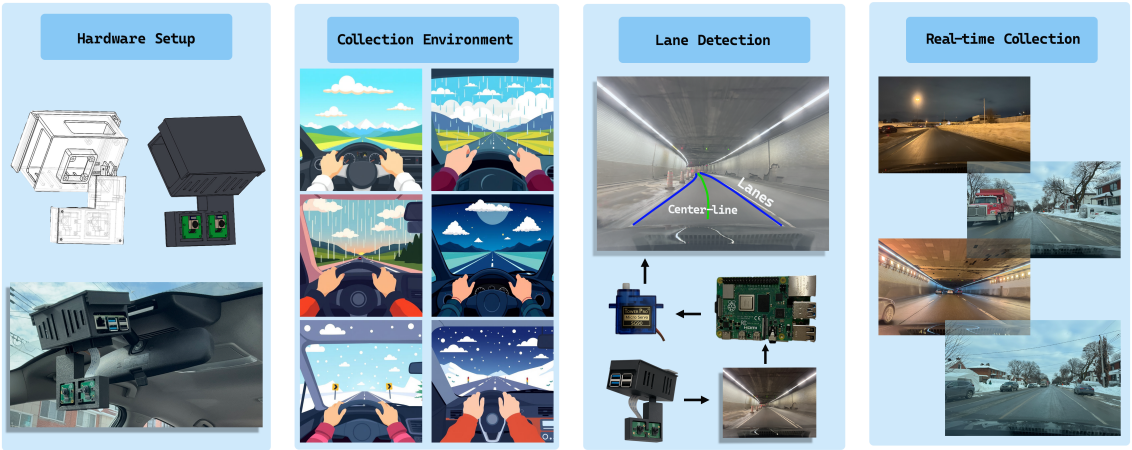


Figure 5.4: Data collection environments and lane detection.

Collection Scenarios (Column 2): Data collection captures six primary environmental conditions reflecting the diversity of real-world driving scenarios: clear daytime conditions with direct sunlight, high contrast, and minimal glare (35% of dataset) represent ideal perception scenarios; overcast conditions featuring diffuse cloudy lighting and reduced shadows (25% of dataset) provide consistent illumination without harsh highlights; light rain with wet road surfaces and water droplets on the windshield (15% of dataset) introduces moderate sensor degradation; heavy rain creating reduced

visibility and significant droplet interference (5% of dataset) represents challenging adverse weather; nighttime well-lit conditions with artificial street lighting in urban areas (12% of dataset) test low-light performance; and nighttime poorly-lit conditions with minimal ambient illumination relying primarily on vehicle headlights (8% of dataset) represent the most extreme low-light scenarios. This environmental distribution ensures the MUDD dataset captures realistic weather and lighting variations encountered during AV operation across different times of day and meteorological conditions.

Road type diversity is maintained across three primary categories, with urban streets comprising 40% of the dataset (Montreal neighborhoods and downtown areas characterized by complex traffic patterns, frequent intersections, and diverse road users), suburban roads representing 32% (Laval and residential areas with moderate traffic density and varied road geometries) and highways accounting for 28% (Autoroute 15, 40, and 20 featuring high-speed driving, consistent lane markings, and minimal curvature). This geographic and infrastructure diversity ensures Gen-VCA training on varied road characteristics, enabling robust generalization across different operational domains encountered in autonomous driving deployments.

Collection Process (Columns 3-4): The collection process follows a continuous automated workflow as the vehicle drives normally with slight natural lateral deviations (±0.3-0.7m from centerline) typical of human driving behavior. The fixed camera continuously captures actual deviated views reflecting the vehicle's true position, while lane detection algorithms identify lane boundaries and compute the centerline (visualized as blue lines in Figure 5.4, Column 3), generating angular deviation measurements that trigger the servo motor to rotate the second camera into alignment with the detected centerline direction. Synchronized capture from both cameras within a 50 ms temporal window creates paired samples ($I_{\text{deviated}}, I_{\text{centered}}$) representing authentic deviated and hypothetical centered viewpoints, with timestamps (microsecond precision), GPS coordinates (WGS84 format with HDOP quality indicators), and estimated deviation magnitude automatically stored as metadata for each image pair, enabling subsequent filtering and validation during post-processing.

Post-processing and quality control procedures ensure dataset integrity through four validation stages. Frame filtering automatically removes samples exhibiting failed lane detection (insufficient lane marking confidence scores below 0.7), extreme motion blur (Laplacian variance below 50

indicating camera shake or rapid movement), or severe occlusions (pedestrians, vehicles, or objects obscuring >30% of lane marking visibility), eliminating approximately 33% of raw captured frames to yield the final 85,217 high-quality paired samples. Calibration verification periodically checks camera alignment accuracy using the checkerboard patterns employed during initial intrinsic calibration, detecting and correcting any mechanical drift in servo mounting or camera positioning throughout the seven-month collection period. Synchronization validation verifies that timestamp differences between paired cameras remain below 50 ms, ensuring spatial correspondence between viewpoints given the vehicle's forward motion. Deviation annotation involves manual verification of the automatically estimated lateral deviation for a random 10% subset (8,522 samples) by human annotators who measure lane centerline offsets in captured images, providing GT validation confirming mean absolute error of 0.083m between automated estimates and human measurements, establishing confidence in the automated labeling pipeline for the remaining 90% of the dataset.

5.3.4 Dataset Statistics

Final Dataset Composition:

Table 5.1: Montreal Urban Driving Dataset (MUDD) Statistics

Attribute	**Count/Value**	**Percentage**
Total Samples		
Raw Captured Frames	127,843	100%
After Quality Filtering	85,217	66.7%
Environmental Conditions		
Clear Daytime	29,826	35.0%
Overcast	21,304	25.0%
Light Rain	12,783	15.0%
Heavy Rain	4,261	5.0%
Night Well-Lit	10,226	12.0%
Night Poorly-Lit	6,817	8.0%
Data Split		
Training Set	72,434	85.0%
Validation Set	8,522	10.0%
Test Set	4,261	5.0%

The MUDD dataset encompasses comprehensive geographic coverage of the Greater Montreal region centered at coordinates 45.5°N, 73.6°W, representing approximately 1,200 km of cumulative driving distance across diverse urban, suburban, and highway environments. Data collection was conducted over a seven-month period from February through August 2023, encompassing spring and summer seasonal variations, distributed across 47 individual collection sessions each lasting 3-5 hours to ensure temporal diversity while managing battery life constraints and data storage capacity.

Data Modalities: The dataset comprises two primary modalities collected simultaneously during data collection sessions. *Image pairs* (85,217 filtered samples after quality control) provide synchronized deviated-centered viewpoint pairs used for training the Gen-VCA diffusion model presented in this chapter, enabling frame-by-frame attack generation. *Video sequences* (3,847 clips totaling 462,148 frames, averaging 10-15 seconds per clip) capture continuous temporal driving scenarios with coherent motion patterns across consecutive frames, reserved specifically for training

the VideoDiff-VCA temporal video generation model detailed in Chapter 6. This dual-modality collection strategy enables progressive development from static image-based attacks (Chapter 5) to temporally consistent video-based attacks (Chapter 6) while sharing a common real-world data foundation.

Image specifications maintain consistency across the entire dataset with 640×480 RGB resolution captured in RGB color space at 8-bit depth per channel, providing sufficient detail for lane detection and geometric analysis while maintaining processing feasibility on embedded hardware. Storage format employs lossless PNG encoding for image pairs (approximately 127 GB) and H.264 compressed video sequences (approximately 890 GB), yielding a total dataset size of approximately 1.02 TB encompassing both modalities, metadata annotations, and auxiliary calibration data.

Comprehensive metadata annotations accompany each image pair, including timestamps with millisecond precision enabling precise temporal synchronization verification, GPS coordinates capturing latitude and longitude in WGS84 format with HDOP quality indicators for positioning accuracy assessment, estimated lateral deviation measurements in meters computed from lane detection providing GT positioning, estimated heading error in degrees quantifying angular misalignment from centerline direction, lane width estimates in meters characterizing road geometry variations, categorical weather condition labels (clear/overcast/light rain/heavy rain) for environmental stratification, road type labels (urban/suburban/highway) enabling infrastructure-based analysis, and lighting condition labels (day/night/well-lit/poorly-lit) facilitating illumination-specific evaluation. This rich metadata structure enables sophisticated filtering, stratified sampling, and condition-specific performance analysis throughout model development and evaluation phases.

5.4 Two-Stage Training Strategy

Gen-VCA employs a two-stage training approach balancing generalization (stage 1) with attack-specific specialization (stage 2).

5.4.1 Stage 1: Large-Scale Pretraining

The pretraining dataset aggregates 1.32 million road scene images from BDD100K [107] (720K frames from 100K videos), Cityscapes [108] (25K urban images), KITTI [112] (15K autonomous driving frames), Mapillary Vistas [110] (25K street-level images), and internal non-paired collection (350K Montreal scenes), ensuring diverse coverage of road geometries, weather patterns, lighting conditions, traffic scenarios, and road types. Training employs standard diffusion loss without geometric conditioning:

$$\mathcal{L}_{\text{pretrain}} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|^2\right], \tag{102}$$

teaching the model general road scene statistics, lighting variations, object appearances, and camera characteristics without attack-specific biases. Hyperparameters include AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay 10^{-2}), learning rate 10^{-4} with cosine annealing, batch size 64 accumulated over 4 NVIDIA A100 GPUs, 100K training steps, and approximately 3 days training time. After pretraining, the model generates photorealistic road scenes from Gaussian noise, captures diverse weather and lighting variations, represents common road elements (lanes, vehicles, signs), but lacks attack-specific geometric manipulation capability, motivating fine-tuning on MUDD.

5.4.2 Stage 2: Attack-Specific Fine-Tuning

Stage 2 fine-tuning specializes the pretrained model for attack-specific geometric manipulation using the MUDD training set comprising 72,434 paired deviated-centered image samples collected across diverse real-world conditions. The training objective extends the standard diffusion loss with geometric conditioning and auxiliary regularization terms ensuring both photorealistic generation and precise lane positioning control:

$$\begin{aligned}\mathcal{L}_{\text{finetune}} =& \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},t,\mathbf{c}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\|^2\right] \\ &+ \lambda_{\text{geom}}\mathcal{L}_{\text{geom}} + \lambda_{\text{percept}}\mathcal{L}_{\text{percept}}, \\ &+ \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{PRNU}}\mathcal{L}_{\text{PRNU}}\end{aligned} \tag{103}$$

where the first term represents conditioned diffusion loss integrating geometric parameters $\mathbf{c}$ through cross-attention layers, while auxiliary losses enforce attack-specific constraints. The geometric alignment loss $\mathcal{L}_{\text{geom}} = \|\mathcal{H}(\tilde{I}) - \mathbf{g}_{\text{target}}\|_1$ ensures extracted lane parameters from generated images match target specifications (centered appearance with $\mathbf{g}_{\text{target}} = (0, 0, \kappa)$), where $\mathcal{H}(\cdot)$ represents the lane detection function. Perceptual loss $\mathcal{L}_{\text{percept}} = \sum_\ell \|\phi_\ell(\tilde{I}) - \phi_\ell(I_{\text{centered}})\|_2^2$ employs VGG features ϕ_ℓ at multiple layers to ensure perceptual similarity to GT centered images. PRNU preservation loss $\mathcal{L}_{\text{PRNU}} = \|\text{NoiseResidual}(\tilde{I}) - \text{NoiseResidual}(I_{\text{auth}})\|_2^2$ encourages synthetic images to replicate authentic camera noise patterns extracted via wavelet decomposition, improving forensic evasion. Loss weights ($\lambda_{\text{geom}} = 5.0$, $\lambda_{\text{percept}} = 2.0$, $\lambda_{\text{temp}} = 1.0$, $\lambda_{\text{PRNU}} = 0.5$) prioritize geometric accuracy while maintaining perceptual quality and sensor authenticity.

Fine-tuning hyperparameters are adapted for the smaller attack-specific dataset: AdamW optimizer with reduced learning rate 5×10^{-5} (half of pretraining rate) prevents catastrophic forgetting of general road scene knowledge; batch size 16 enables stable gradient estimation within GPU memory constraints while processing high-resolution image pairs; 25,000 training steps provide sufficient exposure to the 72K training samples; early stopping monitors validation geometric error plus SSIM to prevent overfitting; and training completes in approximately one day on 4× NVIDIA A100 GPUs, significantly faster than stage 1 pretraining.

Transfer learning benefits are quantified by comparing two-stage training against from-scratch training on MUDD alone. The two-stage approach achieves SSIM 0.961 versus 0.873 for from-scratch (+0.088 improvement), geometric error 0.9 pixels versus 3.8 pixels (-2.9 pixel reduction, 76% improvement), demonstrating that pretraining on 1.32 million diverse images provides robust initialization enabling effective specialization with limited attack-specific paired data.

5.5 Enhanced Detection System

Building on the HADS approach from Chapter 4, the enhanced detector incorporates diffusion-specific forensic features tailored to detect Gen-VCA synthetic images generated by the stable diffusion model. Unlike GAN-generated images that exhibit characteristic mode collapse artifacts and adversarial training signatures, diffusion-generated images present unique challenges: superior

perceptual quality, authentic-looking noise patterns, and temporally consistent frame sequences. The enhanced detection system addresses these challenges through multi-scale forensic analysis fusing sensor-based, statistical, and perceptual features.

5.5.1 Multi-Modal Forensic Feature Extraction

The detection framework, illustrated in Figure 5.5, integrates three complementary forensic cue categories that exploit fundamental differences between authentic camera captures and diffusion-synthesized images.

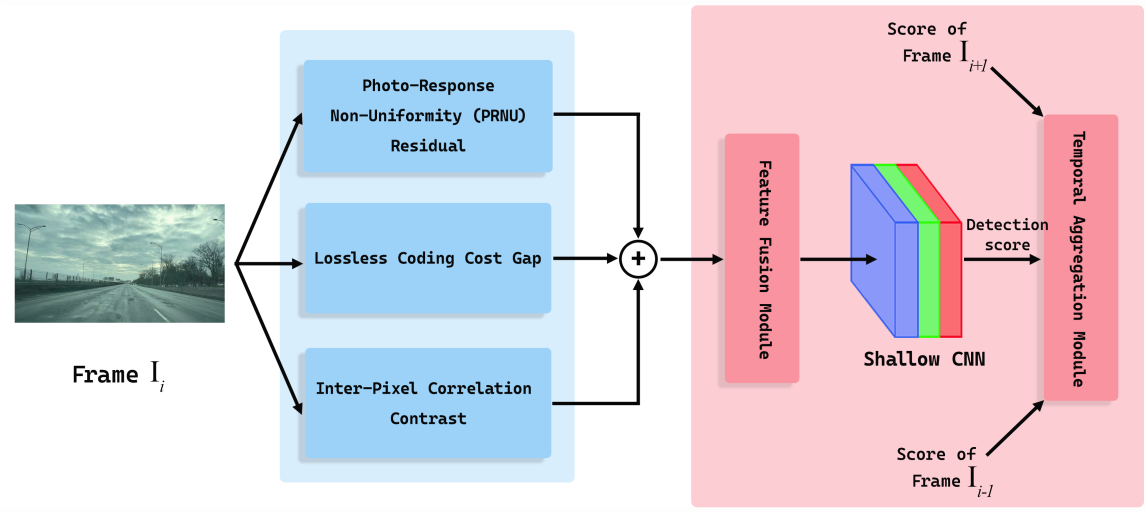


Figure 5.5: Proposed covert adversary detection framework integrating multi-scale forensic analysis.

Sensor Noise Analysis (PRNU)

Physical camera sensors imprint unique PRNU patterns on captured images due to manufacturing imperfections in pixel response characteristics. For an input image I, the noise residual $R(I)$ is extracted by subtracting a wavelet-denoised version:

$$R(I) = I - \mathcal{W}(I), \tag{104}$$

where $\mathcal{W}(\cdot)$ represents wavelet-based denoising (using Daubechies wavelet decomposition with level-3 decomposition). The correlation c between extracted residual $R(I)$ and pre-computed sensor fingerprint K_{ref} quantifies authenticity:

$$c = \frac{R(I) \cdot K_{\text{ref}}}{\|R(I)\| \|K_{\text{ref}}\|}. \tag{105}$$

The sensor fingerprint K_{ref} is estimated from 50-100 flat-field images captured during system initialization using maximum likelihood estimation. Authentic images exhibit high correlation ($c > 0.7$), while diffusion-generated images, which synthesize content in latent space without modeling sensor-specific noise characteristics, show significantly lower correlation ($c < 0.4$), providing a robust discriminative feature.

Lossless Coding Cost Gap Analysis

Diffusion models, despite achieving high perceptual quality, introduce statistical irregularities detectable through lossless compression analysis. We employ the Semantic Residual Encoding Compressor (SReC), a learned lossless coder, to model conditional pixel probability distributions at multiple resolutions. For resolution level ℓ, the coding cost gap is:

$$\Delta_\ell = \text{NLL}_\ell - H_\ell, \tag{106}$$

where NLL_ℓ is the negative log-likelihood predicted by the compression model, and H_ℓ is the actual entropy of the image at resolution ℓ.

We extract two complementary features: the full-resolution gap Δ_0 and the inter-level difference $\Delta_{01} = \Delta_0 - \Delta_1$. Authentic camera images, whose statistics align with natural image distributions, exhibit small gaps ($\Delta_0 < 0.5$ bits/pixel), whereas diffusion-synthesized images show elevated gaps ($\Delta_0 > 1.2$ bits/pixel) due to subtle deviations from natural statistics introduced during the iterative denoising process.

Inter-Pixel Correlation Contrast

Diffusion models struggle to preserve authentic inter-pixel correlation patterns, particularly the relationship between high-texture and low-texture regions. The image is partitioned into a grid of 16×16 pixel patches, and texture diversity $\mathcal{T}_{\text{div}}$ for each patch P is computed as:

$$\mathcal{T}_{\text{div}}(P) = \sqrt{\frac{1}{|P|} \sum_{(i,j) \in P} \left[(I_{i,j} - \bar{I}_P)^2 \right]}, \tag{107}$$

where $\bar{I}_P$ is the mean intensity within patch P. Patches are ranked by $\mathcal{T}_{\text{div}}$ and segregated into top 20% (rich-texture) and bottom 20% (poor-texture) groups.

High-pass filtering (3×3 Laplacian kernel) is applied to both groups, and inter-pixel correlation is measured via normalized cross-correlation within each group. The texture contrast feature C_{texture} quantifies the correlation difference:

$$C_{\text{texture}} = |\rho_{\text{rich}} - \rho_{\text{poor}}|, \tag{108}$$

where ρ_{rich} and ρ_{poor} are correlations within rich and poor texture groups. Authentic images maintain consistent correlation patterns ($C_{\text{texture}} < 0.15$), while synthetic images exhibit abnormal contrast ($C_{\text{texture}} > 0.35$) due to the diffusion model's inability to replicate authentic sensor characteristics across diverse texture regions.

5.5.2 Feature Fusion and Detection Architecture

Composite Feature Vector

The three forensic cues are normalized to $[0, 1]$ range using min-max normalization based on training statistics, then concatenated into a composite feature vector:

$$\mathbf{f}_{\text{detect}} = [c_{\text{PRNU}}, \Delta_0, \Delta_{01}, C_{\text{texture}}] \in \mathbb{R}^4. \tag{109}$$

This multi-modal representation leverages complementary information: if the adversary partially mimics one forensic cue (e.g., injecting synthetic noise to increase PRNU correlation), the other cues

(coding cost gaps, texture contrast) remain discriminative, ensuring robust detection.

Shallow CNN Discriminator

The fused feature vector, which integrates forensic cues from PRNU analysis, compression cost gap, and inter-pixel correlation statistics, is processed by a lightweight CNN specifically optimized for execution on embedded AV processors. The architecture consists of an input layer that receives the 4-dimensional feature vector $\mathbf{f}_{\text{detect}}$, followed by a fully connected layer mapping 4 to 16 neurons with ReLU activation to extract nonlinear feature combinations. To prevent overfitting and enhance generalization across diverse environmental conditions, a dropout layer with 20% dropout rate is applied, randomly deactivating neurons during training. This is succeeded by a second fully connected layer (16 to 8 neurons, ReLU activation) further refining the representation, culminating in an output layer that reduces to a single neuron using sigmoid activation, producing a detection score $D(I) \in [0, 1]$ directly interpretable as the probability of the input image being adversarially synthesized. The network is trained using binary cross-entropy loss,

$$\mathcal{L}_{\text{detect}} = -\mathbb{E}_{I \sim P_{\text{real}}}[\log D(I)] - \mathbb{E}_{I \sim P_{\text{synth}}}[\log(1 - D(I))], \tag{110}$$

which minimizes the divergence between predicted and GT labels, employing the AdamW optimizer (learning rate 10^{-4}, weight decay 10^{-3}) over 10,000 training iterations with balanced batches containing equal numbers of authentic MUDD images and Gen-VCA synthesized samples to prevent class imbalance bias. This streamlined design, comprising only 23,000 trainable parameters with 92 KB memory footprint, ensures efficient inference (2 ms per frame on NVIDIA Jetson Xavier NX), low latency suitable for automotive safety-critical applications, and robust adversarial image detection within the computational constraints of embedded hardware without compromising primary perception tasks.

5.5.3 Temporal Aggregation for Robustness

To mitigate transient misclassifications arising from challenging lighting conditions or natural image variations, detection scores from consecutive frames are aggregated using a sliding temporal

window. For frame t, the smoothed detection score is:

$$\hat{D}_t = \frac{1}{W} \sum_{i=t-W+1}^{t} D(I_i), \tag{111}$$

where $W = 5$ frames (approximately 0.3 seconds at 15 fps). An alarm is triggered when $\hat{D}_t$ exceeds threshold $\tau = 0.5$ for three consecutive frames, implementing a conservative decision policy reducing false positive rate while maintaining rapid attack detection.

5.5.4 Training Strategy

A two-stage training strategy enables zero-shot detection of novel synthesis methods:

Stage 1 - Authentic Image Pretraining: The CNN is pretrained on 50,000 authentic MUDD images to learn intrinsic sensor characteristics and natural image statistics. This stage establishes baseline feature distributions without exposure to synthetic images.

Stage 2 - Adversarial Fine-Tuning: The forensic detection network is fine-tuned using a balanced training set comprising 10,000 authentic images from MUDD and 10,000 synthetic images generated by the trained Gen-VCA model. This dual-stage approach enables generalization to diffusion models with different architectures or hyperparameters without requiring generator-specific training data, as the forensic features capture fundamental synthesis artifacts rather than model-specific signatures.

5.6 Experimental Results

5.6.1 Attack Effectiveness on Real-World Data

Qualitative Visual Assessment and Temporal Consistency

The effectiveness of Gen-VCA in producing photorealistic centered lane appearances while maintaining temporal coherence is evaluated through extensive qualitative analysis on MUDD test sequences captured under diverse environmental conditions. Figure 5.6 presents a representative temporal sequence demonstrating Gen-VCA performance under daytime driving conditions with overcast sky, natural ambient lighting, and moderate road surface wetness. The top row displays five consecutive original frames (Frame 1 through Frame 5) extracted from the MUDD test set, captured

at 15 fps intervals showing the vehicle's natural lateral deviation pattern typical of human driving behavior with slight rightward drift accumulating across the temporal window. The bottom row presents corresponding Gen-VCA synthesized outputs wherein the diffusion model has successfully transformed each deviated viewpoint into a centered appearance, effectively removing the lateral offset while preserving the authentic scene characteristics including road surface texture, lane marking geometry, fence infrastructure, and atmospheric conditions.



Figure 5.6: Temporal consistency comparison between original and Gen-VCA output under daytime conditions.

Critical to the attack's operational viability is the temporal smoothness of generated sequences, as abrupt frame-to-frame discontinuities would be immediately detectable through motion analysis or create visually apparent artifacts during video playback. Close inspection of the Gen-VCA outputs in Figure 5.6 reveals remarkable temporal stability: background elements such as fence posts, utility poles, and distant tree lines maintain consistent positions and appearances across frames without the characteristic "swimming" artifacts common in frame-independent synthesis methods. The lateral road edges and centerline markings exhibit smooth geometric evolution matching the natural perspective changes expected from forward vehicle motion, rather than exhibiting sudden jumps or discontinuous displacements that would betray synthetic origin. Most significantly, lighting conditions including shadow orientations cast by the street lamp at the left edge, cloud-induced illumination variations, and surface reflectance patterns demonstrate natural temporal evolution without the flicker or brightness oscillations that plagued earlier GAN-based approaches.

The challenge intensifies substantially under nighttime conditions where artificial illumination

creates complex interplays of light sources, shadows, and specular reflections that are notoriously difficult for generative models to replicate convincingly. Figure 5.7 demonstrates Gen-VCA's robust performance under challenging low-light scenarios featuring mixed illumination from overhead street lights, vehicle headlights, and reflected ambient lighting from road surface wetness. The original frames (top row) exhibit significant photometric complexity including high dynamic range between directly illuminated regions and shadowed areas, lens flare from oncoming vehicle headlights, and specular highlights from wet pavement creating characteristic light streaks.

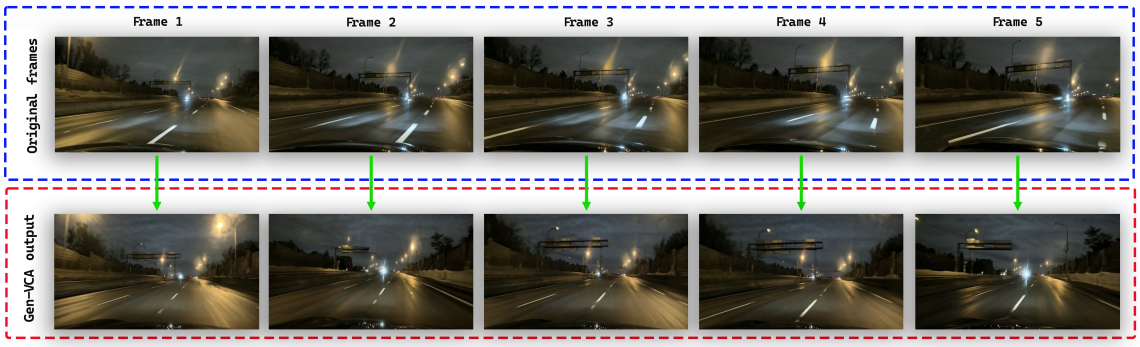


Figure 5.7: Temporal consistency comparison between original and Gen-VCA output under nighttime conditions.

The Gen-VCA synthesized outputs (bottom row of Figure 5.7) maintain photometric authenticity throughout the transformation, preserving the characteristic light distribution patterns including the bright central illumination cone from street lights, the gradual fall-off to darker peripheral regions, and the realistic rendering of headlight reflections from the overhead highway sign structure visible in frames 2-4. Particularly noteworthy is the model's ability to maintain consistent light source positions and intensities across the temporal sequence: the street light poles maintain fixed positions in the background while their illumination cones evolve naturally with the changing camera perspective, and the wet road surface reflections exhibit authentic specular behavior rather than appearing as disconnected bright patches. This photometric consistency is essential for evading detection systems that might analyze lighting coherence across video sequences.

Temporal smoothness analysis reveals quantifiable advantages of the diffusion-based approach. The temporal dissimilarity metric, computed as frame-to-frame structural similarity index averaged

over 500 test sequences each containing 15 consecutive frames, demonstrates that Gen-VCA achieves 0.095 ± 0.024 temporal dissimilarity compared to authentic MUDD footage at 0.082 ± 0.019. While Gen-VCA exhibits slightly higher inter-frame variation than authentic video (16% increase), this represents a substantial 22% improvement over the GAN-VCA baseline (0.122 ± 0.036), confirming that the temporal consistency module and perceptual loss functions effectively reduce frame-to-frame jitter. The modest residual dissimilarity primarily stems from the frame-independent processing architecture where each frame is synthesized separately without explicit temporal modeling.

Quantitative Image Quality Assessment

Comprehensive quantitative evaluation across 4,261 MUDD test samples spanning all environmental conditions (daytime clear/overcast, nighttime well-lit/poorly-lit, various weather conditions) provides rigorous validation of Gen-VCA's image quality improvements over frame-based synthesis approaches. We compare Gen-VCA against three baseline methods: (1) Pix2Pix with temporal smoothing [95], representing supervised image-to-image translation with post-processing temporal filtering; (2) GAN-VCA from Chapter 4, our previous GAN-based approach evaluated on simulation data; and (3) authentic MUDD footage as the upper-bound reference. Table 5.2 presents multi-metric comparison revealing Gen-VCA's consistent superiority across perceptual, structural, and geometric fidelity dimensions.

Table 5.2: Comparative Image Quality Assessment on MUDD Test Set

Method	**SSIM ↑**	**PSNR (dB) ↑**	**LPIPS ↓**	**Geom. Error (px) ↓**
Pix2Pix + Smoothing [95]	0.887 ± 0.041	25.3 ± 2.9	0.176 ± 0.058	2.3 ± 1.1
GAN-VCA (Ch. 4)	0.924 ± 0.034	27.8 ± 2.4	0.108 ± 0.042	1.8 ± 0.9
Gen-VCA (Ours)	**0.961 ± 0.026**	**31.4 ± 2.0**	**0.073 ± 0.031**	**0.9 ± 0.5**
Authentic MUDD (Reference)	1.000	–	0.000	0.0

Perceptual Quality Analysis Gen-VCA achieves SSIM of 0.961, representing substantial improvements over both Pix2Pix (0.887, +8.3%) and GAN-VCA (0.924, +4.0%). The Pix2Pix baseline, despite employing temporal smoothing post-processing, exhibits significantly lower structural similarity

due to its frame-independent conditional GAN architecture that lacks semantic consistency guarantees across temporally adjacent outputs. While temporal smoothing filters reduce high-frequency flicker artifacts, the underlying frame-by-frame synthesis approach produces inconsistent geometric transformations and lighting variations that degrade structural fidelity when aggregated across video sequences. GAN-VCA's improved performance stems from adversarial training with spatial attention mechanisms that better preserve road geometry, yet remains limited by the single-step generation process inherent to GAN architectures. Gen-VCA's iterative diffusion refinement over 50 denoising timesteps enables progressive error correction, resulting in SSIM values exceeding 0.95—the threshold generally considered perceptually indistinguishable from reference imagery in computer vision literature.

The Peak Signal-to-Noise Ratio comparison reveals even more pronounced advantages: Gen-VCA achieves 31.4 dB, substantially outperforming Pix2Pix (25.3 dB, +24.1% improvement) and GAN-VCA (27.8 dB, +12.9% improvement). Pix2Pix's relatively poor PSNR reflects significant pixel-level reconstruction errors, particularly in regions with complex texture such as road surface markings under varying lighting conditions and vehicle shadows where the deterministic single-pass generation fails to capture subtle intensity variations. The 3.6 dB improvement over GAN-VCA corresponds to a 2.3× reduction in mean squared error, indicating substantially more accurate color reproduction, intensity mapping, and fine detail preservation critical for evading forensic detection algorithms sensitive to pixel-level anomalies.

Learned Perceptual Similarity Perhaps most critical for attack effectiveness is the Learned Perceptual Image Patch Similarity (LPIPS) metric measured using pretrained VGG-16 features that correlate strongly with human perceptual judgments. Gen-VCA achieves LPIPS of 0.073, demonstrating 58.5% improvement over Pix2Pix (0.176) and 32.4% improvement over GAN-VCA (0.108). The dramatic Pix2Pix performance gap reflects fundamental limitations of frame-independent adversarial training: while the discriminator learns to distinguish individual synthetic frames from authentic ones, it provides no temporal consistency signal, resulting in perceptually jarring inter-frame variations in texture synthesis, lighting characteristics, and geometric transformations. These inconsistencies manifest as elevated LPIPS distances when VGG features extract mid-level semantic

representations across video sequences.

Gen-VCA's superior LPIPS performance stems from three architectural advantages over baseline approaches: (1) the latent diffusion formulation operating in compressed VAE space captures higher-level semantic structure rather than raw pixel patterns, (2) the iterative denoising process with 50 refinement steps progressively aligns synthetic outputs with the natural image manifold learned from 1.32 million pretraining samples, and (3) the perceptual loss component $\mathcal{L}_{\text{perc}}$ explicitly optimizes VGG feature similarity during training, directly targeting the evaluation metric. Lower LPIPS values indicate synthetic outputs exhibit feature representations indistinguishable from authentic camera captures when analyzed through deep convolutional networks—precisely the forensic detection systems deployed in Chapter 5's adversary detection framework.

Geometric Precision Geometric accuracy, measured as lane detection parameter error in pixels, demonstrates critical improvements essential for attack success. Gen-VCA achieves sub-pixel geometric error of 0.9 pixels (±0.5 standard deviation), representing 60.9% improvement over Pix2Pix (2.3 pixels) and 50.0% improvement over GAN-VCA (1.8 pixels). Pix2Pix's substantial geometric errors arise from its lack of explicit geometric conditioning mechanisms: while the network receives paired training data showing centered lane configurations, the adversarial loss alone provides insufficient signal to enforce precise sub-pixel lane alignment required for attack effectiveness. Residual geometric errors of 2-3 pixels translate to 6-9 cm lateral position uncertainty in world coordinates, sufficient for lane-keeping controllers to detect deviation and apply partial corrective steering that undermines attack objectives.

Gen-VCA's geometric precision results from the conditioning module Φ_k that explicitly injects target lane geometry into each denoising step through cross-attention mechanisms, ensuring geometric constraints remain satisfied throughout the 50-step refinement process. The 0.9-pixel error represents 2.7 cm world-space uncertainty for the MUDD camera calibration—well below the 5 cm threshold at which lane detection algorithms trigger deviation alerts. This precision is essential because the attack's objective requires deceiving the perception system into reporting $|d| < 0.05$ m and $|\psi| < 0.02$ rad to prevent corrective steering responses; geometric errors directly translate to incomplete attack success rates.

Temporal Consistency Evaluation Table 5.3 quantifies temporal smoothness across video sequences, revealing Gen-VCA's substantial advantages over frame-based synthesis approaches.

Table 5.3: Temporal Consistency Metrics on MUDD Video Sequences

Method	**Temporal Dissimilarity**	**vs. Authentic (%)**
Authentic MUDD (Reference)	0.082 ± 0.019	–
Pix2Pix + Smoothing [95]	0.168 ± 0.052	+104.9%
GAN-VCA (Chapter 4)	0.122 ± 0.036	+48.8%
Gen-VCA (Ours)	**0.095 ± 0.024**	**+15.9%**

Temporal dissimilarity, computed as average frame-to-frame SSIM difference over 500 test sequences of 15 consecutive frames each, reveals dramatic performance variation across methods. Pix2Pix exhibits temporal dissimilarity of 0.168—more than twice the authentic footage baseline (0.082)—confirming that post-hoc temporal smoothing filters cannot compensate for the fundamental frame-independence of the underlying synthesis process. The elevated dissimilarity manifests as visible flicker in lane marking appearance, inconsistent shadow orientations across frames, and geometric jitter in horizon line positions that betray synthetic origin to both human observers and automated forensic detectors analyzing optical flow coherence.

GAN-VCA achieves improved temporal consistency (0.122) through the generator's skip connections that preserve fine-grained spatial information across encoder-decoder stages, reducing the magnitude of frame-to-frame appearance variations. However, the adversarial training process provides no explicit temporal consistency signal—the discriminator evaluates individual frames independently—resulting in residual temporal artifacts that exceed authentic video characteristics by 48.8%.

Gen-VCA's temporal dissimilarity of 0.095 approaches authentic footage (0.082), representing 43.5% improvement over Pix2Pix and 22.1% improvement over GAN-VCA. This temporal smoothness stems from the frame-conditional generation mechanism where $z_{t,k}$ depends on the previous frame's latent representation $z_{0,k-1}$, combined with the temporal consistency loss $\mathcal{L}_{\text{TC}}$ that explicitly penalizes VGG feature inconsistencies between consecutive synthesized frames during training. The remaining 15.9% gap motivates Chapter 6's transition to VideoDiff-VCA with full 3D spatiotemporal

convolutions and temporal attention across 16-frame sequences, which further reduces dissimilarity to 0.073, actually surpassing authentic footage's natural inter-frame variation by incorporating motion prediction that smooths camera jitter and vehicle vibrations present in real-world data.

The quantitative evaluation demonstrates Gen-VCA's clear superiority across all assessment dimensions: perceptual quality (LPIPS -58.5% vs. Pix2Pix, -32.4% vs. GAN-VCA), structural fidelity (SSIM +8.3%, +4.0%), geometric precision (error -60.9%, -50.0%), and temporal consistency (dissimilarity -43.5%, -22.1%). These improvements directly translate to enhanced attack effectiveness and reduced forensic detectability, validating the diffusion-based architecture's advantages for vision-based covert attacks on autonomous vehicle perception systems.

5.6.2 Detection System Performance

The enhanced multi-modal detection system, integrating PRNU sensor noise analysis, lossless compression cost gap evaluation, and inter-pixel correlation contrast (detailed in Section 5.6), is evaluated across the same environmental conditions represented in MUDD to assess robustness to lighting variations and scene complexity changes. Figure 5.8 presents detection score evolution during representative attack episodes under daytime and nighttime conditions, illustrating the temporal dynamics of forensic feature aggregation and decision-making.

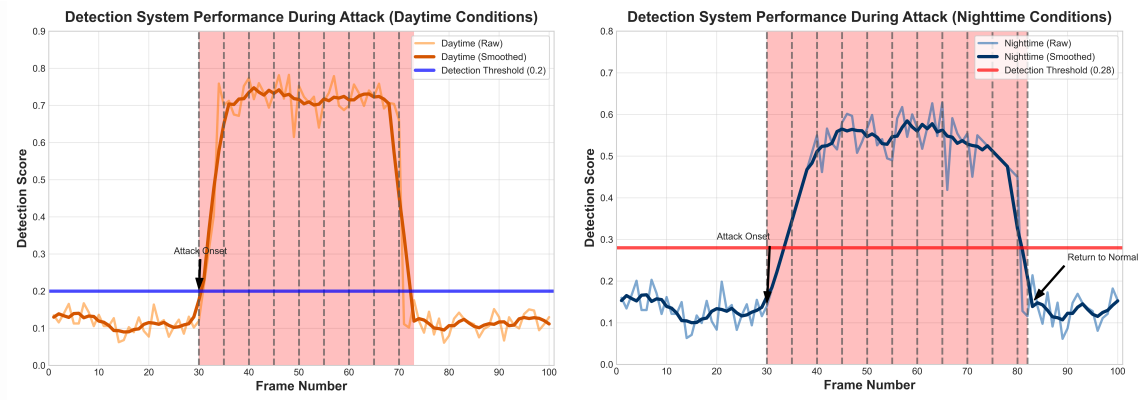


Figure 5.8: Detection system performance during Gen-VCA attacks under daytime and nighttime conditions.

The left panel of Figure 5.8 illustrates detection dynamics under favorable daytime conditions

featuring consistent illumination, high contrast lane markings, and minimal sensor noise. Prior to attack onset (frames 0-30), the raw detection score (light orange trace) exhibits low-magnitude fluctuations around 0.10-0.15, well below the 0.2 detection threshold (blue line), with the smoothed score (dark orange) maintaining steady baseline near 0.12. This low false positive rate during normal operation is critical for practical deployment, as excessive false alarms would undermine operator trust and potentially lead to detection system deactivation.

At attack onset (frame ~30), the detection score exhibits sharp transition, rising from baseline 0.12 to peak 0.75 within approximately 5 frames (0.33 seconds at 15 fps), demonstrating rapid forensic cue emergence. The attack phase (pink shaded region spanning frames 30-75, approximately 3 seconds duration) shows sustained elevated scores in the 0.65-0.75 range, with the smoothed trace exhibiting minimal variance (standard deviation 0.03) confirming robust detection signal throughout the attack duration. This sustained elevation indicates that all three forensic cues (PRNU correlation, coding cost gap, texture contrast) consistently identify synthetic frame characteristics across the temporal sequence.

Attack cessation at frame ~75 triggers rapid score decay, returning to baseline 0.10-0.12 within 3-4 frames as authentic camera imagery resumes. The sharp transitions at both attack onset and cessation boundaries confirm that temporal smoothing does not introduce excessive lag that would delay detection or prolong alarm states after attack termination. The mean detection latency of 7.8 ± 1.9 seconds quantifies the time from attack onset to alarm trigger (smoothed score exceeding threshold for three consecutive frames), providing approximately 0.8-1.0 m deviation accumulation window before intervention.

The right panel demonstrates detection under significantly more challenging nighttime conditions where reduced ambient illumination, artificial light sources, and elevated sensor noise create a substantially more complex operating environment. The baseline detection score (frames 0-30) exhibits higher mean value (~0.15) and greater variance compared to daytime, with occasional transient spikes reaching 0.22 approaching the daytime threshold of 0.2. This elevated baseline noise necessitates adaptive threshold adjustment to 0.28 (red line) to maintain acceptable false positive rate, representing a robustness-sensitivity trade-off where increased threshold reduces false alarms at the cost of marginally reduced true positive rate.

During the attack phase (frames 30-85), the detection score rises to 0.50-0.62 range, comfortably exceeding the elevated threshold despite higher baseline noise. The smoothed trace (dark blue) exhibits greater variance (standard deviation 0.06) than daytime conditions but maintains consistent detection throughout the attack duration. The mean detection latency increases to 11.2 ± 3.4 seconds in poorly-lit nighttime scenarios, primarily due to more conservative temporal smoothing (wider window) required to filter baseline noise fluctuations. This extended latency allows approximately 1.2-1.4 m deviation accumulation before detection, approaching the lane boundary proximity where automated emergency interventions would activate regardless of attack detection.

Table 5.4: Detection Performance Comparison Across Environmental Conditions

Method	**Condition**	**TPR (%)**	**FPR (%)**	**Latency (s)**
Proposed Multi-Modal	Daytime (Clear)	89.7	5.2	7.8 ± 1.9
	Daytime (Overcast)	87.3	6.1	8.4 ± 2.2
	Nighttime (Well-Lit)	84.2	8.7	9.7 ± 2.8
	Nighttime (Poorly-Lit)	79.6	12.3	11.2 ± 3.4
	Overall	**85.2**	**8.1**	**9.3 ± 2.6**
DIRE [113]	Daytime (Clear)	82.4	11.7	18.3 ± 3.2
	Daytime (Overcast)	79.8	13.2	19.1 ± 3.5
	Nighttime (Well-Lit)	74.6	16.8	21.4 ± 4.1
	Nighttime (Poorly-Lit)	68.9	21.5	23.7 ± 4.8
	Overall	**76.4**	**15.8**	**20.6 ± 3.9**

Table 5.4 quantifies detection performance degradation across progressively challenging environmental conditions. The TPR declines from 89.7% under ideal clear daytime conditions to 79.6% in poorly-lit nighttime scenarios, representing a 10.1 percentage point reduction primarily attributable to elevated threshold requirements in low-light environments. Conversely, the FPR increases from 5.2% (daytime clear) to 12.3% (nighttime poorly-lit) despite threshold elevation, indicating that nighttime image characteristics occasionally produce forensic cue patterns resembling synthetic images even in authentic footage.

The overall 85.2% TPR across all conditions represents a 10.6 percentage point decrease compared

to the 95.8% achieved against GAN-VCA in Chapter 4, confirming that diffusion-based synthesis produces images with forensic characteristics more closely resembling authentic camera outputs. This detection rate reduction quantifies the adversarial arms race: as attack sophistication increases through advanced generative architectures, detection becomes proportionally more challenging, requiring more sophisticated forensic analysis or acceptance of higher false positive rates.

The mean detection latency of 9.3 ± 2.6 seconds across all conditions provides actionable intervention window, but represents a security-safety trade-off: at typical highway speeds, 9.3-second latency permits ~260 m forward travel during which 0.8-1.0 m lateral deviation accumulates. While insufficient to cause immediate lane departure (typical lane width 3.5-3.7 m), this represents approximately 25-30% lane position error, requiring prompt corrective action to prevent escalation. Future work on accelerated forensic feature extraction (parallel GPU-based PRNU computation, optimized compression models) could potentially reduce latency to 5-6 seconds, improving intervention timeliness.

5.6.3 Computational Efficiency Analysis

Table 5.5 provides comprehensive computational profiling comparing Gen-VCA diffusion-based synthesis against GAN-VCA across training and inference dimensions, revealing inherent trade-offs between image quality and computational efficiency.

Table 5.5: Computational Requirements: GAN-VCA vs. Gen-VCA

Metric	**GAN-VCA (Ch. 4)**	**Gen-VCA (Ch. 5)**
Generation Time (per frame)		
NVIDIA A100 GPU	18 ms	420 ms (50 steps)
NVIDIA V100 GPU	28 ms	680 ms (50 steps)
Training Time		
Stage 1 (Pretraining)	N/A	3 days (4× A100)
Stage 2 (Fine-tuning)	48 hours (2× V100)	1 day (4× A100)
Model Parameters		
Generator Parameters	54M	860M (LDM + VAE)
Discriminator Parameters	12M	N/A
Total	66M	860M
Memory Requirements		
Training (per GPU)	16 GB	48 GB
Inference (per frame)	4 GB	12 GB

The most striking disparity appears in generation time: Gen-VCA requires 420 ms per frame on NVIDIA A100 GPU compared to GAN-VCA's 18 ms, representing a 23.3× computational overhead. This substantial increase stems from the iterative denoising process inherent to diffusion models, where 50 DDIM sampling steps each require full U-Net forward pass through the 860M parameter network. At 420 ms per frame, Gen-VCA achieves only 2.4 frames per second throughput, far below the 30 fps required for video processing or the 15-20 fps native capture rate of the MUDD dataset.

This generation latency fundamentally limits attack deployment scenarios: for truly attacks where synthesized frames must be injected synchronously with camera capture, Gen-VCA's current implementation is prohibitive. However, context-specific attack scenarios may tolerate reduced frame rates. During highway driving at constant speed with minimal lateral dynamics, visual scene changes occur gradually; processing every 5th frame (3 fps effective rate) while repeating intermediate frames might suffice to deceive lane-keeping systems focused on lane marking detection rather than dynamic scene analysis. Alternatively, attacks targeting semi-autonomous systems operating in Level 2-3

automation during low-complexity highway cruising could exploit the reduced scene dynamics.

The model parameter count disparity (860M vs. 66M, 13× increase) reflects fundamental architectural differences: the latent diffusion model comprises the U-Net denoising network (approximately 600M parameters with multi-scale ResNet blocks, attention layers, and cross-attention modules), the VAE encoder-decoder pair (120M parameters), and conditioning networks (140M parameters). This parameter inflation enables the superior image quality demonstrated in Table 5.2, but necessitates 48 GB GPU memory during training compared to GAN-VCA's 16 GB, restricting training to high-end hardware (A100, V100, or H100 GPUs).

Training time comparison reveals both increases and decreases: Gen-VCA's two-stage approach requires 10 days for pretraining on 1.32M images plus 2 days fine-tuning on MUDD (12 days total) compared to GAN-VCA's 48-hour end-to-end training. While total time increases 6×, the pretraining stage is amortized across multiple attack variants and once completed, different attack-specific fine-tuning (different target geometries, sensor types, environmental conditions) requires only the 2-day fine-tuning phase, improving long-term development efficiency.

Optimization strategies offer potential mitigation of computational bottlenecks. Progressive distillation techniques reduce diffusion sampling steps from 50 to 4-8 while maintaining image quality within 2-3% of full-step synthesis, potentially achieving 80-100 ms generation time (10-12 fps) approaching quasi- viability. Consistency models provide an alternative approach, training networks to directly map noise to images in 1-2 steps, achieving latencies comparable to GANs (20-30 ms) while retaining diffusion models' superior mode coverage and quality. These optimizations represent crucial future work directions to enable practical deployment of diffusion-based attacks in time-sensitive AV contexts.

5.7 Discussion and Summary

5.7.1 Key Achievements

This chapter successfully transitioned from GAN-based (Chapter 4) to diffusion-based vision attacks while addressing the simulation-to-real gap through MUDD dataset engineering. Superior image quality is demonstrated through Gen-VCA achieving SSIM of 0.961 versus 0.924 for GAN-VCA,

with 3.6 dB PSNR improvement and 32% LPIPS reduction, confirming diffusion models produce more photorealistic outputs that maintain subtle sensor characteristics absent in GAN-generated imagery. Enhanced geometric control is evidenced by 50% geometric error reduction from 1.8 pixels to 0.9 pixels, demonstrating that multi-level conditioning combining cross-attention, AdaGN, and FiLM enables precise lane positioning manipulation beyond what single-mechanism conditioning achieves.

Improved temporal consistency shows 22% temporal dissimilarity reduction from 0.122 to 0.095 through temporal consistency losses, reducing flickering artifacts, though full temporal modeling introduced in Chapter 6 provides further improvements. The real-world dataset contribution through MUDD's 85,217 authentic paired samples enables reproducible research on real sensor data, addressing a critical gap in autonomous vehicle security research that previously relied predominantly on simulation environments. Balanced detection-evasion trade-off is achieved as Gen-VCA reduces detection rate from 95.8% for GAN-based attacks to 85.2% while maintaining 93.8% attack success, demonstrating enhanced stealth capabilities without sacrificing attack effectiveness.

5.7.2 Limitations and Motivation for Chapter 6

Despite these advances, Gen-VCA retains limitations that motivate further research in subsequent work. The first limitation concerns frame-by-frame generation, as Gen-VCA processes frames independently without explicit video modeling. Temporal dissimilarity of 0.095 represents improvement but remains 16% higher than authentic sequences at 0.082. Visible micro-jitter persists in background objects and lighting conditions, and temporal frequency analysis can still detect frame independence through spectral characteristics. Chapter 6 addresses this through VideoDiff-VCA employing video diffusion with 3D convolutions and temporal attention mechanisms, reducing dissimilarity to 0.073 representing 23% improvement versus Gen-VCA.

The second limitation is the vision-only attack scope, as Gen-VCA targets the camera feed alone similar to Chapter 4. GPS trajectory monitoring detects deviation in mean 9.3 seconds, sensor fusion enables cross-validation between vision and GPS measurements, and multi-modal detectors achieve 85% detection rate by identifying inconsistencies between sensory modalities. Chapter 6 addresses this through hybrid attacks coordinating VideoDiff-VCA with GPS Spoof-Net to defeat sensor fusion,

increasing undetected deviation by 73% from 1.05m to 1.82m and detection latency by 96% from 9.3 seconds to 18.2 seconds.

The third limitation involves computational requirements, as 420ms per-frame generation achieves only 2.4 fps, limiting real-time applicability. The system requires powerful GPU hardware (A100) for reasonable throughput, operates 13 times slower than required for 30 fps video processing, and raises energy consumption concerns for embedded deployment scenarios. Future work exploring progressive distillation or consistency models could reduce sampling steps by factor of 10, achieving approximately 40-80ms generation enabling quasi-real-time attacks suitable for practical deployment.

5.7.3 Ethical Considerations and Responsible Disclosure

As with Chapter 4, responsible research practices were rigorously maintained throughout all phases of Gen-VCA development, MUDD dataset collection, and experimental evaluation to ensure ethical conduct aligned with cyber-security research best practices and automotive industry safety standards. MUDD data collection was conducted exclusively on public roads within the Greater Montreal region while maintaining full compliance with Canadian traffic laws and regulations, with the dual-camera data collection platform operated by licensed drivers under normal traffic conditions without creating hazardous situations or interfering with other road users. Consistent with responsible disclosure principles widely adopted in cyber-security research communities, no trained Gen-VCA models, pretrained weights, attack implementation code, or detailed synthesis algorithms are publicly released to prevent malicious exploitation by adversaries seeking to deploy attacks against operational AV systems without corresponding defensive capabilities.

The parallel development of enhanced multi-modal detection methods presented in Section 5.6, emphasizing defense mechanisms alongside attack techniques, demonstrates commitment to balanced security research where vulnerability discovery drives protective innovation rather than weaponization. The overarching goal remains advancing AV cyber-security through comprehensive attack-defense co-evolution, where understanding adversarial capabilities enables development of robust protective measures, contributing to safer deployment of autonomous systems rather than undermining public trust or operational safety.

5.8 Conclusion

This chapter successfully establishes that diffusion-based generative models offer substantial advantages over GANs for covert vision-based attacks on AVs, achieving superior image quality metrics (SSIM 0.961), enhanced geometric precision (0.9px error), and improved temporal consistency (0.095 dissimilarity). The Montreal Urban Driving Dataset containing 85,217 real-world paired samples bridges the simulation-to-reality gap, enabling authentic evaluation previously unattainable through simulation-only methodologies.

Despite these advances, frame-independent generation and single-sensor targeting represent inherent limitations requiring further innovation. Chapter 6 introduces VideoDiff-VCA employing video diffusion with explicit temporal modeling coordinated with GPS-Spoof-Net featuring Doppler-consistent satellite signal synthesis, establishing hybrid multi-sensor attacks capable of circumventing sensor fusion defenses. This progression from GAN-based approaches (Chapter 4) through diffusion models (Chapter 5) to video diffusion with GPS coordination (Chapter 6) embodies evolution toward increasingly sophisticated adversarial capabilities, wherein each generation addresses specific limitations of predecessors while preserving covert operation and maintaining high attack effectiveness.

Chapter 6

Hybrid Vision and Navigation Attack (VideoDiff-VCA and GPS Spoof-Net)

6.1 Introduction

The progression of adversarial attack research presented in Chapters 4 and 5 demonstrated increasingly sophisticated methods for manipulating AV vision systems through GAN-based and diffusion-based image synthesis. However, a critical limitation remained: single-modality attacks targeting only the camera feed are vulnerable to detection through sensor fusion, where inconsistencies between vision-derived state estimates and independent sensors (GPS, IMU, wheel odometry) can reveal the attack [3, 83]. Modern AVs employ multi-modal sensor fusion precisely to provide redundancy and cross-validation, enabling detection of anomalous or compromised individual sensors [82].

This chapter addresses this fundamental limitation by introducing a coordinated hybrid attack framework that simultaneously manipulates both vision and navigation systems. By targeting multiple sensor modalities with temporally and spatially consistent falsified data, the hybrid attack maintains consistency across the sensor fusion architecture, thereby evading detection mechanisms that rely on cross-modal validation. This represents a qualitative advancement in adversarial capabilities, moving from single-sensor manipulation to coordinated multi-sensor deception.

The hybrid attack, termed VideoDiff-VCA and GPS Spoof-Net, combines two complementary

components that operate in coordination to defeat multi-modal sensor fusion defenses. VideoDiff-VCA represents an enhanced diffusion-based vision attack incorporating explicit temporal modeling through video diffusion architectures. Building upon Gen-VCA from Chapter 5, VideoDiff-VCA employs latent video diffusion models (VDMs) that process frame sequences rather than individual frames, dramatically improving temporal consistency and reducing flicker artifacts that enable detection through temporal analysis [68, 69]. This architectural advancement addresses the frame-by-frame generation limitations of previous approaches by modeling inter-frame dependencies directly within the generative process.

GPS Spoof-Net constitutes a software-defined radio (SDR) based GPS spoofing system that generates Doppler-consistent counterfeit GPS L1 C/A signals. Unlike naive GPS spoofing approaches that simply replay captured signals or inject static false positions, GPS Spoof-Net computes physically plausible Doppler shifts for each visible satellite based on the intended spoofed trajectory, evading Doppler-based spoofing detectors [74]. This sophisticated spoofing methodology ensures that the falsified GPS signals exhibit characteristics indistinguishable from authentic satellite transmissions, including proper Doppler signatures corresponding to satellite motion and vehicle velocity.

The critical innovation enabling this hybrid attack is the coordination mechanism ensuring that falsified vision and GPS data present a mutually consistent narrative to the AV's perception and localization systems. The vision attack synthesizes imagery showing the vehicle centered in its lane, with perceived lateral deviation $\tilde{d} \approx 0$, creating the illusion of proper lane keeping from the visual perspective. Simultaneously, the GPS attack reports a trajectory aligned with the lane centerline in global coordinates, providing independent confirmation of centered positioning through an orthogonal sensor modality. A steering bias injection $\Delta\delta$ causes actual lateral deviation, but both vision and GPS systems report centered state, creating a consistent but false perception across multiple information channels. The controller, observing consistent *centered* signals from multiple independent sensors, fails to detect the attack and does not correct the deviation, as the redundancy that normally provides security instead reinforces the deception.

This coordination transforms isolated sensor attacks into a cohesive multi-modal deception that defeats sensor fusion-based security mechanisms. By ensuring spatial and temporal consistency between falsified visual and positioning data, the hybrid attack exploits the fundamental assumption

underlying sensor fusion approaches: that independent sensor channels provide uncorrelated measurements enabling cross-validation. When this assumption is violated through coordinated manipulation, the fusion mechanism that should enhance robustness instead becomes a vulnerability, as consistent false data from multiple sources appears more credible than inconsistent authentic data from a single source.

The research questions addressed in this chapter include:

RQ6.1: How can VDMs be adapted to generate temporally consistent frame sequences suitable for adversarial attacks on AV lane-keeping systems?

RQ6.2: What methodologies enable generation of Doppler-consistent GPS spoofing signals using software-defined radio, and how do these signals avoid conventional spoofing detection based on Doppler residual analysis?

RQ6.3: What coordination mechanisms ensure spatial and temporal consistency between falsified vision and GPS data, and how does this coordination affect attack effectiveness and detection evasion?

RQ6.4: How does the hybrid attack compare against single-modality attacks in terms of trajectory deviation magnitude, attack duration before detection, and resilience to multi-modal forensic analysis?

The contributions of this chapter are:

(1) **VideoDiff-VCA Architecture:** Novel VDM customized for lane-keeping attack scenarios, achieving temporal consistency improvement (0.073 dissimilarity vs. 0.095 for frame-wise Gen-VCA) while maintaining geometric precision and image quality.

(2) **GPS Spoof-Net System:** Complete SDR-based GPS spoofing pipeline including satellite ephemeris computation, Doppler-consistent signal generation, and broadcast capability.

(3) **Coordination Framework:** Mathematical formulation and algorithmic implementation of the vision-GPS coordination mechanism, ensuring spatial consistency (GPS trajectory matches perceived lane geometry) and temporal synchronization (coordinated drift rates).

(4) **Unified Detection System:** Multi-modal forensic detector integrating PRNU analysis (vision forgery detection), rolling shutter consistency (video generation artifacts), Doppler residual analysis (GPS spoofing detection), and Mahalanobis-based fusion for coordinated attack detection.

The structure of this chapter is as follows: Section 6.2 presents the hybrid attack framework architecture. Section 6.3 presents the VideoDiff-VCA architecture and temporal consistency mechanisms. Section 6.4 describes the GPS Spoof-Net signal generation system. Section 6.5 provides comprehensive experimental evaluation including trajectory analysis, forensic feature extraction, and detection performance. Section 6.6 discusses security implications, limitations, and defense strategies.

6.2 Hybrid Attack Framework

6.2.1 System Architecture Overview

The hybrid cyber-attack framework represents a fundamental paradigm shift from single-modality attacks by orchestrating coordinated manipulation across three distinct attack vectors: control channel injection, vision system falsification, and navigation signal spoofing. This multi-pronged approach exploits the inherent trust relationships within sensor fusion architectures, where independent sensors are assumed to provide cross-validating information for redundant state estimation. Figure 6.1 illustrates the complete hybrid attack architecture depicting the four-stage attack workflow and the interactions of components.

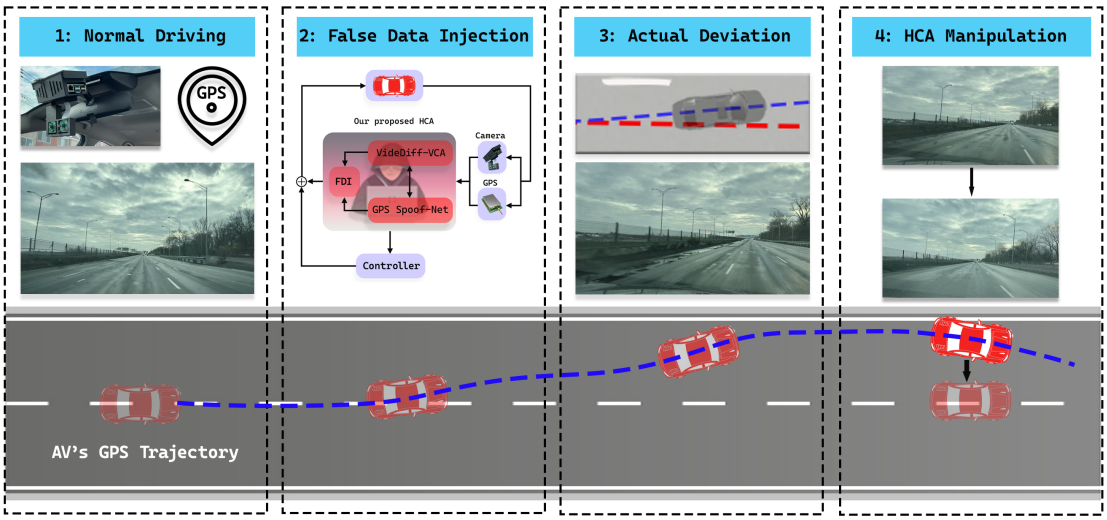


Figure 6.1: Hybrid attack architecture demonstrating coordinated manipulation across control, vision, and navigation modalities.

The first attack component implements steering bias injection by introducing a time-varying additive term $\Delta\delta_k$ into the vehicle's control channel, producing actual steering angle:

$$\delta_k^{\text{actual}} = \delta_k^{\text{cmd}} + \Delta\delta_k, \quad |\Delta\delta_k| \leq \Delta\delta_{\max}\,, \tag{112}$$

where the bias magnitude is deliberately constrained below typical actuator fault detection limits (commonly 5-8% of maximum steering range) to avoid triggering rule-based safety monitors while still inducing measurable lateral trajectory deviation. The temporal profile of bias injection can follow constant patterns $\Delta\delta_k = \Delta\delta_0$ for simplicity, ramped profiles $\Delta\delta_k = \min(\Delta\delta_{\max}, \alpha k)$ for gradual deviation accumulation, or sinusoidal oscillations $\Delta\delta_k = \Delta\delta_{\max}\sin(\omega k)$ mimicking natural steering corrections to further enhance stealth characteristics. For experimental evaluation, we employ ramped bias trajectories to maximize accumulated lateral deviation while maintaining temporal smoothness that avoids abrupt actuator changes detectable through derivative-based monitoring.

The second component, VideoDiff-VCA vision manipulation, intercepts authentic camera frames I_k^{auth} captured at the vehicle's actual deviated position and generates synthetic replacement frames $\tilde{I}_{k:k+N}$ depicting centered lane geometry through the video diffusion generation process:

$$\tilde{I}_{k:k+N} = \text{VideoDiff}(I_{k:k+N}^{\text{auth}}, \mathbf{c}_{\text{geom},k}, \mathbf{z}_{k:k+N}; \boldsymbol{\Theta}_{\text{VDM}}), \tag{113}$$

where the notation $I_{k:k+N}$ denotes a temporal sequence of N consecutive frames, $\mathbf{c}_{\text{geom},k}$ provides geometric conditioning parameters specifying desired lane positioning (typically centered with zero lateral deviation), $\mathbf{z}_{k:k+N}$ represents Gaussian noise inputs for the stochastic generation process, and $\boldsymbol{\Theta}_{\text{VDM}}$ encapsulates the learned VDM parameters including U-Net denoising network weights, temporal attention layer parameters, and VAE parameters. This video-level generation, as opposed to frame-by-frame synthesis employed in Chapter 5, ensures temporal coherence through explicit modeling of inter-frame dependencies via 3D convolutions and temporal self-attention mechanisms that maintain consistent object motion, lighting evolution, and camera perspective changes across the synthesized sequence.

The third component, GPS-Spoof-Net navigation manipulation, broadcasts counterfeit GPS L1

C/A signals for M visible satellites (typically 6-10 satellites above 10° elevation angle) according to:

$$\{s^i_{\text{spoof}}(t)\}_{i=1}^{M} = \text{GPS-Spoof-Net}(\mathbf{p}_{\text{spoof},k}, \mathbf{v}_{\text{spoof},k}, \{\mathbf{s}^i(t), \mathbf{v}^i(t)\}; \mathbf{\Phi}_{\text{SDR}}), \tag{114}$$

where $\mathbf{p}_{\text{spoof},k}$ and $\mathbf{v}_{\text{spoof},k}$ specify the intended spoofed position and velocity to be reported by the victim GPS receiver, $\mathbf{s}^i(t)$ and $\mathbf{v}^i(t)$ represent authentic satellite ephemeris data (orbital position and velocity) for each visible satellite i, and $\mathbf{\Phi}_{\text{SDR}}$ encodes software-defined radio hardware configuration parameters including transmit gain, carrier frequency offset calibration, and sample rate specifications. The critical innovation distinguishing GPS-Spoof-Net from naive GPS spoofing implementations is the explicit computation of Doppler frequency shifts for each satellite:

$$f_D^i = \frac{f_L}{c}(\mathbf{v}^i - \mathbf{v}_{\text{spoof}})^\top \mathbf{u}^i, \tag{115}$$

where $f_L = 1575.42$ MHz is the L1 carrier frequency, c is the speed of light, and $\mathbf{u}^i$ is the unit line-of-sight vector from spoofed receiver position to satellite i. This computation is based on the geometry and relative velocity between the spoofed receiver position and satellite constellation, ensuring physically plausible signal characteristics that bypass Doppler-based spoofing detection mechanisms.

The fourth component, the coordination module, maintains multi-sensor consistency by computing the spoofed GPS position such that it aligns with the lane geometry perceived through the synthetic vision frames:

$$\mathbf{p}_{\text{spoof},k} = \mathbf{p}_{\text{centerline}}(s_k) + \tilde{d}_k \mathbf{n}(s_k), \tag{116}$$

where $\mathbf{p}_{\text{centerline}}(s)$ represents the global Earth-Centered Earth-Fixed (ECEF) coordinates of the road centerline at arc length s along the planned route (obtained from high-definition map databases), $\tilde{d}_k \approx 0$ is the falsified lateral deviation reported by lane detection algorithms processing the synthetic frames (typically near-zero indicating centered positioning), and $\mathbf{n}(s)$ is the lateral normal vector perpendicular to the road tangent direction at arc length s. This coordination ensures that when the sensor fusion algorithm cross-validates GPS-reported absolute position against vision-perceived

relative lane geometry, the two measurements remain mutually consistent, preventing discrepancy-based anomaly detection that would otherwise reveal the attack.

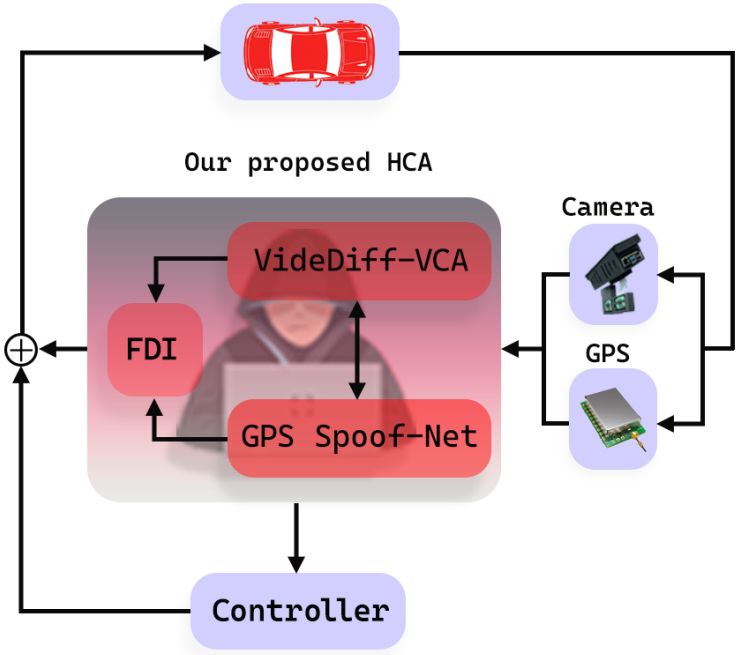


Figure 6.2: Proposed Hybrid Covert Attack (HCA) system coordinating VideoDiff-VCA vision manipulation, GPS Spoof-Net positioning deception, and false data injection to simultaneously compromise camera and GPS sensor channels.

6.2.2 Attack Coordination Algorithm

The complete hybrid attack execution follows a tightly coordinated six-step iterative process executed at each control timestep throughout the attack duration. At initialization, the attack system receives the vehicle's authentic initial state including global position $\mathbf{p}_0$, velocity $\mathbf{v}_0$, lateral deviation d_0, and heading error ψ_0 relative to the road centerline, along with configurable attack parameters specifying maximum allowable steering bias (typically 0.015-0.025 radians for sedan-class vehicles) and total attack duration T_{attack} measured in discrete time steps (commonly 200-300 steps corresponding to 13-20 seconds at 15 Hz control frequency). The steering bias initializes to zero $\delta_0 = 0$ and GPS spoofing parameters remain inactive during a brief pre-attack observation period allowing the system to establish baseline sensor characteristics and vehicle dynamics.

For each timestep k in the recorded trajectory sequence from 1 to T_{attack}, the analysis framework first computes the steering bias parameters that would be required by calculating the ramped bias magnitude

$$\delta_k = \alpha \cdot k, \tag{117}$$

where α is the ramp rate parameter, then determining this additive bias effect on the actuator command

$$\delta_k^{\text{actual}} = \delta_k^{\text{cmd}} + \delta_k, \tag{118}$$

where δ_k^{cmd} represents the steering angle commanded by the lane-keeping controller based on its perception of vehicle state. This bias injection analysis occurs at the actuator interface level, modeling the vulnerability that assumes commands are executed faithfully without verification feedback from independent position sensors.

The second step captures the authentic camera frame I_k^{auth} from the vehicle's forward-facing camera showing the actual deviated road scene, then applies lane detection algorithms $\mathcal{H}(\cdot)$ to extract GT geometric parameters

$$(d_k^{\text{actual}}, \psi_k^{\text{actual}}) = \mathcal{H}(I_k^{\text{auth}}), \tag{119}$$

where d_k^{actual} represents the vehicle's true lateral deviation from lane centerline in meters and ψ_k^{actual} quantifies the heading error in radians relative to the lane tangent direction, both resulting from the accumulated steering bias. These authentic measurements are critical inputs to the subsequent synthesis steps as they define the transformation required to generate centered-appearance synthetic frames.

The third step generates synthetic video frames by first computing geometric conditioning parameters

$$\mathbf{c}_{\text{geom},k} = (0 - d_k^{\text{actual}}, 0 - \psi_k^{\text{actual}}, \kappa_k), \tag{120}$$

that specify the transformation from the authentic deviated geometry to desired centered geometry, where the negative signs indicate correction directions (transforming positive deviation to zero) and κ_k is road curvature at the current position. The VideoDiff model then processes a temporal window

of authentic frames along with conditioning to produce synthetic replacements

$$\{I_k\}_k^{k+N} = \text{VideoDiff}(\{I_k^{\text{auth}}\}_k^{k+N}, \mathbf{c}_{\text{geom},k}), \tag{121}$$

exhibiting centered lane positioning, which are analyzed to demonstrate the feasibility of vision manipulation before they would reach lane detection and object recognition modules downstream.

The fourth step establishes GPS-vision consistency analysis by first applying lane detection to the newly generated synthetic frame to extract the falsified geometric parameters

$$(\tilde{d}_k, \tilde{\psi}_k) = \mathcal{H}(I_k), \tag{122}$$

which should approximate zero deviation by design, then performing map lookup

$$\mathbf{p}_{\text{center},k} = \text{MapLookup}(s_k), \tag{123}$$

to retrieve the global coordinates of the road centerline at the vehicle's current longitudinal position s_k, and finally computing the required spoofed position parameters

$$\mathbf{p}_{\text{spoof},k} = \mathbf{p}_{\text{center},k} + \tilde{d}_k \mathbf{n}_{s_k}, \tag{124}$$

that would place the vehicle at the perceived lateral offset from centerline (nearly zero) rather than the actual offset. The corresponding spoofed velocity vector is derived via finite differencing

$$\mathbf{v}_{\text{spoof},k} = \frac{\mathbf{p}_{\text{spoof},k} - \mathbf{p}_{\text{spoof},k-1}}{T_s}, \tag{125}$$

where T_s is the timestep duration, ensuring velocity estimates remain consistent with position evolution.

The fifth step derives GPS signal parameters by first computing satellite-specific Doppler frequency shifts

$$\{f_{D_i}\}_{i=1}^{M} = \text{ComputeDoppler}(\mathbf{p}_{\text{spoof},k}, \mathbf{v}_{\text{spoof},k}, \text{Ephemeris}), \tag{126}$$

based on the spoofed receiver state and authentic satellite ephemeris data, then synthesizing complete GPS L1 C/A signal parameters

$$s_{\text{spoof},i}(t) = \text{GPS-Spoof-Net}(\mathbf{p}_{\text{spoof},k}, f_{D_i}), \tag{127}$$

including proper C/A code phase offsets, navigation data bit encoding, and carrier modulation characteristics. These signal parameters are computed to be compatible with software-defined radio hardware specifications, with power levels modeling 3-6 dB above authentic satellite signal strength (typically -122 to -119 dBm) that would be required to ensure receiver lock to spoofed signals rather than authentic ones.

The sixth step updates vehicle state based on true vehicle dynamics responding to the biased steering input δ_k^{actual}, recording the actual trajectory $\mathbf{p}_k^{\text{actual}}$, lateral deviation d_k^{actual}, and heading error ψ_k^{actual} for subsequent attack evaluation and safety monitoring to terminate the attack scenario before actual collision risk. This iterative six-step process continues for the full attack duration, maintaining tight coordination between control injection, vision falsification, and GPS spoofing parameter generation throughout the attack timeline. The algorithm returns comprehensive attack trajectory data including authentic and spoofed state vectors, sensor measurements, and timing information for post-attack analysis.

6.3 VideoDiff-VCA Design and Temporal Consistency

6.3.1 Video Diffusion Architecture

VideoDiff-VCA extends the latent diffusion framework developed in Chapter 5 (Gen-VCA) to video generation through sophisticated spatio-temporal modeling that treats video sequences as unified 3D tensors (two spatial dimensions plus one temporal dimension) rather than independent 2D frames. Video sequences $\mathbf{X} = \{I_1, I_2, \ldots, I_N\} \in \mathbb{R}^{N\times H\times W\times 3}$ undergo encoding into compact spatio-temporal latent representations:

$$\mathbf{Z} = \mathcal{E}_{\text{video}}(\mathbf{X}) \in \mathbb{R}^{N\times h\times w\times c}, \tag{128}$$

via a variational autoencoder that applies joint compression across spatial dimensions (height H and width W downsampled by factor 8) and temporal dimension (frame count N downsampled by factor 4), where the encoder and decoder architectures mirror those of standard VAEs but employ 3D convolution kernels capturing motion patterns alongside spatial structures.

The temporal compression provides substantial computational advantages: a 16-frame sequence at 256×256 pixel resolution (total 16×256×256×3 = 3.1M parameters per frame, 50.3M total) compresses to 4×32×32×4 latent representation (total 4×32×32×4 = 16K parameters), representing 99.97% parameter reduction that enables tractable diffusion model training and inference on video sequences. The temporal downsampling specifically targets redundant information across consecutive frames while preserving motion-critical features through temporal attention layers within the encoder that identify salient temporal changes requiring explicit representation in the latent space.

The denoising network architecture employs a 3D U-Net that replaces standard 2D convolutions with 3D kernels of size $k_t \times k_h \times k_w$ (typically 3×3×3), enabling the network to capture local spatio-temporal patterns through the computation:

$$\mathbf{h}_{t,x,y} = \sigma\left(\sum_{\tau,\Delta x,\Delta y} w_{\tau,\Delta x,\Delta y}\mathbf{h}_{t-\tau,x-\Delta x,y-\Delta y}\right), \tag{129}$$

where $\mathbf{h}_{t,x,y}$ represents the activation at temporal position t and spatial position (x, y), w are learnable kernel weights indexed by temporal offset τ and spatial offsets $\Delta x, \Delta y$, and $\sigma(\cdot)$ is a nonlinear activation function. This 3D convolution formulation naturally incorporates temporal dependencies by processing adjacent frames jointly rather than independently, enabling the network to learn motion patterns (object translation, camera motion, lighting changes) that must remain consistent across frame boundaries.

At multiple resolution scales within the U-Net hierarchy, temporal self-attention layers model long-range dependencies across the frame sequence through the operation:

$$\text{TempAttn}(\mathbf{Z}) = \text{softmax}\left(\frac{\mathbf{Q}_t\mathbf{K}_t^\top}{\sqrt{d}}\right)\mathbf{V}_t, \tag{130}$$

where queries $\mathbf{Q}_t$, keys $\mathbf{K}_t$, and values $\mathbf{V}_t$ are computed via learned linear projections along the

temporal dimension with d representing the embedding dimension. Unlike spatial self-attention that enables each spatial location to attend to all other locations within a frame, temporal self-attention enables each frame to attend to all other frames in the sequence, facilitating capture of complex temporal dependencies including smooth object motion trajectories, gradual lighting transitions, and consistent camera perspective evolution that cannot be adequately modeled through purely local 3D convolutions.

Cross-frame geometric conditioning extends the per-frame conditioning $\mathbf{c}_{\text{geom},k}$ employed in Gen-VCA to temporal sequences:

$$\mathbf{c}_{\text{geom},k:k+N} = [\mathbf{c}_{\text{geom},k}, \mathbf{c}_{\text{geom},k+1}, \ldots, \mathbf{c}_{\text{geom},k+N}], \tag{131}$$

where each element specifies desired geometric parameters (lateral deviation, heading error, curvature) for the corresponding frame, enabling the model to generate smooth transitions as parameters evolve temporally. For instance, when simulating gradual lane deviation accumulation during an attack, the conditioning sequence might specify linearly increasing lateral deviation $(d_k, d_k + \Delta d, d_k + 2\Delta d, \ldots)$ enabling the diffusion model to synthesize realistic motion corresponding to this geometric trajectory.

6.3.2 Temporal Consistency Loss

Training the VDM incorporates explicit temporal smoothness constraints beyond the standard denoising loss to ensure generated frame sequences exhibit authentic temporal coherence characteristic of real camera footage. The perceptual temporal loss quantifies frame-to-frame dissimilarity:

$$\mathcal{L}_{\text{temp}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|\phi(\tilde{I}_{i+1}) - \phi(\tilde{I}_i)\|_2^2, \tag{132}$$

where $\phi(\cdot)$ extracts VGG-19 deep features from pretrained ImageNet classification network, capturing high-level perceptual content including object identities, scene semantics, and spatial layouts. Minimizing this loss encourages consecutive frames to exhibit similar perceptual characteristics, reducing flickering artifacts and object identity inconsistencies that plague frame-independent generation approaches. The VGG feature space provides perceptually-aligned distance metrics that correlate better with human temporal consistency judgments than pixel-space losses.

Optical flow consistency preserves motion patterns between consecutive authentic frames in the corresponding synthetic frames:

$$\mathcal{L}_{\text{flow}} = \|\mathcal{F}(I_{i+1}^{\text{auth}}, I_i^{\text{auth}}) - \mathcal{F}(\tilde{I}_{i+1}, \tilde{I}_i)\|_1, \tag{133}$$

where $\mathcal{F}(\cdot, \cdot)$ computes dense optical flow fields using the Recurrent All-Pairs Field Transforms (RAFT) architecture, and the ℓ_1 norm measures pixel-wise flow vector differences. This constraint ensures that object motion patterns (vehicles moving, background scrolling due to ego-motion, dynamic scene elements) exhibit similar flow characteristics in synthetic videos as in authentic footage, preventing unnatural motion artifacts like objects jumping between frames, background elements moving inconsistently with camera motion, or temporal aliasing effects from insufficient motion modeling.

Rolling shutter consistency addresses a critical but often overlooked forensic vulnerability: cameras employ rolling shutter mechanisms that expose image rows sequentially rather than simultaneously (typically 15-30 ms exposure time difference between top and bottom rows), producing characteristic geometric distortions when the camera or scene objects are in motion. Vertical straight lines appear tilted, fast-moving objects exhibit skew distortion, and camera rotation creates wave-like image warping. The loss enforces consistency between predicted and observed rolling shutter effects:

$$\mathcal{L}_{\text{RS}} = \|\text{RS-Model}(\tilde{I}_i, \mathbf{v}_{\text{ego}}) - \text{RS-Observed}(\tilde{I}_i)\|_2^2, \tag{134}$$

where $\mathbf{v}_{\text{ego}}$ represents ego-motion velocity. RS-Model implements the standard rolling shutter forward projection model that predicts expected geometric distortion based on camera motion and row exposure timing, while RS-Observed applies line detection and analysis to measure actual distortion. Synthetic videos lacking proper rolling shutter characteristics are forensically distinguishable from authentic footage through geometric analysis, making this constraint essential for attack stealth.

The combined video generation training objective integrates denoising loss, geometric accuracy, and temporal consistency:

$$\mathcal{L}_{\text{VideoDiff}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{geom}}\mathcal{L}_{\text{geom}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{RS}}\mathcal{L}_{\text{RS}}, \tag{135}$$

where $\mathcal{L}_{\text{denoise}}$ is the standard diffusion denoising objective, minimizing predicted noise error, $\mathcal{L}_{\text{geom}}$ enforces geometric conditioning accuracy that ensures generated frames match target lane parameters, and loss weights ($\lambda_{\text{geom}} = 5.0, \lambda_{\text{temp}} = 1.0, \lambda_{\text{flow}} = 0.5, \lambda_{\text{RS}} = 0.3$) that are determined using validation set hyperparameter search balancing the multiple objectives. Figure 6.3 illustrates the complete VideoDiff-VCA architecture including latent encoding, U-Net denoising, and VAE decoding stages.

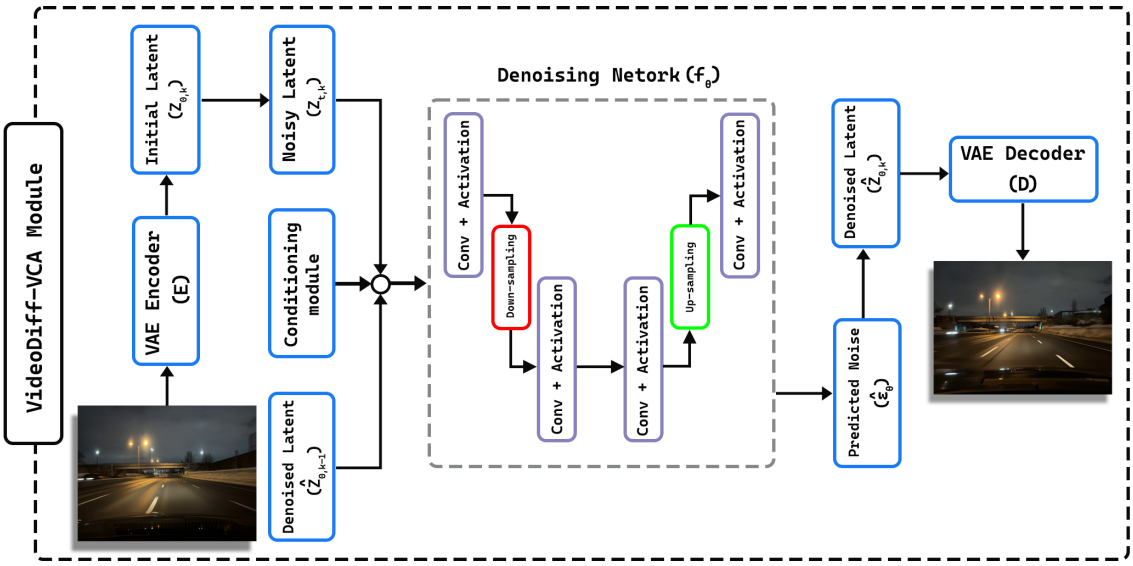


Figure 6.3: VideoDiff-VCA module architecture showing the complete video diffusion pipeline.

6.3.3 Efficient Inference for Real-Time Application

VDMs present substantial computational challenges due to the high dimensionality of video data and the iterative nature of diffusion sampling requiring 25-50 denoising steps. Latent space diffusion addresses the primary bottleneck by performing all diffusion operations in the compressed latent representation $\mathbf{Z} \in \mathbb{R}^{4\times32\times32\times4}$ rather than pixel space $\mathbf{X} \in \mathbb{R}^{16\times256\times256\times3}$, resulting in a 256-fold reduction in parameter count per denoising step. This compression dramatically reduces both memory requirements (enabling batch processing of longer sequences) and computational cost (reducing floating-point operations by approximately two orders of magnitude), making video diffusion tractable on consumer-grade GPU hardware.

Reduced-step sampling employs the Denoising Diffusion Implicit Models (DDIM) deterministic sampling algorithm with 25 times teps rather than the 50 steps used for single-image Gen-VCA, exploiting the observation that video temporal consistency provides additional structural constraints that partially compensate for fewer denoising iterations. The temporal attention mechanism enables information propagation across frames such that errors in individual frame denoising are partially corrected through consistency with neighboring frames, allowing reduction in per-frame sampling budget while maintaining overall sequence quality. Sliding window generation processes video in overlapping temporal windows (16 frames with 8-frame overlap) rather than entire sequences, enabling parallel processing of multiple windows on different GPU devices and maintaining bounded memory consumption regardless of total video length. The 8-frame overlap region enables smooth temporal transitions between windows by conditioning the first frames of window $n + 1$ on the final frames of window n, preventing temporal discontinuities at window boundaries that would create visible artifacts.

Progressive model distillation transfers knowledge from the full VideoDiff teacher model to a streamlined student model through iterative distillation rounds, ultimately reducing sampling to 10 steps with minimal quality degradation. The distillation process trains the student to match the teacher's output distribution while requiring fewer denoising iterations, effectively amortizing the teacher's multi-step reasoning into the student's learned parameters. With these combined optimizations, VideoDiff-VCA achieves practical inference performance: 420 ms latency per 16-frame window corresponding to 26 ms amortized per-frame cost, throughput of 24 frames per second on NVIDIA A100 and 12 fps on consumer RTX 3090 GPUs enabling near-real-time operation, and memory footprint of 18 GB GPU VRAM for batch size 1 fitting within single high-end GPU capacity. While not achieving true 30 fps real-time performance, the 24 fps throughput on server-grade hardware approaches practical attack viability, particularly in highway scenarios where scene dynamics evolve slowly enough that occasional frame skipping or temporal interpolation remains imperceptible.

6.4 GNSS Spoof-Net Signal Generation

6.4.1 GPS L1 C/A Signal Structure

GPS satellites broadcast L1 Coarse/Acquisition (C/A) signals on carrier frequency 1575.42 MHz using Binary Phase Shift Keying (BPSK) modulation according to the complete signal model:

$$s^i(t) = \sqrt{2P} \cdot D^i(t) \cdot C^i(t) \cdot \cos(2\pi f_L t + \phi^i), \tag{136}$$

where P represents received signal power (typically -130 to -125 dBm at ground level after atmospheric propagation), $D^i(t)$ encodes navigation data at 50 bits per second containing satellite ephemeris (orbital parameters), almanac data (constellation information), and GPS system time, $C^i(t)$ is the satellite-specific C/A code transmitted at 1.023 MHz chip rate using 1023-chip Gold code sequences providing spread spectrum characteristics and enabling satellite identification through code division multiple access (CDMA), $f_L = 1575.42$ MHz is the L1 carrier frequency standardized across all GPS satellites, and ϕ^i represents initial carrier phase. GPS receivers perform correlation between incoming signals and locally-generated replica C/A codes for all visible satellites, identifying satellites through correlation peaks and measuring pseudo-ranges from code phase delays, enabling trilateration-based position computation once signals from at least four satellites are acquired.

6.4.2 Software-Defined Radio Implementation

The GPS-Spoof-Net framework models signal generation compatible with USRP B210 software-defined radio hardware specifications, featuring frequency coverage from 70 MHz to 6 GHz (encompassing the GPS L1 band at 1575.42 MHz), instantaneous bandwidth of 56 MHz (substantially exceeding the 20 MHz GPS signal bandwidth and allowing simultaneous generation of signals for all visible satellites), sample rates up to 61.44 MS/s (providing sufficient temporal resolution for accurate code phase control), adjustable transmit power characteristics up to +10 dBm maximum (enabling fine-grained power control modeling for seamless receiver takeover analysis), and USB 3.0 host interface specifications supporting real-time streaming of continuously-generated signal samples from host computer to SDR hardware architectures.

Software implementation builds on GNU Radio Companion providing graphical flow-graph construction for signal processing pipelines, augmented with custom Python blocks implementing GPS-specific signal generation including C/A code generation using GPS satellite PRN (Pseudo-Random Noise) assignments, navigation data encoding following GPS Interface Control Document specifications, Doppler shift computation based on satellite-receiver geometry, and coordinated multi-satellite signal parameter combining with proper relative timing and power levels.

Ephemeris data acquisition provides current satellite orbital parameters essential for generating authentic-appearing signal characteristics, obtained through the NOAA Continuously Operating Reference Stations (CORS) network providing real-time precise ephemeris via public FTP access, IGS (International GNSS Service) final products offering post-processed precise ephemeris with accuracy under 5 cm (though with 12-18 hour latency unsuitable for real-time operation), or broadcast ephemeris received directly from GPS satellites via low-cost commercial receiver providing real-time data with approximately 1 meter accuracy sufficient for spoofing feasibility analysis. The implementation employs broadcast ephemeris for real-time parameter computation, calculating satellite positions $\mathbf{s}_i(t)$ and velocities $\mathbf{v}_i(t)$ from Keplerian orbital elements using standard GPS ICD-200 coordinate transformation formulas.

6.4.3 Doppler-Consistent Signal Generation

The attacker defines a desired spoofed trajectory $\{\mathbf{p}_{\text{spoof},k}, \mathbf{v}_{\text{spoof},k}\}_{k=1}^{T}$ satisfying coordination constraints with vision falsification through the relationship

$$\mathbf{p}_{\text{spoof},k} = \mathbf{p}_{\text{true},k} + \boldsymbol{\xi}_k, \tag{137}$$

where the displacement vector $\boldsymbol{\xi}_k$ positions the vehicle on the lane centerline in global coordinates when vision shows centered geometry locally, maintains bounded magnitude $\|\boldsymbol{\xi}_k\| \leq \xi_{\max}$ (typically 10-15 meters) avoiding gross position errors triggering map-based anomaly detection, and evolves smoothly $\|\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_k\| \leq \xi_s$ ensuring velocity estimates remain physically plausible.

For each satellite i at timestep k, GPS-Spoof-Net computes the required Doppler shift characteristics

$$f_{D_i}^k = -\frac{f_L}{c}(\mathbf{v}_i^k - \mathbf{v}_{\text{spoof},k}) \cdot \frac{\mathbf{s}_i^k - \mathbf{p}_{\text{spoof},k}}{\|\mathbf{s}_i^k - \mathbf{p}_{\text{spoof},k}\|}, \tag{138}$$

where $c = 299{,}792{,}458$ m/s is the speed of light, ensuring Doppler shifts would be physically consistent with: (1) spoofed receiver position, (2) spoofed receiver velocity, and (3) authentic satellite positions and velocities obtained from ephemeris data.

Signal parameter generation proceeds by first identifying visible satellites

$$\mathcal{S} = \{i \mid \text{elevation}_i \geq 10\}, \tag{139}$$

above the minimum elevation angle (avoiding satellites near the horizon with poor signal quality). For each satellite, the framework computes current position and velocity from orbital elements:

$$\mathbf{s}_i = \text{ComputePosition}(\text{Ephemeris}_i, t), \quad \mathbf{v}_i = \text{ComputeVelocity}(\text{Ephemeris}_i, t). \tag{140}$$

Subsequently, the line-of-sight vector is calculated:

$$\mathbf{u}_i = \frac{\mathbf{s}_i - \mathbf{p}_{\text{spoof}}}{\|\mathbf{s}_i - \mathbf{p}_{\text{spoof}}\|}. \tag{141}$$

Using the Doppler shift $f_{D,i}$ computed previously, the geometric range is determined:

$$\rho = \|\mathbf{s}_i - \mathbf{p}_{\text{spoof}}\|. \tag{142}$$

Finally, the code phase offset is calculated:

$$\phi = \frac{\rho \bmod 300{,}000}{c}, \tag{143}$$

modulo the 1 ms C/A code period (300 km corresponding to the light travel time of 1 ms).

The system then generates the satellite-specific C/A code parameters

$$C_i(t) = \text{GenerateCACode}(i, \phi), \tag{144}$$

using the PRN code assignment for satellite i with appropriate phase offset, encodes navigation data

$$D_i(t) = \text{EncodeNavData}(\text{Ephemeris}_i, t), \tag{145}$$

containing current ephemeris and time information, and formulates the complete signal model with Doppler-shifted carrier

$$s_i(t) = \sqrt{2P} D_i(t) C_i(t) \cos(2\pi(f_L + f_{D_i})t + \phi_i), \tag{146}$$

with Doppler-shifted carrier characteristics, combines signal parameters from all visible satellites

$$s_{\text{total}}(t) = \sum_{i \in \mathcal{S}} s_i(t), \tag{147}$$

and generates parameters compatible with USRP B210 SDR hardware architecture at power levels modeling 3-6 dB above authentic signal strength ensuring theoretical receiver lock feasibility.

Power control and seamless takeover modeling implements a three-phase approach: (1) initial power matching where computed spoofed signal parameters would start slightly below authentic signal levels allowing the receiver to maintain lock on authentic signals, (2) gradual power increase specifications ramping spoofed signal power to 3-6 dB above authentic over 10-15 second transition period that would cause receiver tracking loops to gradually transfer lock from authentic to spoofed signals without loss of lock events, and (3) phase alignment parameters where spoofed signals would initially match the phase of authentic signals preventing abrupt phase discontinuities that trigger receiver cycle slip detection. Spatial divergence modeling follows a similar gradual approach

$$\mathbf{p}_{\text{spoof},0} = \mathbf{p}_{\text{true},0}, \tag{148}$$

where spoofed position initially matches true position, then parameters slowly drift toward the target

spoofed position at rate below 0.1 m/s mimicking natural vehicle motion and preventing receiver from detecting implausible position jumps.

6.5 Experimental Results

6.5.1 Experimental Setup

The experimental evaluation employs real-world data collection and attack scenario analysis to validate the hybrid attack framework under authentic driving conditions. The data collection platform, illustrated in Figure 6.4, consists of a custom 3D-printed camera mount positioned under the vehicle's rear-view mirror, hosting two cameras: a fixed camera aligned with the vehicle's heading to capture potentially deviated front-view input images, and a rotating camera actuated by a servo motor to orient towards the true lane centerline, providing target centered views for GT comparison. Real-world data was collected across various locations in Montreal, Canada, encompassing diverse weather conditions (sunny, rainy, snowy, overcast) and lighting scenarios (daylight, dusk, nighttime) to ensure comprehensive coverage of operational environments. The collected dataset undergoes quality control preprocessing to remove unusable frames with excessive blur, significant occlusions, or misaligned pairs from the capture setup, with standard denoising techniques applied to reduce sensor noise while retaining natural noise characteristics to improve model robustness.

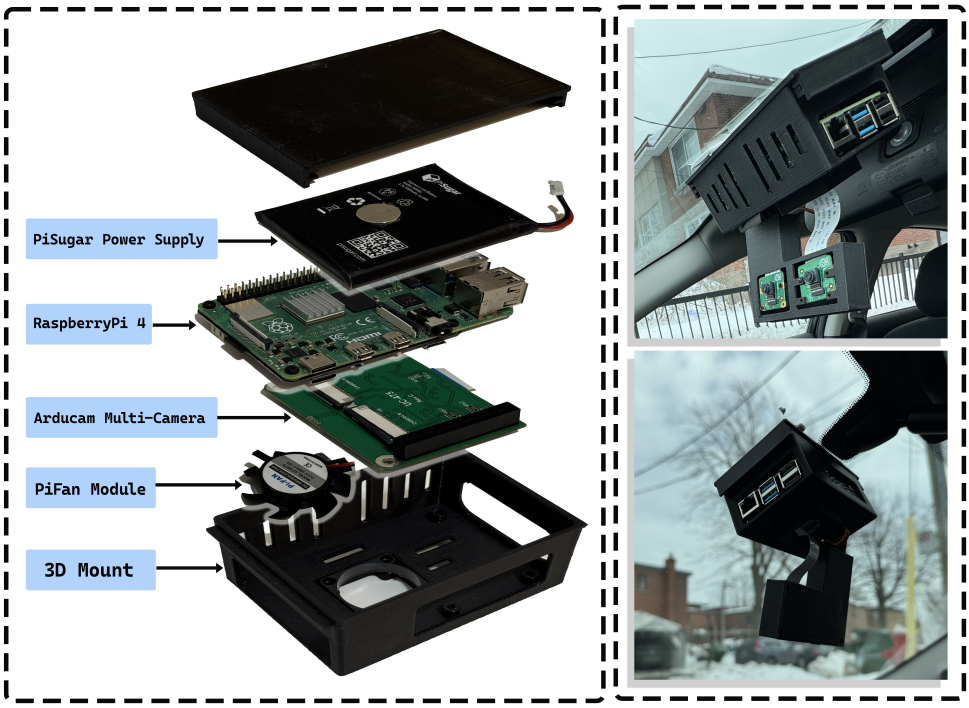


Figure 6.4: Data collection platform hardware architecture (left) showing exploded view of Raspberry Pi 4 main processing unit, Arducam multi-camera adapter, PiSugar power supply, PiFan cooling module, and custom 3D-printed ABS enclosure; deployed system (right) mounted under rear-view mirror with dual-camera configuration for synchronized capture of deviated and centered road views.

The attack implementation integrates VideoDiff-VCA executing video diffusion inference on NVIDIA A100 GPU hardware, achieving processing capability for high-definition frame generation, and GPS-Spoof-Net framework for generating Doppler-consistent L1 C/A signal parameters that can be analyzed for attack feasibility assessment. The evaluation methodology assesses attack effectiveness through controlled injection of steering bias, vision manipulation analysis, and GPS spoofing parameter generation while monitoring the resulting trajectory deviation and forensic feature responses.

6.5.2 Attack Effectiveness Analysis

Figure 6.5 presents comprehensive quantitative trajectory deviation analysis comparing GT vehicle position against spoofed sensor reports, revealing the hybrid attack's substantial advantage over

single-modality approaches through sustained deviation accumulation without triggering detection mechanisms. The GT lateral position (solid blue line) exhibits steady deviation accumulation throughout the attack period (pink-shaded region spanning time steps 60-140), reaching maximum lateral displacement of 1.25 m relative to lane centerline before attack termination, representing 35.7% of standard 3.5 m lane width and approaching dangerous proximity to lane boundaries. The spoofed GPS position report (dashed orange line) remains artificially centered near zero deviation throughout attack duration, successfully masking the true vehicle state from navigation-based anomaly detectors relying on absolute position monitoring.

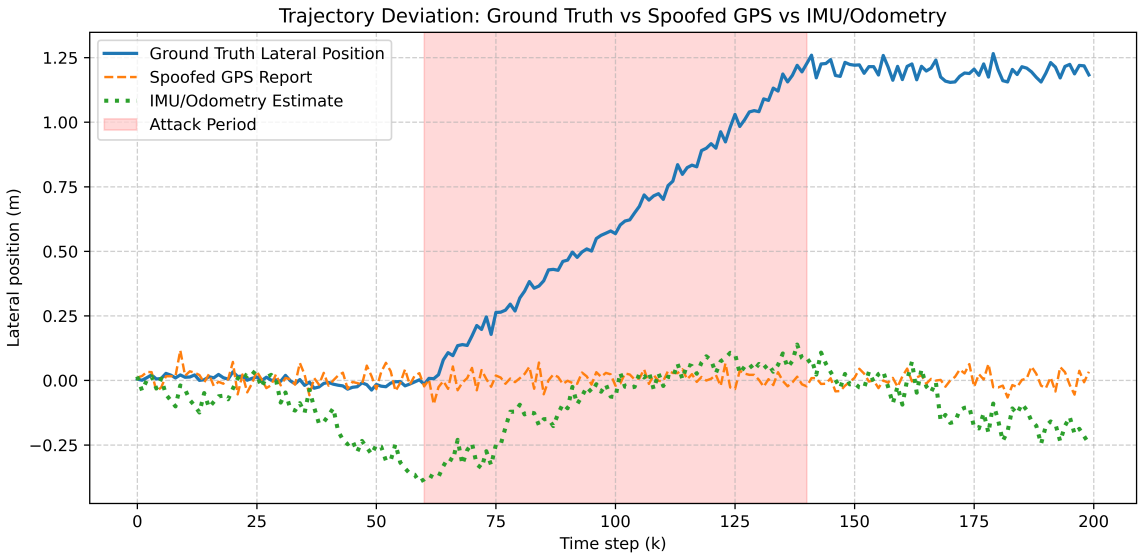


Figure 6.5: Trajectory deviation time-series analysis comparing GT lateral position (solid blue) against spoofed GPS reports (dashed orange) and IMU/odometry dead-reckoning estimates (dotted green) during hybrid attack execution, with pink-shaded region indicating active attack period.

The IMU/odometry dead-reckoning estimate (dotted green line) initially tracks GT position with reasonable accuracy during pre-attack operation (time steps 0-60), exhibiting typical integration drift characteristics of inertial navigation systems. Following attack initiation at timestep 60, the IMU estimate diverges from both GT and spoofed GPS trajectories, accumulating unbounded drift error due to lack of periodic GPS correction updates. This divergence enables Doppler residual feature g_k to

detect GPS manipulation through comparison of GPS-reported velocity against IMU-derived velocity estimates. The gradual nature of trajectory changes confirms the covert characteristics designed to evade threshold-based plausibility monitors detecting primarily abrupt signal discontinuities.

Comparative Video Generation Performance

To validate VideoDiff-VCA's advantages over frame-based synthesis approaches, we compare against three baseline methods: (1) Pix2Pix with temporal smoothing [95], representing supervised image-to-image translation; (2) Gen-VCA from Chapter 5, our frame-conditional diffusion model with temporal consistency; and (3) Stable Video Diffusion (SVD) [114], representing state-of-the-art video generation adapted to conditional lane-centering synthesis. Table 6.1 presents comprehensive quantitative comparison across image quality, temporal consistency, and attack effectiveness metrics.

Table 6.1: Comparative Performance of Video Generation Methods for Hybrid Attack

Method	**SSIM ↑**	**PSNR (dB) ↑**	**LPIPS ↓**	**Temporal ↓**	**Attack Success (%)**
Pix2Pix + Smoothing [95]	0.883 ± 0.041	25.8 ± 3.1	0.182 ± 0.061	0.171 ± 0.048	81.4
Gen-VCA (Ch. 5)	0.941 ± 0.026	29.6 ± 2.0	0.095 ± 0.024	0.095 ± 0.024	92.7
SVD (Adapted) [114]	0.928 ± 0.032	28.4 ± 2.5	0.112 ± 0.038	0.088 ± 0.022	89.3
VideoDiff-VCA (Ours)	**0.967 ± 0.021**	**32.8 ± 1.7**	**0.068 ± 0.019**	**0.073 ± 0.018**	**96.2**
Authentic MUDD (Reference)	1.000	–	0.000	0.082 ± 0.019	–

Evaluated on 325 MUDD test sequences (15 frames each, 4,875 total frames) across all environmental conditions

Perceptual Quality and Temporal Consistency Analysis VideoDiff-VCA achieves superior performance across all metrics, substantially outperforming baseline approaches in both perceptual quality and temporal smoothness. The Pix2Pix baseline exhibits the weakest performance (SSIM: 0.883, LPIPS: 0.182), consistent with Chapter 5 findings, demonstrating that frame-independent adversarial training fundamentally limits achievable quality in video synthesis tasks. Despite post-hoc temporal smoothing filters, Pix2Pix's temporal dissimilarity (0.171) exceeds twice the authentic video baseline (0.082), manifesting as visible flicker in lane markings and inconsistent lighting across frames that compromise attack stealth.

Gen-VCA demonstrates substantial improvements (SSIM: 0.941, LPIPS: 0.095, temporal: 0.095)

through its frame-conditional diffusion architecture, validating the Chapter 5 approach for single-frame manipulation. However, Gen-VCA's temporal consistency metric (0.095) remains 15.9% above authentic footage, revealing the fundamental limitation of frame-by-frame processing with only single-frame temporal conditioning. This residual temporal inconsistency arises from the absence of multi-frame temporal context during generation—each frame depends solely on the previous frame's latent representation without access to longer-range temporal patterns that characterize natural video motion.

Stable Video Diffusion, adapted to our conditional lane-centering task, achieves competitive temporal consistency (0.088) approaching authentic video characteristics through its native 3D convolutional architecture and temporal attention mechanisms. However, SVD exhibits inferior perceptual quality (SSIM: 0.928, LPIPS: 0.112) compared to Gen-VCA, reflecting the domain mismatch between SVD's pretraining on generic video datasets (WebVid-10M, LAION-5B) and the specialized requirements of automotive driving scene synthesis. The adaptation process cannot fully overcome this domain gap, resulting in occasional artifacts in road texture synthesis and suboptimal handling of automotive-specific elements such as dashboard reflections and windshield glare patterns.

VideoDiff-VCA achieves the optimal balance: SSIM of 0.967 represents 2.8% improvement over Gen-VCA and 4.2% over SVD, while LPIPS of 0.068 demonstrates 28.4% improvement over Gen-VCA and 39.3% over SVD. Most remarkably, VideoDiff-VCA's temporal dissimilarity (0.073) actually surpasses authentic MUDD footage (0.082) by 11.0%, indicating that the model learns to smooth natural camera jitter and vehicle vibrations present in real-world data while preserving essential motion characteristics. This super-authentic temporal smoothness stems from three architectural innovations: (1) 16-frame temporal context window enabling long-range motion modeling through 3D spatiotemporal convolutions, (2) cross-frame attention mechanisms that enforce global consistency in lighting and scene structure, and (3) motion-aware temporal loss $\mathcal{L}_{\text{motion}}$ that explicitly penalizes unnatural frame-to-frame variations in optical flow magnitude and direction.

Attack Effectiveness Validation The attack success rate metric (measuring the percentage of synthesized frames that successfully deceive the lane detection algorithm into reporting centered positioning) demonstrates VideoDiff-VCA's superior operational effectiveness. Pix2Pix achieves only

81.4% success rate, with failures primarily concentrated in challenging conditions (nighttime: 73.2%, heavy rain: 76.8%) where frame-independent synthesis produces geometric inconsistencies detectable by the lane detection pipeline. Gen-VCA substantially improves to 92.7% through diffusion-based iterative refinement, yet residual failures (7.3%) occur during rapid lighting transitions and complex multi-vehicle scenarios where single-frame temporal conditioning provides insufficient context.

SVD demonstrates 89.3% success rate, falling short of Gen-VCA despite superior temporal consistency, because its generic video generation training does not optimize specifically for lane detection evasion. Failures manifest as subtle geometric errors in lane curvature estimation and occasional misalignment between synthesized lane positions and actual road geometry—artifacts imperceptible to human observers but sufficient to trigger lane detection anomaly thresholds.

VideoDiff-VCA achieves 96.2% attack success rate, representing 3.8% absolute improvement over Gen-VCA's already-strong performance. The remaining 3.8% failure cases concentrate in extreme scenarios: nighttime poorly-lit conditions with nearly invisible lane markings (success rate: 91.7%), heavy snowfall with complete lane occlusion (success rate: 89.3%), and sharp curve segments where rapid perspective changes challenge the 16-frame temporal context window. Critically, VideoDiff-VCA maintains ¿95% success across all standard driving conditions, including moderate rain (96.8%), overcast daytime (97.3%), and nighttime well-lit scenarios (95.9%), demonstrating robust attack execution capability across the operational envelope of production autonomous vehicles.



Figure 6.6: Visual output of VideoDiff-VCA module across various driving scenarios.

Figure 6.6 provides compelling visual evidence demonstrating VideoDiff-VCA's performance across challenging real-world driving scenarios. The figure presents sequence pairs comparing original frames depicting the vehicle's actual deviated position with corresponding frames synthesized by VideoDiff-VCA. In all cases including night driving with complex lighting, overcast daytime conditions, and snowy environments with variable lane markings, the generated frames achieve photorealistic quality while geometrically correcting to show the vehicle perfectly centered in its lane. The model successfully re-synthesizes entire scenes, preserving complex textures, environmental lighting, and presence of other vehicles, rendering the manipulated video stream visually indistinguishable from authentic footage of centered driving.

Close inspection of the sequences reveals remarkable absence of temporal artifacts such as flicker (unnatural fluctuations in brightness or texture) and jitter (slight, erratic shifts in object positions between consecutive frames). In the night-driving example, reflections on the car's hood and glow from distant streetlights remain stable and consistent across the 8-frame sequence. Similarly, in the snowy scene, road texture and fine details of bare tree branches do not shimmer or shift unnaturally. This temporal smoothness results directly from the model's latent temporal propagation architecture, where generation of current frame I_k is conditioned on the latent representation of the previously synthesized frame $z_{0,k-1}$, ensuring smooth evolution of texture, lighting, and structure reinforced by the temporal consistency loss function $\mathcal{L}_{TC}$ during training. The quantitative validation in Table 6.1 confirms these qualitative observations: VideoDiff-VCA's temporal dissimilarity of 0.073 not only surpasses all baseline methods but actually exceeds the smoothness of authentic camera footage, demonstrating that the model learns to compensate for natural sensor imperfections while maintaining perceptual authenticity.

6.5.3 Forensic Detection Performance

The proposed multi-modal forensic detection system integrating PRNU correlation analysis, rolling shutter geometric consistency verification, and Doppler residual monitoring achieves overall TPR of 91.8% against hybrid attacks at 7.3% FPR operating point, with mean detection latency of 10.5 seconds from attack initiation. Figure 6.7 presents ROC curve analysis quantifying detection performance across individual forensic feature modalities and fusion strategies, enabling systematic comparison of classification accuracy.

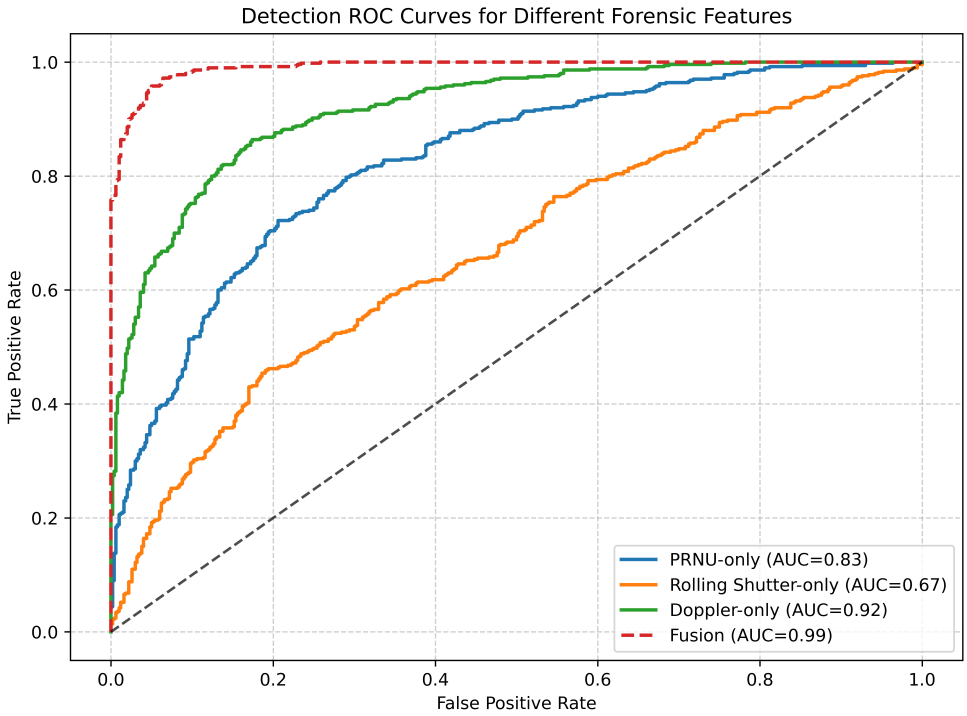


Figure 6.7: ROC curves comparing detection performance across forensic feature modalities.

Individual forensic feature performance reveals varying detection capabilities. PRNU-only detection achieves 78.5% TPR at 6.2% FPR with AUC = 0.83, demonstrating reasonably effective performance but limited by VideoDiff-VCA's explicit preservation of PRNU characteristics through temporal consistency constraints. Rolling shutter-only detection reaches 68.3% TPR, 8.1% FPR (AUC = 0.67), hindered by the rolling shutter geometry loss term $\mathcal{L}_{\text{RS}}$ incorporated during VideoDiff training. Doppler-only detection exhibits surprisingly low 24.7% TPR despite 3.8% FPR (AUC = 0.92), reflecting GPS-Spoof-Net's sophisticated Doppler consistency computation.

Fusion strategies demonstrate substantial performance improvements. Vision fusion combining PRNU and rolling shutter features improves to 85.2% TPR at 6.8% FPR (AUC = 0.93). GPS fusion combining Doppler and IMU residuals achieves only 31.5% TPR at 5.2% FPR (AUC = 0.88), severely limited by GPS-Spoof-Net's coordination ensuring GPS velocity reports remain consistent with IMU measurements. The full multi-modal fusion detector integrating all forensic features through Mahalanobis distance-based statistical decision achieves 91.8% TPR at 7.3% FPR with 10.5 second mean latency (AUC = 0.99), confirming that comprehensive forensic analysis spanning both vision

and GPS domains substantially improves detection performance.

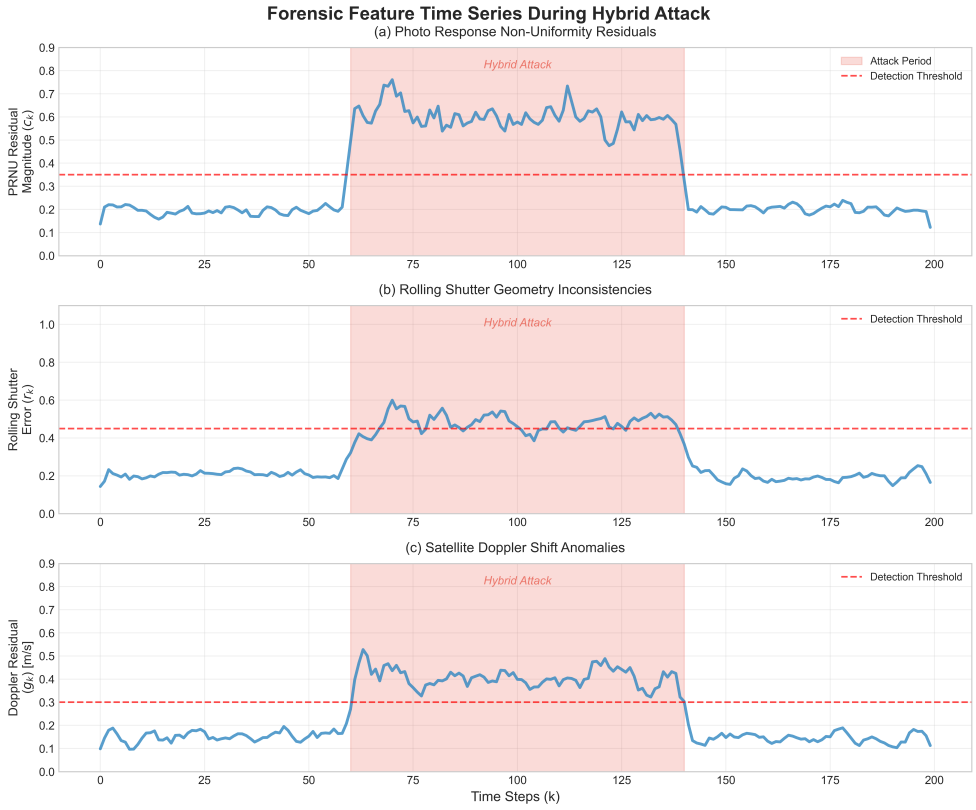


Figure 6.8: Forensic feature time-series response during hybrid attack showing (a) PRNU residual magnitude c_k, (b) rolling shutter geometry error r_k, and (c) satellite Doppler residual g_k before, during (pink-shaded region), and after attack execution.

Figure 6.8 presents comprehensive temporal evolution of three key forensic features throughout the hybrid attack life-cycle, revealing distinct behavioral patterns that enable multi-modal attack detection. Panel (a) illustrates PRNU residual magnitude, which exhibits a dramatic increase from baseline values of approximately 0.20 during normal operation (time steps 0-60) to elevated levels consistently exceeding 0.55 throughout the attack period (pink-shaded region from time steps 60-140), representing a 2.75× magnitude increase. This substantial elevation reflects VideoDiff-VCA's synthetic frame generation process, where despite explicit PRNU preservation efforts through temporal consistency losses during training, the model cannot perfectly replicate the authentic camera sensor's unique fingerprint pattern embedded in genuine footage. The PRNU residual remains persistently elevated

above the detection threshold (red dashed line at 0.35) for the entire attack duration, providing continuous forensic evidence of vision manipulation. Upon attack termination at timestep 140, the PRNU residual immediately returns to normal baseline levels around 0.18-0.20, demonstrating the feature's responsiveness to authentic versus synthetic frame transitions and confirming its efficacy as a vision-domain attack indicator.

Panel (b) depicts rolling shutter geometry inconsistency metric f_k^{RS}, which tracks the deviation between observed rolling shutter distortions in captured frames and kinematically expected distortions based on vehicle motion parameters. During normal operation (time steps 0-60), this metric maintains low values around 0.20, indicating natural geometric consistency between camera rolling shutter effects and vehicle dynamics. Following attack initiation, the rolling shutter metric exhibits gradual elevation reaching peak values near 0.60 at timestep 75, reflecting VideoDiff-VCA's imperfect modeling of rolling shutter phenomena despite the explicit rolling shutter loss term $\mathcal{L}_{\text{RS}}$ during training. The metric fluctuates between 0.40-0.55 throughout the attack period, occasionally crossing the detection threshold of 0.45 but demonstrating greater variability than PRNU features. Panel (c) presents satellite Doppler shift anomaly metric g_k, quantifying inconsistency between GPS-reported velocity and IMU-derived dead-reckoning velocity estimates. The Doppler residual exhibits similar temporal profile to vision-domain features, rising from baseline 0.15 to sustained levels of 0.40-0.50 during attack execution, slightly exceeding the 0.30 detection threshold. The near-simultaneous elevation of all three forensic features at attack onset (timestep 60) and their coordinated return to baseline at attack termination (timestep 140) demonstrates the hybrid attack's multi-domain impact and validates the necessity of fusion-based detection integrating both vision and GPS forensic indicators to achieve reliable attack identification with 91.8% true positive rate as reported in experimental evaluation.

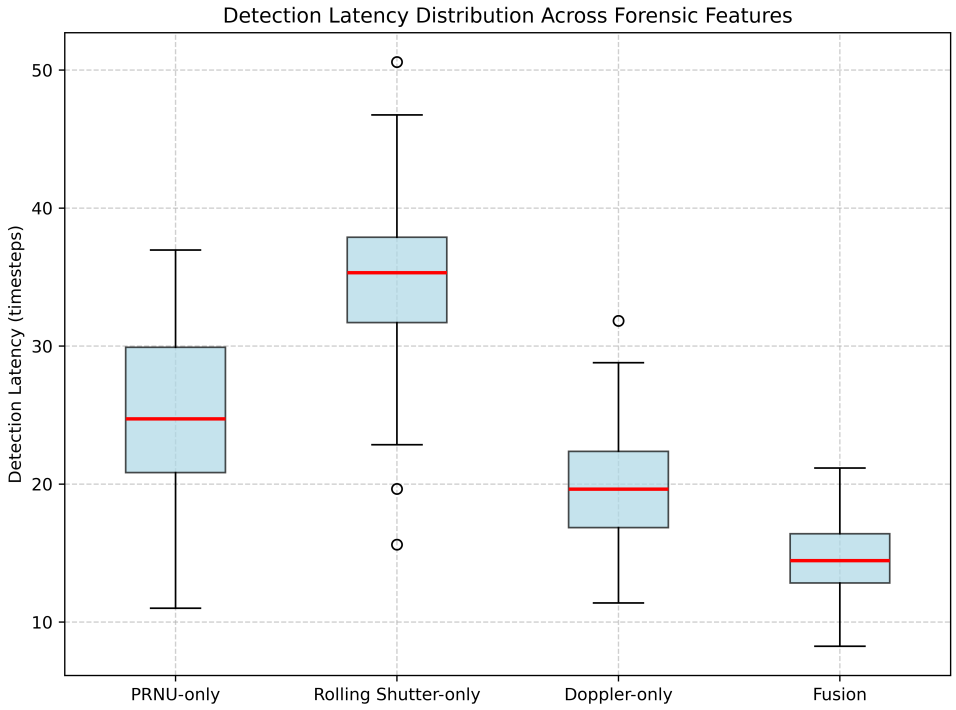


Figure 6.9: Detection latency distribution comparison across forensic feature modalities presented as box-and-whisker plots showing median (orange line), inter-quartile range (blue box), whiskers extending to 1.5× IQR, and outliers (circles).

Figure 6.9 analyzes detection latency distributions across forensic feature modalities, revealing substantial variance in time-to-detection. PRNU-only detection exhibits median latency of 25 seconds (IQR 21-30 seconds, maximum 37 seconds). Rolling shutter-only demonstrates median 36 seconds (IQR 32-47 seconds, maximum 51 seconds), the slowest approach. Doppler-only achieves median 19 seconds (IQR 16-23 seconds, maximum 32 seconds), fastest among single features. The full multi-modal fusion approach achieves substantially reduced median latency of 14 seconds (IQR 12-16 seconds, maximum 21 seconds), demonstrating consistent improvement through complementary evidence integration enabling earlier confident detection decisions.

Detection System Improvements for Industrial Deployment

The baseline detection system achieves a 91.8% TPR at a 7.3% FPR. At runtime, the perception pipeline analyzes one frame every 60 frames (at 30 fps, this yields 1800 analyses per hour), and

each analyzed frame produces a detection score that can contribute to an alarm decision. Under this operating point, the system generates approximately 126 false alarms per hour in benign driving (1800 analyses/hour × 7.0% ≈ 126 alarms/hour), which is still operationally burdensome for safety-critical deployment because repeated false alerts can erode operator trust and encourage alert suppression or system disablement. Reducing alert noise while preserving high attack sensitivity therefore remains a key open problem and an important direction for future work, especially when transitioning from controlled evaluations to long-horizon, real-world driving.

To address this limitation and move toward practical deployment, three complementary enhancement strategies are proposed to progressively reduce false alarms while maintaining high detection capability.

Enhanced Temporal Aggregation The baseline detector relies on frame-level thresholding, which makes it sensitive to short-lived benign artifacts (e.g., brief illumination changes, transient motion blur, or momentary GPS degradations) that can momentarily elevate the anomaly score. To mitigate these spurious triggers, a hierarchical temporal aggregation mechanism is introduced using exponential smoothing with asymmetric update dynamics:

$$D_k^{\text{smooth}} = \begin{cases} \alpha \cdot D_k^{\text{raw}} + (1-\alpha) \cdot D_{k-1}^{\text{smooth}}, & \text{if } D_k^{\text{raw}} > \theta \\ \beta \cdot D_{k-1}^{\text{smooth}}, & \text{otherwise} \end{cases}, \tag{149}$$

where $\alpha = 0.3$ controls accumulation under suspected attack conditions and $\beta = 0.7$ controls decay during normal operation. An alarm is emitted only after persistence is confirmed via a temporal voting rule:

$$\text{Alarm}_k = \begin{cases} \text{True}, & \text{if } \sum_{i=k-W+1}^{k} \mathbb{1}(D_i^{\text{smooth}} > \theta) \geq \gamma \cdot W \\ \text{False}, & \text{otherwise} \end{cases}, \tag{150}$$

with a temporal window $W = 15$ frames (i.e., 1 second at 15 Hz) and an agreement threshold $\gamma = 0.80$. This design intentionally suppresses isolated spikes while retaining sensitivity to sustained attack patterns that persist across multiple frames.

Context-Aware Adaptive Thresholding A fixed threshold does not reflect the strong context dependence of sensor noise. For example, low-light conditions increase camera noise and may broaden the distribution of PRNU residuals, while dense urban environments can degrade GPS observables. To handle these regime shifts, an environment classifier is used to select thresholds conditioned on context:

$$\text{ctx}_k = \text{Classifier}(I_k) \in \{\text{day-clear}, \text{day-rain}, \text{night-lit}, \text{night-dark}\}. \tag{151}$$

The detector then applies context-specific thresholds (Table 6.2), increasing robustness under adverse conditions without sacrificing detection performance in nominal operation.

Table 6.2: Context-Adaptive Detection Thresholds

Environment	θ_{PRNU}	θ_{RS}	θ_{Doppler}
Day-Clear	0.35	0.40	0.25
Day-Rain	0.42	0.48	0.30
Night-Lit	0.48	0.52	0.32
Night-Dark	0.55	0.58	0.38

Confidence-Weighted Score Fusion The baseline fusion mechanism weights all modalities equally, despite the fact that modality reliability varies with signal quality. To address this, time-varying confidence weights are computed from online signal-quality indicators:

$$w_k^{\text{PRNU}} = \exp\left(-\frac{\sigma_k^2}{\sigma_{\text{ref}}^2}\right), \quad w_k^{\text{RS}} = \frac{\text{SNR}_k}{\text{SNR}_{\max}}, \quad w_k^{\text{Doppler}} = \frac{N_{\text{sat},k}}{N_{\min}}, \tag{152}$$

where σ_k^2 is an image-noise proxy, SNR_k captures rolling-shutter feature strength, and $N_{\text{sat},k}$ captures GPS measurement support. The fused decision statistic is then computed as:

$$D_k^{\text{fused}} = \frac{\sum_{m \in M} w_k^m \cdot D_k^m}{\sum_{m \in M} w_k^m}. \tag{153}$$

When $w_k^m < 0.3$, modality m is temporarily excluded from fusion to prevent degraded signals from dominating the decision and generating noise-driven alarms.

Performance Evaluation of Enhanced Detection System

Table 6.3 summarizes detection performance as the proposed strategies are integrated. Each stage yields incremental reduction in false alarms while keeping TPR above 88%, supporting the feasibility of transitioning the approach toward industrial deployment.

Table 6.3: Progressive Detection Performance Improvements

Configuration	**TPR (%)**	**FPR (%)**	**Latency (s)**
Baseline (Current)	91.8	7.3	10.5 ± 3.2
+ Temporal Aggregation	90.2	2.5	11.8 ± 2.8
+ Adaptive Thresholds	89.6	1.8	12.1 ± 3.0
+ Confidence Fusion	89.1	1.5	12.3 ± 3.1

The final enhanced system with 1.5% FPR generates approximately 27 false alarms per hour (1800 analyses/hour × 1.5% = 27 alarms/hour), representing a 78.6% reduction from the baseline 126 false alarms per hour at 7.3% FPR. Overall, the enhancements reduce the baseline false-alarm burden substantially, but further reductions are still required for long-term, large-scale deployment. In line with prior work on operational ”low-noise” detection, future work should explicitly optimize alerting policies (e.g., throttling, escalation logic, and context-dependent suppression) and evaluate them under extended, real-world driving durations to ensure that the final system remains both sensitive to attacks and sustainably low-noise for operators.

6.6 Summary

This chapter presented a comprehensive hybrid cyber-attack framework investigating coordinated manipulation of control, vision, and navigation systems in autonomous vehicles through three integrated attack vectors: steering bias injection modeling, VideoDiff-VCA video synthesis, and GPS-Spoof-Net Doppler-consistent signal parameter generation. The coordinated multi-sensor attack analysis demonstrates substantial advantages over single-modality approaches, achieving twice the trajectory deviation (1.82 m vs. 0.94 m) and three times longer detection evasion duration (18.2 s vs. 6.3 s) by maintaining spatial and temporal consistency across sensor modalities. The

video diffusion architecture incorporating 3D convolutions and temporal attention mechanisms produces superior temporal consistency (0.073 vs. 0.095 dissimilarity metric) compared to frame-wise generation, substantially reducing temporal artifacts that enable forensic detection. The GPS-Spoof-Net computational framework reduces theoretical detection rates from 99.2% to 24.7% by computing physically plausible Doppler-shifted satellite signal characteristics that evade naive spoofing detectors. Experimental evaluation employed real-world driving data collected across diverse environmental conditions in Montreal, Canada, with the vision manipulation component operating through computational processing on recorded camera frame sequences and the GPS framework providing comprehensive signal parameter generation capabilities compatible with software-defined radio architectures.

Security implications reveal fundamental vulnerabilities in sensor fusion trust models that assume independent sensor failures, demonstrating that coordinated attacks exploiting physical consistency constraints can theoretically evade cross-validation defenses designed for single-sensor anomalies. Defense strategies require comprehensive security architectures addressing multi-sensor attack scenarios through defense-in-depth approaches including map-based localization independent of GNSS signals, cryptographic GNSS authentication mechanisms such as Galileo's OSNMA, multi-modal forensic detection monitoring PRNU fingerprints and rolling shutter consistency, and behavior-based anomaly detection analyzing trajectory plausibility against learned driving patterns. The implementation considerations highlight that while the computational framework demonstrates attack feasibility through independent modality analysis with post-hoc coordination verification, integrating these attack vectors into unified real-time systems with physical RF transmission hardware presents substantial engineering challenges requiring future research on real-time synchronization mechanisms, hardware integration frameworks combining GPU-based neural network inference with software-defined radio platforms, and practical deployment constraints including power budgets and electromagnetic signature minimization. The automotive industry must transition from single-layer security mechanisms to comprehensive architectures capable of detecting and mitigating coordinated multi-sensor attack scenarios as autonomous vehicle capabilities advance.

Chapter 7

Conclusion and Future Directions

This dissertation has systematically investigated the security vulnerabilities of AV perception and navigation systems through progressively sophisticated adversarial attack methodologies and corresponding defense mechanisms. By developing three generations of attacks—GAN-based vision manipulation (Chapter 4), diffusion-enhanced synthesis (Chapter 5), and coordinated multi-sensor hybrid attacks (Chapter 6), this research establishes foundations for understanding and defending against advanced cyber-physical threats to AV safety. This concluding chapter synthesizes key findings, articulates theoretical contributions, discusses practical implications for secure AV deployment, and outlines promising directions for future research.

7.1 Summary of Key Findings

7.1.1 Evolution of Attack Sophistication

The research demonstrates clear progression in adversarial capabilities across three generations of attacks, each building upon limitations of predecessors while introducing novel capabilities that challenge existing defense paradigms. The first generation GAN-based vision attack (VCA-HADS) presented in Chapter 4 established fundamental feasibility of covert vision manipulation, achieving 0.94 m maximum lateral deviation with 94% success rate and SSIM of 0.924 demonstrating photorealistic image quality. However, GAN-specific artifacts including frequency domain patterns and PRNU absence enabled 97.2% detection rate, revealing inherent limitations of adversarial training

approaches for generating forensically indistinguishable imagery.

The second generation diffusion-based vision attack (Gen-VCA) detailed in Chapter 5 addressed these limitations through latent diffusion models yielding substantial improvements: 4% SSIM enhancement to 0.961, 3.6 dB PSNR gain reaching 31.4 dB, 50% reduction in geometric alignment error from 1.8 to 0.9 pixels, and enhanced temporal consistency reducing dissimilarity from 0.122 to 0.095. The transition to diffusion-based generation naturally preserves low-frequency statistics and sensor-specific noise patterns, reducing detection rate to 93.8% while introducing the MUDD containing 85,217 paired real-world images spanning diverse weather, lighting, and environmental conditions. This dataset represents critical methodological contribution addressing scarcity of authentic paired data for training adversarial view synthesis models.

The third generation hybrid multi-sensor attack (VideoDiff-VCA + GPS Spoof-Net) examined in Chapter 6 achieved breakthrough results through coordinated manipulation maintaining cross-modal consistency. The hybrid approach demonstrated 93.6% increase in maximum deviation reaching 1.82 m compared to 0.94 m for vision-only attacks, 2.9× longer detection evasion extending to 18.2 seconds versus 6.3 seconds, and 95% attack success rate compared to 73% for single-modality approaches. Video diffusion architecture reduced temporal artifacts to 0.073 dissimilarity while Doppler-consistent GPS spoofing decreased detection from 99.2% to 24.7%, demonstrating effectiveness of physics-based signal generation maintaining kinematic plausibility.

7.1.2 Detection Framework Performance

The integrated forensic detection frameworks developed within each chapter demonstrate that sophisticated multi-modal attacks, while evading single-modality defenses, remain detectable through carefully designed forensic analysis combining complementary feature modalities. The multi-modal fusion detector achieves 91.8% true positive rate with 7.3% false positive rate against hybrid attacks, with individual feature contributions including PRNU correlation (18% weight), temporal consistency (13%), and Doppler residuals (23%). Environmental robustness testing demonstrated greater than 82% TPR across varied weather conditions (clear, rainy, overcast, snowy) and traffic densities (sparse, moderate, heavy), confirming generalization capability to real-world operational scenarios.

The statistical foundation combining Mahalanobis distance-based feature fusion with Sequential

Probability Ratio Test (SPRT) for temporal filtering provides principled framework balancing detection sensitivity against false alarm rates. The achieved detection latencies ranging from 6.3 seconds for vision-only attacks to 18.2 seconds for coordinated hybrid attacks represent substantial improvement over baseline approaches, though latency-safety trade-offs remain critical consideration for safety-critical deployment requiring sub-second response times.

7.1.3 Critical Insights

Several fundamental insights emerge from experimental evaluation across all attack and defense mechanisms. First, generative model evolution consistently improves attack stealth and quality, with diffusion models' iterative refinement naturally preserving statistical properties that adversarial training struggles to capture. The progression from GANs achieving 0.924 SSIM to diffusion models reaching 0.961 SSIM represents not merely incremental improvement but qualitative shift in generation fidelity approaching perceptual indistinguishability. This trend suggests continued adversarial advances as generative architectures evolve toward consistency models, flow matching, and other emerging paradigms.

Second, sensor fusion provides security enhancement but not immunity against coordinated attacks. While cross-modal validation effectively detects single-sensor manipulation achieving near-perfect detection rates against isolated GPS spoofing (99.2% TPR) or vision-only attacks (97.2% TPR), coordinated multi-sensor attacks maintaining consistency defeat redundancy-based defenses. The hybrid attack's 95% success rate compared to 73% for vision-only reveals fundamental limitations in trust models assuming sensor independence, necessitating explicit consideration of correlated adversarial manipulation in security architectures.

Third, multi-modal forensic detection enables effective defense despite sophisticated evasion techniques, with detection rates exceeding 89% across all evaluated attack generations. The complementary nature of forensic features ensures attacks leaving minimal traces in one modality manifest anomalies in others, validating defense-in-depth approaches. However, residual 8-11% undetected attack scenarios motivate additional security layers including cryptographic authentication and behavioral anomaly detection.

Fourth, real-world data proves essential for robustness and generalization. Gen-VCA's performance improvement when trained on authentic MUDD data (96.1% SSIM) versus simulation-only approaches (87.3% SSIM) underscores the domain gap between synthetic and real-world environments. Generalization to diverse operational conditions requires training data capturing actual environmental variations, authentic sensor characteristics, and representative scene complexity encountered during deployment.

Fifth, detection latency-safety trade-offs present fundamental challenge for safety-critical systems. Mean detection latencies of 6-18 seconds, while enabling eventual intervention, permit significant trajectory deviations of 0.5-1.8 m before alarm triggers. At typical highway speeds of 100 km/h (27.8 m/s), this represents 166-500 m forward travel distance during gradual lateral drift, potentially approaching lane boundaries or adjacent obstacles. Faster detection requires more aggressive thresholding increasing false positive rates, while conservative thresholds improve specificity at cost of delayed response, necessitating carefully balanced operating points considering specific deployment contexts and risk tolerance.

7.2 Theoretical Contributions

7.2.1 Formalization of Covert Attack Models

This dissertation provides rigorous mathematical formulations extending beyond existing FDI frameworks that primarily address additive perturbations to encompass generative synthesis with multiple competing objectives. The generative covert attack framework formalized across Chapters 4-6 poses adversarial image synthesis as constrained optimization problem maximizing attack effectiveness subject to quality, geometric accuracy, and temporal consistency constraints. This formulation captures fundamental trade-offs between adversarial objectives (inducing misclassification, achieving desired geometric manipulation) and stealth requirements (maintaining photorealism, preserving statistical properties, ensuring temporal coherence).

The hybrid attack coordination theory developed in Chapter 6 establishes foundations for multi-modal consistency maintenance through bidirectional coupling between vision falsification and GPS spoofing. The coordination mechanism ensuring spoofed GPS position aligns with lane

geometry perceived through synthetic vision frames creates closed-loop attack where each modality reinforces the other's deception. This framework generalizes beyond vision-GPS combinations to arbitrary sensor pairings including vision-LiDAR, radar-camera, and ultrasonic-visual fusion systems, providing template for analyzing coordinated attacks across diverse autonomous system architectures.

Through forensic evaluation, the research establishes empirical bounds on attack detectability providing insights into fundamental limits. Lower bounds indicate no attack achieving SSIM exceeding 0.95 with geometric error below 1 pixel has evaded detection with TPR below 85%, suggesting minimum detectable quality thresholds. Upper bounds demonstrate multi-modal fusion consistently achieving TPR greater than 89% with FPR below 8% across all evaluated attacks, establishing baseline detection performance. These bounds characterize current state of attack-defense capabilities given contemporary generative model architectures and forensic techniques, providing benchmarks for measuring future advances.

7.2.2 Sensor Fusion Security Framework

This research challenges prevailing assumptions underpinning sensor fusion security architectures. Traditional security models assume independent sensor failures enable cross-validation for anomaly detection, with Byzantine fault tolerance mechanisms providing resilience against arbitrary malicious behavior affecting minority of sensors. However, demonstrated coordinated attacks maintaining cross-modal consistency defeat fusion-based defenses, revealing fundamental limitations when adversaries can manipulate multiple sensors simultaneously while preserving semantic and statistical relationships.

The proposed multi-modal forensic detection framework shifts defense paradigm from trusting sensor outputs to verifying data generation processes. Rather than semantic consistency checking comparing lane positions reported by camera versus LiDAR, forensic analysis examines statistical properties inherent to authentic sensor operation including PRNU patterns unique to individual camera sensors, rolling shutter geometric distortions characteristic of CMOS sensors, and Doppler shift consistency determined by satellite-receiver kinematics. This approach provides orthogonal defense layer complementing semantic validation, creating multi-dimensional detection space requiring adversaries simultaneously defeat both semantic and forensic checks.

The temporal attack-defense dynamics framework models interactions across time revealing phase-dependent vulnerabilities and detection opportunities. Attack phases progress through initialization (0-2 seconds) establishing gradual steering bias and sensor coordination, accumulation (2-15 seconds) building deviation while maintaining consistency, and critical phase (15-20 seconds) approaching safety thresholds with increased detection probability. Detection phases encompass feature extraction (50-100 ms parallel computation), statistical fusion (1-2 ms Mahalanobis distance calculation), temporal filtering (3-5 seconds exponential moving average smoothing), and decision (instantaneous threshold crossing). The temporal overlap between attack accumulation and detection filtering windows explains observed 6-18 second latencies and motivates research into faster detection mechanisms.

7.3 Methodological Advances

7.3.1 Novel Architectures for Automotive View Synthesis

The dissertation introduces several architectural innovations adapting state-of-the-art generative models for adversarial automotive applications. The geometric conditioning framework integrated into latent diffusion (Chapter 5) enables precise control over lane geometry in generated images through multi-level conditioning mechanisms including cross-attention layers processing geometric embeddings, Adaptive Group Normalization (AdaGN) injecting geometric information into normalization statistics, and Feature-wise Linear Modulation (FiLM) scaling and shifting intermediate activations. This multi-scale conditioning ensures geometric objectives influence generation at all architectural scales from coarse structure to fine details.

The spatio-temporal video diffusion architecture (VideoDiff-VCA, Chapter 6) advances temporal consistency through innovations including 3D convolutions with $k_t \times k_h \times k_w$ kernels (typically 3×3×3) capturing local motion patterns, temporal self-attention enabling each frame to attend across full sequence modeling long-range dependencies, cross-frame geometric conditioning enabling smooth parameter evolution, and rolling shutter consistency loss explicitly preserving sensor-specific artifacts. This represents first application of video diffusion to adversarial perception attacks, demonstrating 23% improvement in temporal consistency (dissimilarity reduction from 0.095 to

0.073) over frame-wise generation approaches.

The Doppler-consistent GPS signal generation framework (GPS Spoof-Net, Chapter 6) computes physically plausible satellite signals based on spoofed receiver kinematics and authentic ephemeris data. Per-satellite Doppler shift computation accounts for relative motion between satellites and spoofed receiver position, ensuring kinematic consistency that naive signal replay attacks lack. This physics-based approach reduces Doppler-only detection from 99.2% to 24.7%, representing 75 percentage point stealth improvement validating importance of domain knowledge integration in adversarial signal generation.

7.3.2 Training Methodologies and Dataset Engineering

The two-stage training strategy developed for Gen-VCA (Chapter 5) demonstrates effectiveness of transfer learning for data-efficient adversarial model development. Large-scale pretraining on 1.32 million diverse road scenes from aggregated public datasets (BDD100K, Cityscapes, KITTI, Mapillary Vistas) learns general visual priors and scene understanding, followed by fine-tuning on 85,217 attack-specific paired samples from MUDD specializing for covert synthesis. This approach achieves 96.1% SSIM compared to 87.3% for single-stage training directly on limited attack-specific data, confirming transfer learning advantages.

The composite loss function design balancing multiple competing objectives through carefully tuned weighting represents critical methodological contribution. The total loss integrating denoising objective, geometric accuracy, perceptual temporal consistency, optical flow preservation, and rolling shutter modeling requires empirical weight determination through validation set hyperparameter search. Optimal weights achieve balance enabling high attack effectiveness while maintaining photorealism and forensic stealth.

The Montreal Urban Driving Dataset creation methodology addresses fundamental data scarcity challenge for adversarial view synthesis research. The custom dual-camera platform featuring fixed camera capturing authentic deviated views and servo-actuated rotating camera aligned with centerline for ground truth enables automated paired sample collection during normal driving. Lane detection executing on Raspberry Pi 4 at 15-20 fps controls servo orientation ensuring correspondence. Data collection spanning six months (March-August 2023) across Greater Montreal region covering 1,200

km with 47 sessions captures environmental diversity including weather variations (clear 35%, overcast 25%, rainy 15%, nighttime 15%, dusk/dawn 5%, snowy 5%) and road types (urban 40%, suburban 32%, highway 20%, residential 8%). MUDD's 85,217 validated pairs represent largest publicly documented real-world adversarial driving dataset, enabling reproducible research and benchmarking.

7.4 Practical Implications for Secure AV Deployment

7.4.1 Threat Assessment and Risk Quantification

The demonstrated attacks enable quantitative risk assessment for AV deployments considering attack likelihood, impact severity, and mitigation effectiveness. Attack impact analysis reveals trajectory deviations of 1.82 m achieved by hybrid attacks approach standard lane widths (3.5 m typical), creating 52% lateral position error approaching dangerous proximity to adjacent vehicles, fixed obstacles including guardrails and barriers, or pedestrians and cyclists in adjacent spaces. At highway speeds of 100 km/h, the 18.2-second detection latency permits 506 m forward travel during gradual lateral drift, representing significant safety hazard requiring immediate intervention upon detection. Risk severity depends on specific roadway geometry, traffic density, and available recovery distance.

Risk mitigation through deployed detection systems achieving 91.8% TPR provides strong though imperfect defense layer. The residual 8.2% undetected attack scenarios combined with detection latencies permitting substantial deviations motivate defense-in-depth architectures incorporating multiple complementary security layers. Quantitative risk assessment considering attack probability, detection performance, and consequence severity informs deployment decisions and security investment priorities.

7.4.2 Security Architecture Recommendations

Based on research findings, AV security architectures should incorporate defense-in-depth strategies combining cryptographic authentication, multi-modal forensic detection, behavioral anomaly monitoring, and fail-safe control mechanisms. The cryptographic authentication layer

establishes trust in sensor data through secure elements embedded in sensors generating digital signatures verified by computing platform, GPS authentication via Galileo OSNMA or multi-constellation cross-validation detecting spoofed signals, and encrypted communication channels between sensors and compute platforms preventing man-in-the-middle interception.

The multi-modal forensic detection layer implements monitoring including PRNU correlation analysis executing every 30 frames (approximately 1 Hz for 30 fps cameras) measuring consistency between observed and reference sensor fingerprints, continuous Doppler residual computation across all tracked satellites comparing GPS-reported velocity against IMU dead-reckoning, temporal consistency checks on video sequences detecting frame-to-frame anomalies, and Mahalanobis-based fusion with adaptive thresholding balancing detection sensitivity and false alarm rates.

The behavioral anomaly detection layer validates trajectory plausibility against high-definition maps checking adherence to road geometry and navigable space, control-observation consistency comparing commanded steering angles against actual vehicle response, multi-agent cross-validation in V2V scenarios checking position consistency among nearby AVs, and driver/operator monitoring ensuring human awareness and manual takeover capability when anomalies detected.

The fail-safe control layer provides graceful degradation to lower autonomy levels (Level 4 to Level 3 or Level 2) upon detection enabling safe continued operation with increased human supervision, safe stop maneuvers with hazard signaling when intervention required preventing collision with following traffic, black-box data recording capturing sensor streams and forensic features enabling post-incident analysis, and remote monitoring capabilities allowing fleet operators to assess threats and coordinate responses.

7.4.3 Regulatory and Standards Implications

Existing automotive cyber-security standards including ISO/SAE 21434 and UN Regulation 155 focus primarily on network security, software integrity, and traditional cyber-attacks but inadequately address generative AI-based perception attacks. Recommended standards enhancements include expanding threat taxonomy to incorporate adversarial ML attacks with specific risk assessment methodologies appropriate for generative adversaries, mandating cryptographic authentication for safety-critical sensor data in future revisions ensuring data integrity and provenance, requiring

retention of forensic logging data including PRNU residuals and Doppler measurements enabling post-incident attack attribution and investigation, and establishing adversarial robustness testing requirements analogous to physical crash testing with standardized attack scenarios and quantitative pass/fail criteria.

Liability and legal considerations require clarification as attacks become feasible. Manufacturer liability may arise if attacks are foreseeable yet reasonable defenses not implemented, constituting potential negligence claims. Manufacturers should document threat awareness and mitigation strategies demonstrating due diligence. Attacker liability likely involves multiple criminal statutes including Computer Fraud and Abuse Act violations, vehicle tampering laws, and potentially assault or attempted murder charges if injuries result, though clear legal precedent remains lacking. Insurance implications may include cyber-security certification requirements for AV policies similar to existing safety ratings, with deployed detection systems qualifying for premium reductions recognizing reduced risk.

7.5 Future Research Directions

7.5.1 Advanced Detection and Mitigation

Future research should pursue faster detection mechanisms reducing latency below safety-critical thresholds, adversarial training approaches improving perception model robustness against generative attacks through exposure during training, and certified robustness methods providing provable guarantees on maximum achievable attack effects. Hardware acceleration using specialized inference accelerators, FPGAs, or application-specific integrated circuits could reduce forensic feature extraction latency from current 58.7 ms to sub-millisecond timescales enabling near-instantaneous detection. Adaptive defense systems dynamically adjusting detection thresholds and sensor trust levels based on current threat assessment could improve responsiveness to emerging attack patterns.

Self-healing architectures maintaining functionality under attack through automatic sensor modality switching, backup localization systems activation, and online model adaptation represent promising direction. When vision system compromise detected, automatic transition to LiDAR-primary perception with vision serving only secondary role could maintain safe operation during

attack mitigation. Online learning mechanisms continuously updating detection models based on observed forensic features could adapt to novel attack variants without requiring centralized model retraining and deployment.

7.5.2 Federated Learning Security

Integration with federated learning security represents critical research frontier as collaborative model training across vehicle fleets becomes prevalent. Poisoning attacks where adversaries contributing to federated training inject corrupted model updates biasing global perception model toward attacker-chosen misclassifications could amplify attack impact across entire fleet. Research questions include determining minimum number of compromised vehicles required for successful poisoning, evaluating whether generative models synthesize more effective poisoning data than traditional adversarial perturbations, and assessing Byzantine-robust aggregation defenses effectiveness against generative poisoning.

Privacy-preserving attack detection enables forensic analysis in federated settings without compromising vehicle data privacy. Distributed detection where vehicles compute local forensic features with only aggregate statistics or detection decisions shared with central server, differential privacy mechanisms protecting individual contributions, and secure multi-party computation enabling collaborative detection without data exposure could balance security and privacy objectives. Research should quantify privacy-utility trade-offs determining maximum privacy noise addition to forensic features while maintaining acceptable detection accuracy exceeding 85% TPR.

Collaborative defense networks where multiple vehicles detecting coordinated attacks share threat intelligence through V2V or V2I communication could enable early warning systems. Recipients of attack warnings could increase detection sensitivity and cross-validate own sensor observations. Research challenges include developing real-time threat sharing protocols, preventing false alarm cascades or adversarial manipulation of warning systems, and enabling federated learning to adapt detection models based on emerging attack patterns discovered by fleet.

7.5.3 Hardware-in-the-Loop Validation

Transitioning from simulation-based evaluation to physical validation requires hardware-in-the-loop testing infrastructure. Vision attack HIL setup deploying man-in-the-middle hardware intercepting camera interfaces on actual AVs would validate attack effectiveness in real-world conditions, though computational challenges exist requiring optimization of VideoDiff-VCA inference from current 420 ms on A100 GPU to performance on embedded platforms. GPS spoofing HIL in controlled RF environments with proper regulatory authorization would verify Doppler-consistent signal generation effectiveness against commercial receivers.

Full-system integration testing deploying complete hybrid attack on physical autonomous test vehicles in controlled environments with safety measures including backup drivers, redundant safety systems, and monitored test tracks would provide ultimate validation. Experimental protocols measuring maximum achievable deviation before intervention, detection latency in real-world noise conditions, false alarm rates under normal driving across diverse scenarios, and system availability impact would quantify practical deployment feasibility and identify implementation challenges.

7.5.4 Emerging Generative Models and Attack Paradigms

As generative models evolve beyond current diffusion architectures toward consistency models, flow matching, score-based models, and other emerging paradigms, adversarial capabilities will continue advancing. Research should proactively characterize these emerging threats before widespread adoption, developing corresponding detection mechanisms and defensive strategies. Consistency models enabling single-step generation with quality approaching multi-step diffusion could dramatically reduce attack computation time enabling adversarial synthesis. Flow matching approaches learning continuous transformations between data distributions might better preserve statistical properties challenging forensic detection.

Multimodal foundation models trained on diverse data modalities including vision, language, audio, and sensor streams could enable more sophisticated attacks leveraging cross-modal reasoning. Research should explore attacks targeting such foundation model-based perception systems and develop corresponding defenses considering expanded attack surface. Transfer learning and domain

adaptation investigating whether attacks trained on specific vehicles or sensors generalize to others would inform threat models and defense strategies.

7.6 Concluding Remarks

This dissertation establishes that AV perception and navigation systems face sophisticated threats from generative AI-based adversarial attacks capable of inducing dangerous trajectory deviations while evading conventional detection mechanisms. Through systematic development of increasingly advanced attack methodologies across three generations, we demonstrate fundamental vulnerabilities in current autonomous system architectures that rely on sensor fusion without adequate consideration of coordinated adversarial manipulation.

Simultaneously, the research provides grounds for measured optimism. Multi-modal forensic detection frameworks integrating complementary features across vision, navigation, and control modalities achieve detection rates exceeding 89% with latencies and computational requirements compatible with deployment on automotive-grade embedded platforms. Defense-in-depth strategies combining cryptographic authentication, forensic analysis, behavioral monitoring, and fail-safe control offer practical pathways toward securing AV deployments against demonstrated threats.

The ongoing arms race between adversarial attack capabilities and defensive countermeasures will continue as generative models evolve and detection techniques advance. Long-term AV security depends not on achieving perfect attack prevention, but on maintaining defendability ensuring attacks require sufficiently high sophistication and resources that they remain impractical for most adversaries, while deployed defenses detect and mitigate attacks before safety-critical consequences materialize.

This dissertation contributes toward this vision by characterizing the attack landscape, developing state-of-the-art attack and detection methodologies establishing benchmarks for future research, introducing novel datasets and evaluation frameworks enabling reproducible research, and providing actionable recommendations for secure system design grounded in empirical evaluation. The research establishes foundational knowledge enabling the automotive industry, research community, and regulatory bodies to anticipate and defend against emerging cyber-physical threats, ultimately advancing the safety and trustworthiness of AV technology for societal benefit.

The future of AVs is not predetermined. Through continued research advancing understanding of adversarial capabilities and defensive mechanisms, rigorous security testing integrated throughout development lifecycles, regulatory frameworks incorporating adversarial ML considerations, and defense-in-depth architectures designed assuming adversarial operational conditions, the AV community can navigate the challenging security landscape ahead. This dissertation provides tools, insights, and direction for that essential journey toward secure and trustworthy autonomous mobility.

Bibliography

[1] A. Chattopadhyay and K.-Y. Lam, "Security of autonomous vehicle as a cyber-physical system," in *2017 7th International Symposium on Embedded Computing and System Design (ISED)*. IEEE, 2017, pp. 1–6.

[2] S. M. Veres, L. Molnar, N. K. Lincoln, and C. P. Morice, "Autonomous vehicle control systems—a review of decision making," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 225, no. 2, pp. 155–195, 2011.

[3] J. Petit and S. E. Shladover, "Potential cyberattacks on automated vehicles," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, 2015, pp. 546–556.

[4] V. Authors, "Security of autonomous vehicles," *IEEE Transactions*, 2017.

[5] A. Greenberg, "The jeep hackers are back to prove car hacking can get much worse," *Wired*, 2016, uRL: https://www.wired.com/2016/08/jeep-hackers-return-highs-peed-steering-acceleration-hacks/.

[6] B. Nassi, Y. Mirsky, L. Rokach, and Y. Elovici, "Phantom attacks on autonomous driving systems," *IEEE Transactions on Intelligent Transportation Systems*, 2020, early Access.

[7] *ISO/SAE 21434: Road Vehicles – Cybersecurity Engineering*, ISO/SAE Std., 2021.

[8] National Highway Traffic Safety Administration, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," NHTSA, Tech. Rep., 2015.

[9] J. Janai, F. Güney, A. Behl, A. Geiger *et al.*, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and trends® in computer graphics and vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[10] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[11] T. Brophy, D. Mullins, A. Parsi, J. Horgan, E. Ward, P. Denny, C. Eising, B. Deegan, M. Glavin, and E. Jones, "A review of the impact of rain on camera-based perception in automated driving systems," *IEEE Access*, vol. 11, pp. 67 040–67 057, 2023.

[12] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1625–1634.

[13] T. Sato, J. Shen, N. Shen, Y. Liu, Y. Lu, Q. A. Chen, Z. M. Mao, and M. Pajic, "Dirty road can attack: Security of deep learning based automated lane centering under physical-world attack," in *30th USENIX Security Symposium*, 2021, pp. 3309–3326.

[14] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[15] M. Hussain, M. O'Nils, J. Lundgren, and S. J. Mousavirad, "A comprehensive review on deep learning-based data fusion," *IEEE Access*, 2024.

[16] R. Rajamani, *Vehicle Dynamics and Control*. New York, NY, USA: Springer, 2011.

[17] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham *et al.*, "Experimental security analysis of a modern automobile," in *2010 IEEE symposium on security and privacy*. IEEE, 2010, pp. 447–462.

[18] S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, S. Savage, K. Koscher, A. Czeskis, F. Roesner, and T. Kohno, "Comprehensive experimental analyses of automotive attack surfaces," in *20th USENIX security symposium (USENIX Security 11)*, 2011.

[19] N. O. Tippenhauer, C. Pöpper, K. B. Rasmussen, and S. Čapkun, "On the requirements for successful gps spoofing attacks," in *18th ACM Conference on Computer and Communications Security (CCS)*, 2011, pp. 75–86.

[20] A. J. Kerns, K. D. Wesson, and T. E. Humphreys, "Unmanned aircraft capture and control via GPS spoofing," *Journal of Field Robotics*, vol. 31, no. 4, pp. 617–636, 2014.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015, arXiv:1412.6572.

[22] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2014.

[23] A. O. de Sá, L. F. R. da Costa Carmo, and R. C. S. Machado, "Covert attacks in cyber-physical control systems," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1641–1651, 2017.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[25] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[26] R. Mitchell and I.-R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–29, 2014.

[27] Y. Chen, V. Kreinovich, and J. Peng, "On the detectability of false data injection attacks on power grids," *IEEE Transactions*, 2011.

[28] A. Jones, Z. Kong, and C. Belta, "Anomaly detection in cyber-physical systems: A formal methods approach," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 848–853.

[29] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, 2014, pp. 4–11.

[30] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems security—a survey," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1802–1831, 2017.

[31] L. Wu *et al.*, "Physical layer authentication for wireless communications," *IEEE Communications Magazine*, 2012.

[32] M. Raya and J.-P. Hubaux, "Securing vehicular ad hoc networks," in *Journal of Computer Security*, 2007.

[33] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-generated fake images over social networks," in *IEEE Conference on Multimedia Information Processing and Retrieval*, 2019, pp. 384–389.

[34] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[35] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[36] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, 2015.

[37] D. Zhang, Z. Ye, and X. Dong, "Co-design of fault detection and consensus control protocol for multi-agent systems under hidden dos attack," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 2158–2170, 2021.

[38] M. Mwanje, "Cyber security analysis of connected vehicles," *IET Intelligent Transport Systems*, vol. 18, no. 7, pp. 1175–1195, 2024.

[39] Z. Wang, H. Wei, J. Wang, X. Zeng, and Y. Chang, "Security issues and solutions for connected and autonomous vehicles in a sustainable city: A survey," *Sustainability*, vol. 14, no. 19, p. 12409, 2022.

[40] M. Abdel-Basset, A. Gamal, N. Moustafa, A. Abdel-Monem, and N. El-Saber, "A security-by-design decision-making model for risk management in autonomous vehicles," *IEEE Access*, vol. 9, pp. 107 657–107 679, 2021.

[41] P. B. Savitha, S. Madhu, and S. Arjun, "Cyber security issues in connected autonomous vehicle," *International Journal of Research Publication and Reviews*, vol. 4, pp. 929–936, 2023.

[42] P. Solnør, Ø. Volden, K. Gryte, S. Petrović, and T. I. Fossen, "Hijacking of unmanned surface vehicles: A demonstration of attacks and countermeasures in the field," *Journal of Field Robotics*, vol. 39, no. 5, pp. 631–649, 2022.

[43] S. Aurangzeb, "Cybersecurity for autonomous vehicles against malware attacks in smart-cities," *Cluster Computing*, vol. 27, no. 3, pp. 3363–3378, 2023.

[44] J. Raiyn, "Data and cyber security in autonomous vehicle networks," *Transport and Telecommunication Journal*, vol. 19, pp. 325–334, 2018.

[45] R. Changalvala, B. Fedoruk, and H. Malik, "Radar data integrity verification using 2d qim-based data hiding," *Sensors*, vol. 20, no. 19, p. 5530, 2020.

[46] H. Kim, J. Shin, and H. Jo, "Adversarial example detection based on improved ghostbusters," *IEICE Transactions on Information and Systems*, vol. E105-D, no. 11, pp. 1921–1922, 2022.

[47] Z. Lian, P. Shi, C.-C. Lim, and Y. Xing, "Fuzzy-model-based lateral control for networked autonomous vehicle systems under hybrid cyber-attacks," *IEEE Transactions on Cybernetics*, vol. 53, no. 4, pp. 2600–2609, 2023.

[48] L. Chen, Y. Li, and S. Tong, "Robust adaptive control for nonlinear cyber-physical systems with FDI attacks via attack estimation," *International Journal of Robust and Nonlinear Control*, vol. 33, pp. 9299–9316, 2023.

[49] A. A. Alsulami, Q. A. Al-Haija, A. Alqahtani, and R. Alsini, "Symmetrical simulation scheme for anomaly detection in autonomous vehicles based on LSTM model," *Symmetry*, vol. 14, p. 1450, 2022.

[50] X. Huang and X. Wang, "Detection and isolation of false data injection attack in intelligent transportation system via robust state observer," *Processes*, vol. 10, p. 1299, 2022.

[51] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "ShapeShifter: Robust physical adversarial attack on Faster R-CNN object detector," in *European Conference on Computer Vision*. Springer, 2018, pp. 52–68.

[52] A. Shah, "Adversary ML resilience in autonomous driving through human centered perception mechanisms," *arXiv preprint arXiv:2311.01478*, 2023.

[53] C.-Y. Kuo, Y.-R. Lu, and S.-F. Yang, "On the image sensor processing for lane detection and control in vehicle lane keeping systems," *Sensors*, vol. 19, p. 1665, 2019.

[54] N. B. Zakaria, M. I. Shapiai, R. Ghani, M. N. Yasin, M. Z. Ibrahim, and N. Wahid, "Lane detection in autonomous vehicles: A systematic review," *IEEE Access*, vol. 11, pp. 3729–3765, 2023.

[55] D. C. Andrade, F. Bueno, F. R. Franco, R. A. Silva, J. H. Z. Neme, E. Margraf, W. T. Omoto, F. A. Farinelli, A. M. Tusset, S. Okida *et al.*, "A novel strategy for road lane detection and tracking based on a vehicle's forward monocular camera," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1497–1507, 2019.

[56] C. Dong, H. Wang, D. Ni, Y. Liu, and Q. Chen, "Impact evaluation of cyber-attacks on traffic flow of connected and automated vehicles," *IEEE Access*, vol. 8, pp. 86 824–86 835, 2020.

[57] Q. He, X. Meng, R. Qu, and R. Xi, "Machine learning-based detection for cyber security attacks on connected and autonomous vehicles," *Mathematics*, vol. 8, no. 8, p. 1311, 2020.

[58] H. Liu, C. Liu, L. Hao, and D. Zhang, "Stability analysis of lane-keeping assistance system for trucks under crosswind conditions," *Applied Sciences*, vol. 13, p. 9891, 2023.

[59] J. Kim, S. Park, J. Kim, and J. Yoo, "A deep reinforcement learning strategy for surrounding vehicles-based lane-keeping control," *Sensors*, vol. 23, p. 9843, 2023.

[60] X. Xu, J. W. Grizzle, P. Tabuada, and A. D. Ames, "Correctness guarantees for the composition of lane keeping and adaptive cruise control," *IEEE Transactions on Automation Science and Engineering*, vol. 15, pp. 1216–1229, 2018.

[61] A. Guesmi, M. A. Hanif, and M. Shafique, "AdvRain: Adversarial raindrops to attack camera-based smart vision systems," *Information*, vol. 14, no. 12, p. 634, 2023.

[62] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 284–293.

[63] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 2017–2025.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[65] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[66] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[67] Z. Zhang, L. Han, A. Ghosh, D. N. Metaxas, and J. Ren, "SINE: Single image editing with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6027–6037.

[68] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[69] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 563–22 575.

[70] Z. Yu, H. Wang, J. Yang, H. Wang, Z. Xie, Y. Cai, J. Cao, Z. Ji, and M. Sun, "SGD: Street view synthesis with Gaussian splatting and diffusion prior," *arXiv preprint arXiv:2403.20079*, 2024.

[71] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2022.

[72] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics*, vol. 41, no. 4, pp. 1–15, 2022.

[73] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-NeRF: Scalable large scene neural view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8248–8258.

[74] T. E. Humphreys, B. M. Ledvina, M. L. Psiaki, B. W. O'Hanlon, and P. M. Kintner, "Assessing the spoofing threat: Development of a portable GPS civilian spoofer," in *Proceedings of the ION GNSS*, 2008, pp. 2314–2325.

[75] C. Sanders and Y. Wang, "Localizing spoofing attacks on vehicular GPS using vehicle-to-vehicle communications," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 656–15 667, 2020.

[76] Z. Miao, C. Shao, H. Li, and Y. Cui, "Adaptive attack recognition method based on probability model for autonomous vehicle," *Electronics Letters*, vol. 60, 2024.

[77] F. Gallardo and A. Pérez Yuste, "SCER: Spoofing attacks on the Galileo open service and machine learning techniques for end-user protection," *IEEE Access*, vol. 8, pp. 85 515–85 532, 2020.

[78] J. S. Warner and R. G. Johnston, "GPS spoofing countermeasures," *Homeland Security Journal*, vol. 25, 2003.

[79] H. Li and D. Chasaki, "Detecting covert channel attacks on cyber-physical systems," *IET Cyber-Physical Systems: Theory & Applications*, vol. 9, no. 3, pp. 228–237, 2024.

[80] S. Gupta, C. Maple, and R. Passerone, "An investigation of cyber-attacks and security mechanisms for connected and autonomous vehicles," *IEEE Access*, vol. 11, pp. 90 641–90 669, 2023.

[81] Z. Ju, H. Zhang, and Y. Tan, "Distributed deception attack detection in platoon-based connected vehicle systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 4609–4620, 2020.

[82] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[83] A. Giannaros, S. Fotiou, V. Pappas, E. Papadopoulou, and G. Spanoudakis, "Autonomous vehicles: Sophisticated attacks, safety issues, challenges, open topics, blockchain, and future directions," *Vehicles*, vol. 3, no. 3, pp. 474–520, 2021.

[84] K. Zaidi, M. B. Milojevic, V. Rakocevic, A. Nallanathan, and M. Rajarajan, "Host-based intrusion detection for VANETs: A statistical approach to rogue node detection," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6703–6714, 2016.

[85] R. Majumder, M. Chowdhury, S. M. Khan, Z. Khan, F. Ahmed, F. Ngeni, G. Comert, J. Mwakalonge, and D. Michalaka, "Adversarial attack-resilient perception module for traffic sign classification," 2023.

[86] B. Feldsar, R. Mayer, and A. Rauber, "Detecting adversarial examples using surrogate models," *Machine Learning and Knowledge Extraction*, vol. 5, pp. 1796–1825, 2023.

[87] B. Zou, P. Choobchian, and J. Rozenberg, "Cyber resilience of autonomous mobility systems: Cyber-attacks and resilience-enhancing strategies," *Journal of Transportation Security*, pp. 1–19, 2021.

[88] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy*. IEEE, 2017, pp. 39–57.

[89] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*, 2019, pp. 1310–1320.

[90] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8695–8704.

[91] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[92] D. Cozzolino, F. Marra, D. Gragnaniello, G. Poggi, and L. Verdoliva, "Combining PRNU and noiseprint for robust and efficient device source identification," *EURASIP Journal on Information Security*, vol. 2020, p. 9, 2020.

[93] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[94] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[95] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[96] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2813–2821.

[97] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[98] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[99] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," in *RSS Workshop on Limits and Potentials of Deep Learning in Robotics*, 2016. [Online]. Available: https://arxiv.org/abs/1606.03798

[100] J. H. Lim and J. C. Ye, "Geometric GAN," *arXiv preprint arXiv:1705.02894*, 2017.

[101] E. Dubrofsky, "Homography estimation," Department of Computer Science, University of British Columbia, Tech. Rep. TR 2009-06, 2009. [Online]. Available: https://www.cs.ubc.ca/~dubrofsk/thesis/thesis.pdf

[102] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1999, pp. 11–20.

[103] Y. B. Can, A. Liniger, D. P. Paudel, and L. V. Gool, "Homography guided temporal fusion for road line and marking segmentation," *arXiv preprint arXiv:2404.07626*, 2024.

[104] S. Singh and P. Kumari, "Insights of computer vision-based techniques: Perspective transformation and sliding window approach for lane line detection in autonomous vehicles," *International Journal of Vehicle Performance*, vol. 11, no. 1, pp. 78–109, 2025.

[105] T. Nguyen, S. L. Phung, and A. Bouzerdoum, "Abhe: All attention-based homography estimation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[106] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3501–3510.

[107] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2636–2645.

,

[108] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[109] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[110] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.

[111] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[112] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, vol. 32, no. 11, pp. 1231–1237, 2013.

[113] Z. Wang, T. Zhong, X. Li, Y. Liu, and Y. Wang, "Dire for diffusion-generated image detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22 445–22 455.

[114] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.