Enhancing the Security of Autonomous Vehicles: Detection of Adversarial Attacks on Perception Systems

Farah Sherif

Supervisor: Dr. Muhammad Hataba

Faculty of Informatics and Computer Science

27-01-2025

Outline

- Introduction
- Related Work
- System Models
- Experimental Setup
- Results
- Conclusion

Introduction

Motivation

- **Global Impact:** AVs are transforming transportation with enhanced safety and efficiency.
- **Rapid Market Growth:** Expected market value rise from **$22.22B** (2021) to **$75.95B** (2027) at **22.75% CAGR.**
- **Big-Tech Focus:** Major investments from Google, Apple, and Uber.
- **Vulnerability:** Reliance on machine learning exposes AVs to adversarial attacks.
- **Security Priority:** Widespread adoption makes robust security essential.

Introduction

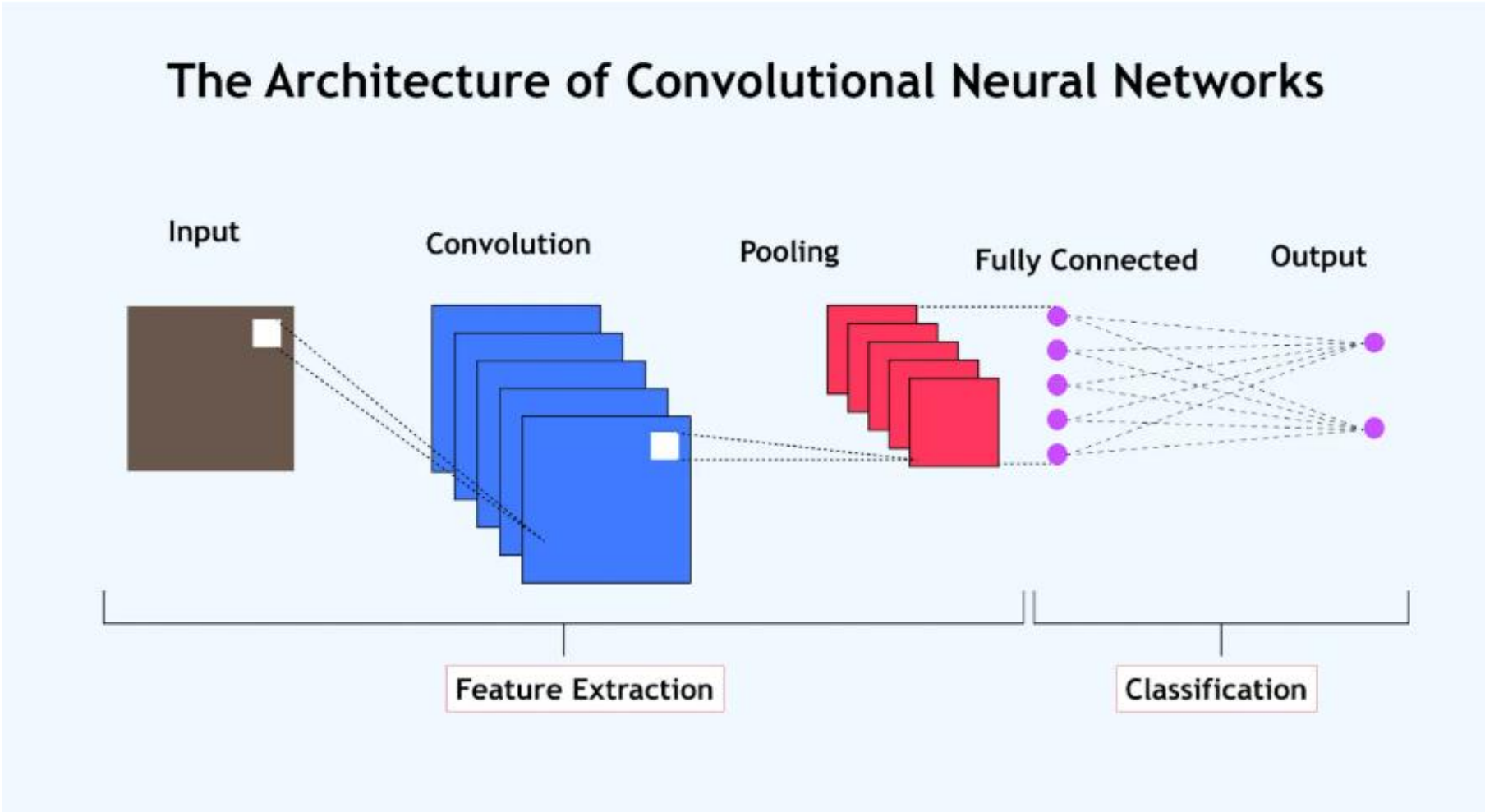
Problem Statement

- **Critical need for security:** Adversarial Attacks targeting the perception system of AVs lead to detrimental effects
- **Real-world Incidents:** Tesla’s 2016 and Uber’s 2018 crashes, and the remote hacking of a Jeep's systems highlight the dangers of such attacks.
- **Research Gaps:** Limited focus on multi-sensor systems and sensor fusion for detecting adversarial attacks.
- **GAN Usage:** Most research involving GANs has focused on developing adversarial attacks rather than defensive strategies.

Thesis Contributions

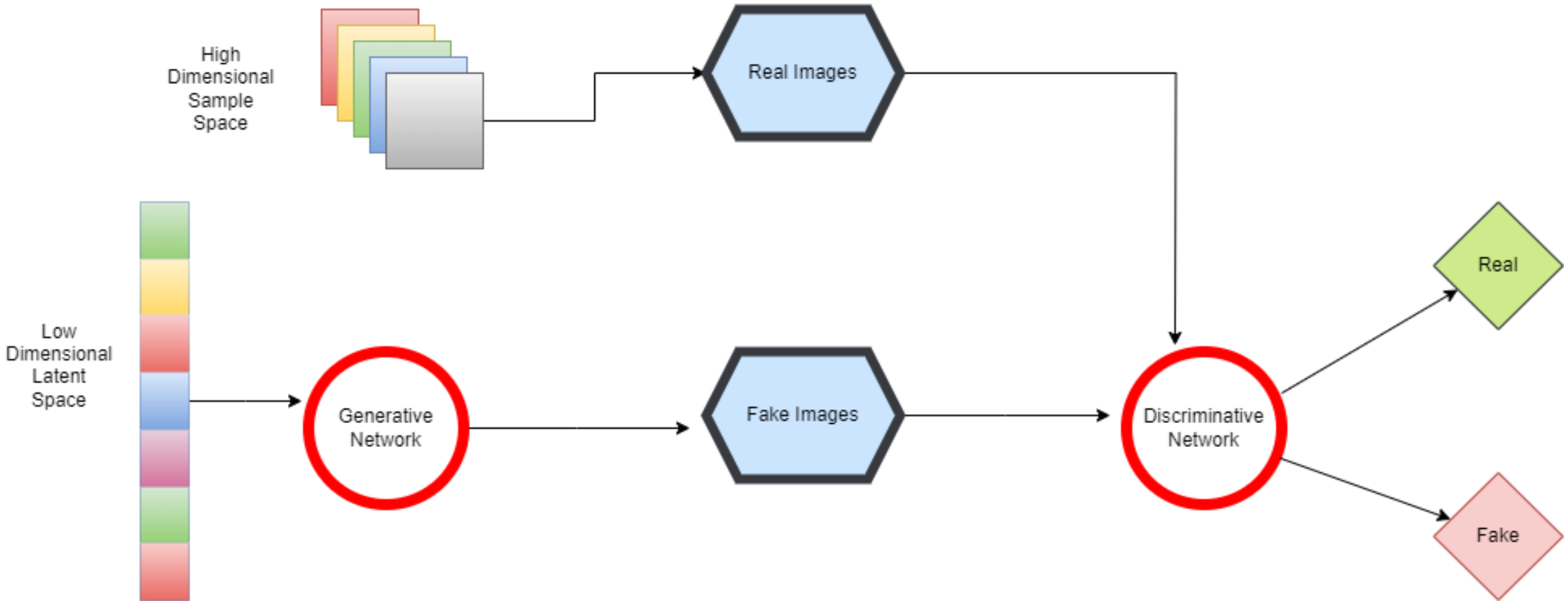
- Dynamic Threat Modeling to Identify and Mitigate AV System Vulnerabilities
- Developing Diverse Adversarial Scenarios to Explore AV Vulnerabilities
- Analyzing Adversarial Impacts and Leveraging Sensor Fusion for Multi-Modal System Security
- Harnessing GANs for Robust Defense of Traffic Sign Recognition Models
- Designing Tailored Detection Mechanisms Using Deep Learning and Adversarial Training

Background


The Architecture of Convolutional Neural Networks
Input
Convolution
Pooling
Fully Connected
Output
Feature Extraction
Classification

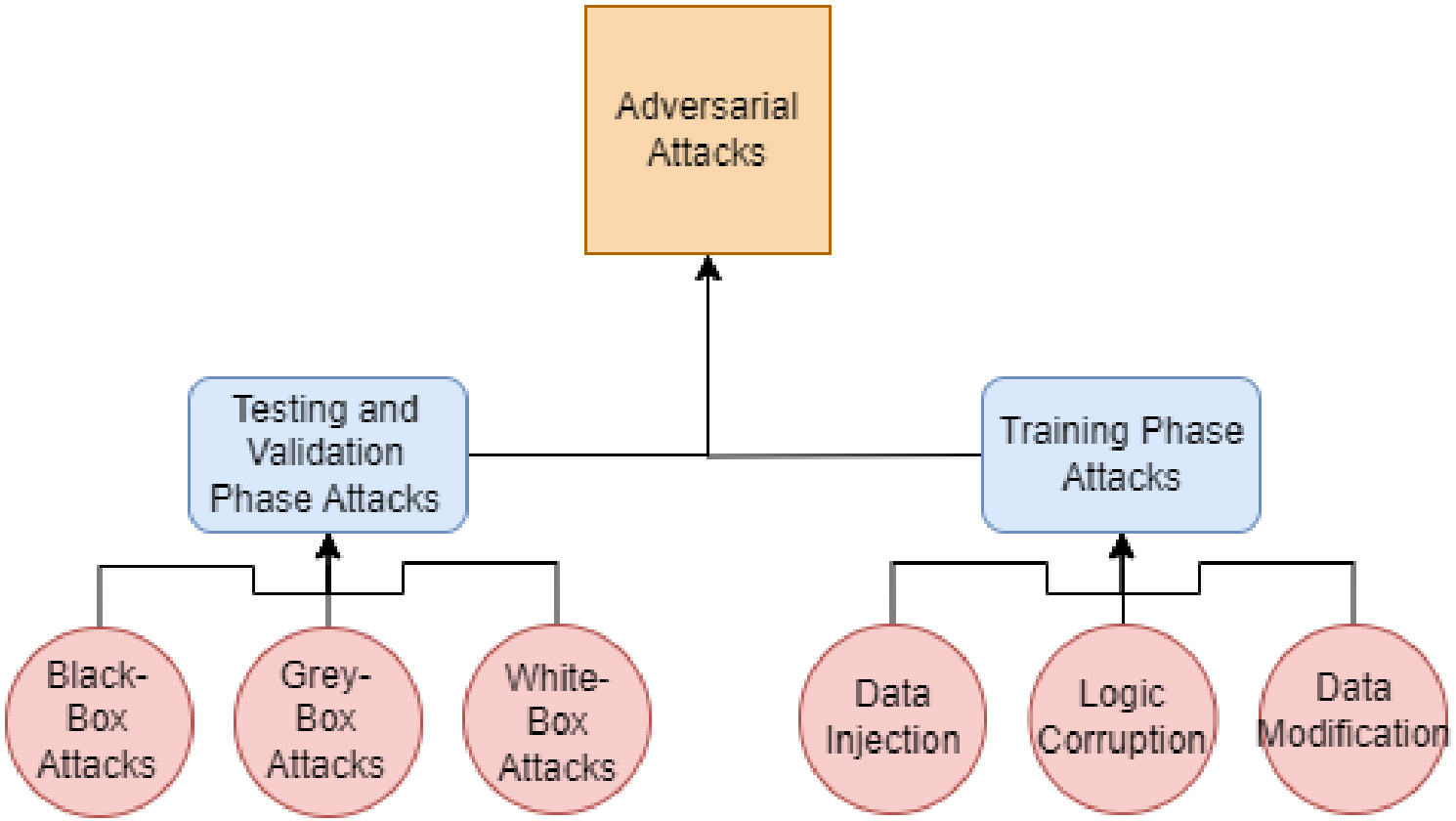
Background

Generative Adversarial Networks



Background

Adversarial Attacks Taxonomy

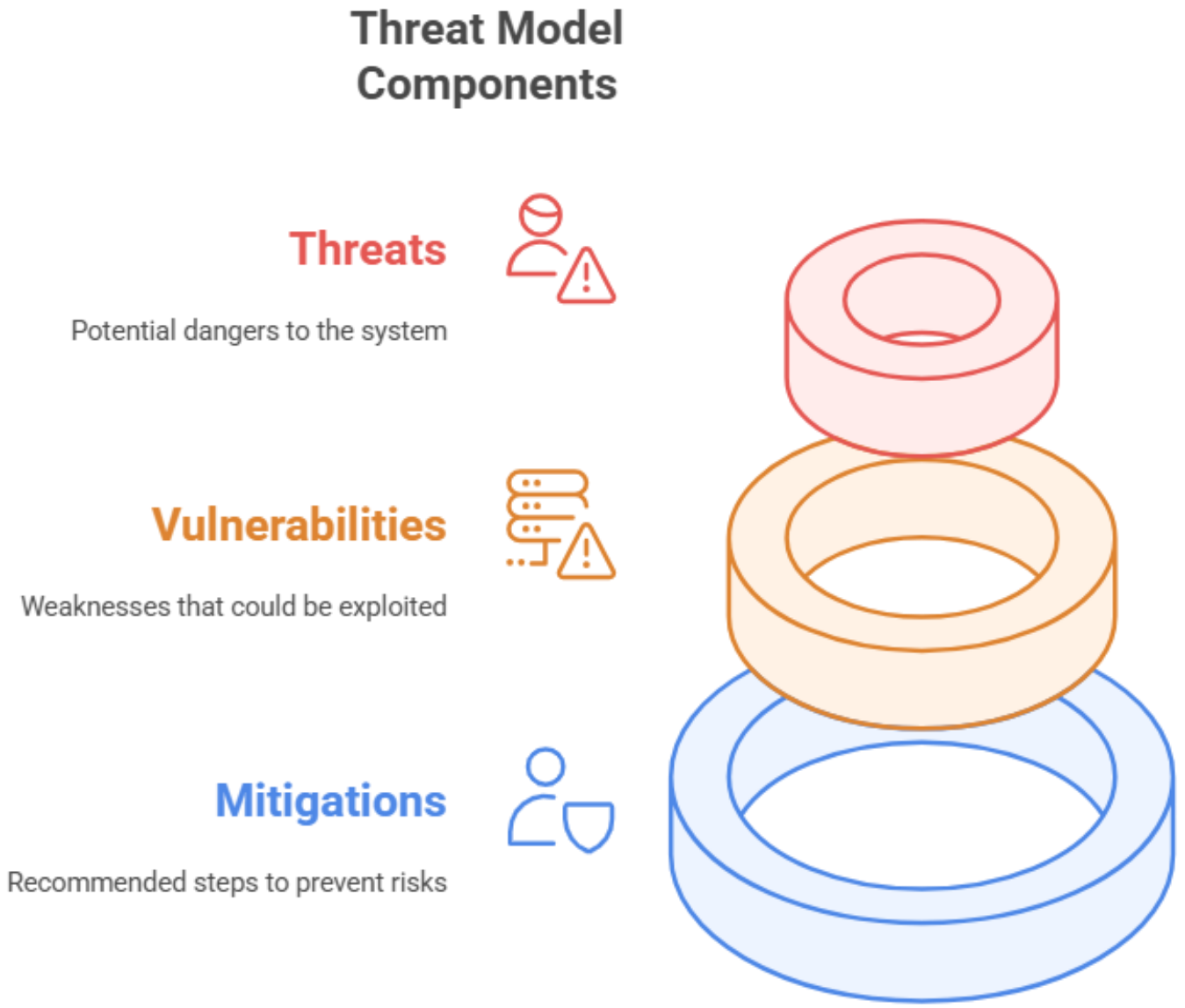


Selected Related Work

	Adaptive Square Attack [1]	Darts: Deceiving autonomous cars with toxic signs [2]	EPMF: Efficient Perception-aware Multi-sensor Fusion for 3D Semantic Segmentation [3]
What it does	Black-box attack targeting a DNN-based traffic sign recognition model	• Out-of-distribution & Lenticular Printing attacks • The findings underscore the need for more robust security measures in ML-driven recognition systems for autonomous vehicles	uses a novel two-stream network and perception-aware losses to align and combine features from LiDAR and camera data
How it relates to this research	Lacks the use of generative models for defense	Did not use detection and sensor fusion-based defense strategies	Inspired the implementations of fusion proposed

Threat Model

- Fake Vehicle Attack
- Traffic Sign Manipulation Attack
- Evasion Attack on Traffic Sign Classifier Using GAN



System Models

Design Goals

Attack Scenario 1

Attack targets camera input adding fake vehicles that OD model will detect

Detection Mechanisms:

- Binary Classifier - Anomaly Detection
- Sensor Fusion

Attack Scenario 2

Attack targets camera input replacing traffic signs; affecting traffic sign recognition

Detection Mechanisms:

- Binary Classifier - Anomaly Detection

Attack Scenario 3

GAN

Evasion attack on traffic sign classifier by integrating GAN-generated traffic signs in test set

Detection Mechanisms:

- Adversarial Training – Anomaly Detection

System Design

Fake Vehicle Attack

Adversarial Scenes Preprocessing:

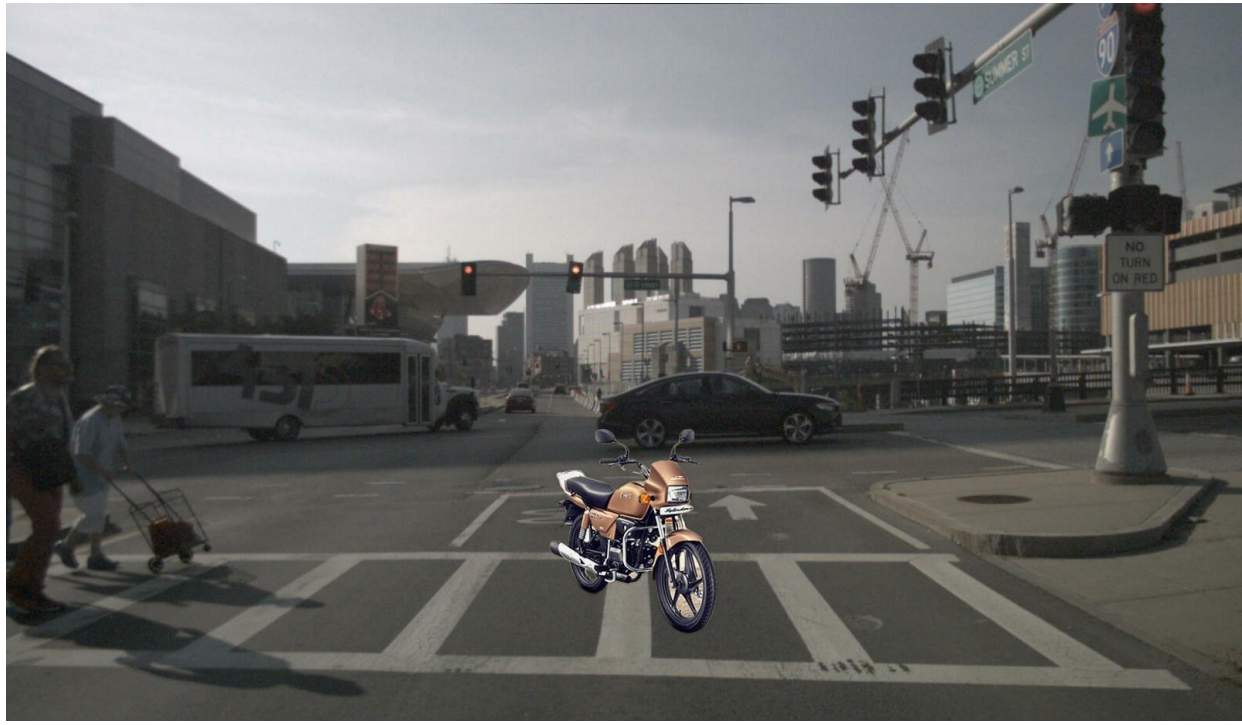
- Dataset of vehicle images

- Only ground-level vehicles are used to ensure compatibility with the NuScenes camera images.

- To ensure realism:
  - The backgrounds of the vehicle images are removed
  - Vehicles resized proportionately
  - Vehicles placed on bottom half of the scene
  - Visual adjustments:
    - Brightness
    - Contrast
    - Shadows
    - Depth Blur (for vehicles placed farther away)

System Design

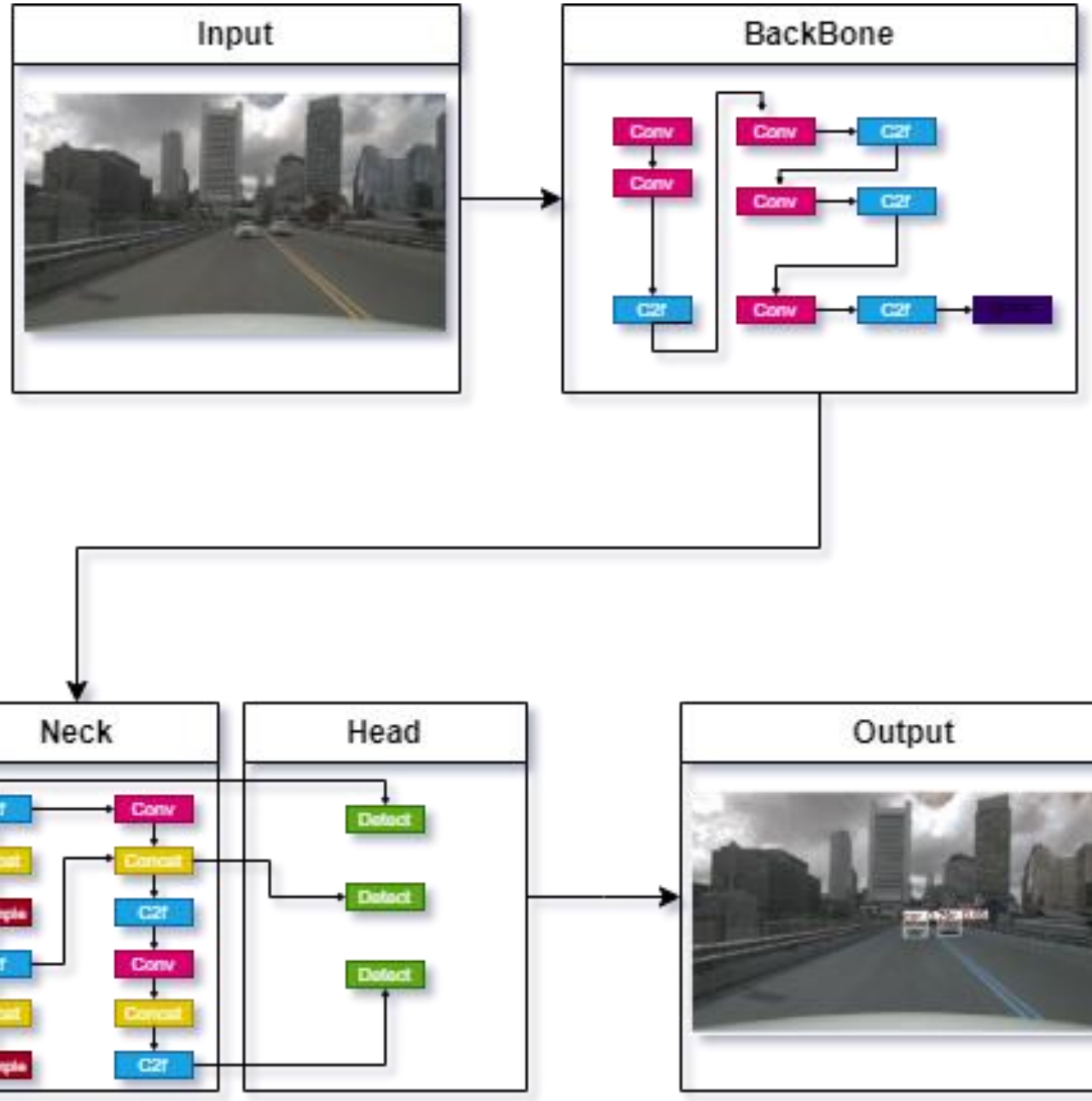
Fake Vehicle Attack



System Design

Fake Vehicle Attack

The object detection model used was the 'n' variant of the YOLOv8 model for its **speed** and **efficiency**



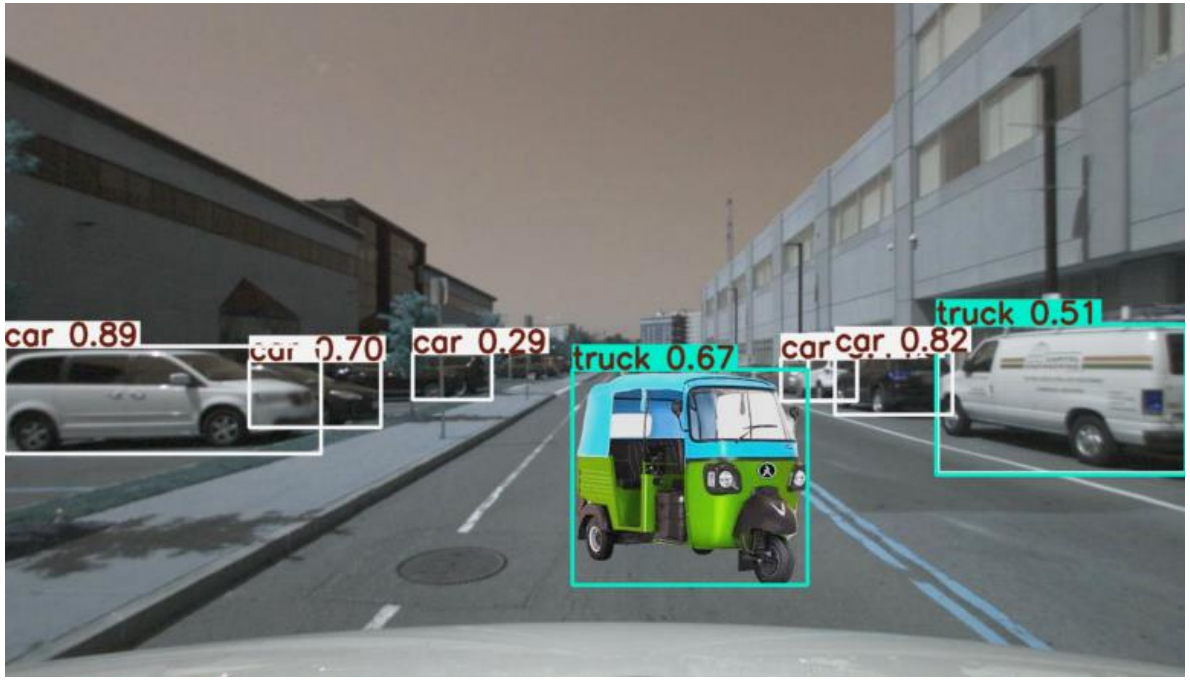
Backbone: Extracts hierarchical features from input (edges, textures).

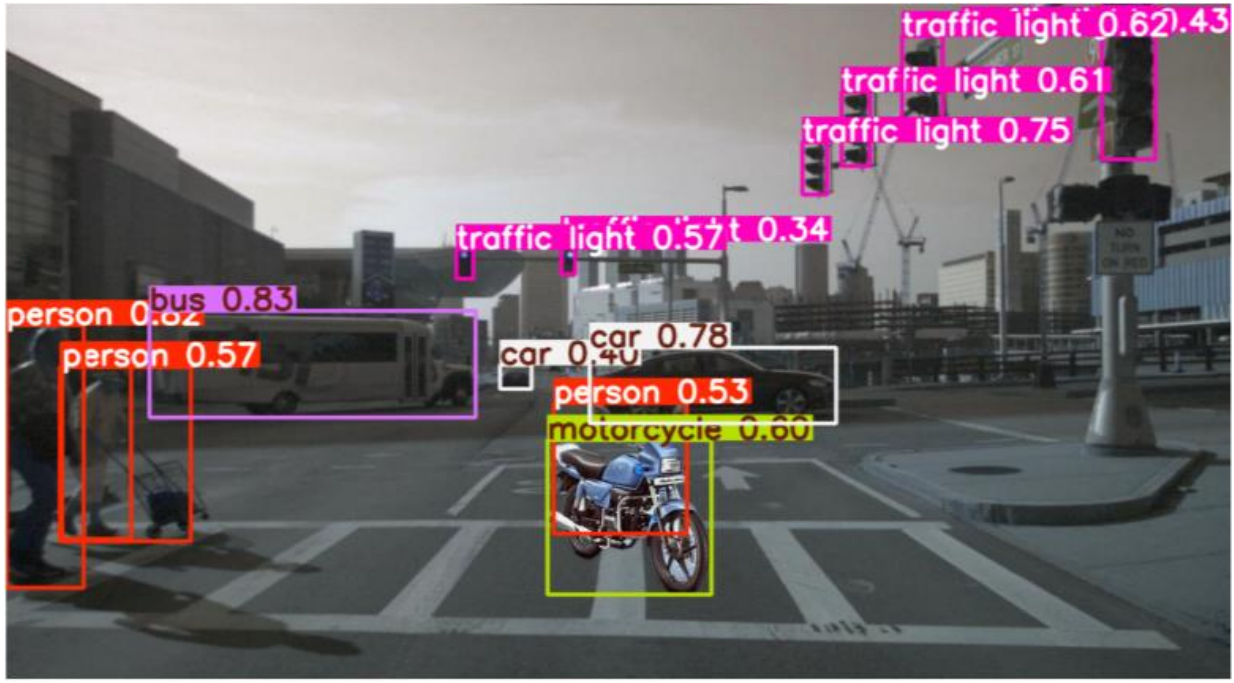
Neck: Fuses feature maps for context and accuracy improvement.

Head: Outputs bounding boxes and confidence scores for detected objects.

System Design

Fake Vehicle Attack





System Design

Fake Vehicle Attack – Detection Mechanisms

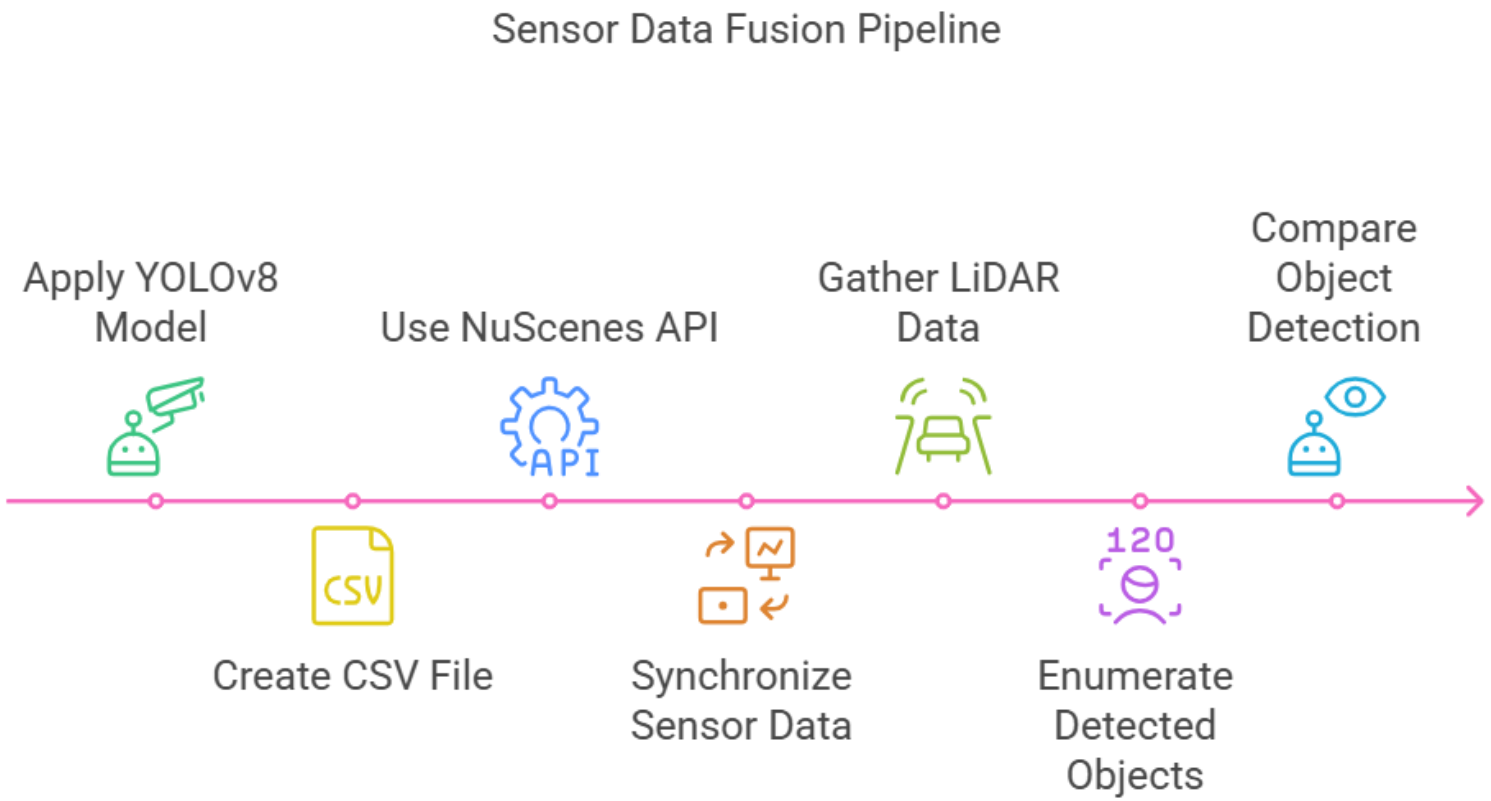
1. Binary Classifier
   - CNN Model Using ResNet-18 model
   - Images underwent transformations to ensure uniformity and compatibility
   - Final layer replaced with a linear layer to adapt to the **binary** classification
   - Sigmoid Activation Function for probability
   - Binary Cross Entropy Loss (BCE Loss)
   - Adam Optimizer

$$F(x) = H(x) - x$$

System Design

Fake Vehicle Attack – Detection Mechanisms

2. Sensor Fusion
   a) Numerical Comparison

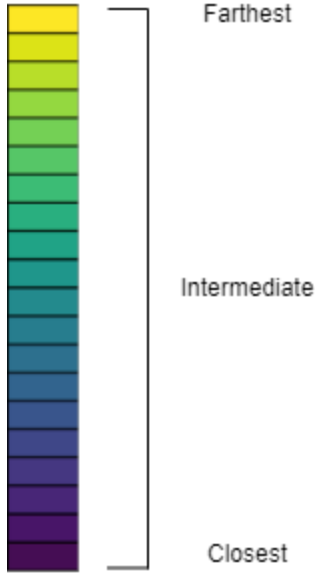


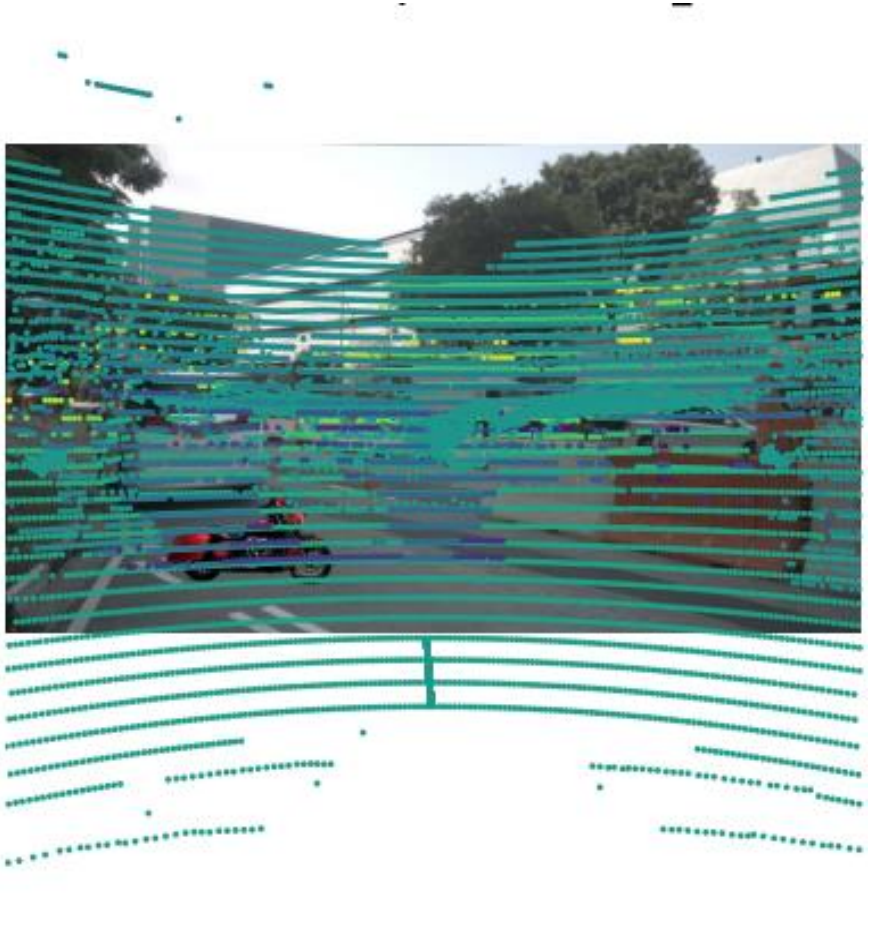
System Design

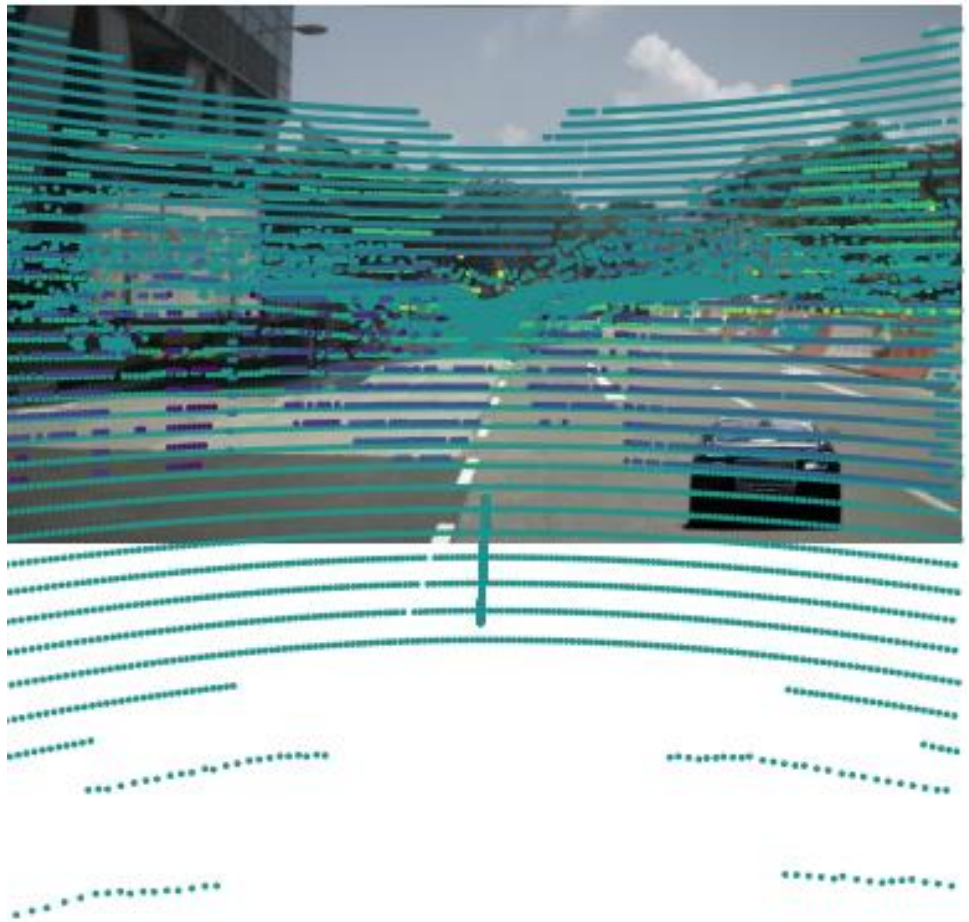
Fake Vehicle Attack – Detection Mechanisms

2. Sensor Fusion
   a) Visual Comparison

- Extrinsic Transformation
- Intrinsic Transformation
- Depth visualized using the "viridis" colormap:



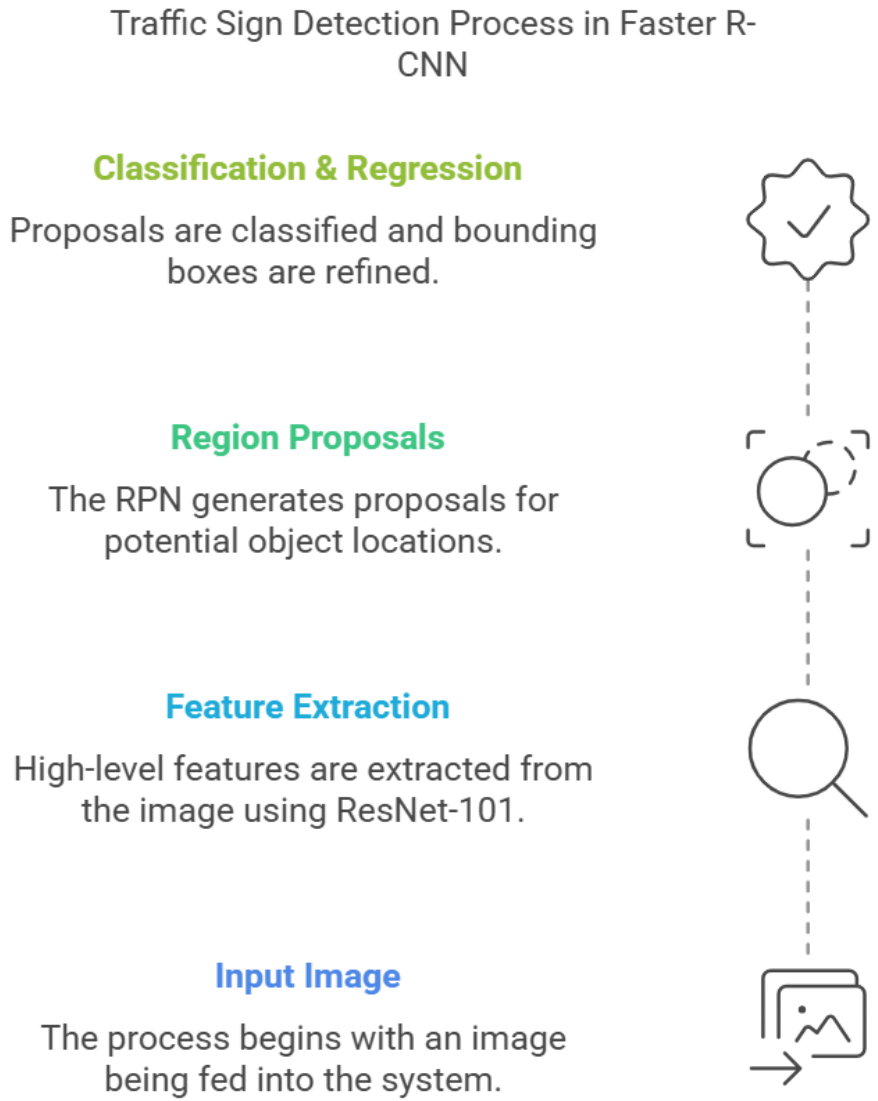




System Design

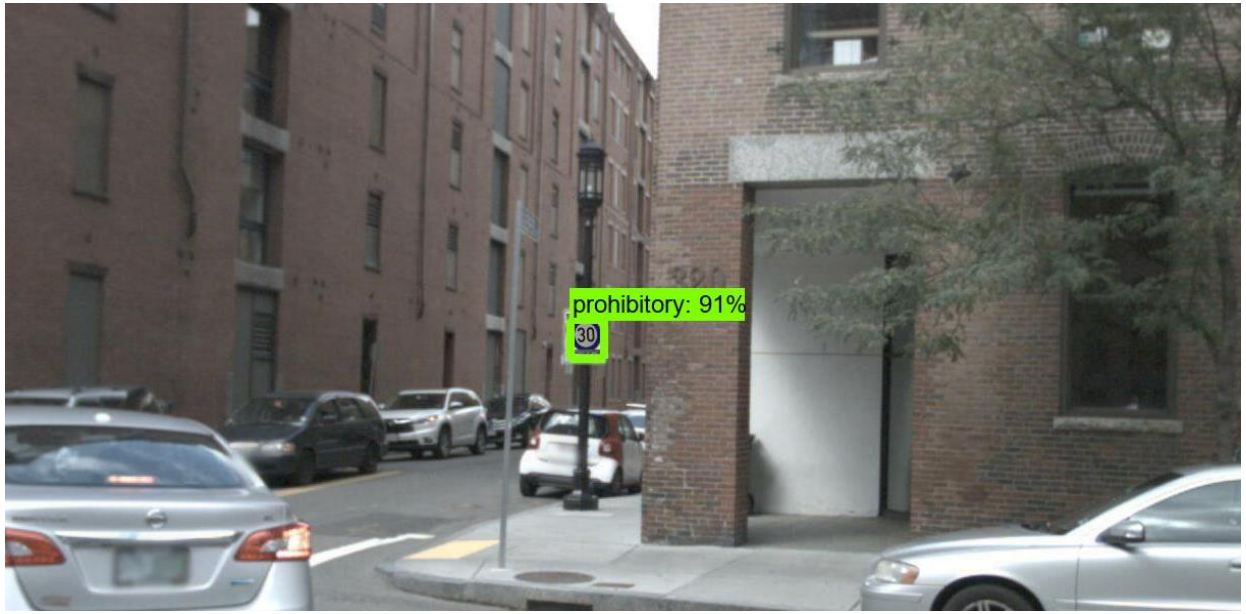
Traffic Sign Manipulation Attack

- Faster-ResNet-101 Model for traffic sign detection
- Each detection is replaced with a randomly chosen traffic sign from the GTSRB - German Traffic Sign Recognition Benchmark
- Leads to the misclassification of traffic signs or not detecting them at all
- **Serious side-effects**: Violating traffic sign laws, and legal or fatal consequences



System Design

Traffic Sign Manipulation Attack



Fake

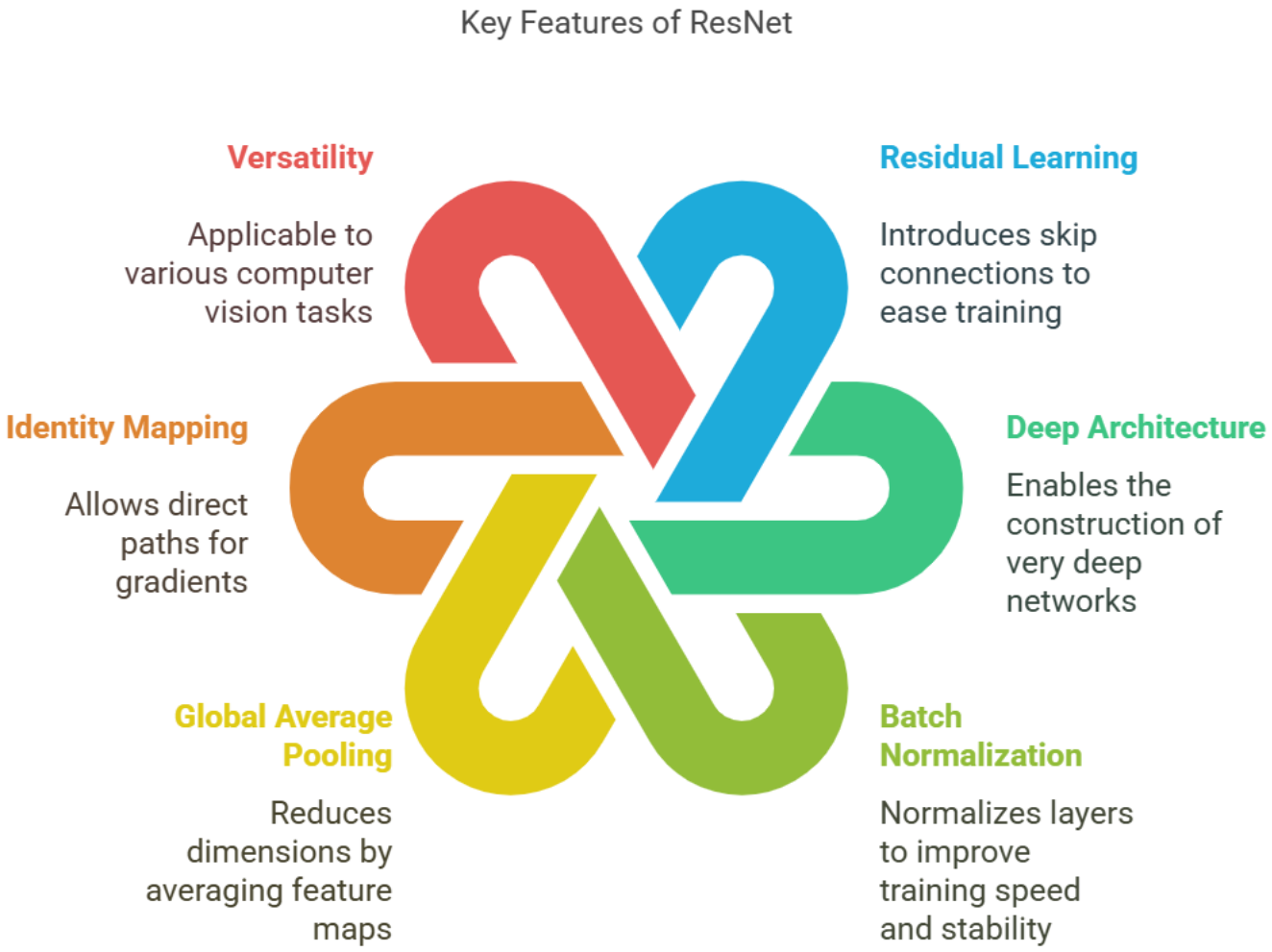
Real

System Design

Traffic Sign Manipulation Attack – Detection Mechanism

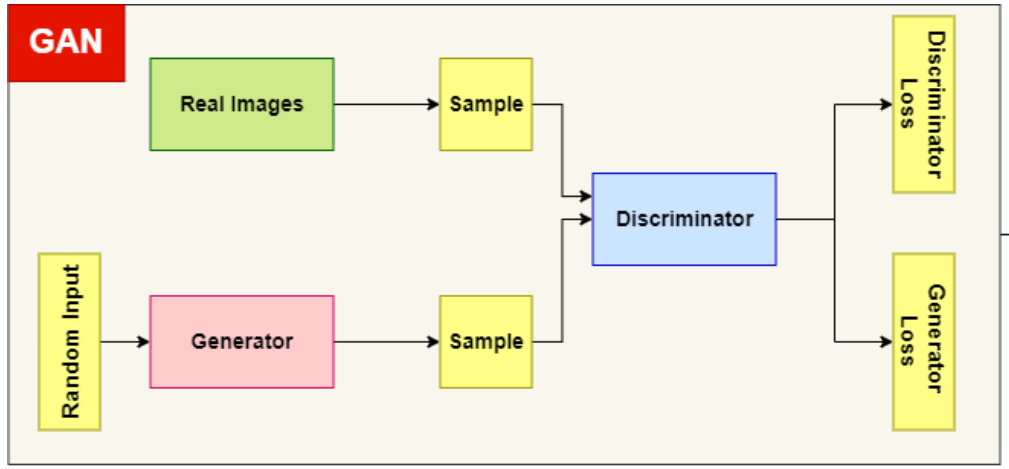
Binary Classifier:

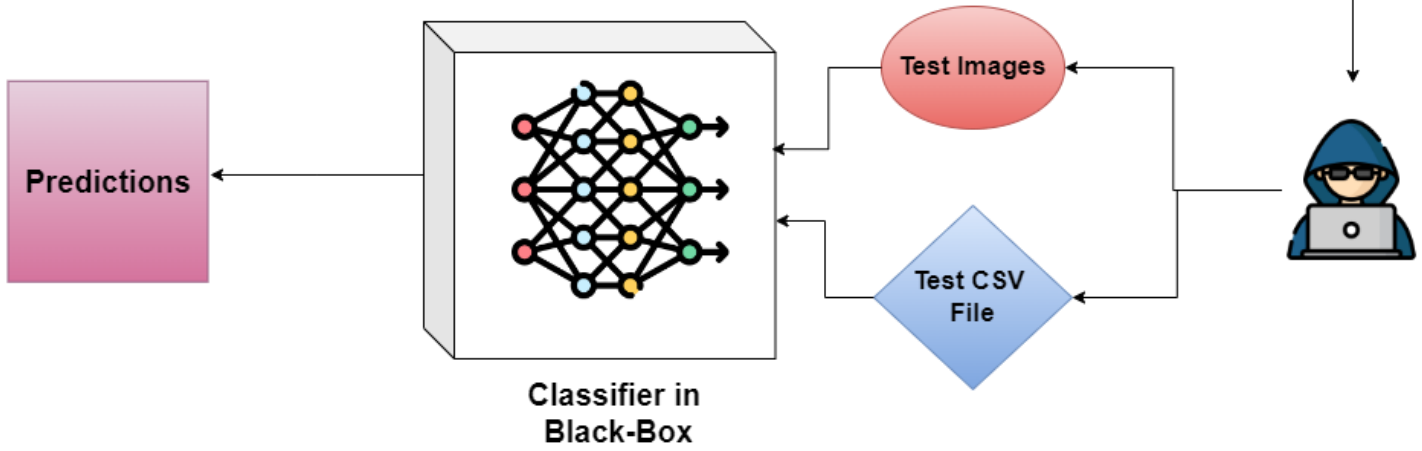
Same model used for the previous attack



System Design

Evasion Attack on Traffic Sign Classifier Using a GAN





- Deep Convolutional GAN produces adversarial traffic signs
- Attack targets the classifier by adding adversarial examples in test set
- It is a black-box attack

System Design

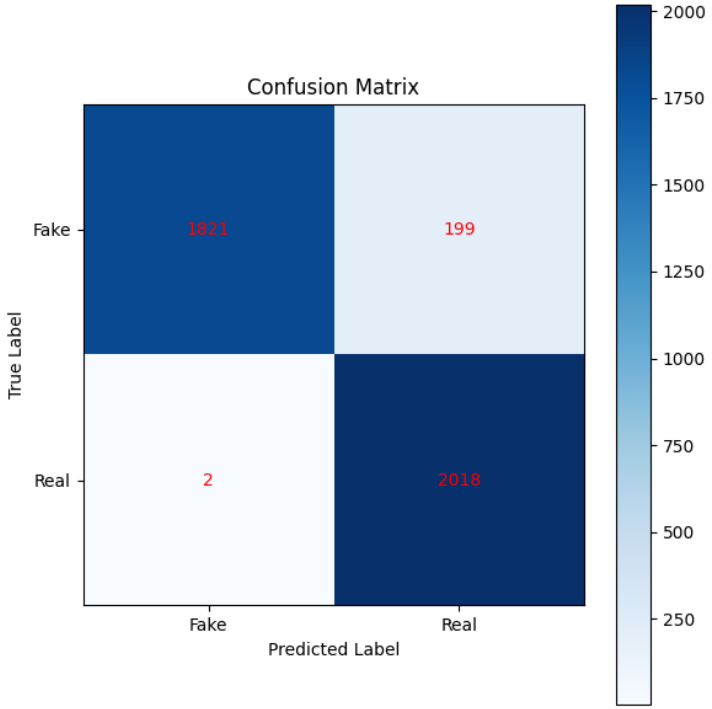
Evasion Attack on Traffic Sign Classifier Using a GAN – Detection Machanism

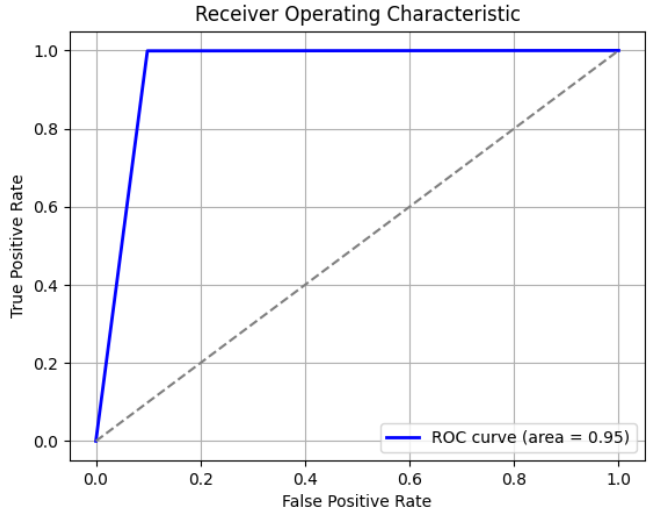
Adversarial Training:

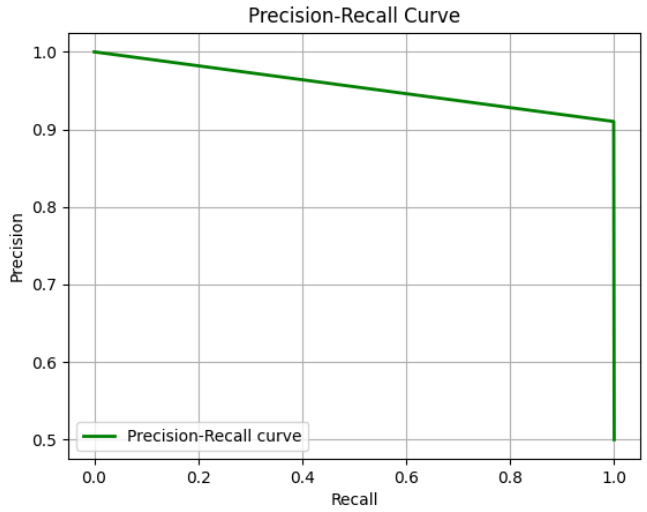
- Additional "Adversarial" class was added to labels in the classifier
- GAN-generated traffic signs added to the training dataset

Results Discussion

Fake Vehicle Attack – Evaluating The Binary Classifier





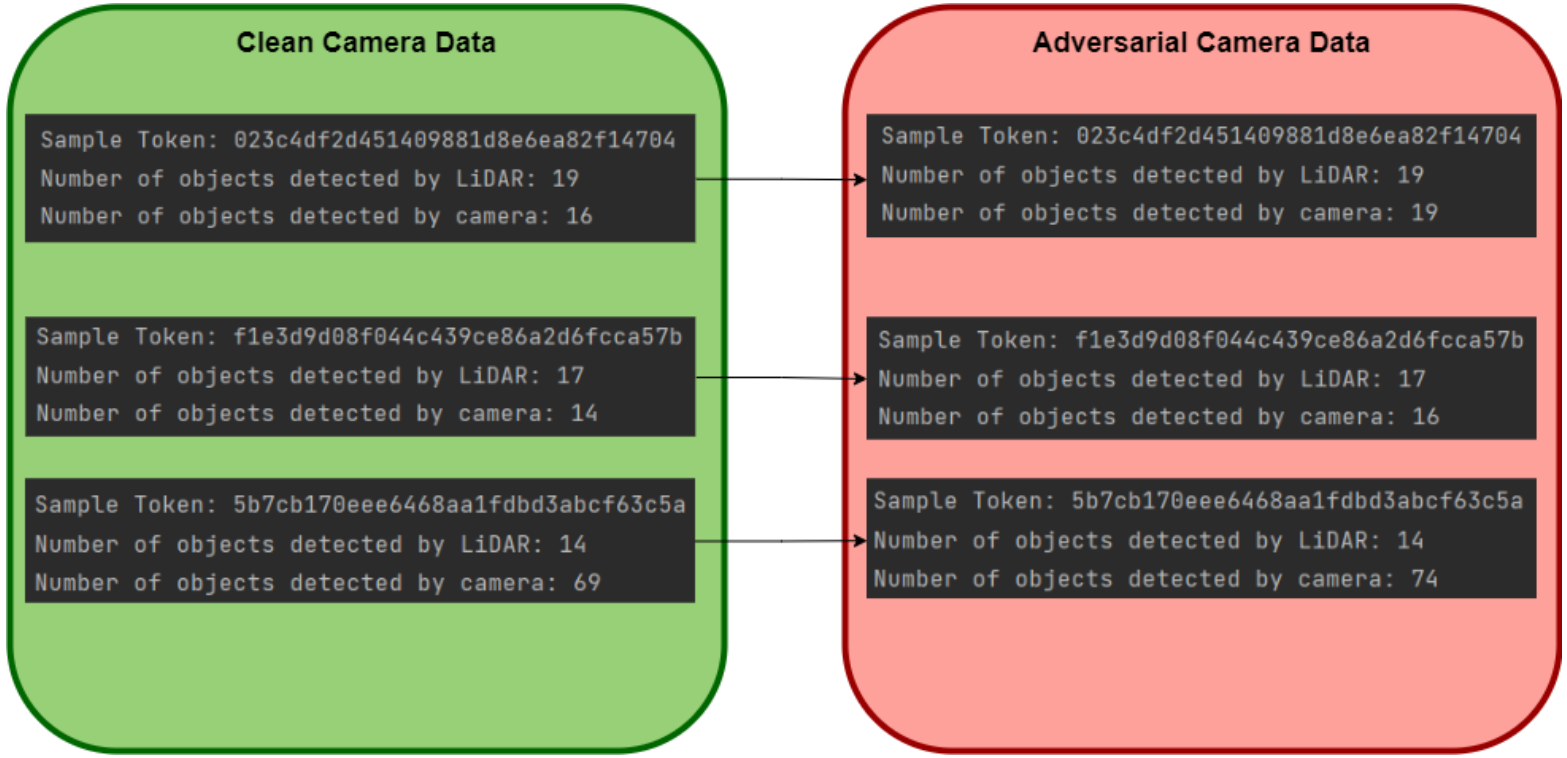


Key Results:

- The model achieved a high accuracy of **95.02%.**
- It successfully identified **90.15%** of adversarial examples, detecting 1,821 out of 2,020 attacks.
- The false negative rate was very low, at just **0.10%.**

Results Discussion

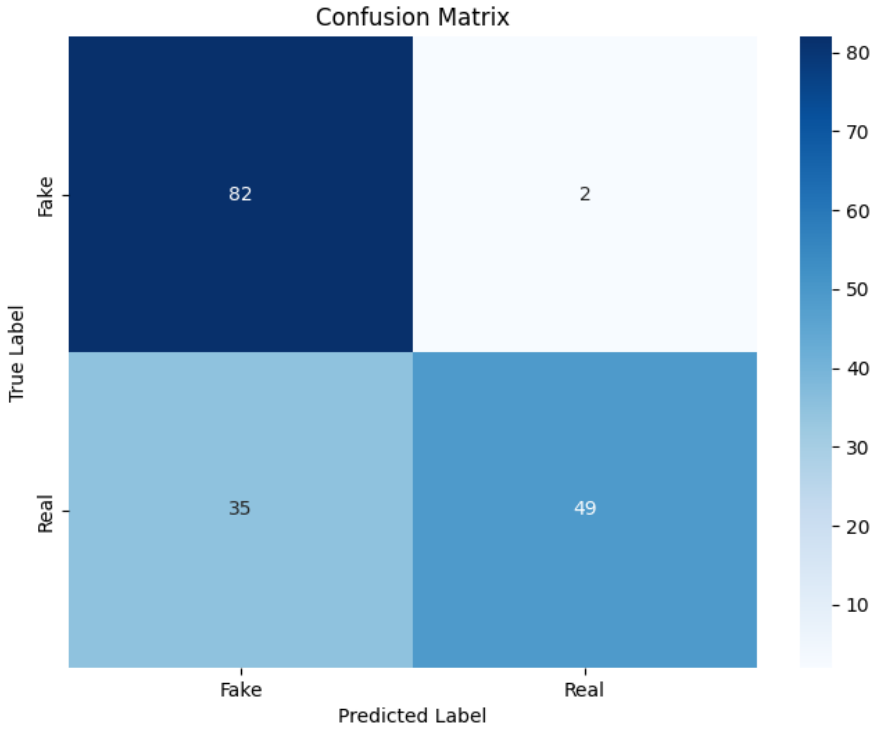
Fake Vehicle Attack – Sensor Fusion

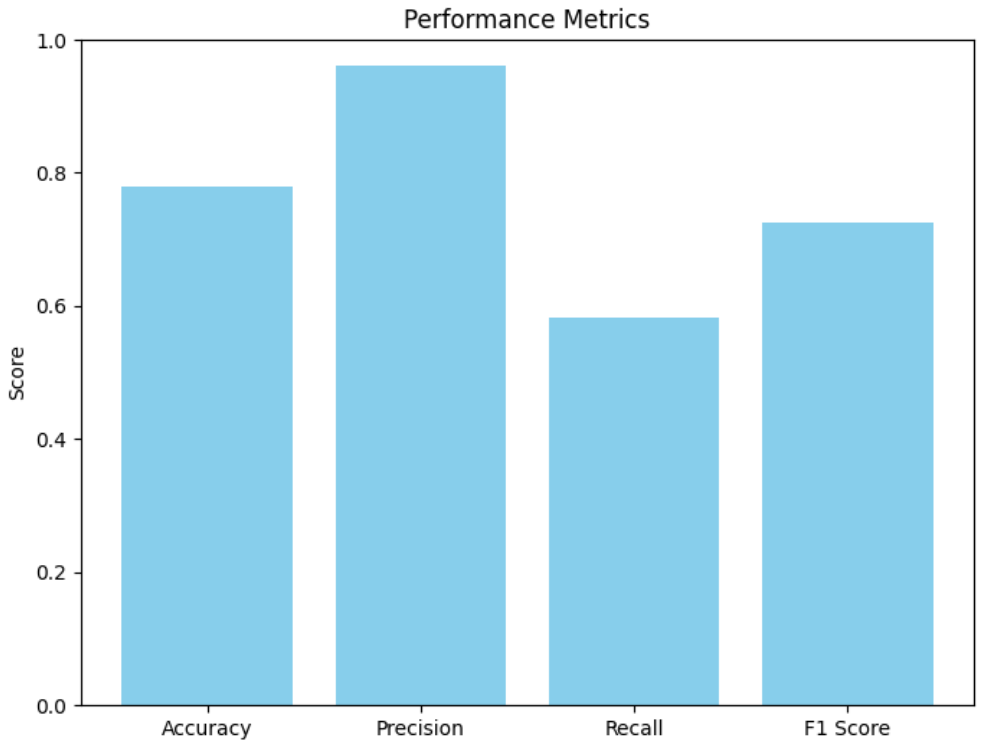


- Adversarial attacks amplify discrepancies between sensor readings
- Integration of additional variables (eg. object position) in the fusion is necessary
- Higlights effectiveness of sensor fusion in detecting adversarial examples

Results Discussion

Traffic Sign Manipulation – Evaluating The Binary Classifier





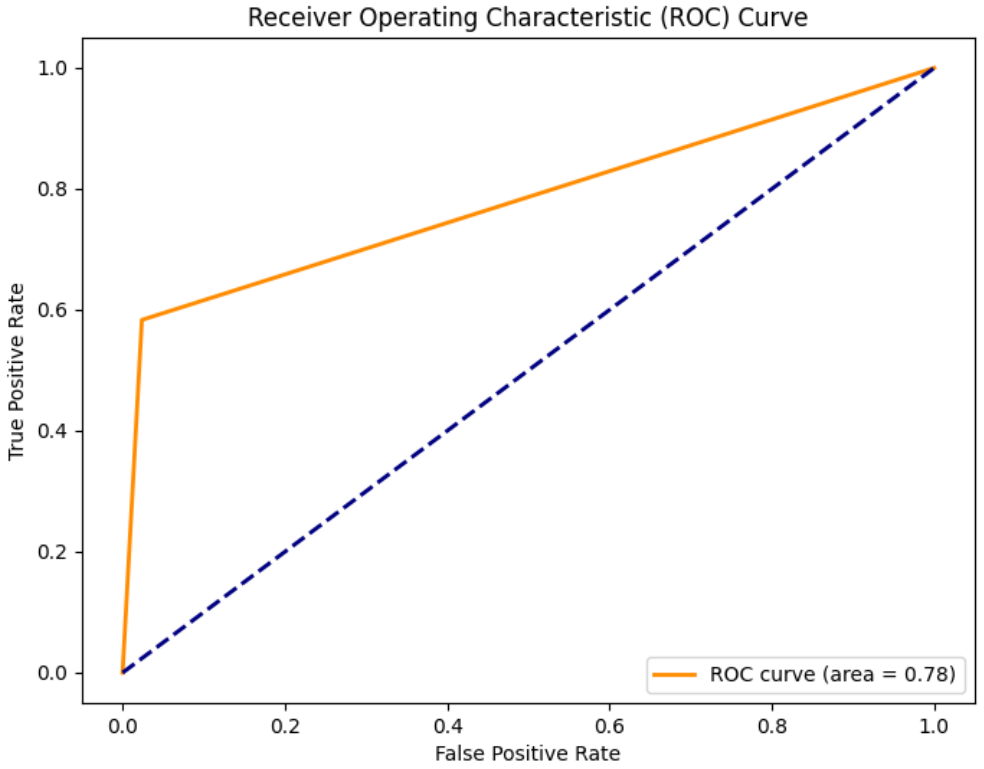
Model excels at identifying 'Fake' traffic signs:

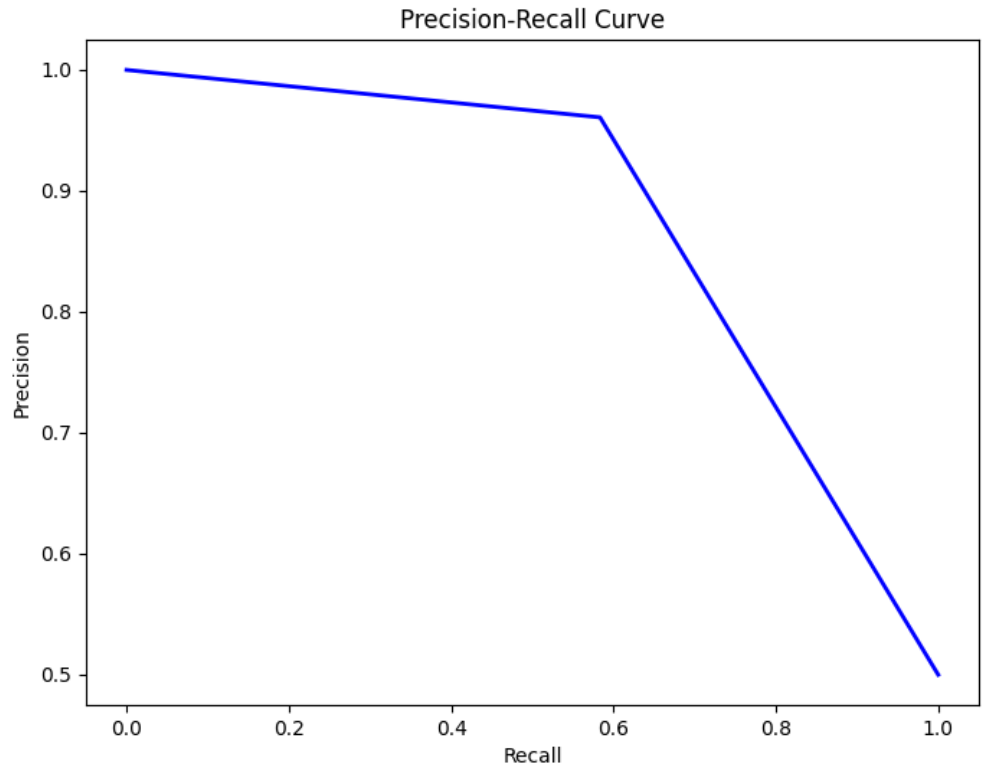
- **High True Negatives**
- **Low False Positives (only 2)**

- Prioritization of precision over recall, aligns with the **safety-first design** for autonomous vehicles.
- Low recall suggests room for improvement in identifying "Real" images to ensure robustness in real-world scenarios

Results Discussion

Traffic Sign Manipulation – Evaluating The Binary Classifier





Accuracy: 78%

Prioritizing high precision over recall

Results Discussion

Evasion Attack on Traffic Sign Classifier Using a GAN – Evaluating Adversarial Training

Metric	No Attack	Attack, No Adversarial Training	Attack, With Adversarial Training
Accuracy %	98.6	81.56	95.09
Weighted Avg Precision	0.98	0.82	0.95
Weighted Avg Recall	0.99	0.82	0.95
Weighted Avg F1-Score	0.99	0.81	0.95

- Evasion attacks drastically degrade model performance
- Adversarial training is an effective defense making AVs more resilient against adversarial examples

Conclusion

- Strengthens the **safety and reliability** of autonomous vehicles against adversarial attacks
- Introduces robust **detection mechanisms** to safeguard autonomous perception systems
- Demonstrates the potential of **sensor fusion** to enhance system resilience
- Demonstrates the utilization of **GANs** in creating defense mechanisms
- Highlights the broader applicability of the findings to other domains:
  - **Drones**
  - **Autonomous submarines**
  - **Medical imaging**
- Supports the development of trustworthy, secure technologies for high-stakes environments.
- A publication has already been submitted in **IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)**

Future Work

1. Extend the binary-classifiers to handle **multi-class classification** of the attack beyond "Real" and "Fake"

2. Optimize real-time performance with **compressed or greyscale image processing**

3. Explore **more advanced sensor fusion techniques** for deeper spatial and semantic insights

4. Validate detection mechanisms in real-world and dynamic simulated environments **(CARLA Simulator)**

5. Investigate locally hosting the system on the **embedded system** of a vehicle

6. Enhance the perfromance of the system by utilizing **GPUs**

7. We plan to submit a **Journal Publication** which will include some of the enhancements to our work

References

[1] Li, Y., Xu, X., Xiao, J., Li, S., & Shen, H. T. (2020). Adaptive square attack: Fooling autonomous cars with adversarial traffic signs. *IEEE Internet of Things Journal*, *8*(8), 6337-6347.

[2] Sitawarin, C., Bhagoji, A. N., Mosenia, A., Chiang, M., & Mittal, P. (2018). Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*.

[3] Tan, M., Zhuang, Z., Chen, S., Li, R., Jia, K., Wang, Q., & Li, Y. (2024). EPMF: Efficient Perception-Aware Multi-Sensor Fusion for 3D Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[4] upGrad. Basic CNN Architecture: A Beginner's Guide to Convolutional Neural Networks. Accessed: 2025-01-07. 2025. url: https : / / www . upgrad . com / blog / basic-cnn-architecture/.

NuScenes Dataset

GTSRB - German Traffic Sign Recognition Benchmark

Thank You!

Any Questions?