ICMI 2025
IEEE 4th International Conference on Computing and Machine Intelligence
April 05-06, 2025
Central Michigan University, USA

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

Farah Sherif, Zyad Mahrous, and Muhammad Hataba

GIU GERMAN INTERNATIONAL UNIVERSITY
الجامعة الألمانية الدولية

NTI National Telecommunication Institute

Outline

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

Introduction

- **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.**
- **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.

Contributions

- Development of an Adversarial Attack on AV Traffic Sign Recognition Systems
- Design of a Tailored Lightweight Binary Classifier for Adversarial Detection in Real-time.
- Safety-Driven Detection Approach to Reduce Critical Misclassifications

Selected Related Work

	Adaptive Square Attack [1]	Darts: Deceiving autonomous cars with toxic signs [2]	Building Robust Deep Neural Networks for Road Sign Detection [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	• Defends against adversarial attacks using an autoencoder with a memory module to retain clean image features. • Evaluated against Hijacking, Vanishing, Fabrication, and Mislabelling attacks
Comparison	• Evaluated on traffic sign images not dynamic driving scenes • Focuses solely on the attack	• Single sign misclassification not modifying traffic signs in full driving scenes. • Targets classification models with optical illusions and dataset shifts • Lacks detection mechanisms	• Focuses on feature reconstruction • Our approach is more lightweight for real-time deployment whereas autoencoders are computationally expensive

Threat Model

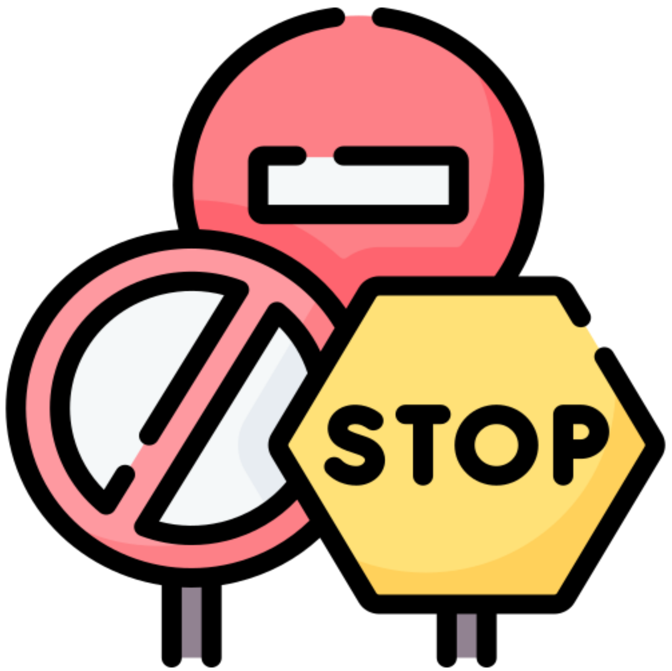
Threats:

- Adversarial Manipulation of Camera Feed
- Traffic Sign Misclassification
- Impacting Decision Making

Vulnerabilities:

- Lack of Input Integrity Checks
- Weak Robustness against adversarial perturbations

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

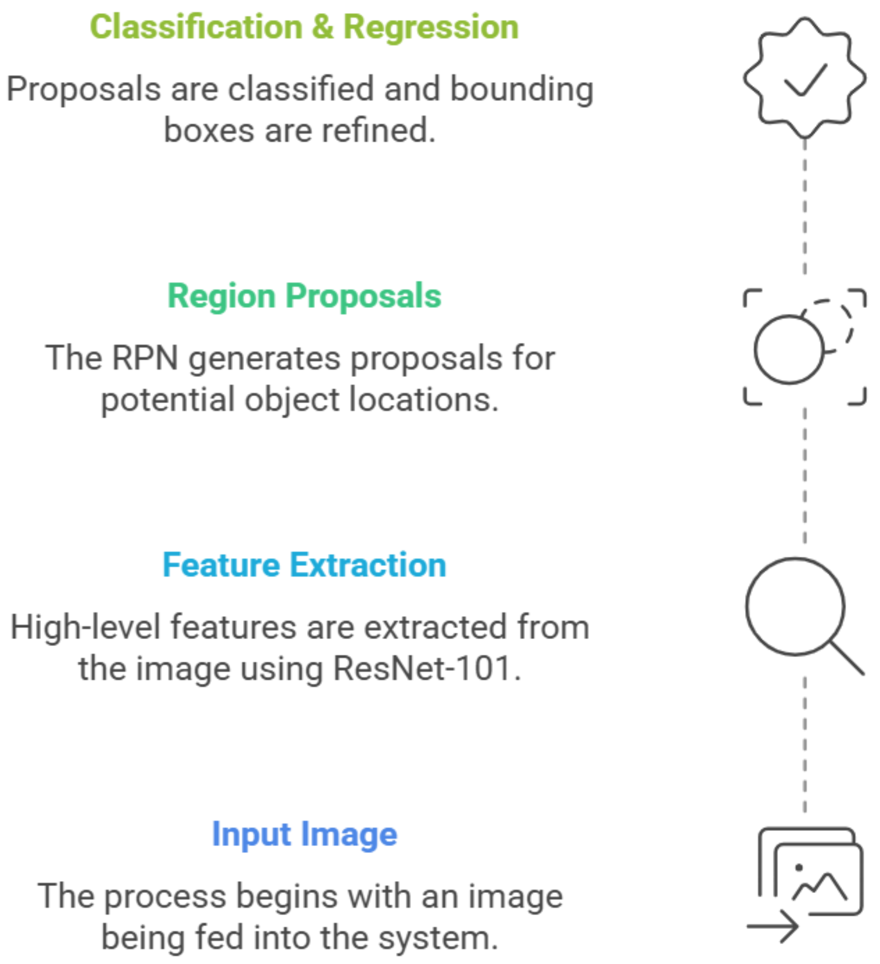


System Design

Traffic Sign Manipulation Attack

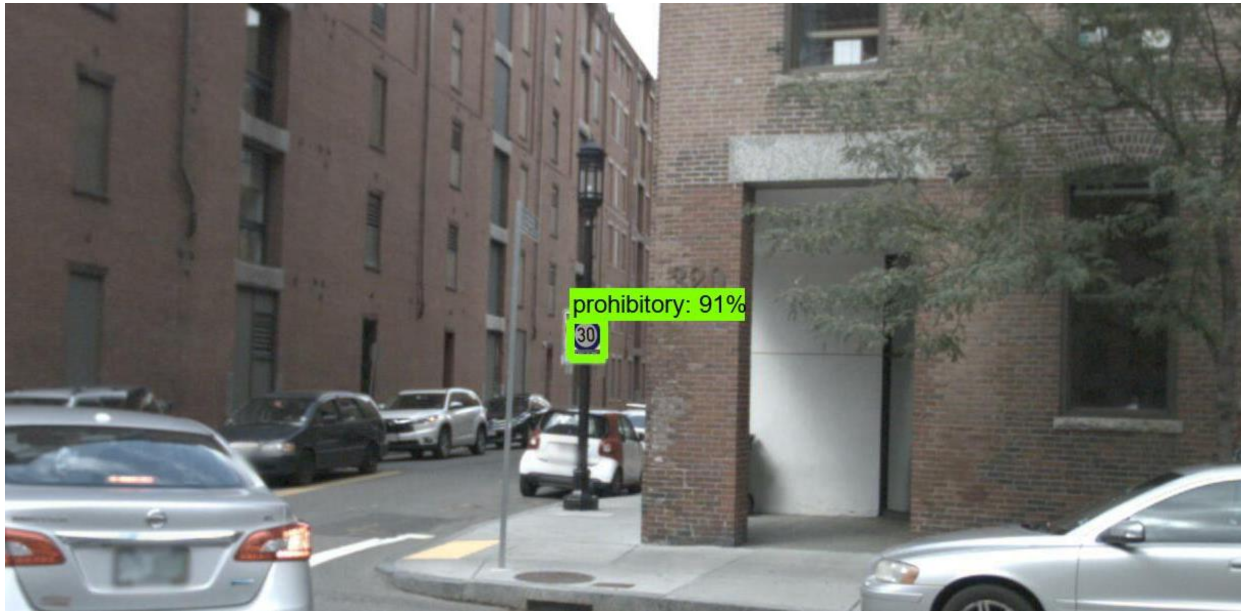
- Faster-ResNet-101 Model for traffic sign detection
- Each detection is replaced with a randomly chosen traffic sign from the GTSRB - German Traffic Sign Recognition Benchmark
- Leads to the misclassification of traffic signs or not detecting them at all
- **Serious side-effects**: Violating traffic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

Traffic Sign Detection Process in Faster R-CNN



System Design

Traffic Sign Manipulation Attack



Fake

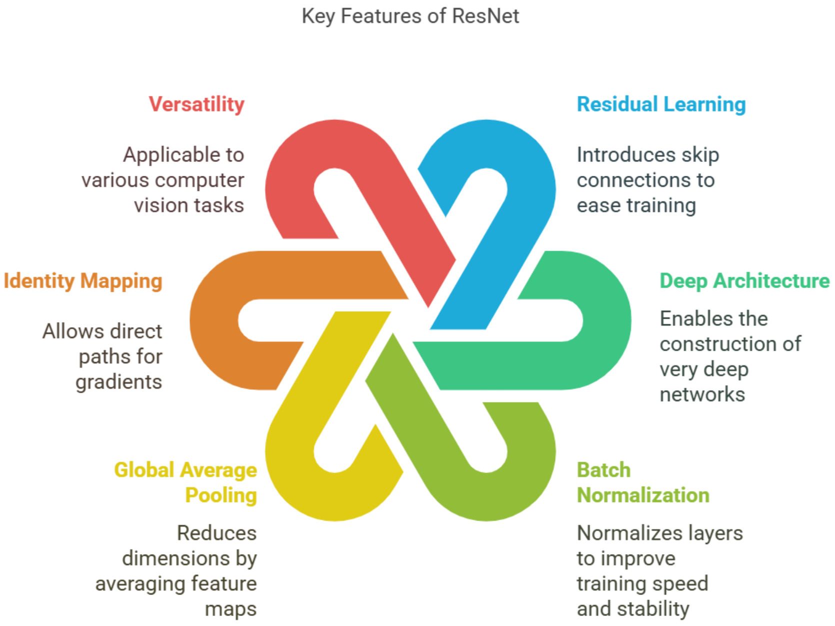
Real

System Design

Traffic Sign Manipulation Attack – Detection Mechanisms

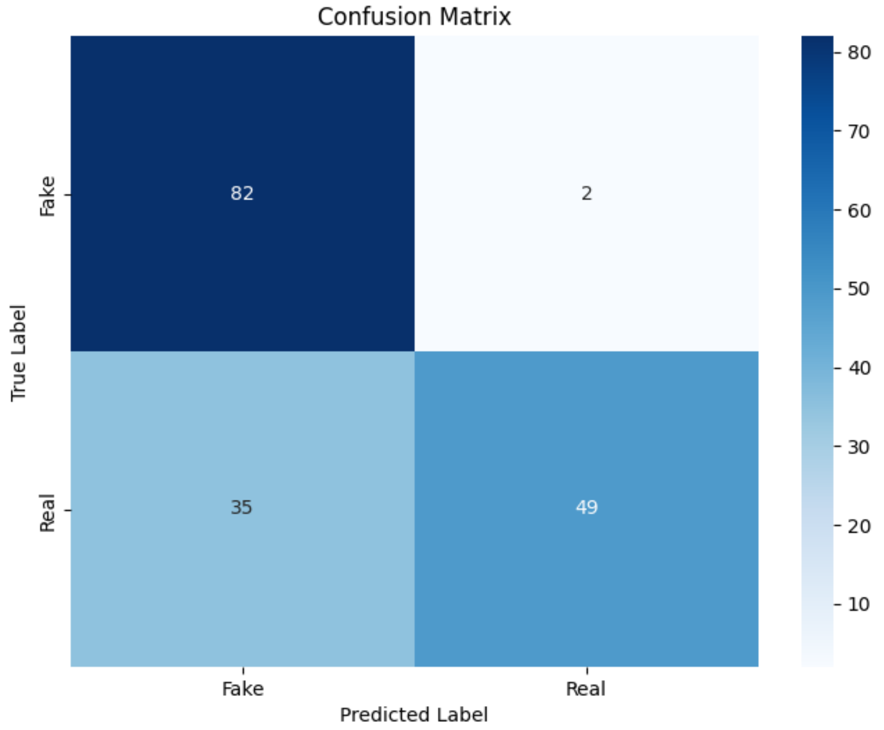
Binary Classifier

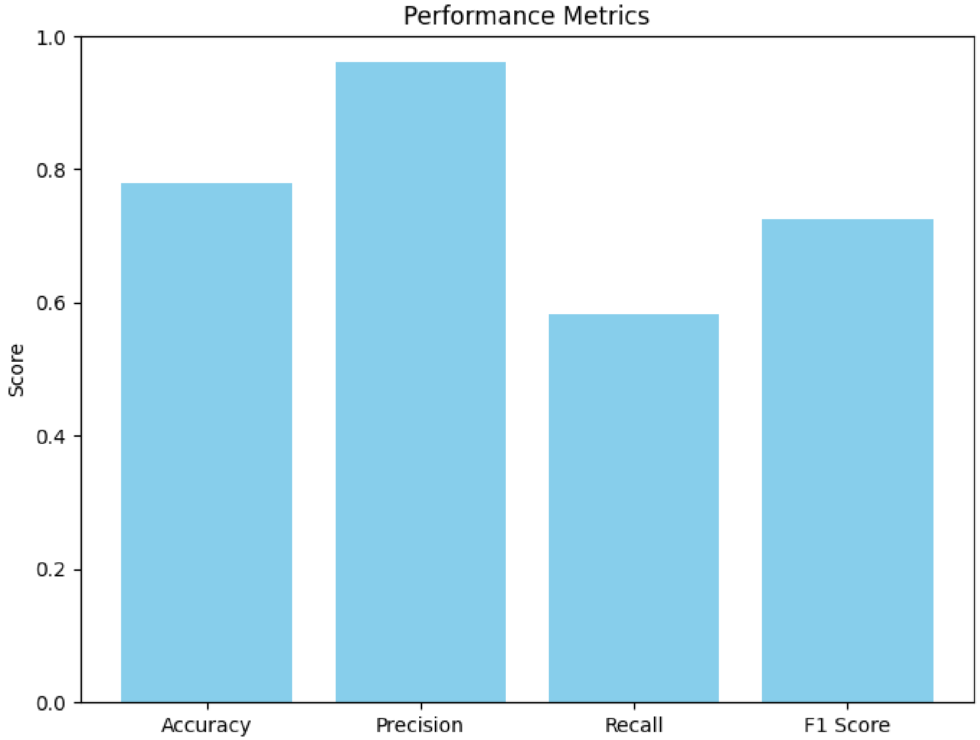
- CNN Model Using ResNet-18 model
- Images underwent transformations to ensure uniformity and compatibility
- Final layer replaced with a linear layer to adapt to the **binary** classification
- Sigmoid Activation Function for probability
- Adam Optimizer



Results Discussion

Traffic Sign Manipulation – Evaluating The Binary Classifier





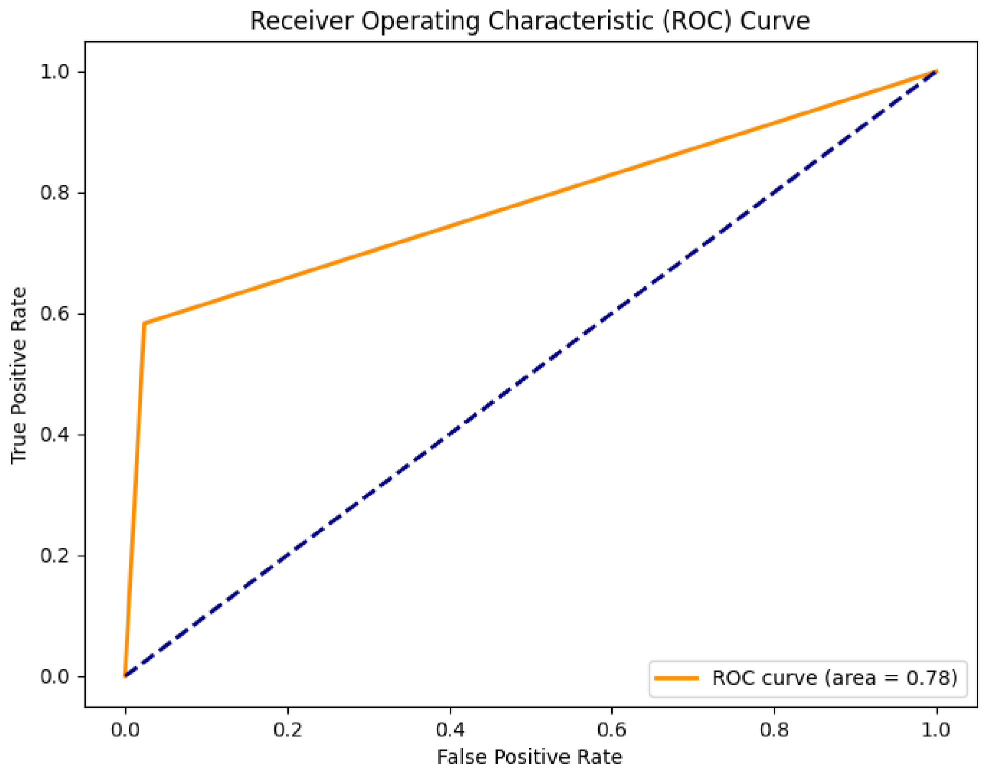
Model excels at identifying 'Fake' traffic signs:

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

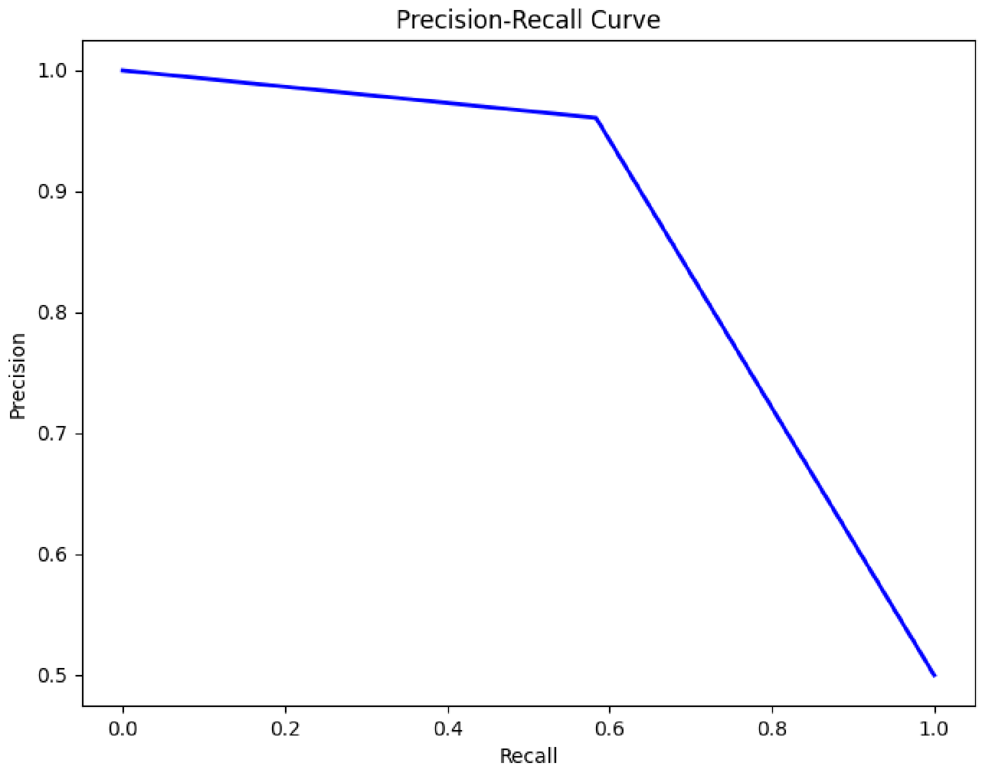
- Prioritization of precision over recall, aligns with the **safety-first design** for autonomous vehicles.
- Low recall suggests room for improvement in identifying "Real" images to ensure robustness in real-world scenarios

Results Discussion

Traffic Sign Manipulation – Evaluating The Binary Classifier



Accuracy: 78%



Prioritizing high precision over recall

Conclusion

- Strengthens the **safety and reliability** of autonomous vehicles against adversarial attacks
- Introduces robust **detection mechanism** to safeguard autonomous perception systems
- Designed and implemented an attack scenario demonstrating traffic sign misclassification.
- 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.

Future Work

1. Explore **Advanced Decision Mechanism** – Implement a safety-driven approach for uncertain traffic sign classifications

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

3. Explore **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 performance of the system by utilizing **GPUs.**

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] A. M. Aung, Y. Fadila, R. Gondokaryono, and L. Gonzalez, “Building robust deep neural networks for road sign detection,” arXiv preprint arXiv:1712.09327, 2017.

[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/.

[5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11 621–11 631, 2020. [Online]. Available: https://www.nuscenes.org/

[6] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “The German traffic sign recognition benchmark: a multi-class classification competition,” in The 2011 international joint conference on neural networks. IEEE, 2011, pp. 1453–1460.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, 2009.

Thank You!

Any Questions?