Group 13: Impact of Adversarial Attacks on Autonomous Car Object Detection

Haani Syed, Viren Tated, Anneth Sivakumar, Sidhardh Alluri

Abstract—As autonomous driving software systems increasingly rely on AI-based object detection models, ensuring their defense capabilities against adversarial attacks has become crucial. Strategically designed adversarial attacks can lead to major failures in detecting key objects in the driving environment, posing significant safety concerns. In this paper, we introduce a novel lightweight anomaly detection framework that combines Principal Component Analysis and Support Vector Machines to detect adversarial ghost attacks. We also investigate the Attack Success Rate and True Positive Rate of state-of-the-art object detection models (e.g., YOLOv7, Faster R-CNN, DETR) under adversarial conditions tailored to autonomous driving scenarios. Additionally, we propose an impactful algorithmic improvement to the Ghostbusters model [1] with rigorous analysis to demonstrate its effectiveness. These evaluations leverage a new benchmark dataset, IPRPAS, *Images and Point Clouds of Resembling Physical Adversarial Samples* [2] for testing. Models are trained on standard autonomous driving datasets such as KITTI and COCO. Our findings emphasize the need for resilient and safety-aware object detection strategies in autonomous driving vehicles.

Index Terms—Group 13, Autonomous Vehicle, Cybersecurity, Physical Adversarial Samples, Adversarial Attacks, Object Detection, Advanced Data Analytics

1 Introduction

As autonomous vehicles increasingly rely on object detection models (ODMs) to interpret their environment, ensuring the robustness of these systems against adversarial attacks has become critical. Among the most concerning threats are ghost attacks and phantom objects, which can cause ODMs to falsely detect key objects, such as road signs or pedestrians. These attacks compromise the safety and reliability of autonomous systems by manipulating input data in subtle but malicious ways.

A major concern is the sensitivity of object detection models to environmental conditions, particularly daylight lighting, which can enhance or suppress the visual artifacts introduced by adversarial perturbations. These effects can change the success rate of attacks, especially on small or shape-dependent objects like traffic signs—which play a critical role in autonomous navigation. Furthermore, current adversarial defense strategies are often based on deep learning, requiring large datasets, and high computational power.

This research addresses the challenge of detecting adversarial ghost and phantom objects using a more interpretable and efficient approach. This study uses 2D images from Queen's University's Reliable Software Technology Laboratory's IPRPAS Dataset to explore the effectiveness of different ODMs and propose a lightweight defense using PCA-SVM.

Main Contributions:

- Applied PCA and SVM to identify adversarial samples and compared results to deep learning models.
- Performed a comparative analysis of various ODMs to assess their vulnerability to ghost and phantom attacks in different lighting condtions.

This project investigates adversarial vulnerabilities in ODMs from three distinct perspectives:

1) **RQ1:** Can adversarial ghost attacks in 2D images be reliably detected using a lightweight anomaly detection framework based on classical machine learning techniques?
2) **RQ2:** Why do different object detection models exhibit varying True Positive Rates and Attack Success Rates when processing images containing phantom objects under daylight conditions?
3) **RQ3:** What improved methods can be used to detect phantom traffic signs more effectively?

2 Related Work

Describe what are the relevant research work here. For each work, briefly describe what the authors did and found.

The paper *Physical Hijacking Attacks against Object Trackers* explores critical vulnerabilities in object detection and tracking systems within the realm of autonomous vehicle security [3]. The authors introduce a novel class of physical hijacking attacks, wherein adversaries exploit real-world adversarial perturbations to manipulate tracking models, leading to object misidentification or loss of tracking.

This research is highly relevant to our project, as it provides a strong foundation in major object detection and tracking algorithms, including YOLO, R-CNN, heatmap generation, and various attack strategies. The discussion on testing models across simulated, test, and real-world datasets aligns closely with our objectives, particularly as we plan to utilize well-established datasets such as COCO

- *Member 1, 2, 3, and 4 are with School of Computing at Queen's University.*
E-mail: 21ahs7@queensu.ca, viren.tated@queensu.ca, 21as221@queensu.ca, 21sva3@queensu.ca

and the IPRPAS Queen's University Reliable Software Technology Laboratory dataset. Furthermore, the paper offers valuable insights into best practices for evaluating object detection performance under adversarial conditions, highlighting the vulnerabilities of existing models and refining our research questions aimed at developing more robust detection methods.

Similarly, the paper *That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using Spatiotemporal Consistency* is highly relevant to our team's work on adversarial ghost attacks and phantom objects in object detection systems [4]. Our focus on phantom object detection and data analytics directly aligns with the paper's exploration of spatiotemporal consistency as a means of detecting adversarial anomalies.

This study provides a comprehensive overview of current object detection techniques, strengthening our domain knowledge. The concept of spatiotemporal consistency presents a promising alternative to traditional methods like PCA-SVM, offering a potentially less computationally intensive approach to detecting phantom objects. Additionally, the paper underscores the importance of contextual validation, which is particularly relevant to our investigation of how varying lighting conditions impact object detection across different models.

Likewise, the paper *Pedestrian Detection at Daytime and Nighttime Conditions Based on YOLO-v5* explores a multispectral approach that integrates RGB and IR images to enhance pedestrian detection in varying lighting conditions. It addresses the limitations of traditional detection methods, particularly in nighttime scenarios, by leveraging YOLO-v5's architecture to improve accuracy. The study provides a rigorous evaluation of YOLO-v5's performance across multiple datasets, using key metrics such as precision, recall, and mAP, establishing benchmarks for pedestrian detection under challenging conditions.

This research is directly relevant to our study, as it offers insights into YOLO's performance in adverse environments, which is crucial for understanding its susceptibility to adversarial attacks. The discussion on multi-spectral data fusion highlights potential strategies for improving detection robustness—an aspect that aligns with our goal of evaluating attack success rates on YOLO-based systems. Additionally, the study's focus on performance variability under different lighting conditions provides valuable context for assessing the resilience of object detection models against adversarial perturbations.

The paper *Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks* addresses vulnerabilities in advanced driver-assistance systems (ADAS) called "split-second phantom attacks.". These manipulate the system by projecting near instantaneous (just a few milliseconds) "phantoms" which simulate obstacles that ADAS-vehicles are meant to avoid, whether by projecting real holograms or sending them as attacks to the system. These are depthless but appear for so short a time that they cause dangerous behavior in ADAS-vehicles.

Our project involves identifying the attacks and deploying strategies against said attacks, and this study provides the entire theoretical basis for the methods we deploy and our analysis of our dataset, and provides an eminently useful benchmark for measuring performance of R-CNNs in these use cases against competing algorithms. In our third research question our intention is also to directly build off of Ghostbusters to improve performance against attacks involving signs, the theoretical basis for which underpins much of the testing with the multiple R-CNNs.

3 Methodology

3.1 RQ1: Can adversarial ghost attacks in 2D images be reliably detected using a lightweight anomaly detection framework based on classical machine learning techniques system?

This study employs a step-by-step approach involving data preprocessing, feature extraction, dimensionality reduction, model selection, and post-processing. The methodology combines classical machine learning techniques—Principal Component Analysis (PCA) and One-Class Support Vector Machine (OCSVM)—to develop a lightweight, interpretable framework for adversarial detection.

To begin, the input data consists of 2D images from the IPRPAS dataset must be preprocessed. Each image is standardized and normalized. A sliding window partitions the image into overlapping 32×32 pixel patches, with a 16-pixel step size to ensure full spatial coverage. This ensures comprehensive spatial coverage while maintaining computational efficiency.

Following data segmentation, feature extraction is performed on each patch to capture visual and structural characteristics essential for distinguishing adversarial patterns. Two types of features are extracted: color histograms and Histogram of Oriented Gradients (HOG). The color histogram is computed using the Blue-Green-Red (BGR) color channels and encodes the distribution of color intensities, which is useful for identifying abnormal textural patterns introduced by adversarial attacks. Simultaneously, HOG features are extracted from the grayscale version of each patch to emphasize edge orientations and local gradient structures, which are sensitive to subtle perturbations. These two sets of features are concatenated into a single feature vector representing each image patch.

Due to high dimensionality of the combined feature vectors, PCA reduces the feature space by projecting data onto principal components that retain the most variance. This reduces noise and enhances feature separability while improving computational efficiency.

The transformed features are then used to train the OCSVM. After considering alternatives like Isolation Forests and Autoencoders, OCSVM was selected for its effectiveness in one-class anomaly detection. The model employs a Radial Basis Function (RBF) kernel to capture complex relationships in high-dimensional feature spaces. The nu parameter determines the proportion of predicted outliers, whereas the gamma parameter adjusts the kernel's flexibility. These hyper-parameters are optimized to maximize performance.

After classification, post-processing techniques are applied to refine the predictions. The anomaly predictions are first transformed into a binary mask, which is then refined using morphological opening and connected component to remove noise and isolate valid detections. Bounding boxes

are then drawn around the remaining irregular regions to localize the detected ghost attacks within the image.

To evaluate the performance of the PCA-OCSVM model, each image is visually inspected to determine whether adversarial regions were correctly identified. If the detection is accurate, the image will be marked as correctly classified, and the process will move on to the next image. This step-by-step methodology, from feature extraction and PCA transformation to OCSVM-based classification and post-processing, provides a clear, interpretable, and computationally efficient framework for detecting adversarial ghost attacks in 2D images.

For comparison, the YOLOv7 (You Only Look Once) object detection model was selected for testing. As a deep learning benchmark, YOLOv7 is able to perform object detection by predicting bounding boxes and class probabilities in a single pass of images. Its ability to operate in real time while maintaining high detection accuracy makes it an appropriate benchmark for evaluating the performance of computationally lighter models such as PCA-OCSVM. Additionally, the same images are evaluated using YOLOv7, focusing on the success rate of detecting ghost objects.

Through this comparative analysis, it assesses the trade-offs between deep learning-based and classical feature-based detection strategies. If YOLOv7 outperforms PCA-OCSVM, it suggests that the complexity and computational demands of deep learning models may offer robustness benefits that outweigh their costs. Conversely, if PCA-OCSVM performs comparably or better, it underscores its potential as a lightweight, viable alternative in environments where computational resources are constrained.

3.2 RQ2: Why do different object detection models exhibit varying True Positive Rates and Attack Success Rates when processing images containing phantom objects under daylight conditions?

This experiment evaluated 2D object detection performance under *daytime* conditions—including both *cloudy* and *sunny* scenarios—using two vision-based models: **Faster R-CNN** and **DEtection TRansformer (DETR)**. Both models were pre-trained on the COCO dataset, a large-scale benchmark containing over 118,000 annotated images across 80 object categories, designed to promote generalization across diverse scenes.

For evaluation, we used 1,639 `.png` images from the `Left_images/` folder of the IPRPAS dataset, which contains sequentially numbered frames (e.g., `000001.png`) captured from a vehicle-mounted left camera. Corresponding ground-truth labels were extracted from the `labels/Label` directory. Each label file (e.g., `000001.txt`) includes bounding box coordinates in absolute pixel format $(x_{min}, y_{min}, x_{max}, y_{max})$ along with an authenticity tag: `R` for real or `F` for phantom.

All experiments were conducted on the Queen's School of Computing Lobot Cluster, using dual RTX A5000 (24GB) GPUs, 256GB RAM, and a 32-core AMD EPYC 2.6/3.3GHz CPU configuration for accelerated inference.

3.2.1 Unified Evaluation Framework

Both DETR and Faster R-CNN were assessed using a consistent evaluation pipeline comprising four core stages:

1) **Image Preprocessing**:
   - All input images were resized to a fixed height of 1000 pixels and converted to tensors suitable for model input.
   - Inference was conducted using GPU acceleration with models running in parallel where supported.
2) **Ground-Truth Parsing**:
   - Bounding boxes and authenticity labels were parsed from the corresponding annotation files.
   - Images with missing labels, malformed entries, or invalid indices were excluded from the evaluation set.
3) **Bounding Box Postprocessing and Matching**:
   - Model predictions were filtered using a confidence threshold of 0.3.
   - For DETR, outputs in normalized format $[x_{center}, y_{center}, w, h]$ were converted to absolute pixel coordinates.
   - Predicted boxes were matched to ground-truth annotations using an Intersection over Union (IoU) threshold of 0.1.
   - A one-to-one matching policy was enforced, ensuring that each ground-truth object could be matched to at most one predicted box.
4) **Metric Computation**:
   - True Positives (TP) and False Negatives (FN) were computed separately for real (`R`) and phantom (`F`) objects.
   - The total number of phantom ground-truth instances was fixed at 3,409 across all evaluations.

3.2.2 Metric Definitions

We used two primary metrics to assess detection performance and vulnerability:

- **True Positive Rate (TPR)** for real objects:

$$TPR_{real} = \frac{TP_{real}}{TP_{real} + FN_{real}}$$

- **Attack Success Rate (ASR)** for phantom objects:

$$ASR = \frac{FN_{fake}}{3409}$$

  This metric captures the proportion of phantom objects that successfully evaded detection.

3.2.3 Model-Specific Implementation Details

Faster R-CNN:

- We used the `fasterrcnn_resnet50_fpn` model from the torchvision library.
- Predictions were extracted directly in absolute pixel format, simplifying post-processing.
- Inference was parallelized across available GPUs.

DETR:

- The `detr_resnet50` model was loaded from the official `facebookresearch/detr` repository.
- A softmax was applied to model logits to compute class probabilities before applying the confidence threshold.
- Bounding boxes were output in normalized format and transformed to match the absolute format of the ground-truth annotations.

3.2.4 Summary

This unified methodology enabled a direct comparison between DETR and Faster R-CNN under identical conditions. By analyzing both TPR and ASR across real and phantom objects, we quantified each model's ability to detect real objects and its susceptibility to phantom objects under real-world daytime scenarios.

3.3 RQ3: What improved methods can be used to detect phantom traffic signs?

In this experiment we aim to utilize the vision-based "Council of Experts"**GhostBusters** model as defined in the landmark paper *"Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks".* [1] on *Phantom Traffic Sign* detection to achieve improved object detection in this class of *Phantom Objects*.

The GhostBusters model needs to first receive a set of real footage from dashcam driving footage containing signs to be able to extract a set of real traffic signs and a set of images from which the signs were removed to demonstrate *good placement*. Thereafter, GhostBusters requires a set of *augmented* driving footage containing Phantom Traffic Signs from which it extracts this set of "Phantom Objects. ".

The dataset that was originally produced to use in the paper is no longer available, so here the **KITTI-STEP** dataset is used. KITTI-STEP works on the KITTI-MOTS dataset to produce dense pixel-wise segmentation labels for every pixel as *Panoptic Maps* stored as images. These allow us to augment a subset of KITTI-STEP (the rest being used for the *real* subset) by heuristically tracking the segments using bounding boxes(since KITTI-STEP only provides unique tracking IDs at for "salient objects" such as cars and pedestrians) tracking fixed positions such as traffic pole objects on which a Phantom Traffic Sign might be imposed.

They are imposed on a fixed object (fixed in real-space) over a burst of successive frames (3-5) to support training of the *optical flow* of GhostBuster's model, implemented within the *Depth* expert. For each augmented frame the imposed sign is matched with the nits of the surrounding scene which in turn assists with the training of the *Light* expert. Each traffic sign sample chosen for insertion can be alpha-blended which in turn assists to train the *Context* expert.

For generating a workable and trainable augmented dataset, the parameters of GhostBuster's dataset as described in "Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks" [1] are followed, though in the absence of their dataset with augmentations, a series of heuristics for injection of Phantom Traffic Signs had to be developed. We leverage the Panoptic Maps and start by processing the average resolution of the traffic sign segments in each Panoptic Map annotation. This sets optional bounds for sizing the injected images, which can be optionally called in the function call. Otherwise we've developed settings for maximum/minimum brightness, which avoids the annotation from making the insertions overly bright (and thereby obvious to the algorithm). We also have a function final scale, which processes more broadly to scale the image based on estimated depth from the previously calculated average of all bounding-boxes and then the height of the given object, which gives us a rough idea of how to scale the image without explicitly processing depth. We then cap this by selectively choosing maximum scale factors, which can effectively control the upper limit of how much images scale based on the heuristic we use to calculate loss. This in turn produces a good dataset for training detection of Phantom Images, along the lines of the directory that isn't present.

To feed the augmented data to the model without losing the Depth Model (requiring video data) or losing fidelity for training through lossy video processing, we employ the lossless*FFV1 codec*. This allows us to still extract and train on the optical flow from image to image without losing frames and pixel counts per frame and thus underfitting, like we would get with the standard *mp4v codec*.

A pretrained RCNN *faster_rcnn_inception_resnet_v2_atrous* is leveraged for sign detection. GhostBuster's script then uses this to extract some feature x^t which is then scaled to 128×128. For the Context Model ("expert") the Context is created by extracting the feature x^t, scaling to 128×128 but then setting the center x^t to 0 so as to get the "Context". For the Surface Model the CNN gets the full color feature x^t at 128×128 and trains on the "texture" so to speak (consistency of the Surface). The Light Model gets the same feature x^t but the features at the Light Model are given by getting the maximum RGB value at each of the feature's pixels. This is given by

$$x[i,j] = argmax \cdot x^t[i,j,k]$$

This model checks the relative irregularity of each sign. The Depth Model takes the feature at x^t and then computes optical flow between x^t and x^{t-1}. This is given by the Gunner Farneback algorithm which gets some 2D field v and converts it to HSV by computing Vector/Magnitude of each feature. This HSV is subsequently reconverted into an RGB image.

Once the data is populated in each by the pretrained RCNN, it is then trained on each "Expert Model" as described above. The "Council of Experts" idea for GhostBusters is implemented in the fifth *Combiner* model which is trained alongside the other four. Whenever the other models pass a predication the activation of this fifth "neuron" is captured and represents the cumulative embedding of the each model's reasoning. The fifth Combiner Model passes the final prediction on each model.

In addition to our novel data-set creation method, which is an extension of the data-set created for GhostBusters, we adjust the decay method given in the initial implementation to fit with newer libraries. Where before in the GhostBuster 2020's implementation decay was adapted by stating a simple

$$lr = 1e - 3$$

New tensorflow.keras libraries instead leverage a distinct decay function for learning which we should be able to use.

4 Dataset

This study utilizes the IPRPAS dataset, "*IPRPAS: A Dataset of Resembling Physical Adversarial Samples to Assess Object Detection in Intelligent Vehicles*", to evaluate how different object detection models perform in the presence of adversarial ghost and phantom objects [2]. Developed by the Queen's University Reliable Software Technology Laboratory, the IPRPAS dataset is explicitly designed to assess the robustness of perception systems under physical adversarial attack scenarios. Access to this dataset is regulated by Professor Mohammad Zulkernine, Principal Investigator of Queen's University Reliable Software Technology Laboratory, Director of Queen's Centre for Security and Privacy, and Tier 1 Canada Research Chair in Cyber-Physical System Security.

The dataset comprises 3,409 annotated 2D RGB images captured in both indoor and outdoor environments. Scenes include common driving elements such as cars, pedestrians, and traffic signs, with adversarial perturbations introduced through printed or projected phantom objects. Images are collected under varying lighting conditions (day and night), camera angles, positions, and scene complexity, providing a realistic benchmark for evaluating detection robustness under adversarial influence. As described by the authors: "IPRPAS includes various common driving objects such as cars, traffic signs, traffic lights, and persons under different weather conditions, positions, distances, lighting conditions, and viewing angles. The dataset contains images, 3D point clouds in multiple formats (.pcd, .bin, .ply), object labels, and calibration files. (2025-01-16)" [2].

The IPRPAS dataset directly supports the three central research questions of this study. For RQ1, it enables the development of a lightweight PCA-OCSVM anomaly detection framework based on patch-wise feature extraction. Images from the Left_Images/ directory were resized to a fixed resolution and divided using a sliding window (32×32, stride of 16). Each patch underwent grayscale conversion, histogram equalization, and feature extraction using Histogram of Oriented Gradients (HOG) and color histograms.

For RQ2, the images of Left_Images folder (daytime left camera images) were used to test two object detection models, both of which were pretrained on the COCO dataset. Each model was assessed using bounding box inference and Intersection over Union (IoU) matching. Ground-truth annotations were extracted from corresponding .txt label files using the authenticity columns. Predicted bounding boxes were converted to absolute coordinates, an IOU threshold of 0.1 was used to match detections against ground truth. To enhance detection, the confidence threshold was set to 0.3.

For RQ3, KITT-STEP dataset was used for training since KITTI-STEP has panoptic-map annotations which track the positions of objects in a dashcam frame using encoded positions in the map which are stored as PNG files [5].A subset of these images were augmented with "phantom" traffic signs so that it can be used to train the ghostbusters model used in this question.

The IPRPAS dataset was released on January 28, 2025, and is publicly available through the Queen's University Dataverse. Its adversarial diversity, structured annotations, and realistic environmental variability make it well suited for benchmarking both classical and deep learning-based object detection models in the context of adversarial threat assessment.

5 Experiments and Results

5.1 RQ1: Can adversarial ghost attacks in 2D images be reliably detected using a lightweight anomaly detection framework based on classical machine learning techniques system?

The PCA-OCSVM model was evaluated on the IPRPAS dataset, which includes images captured during both daytime and nighttime conditions, with adversarial objects presented in printed and projected formats. The goal was to assess the model's effectiveness in detecting these objects across environmental settings and attack modalities.

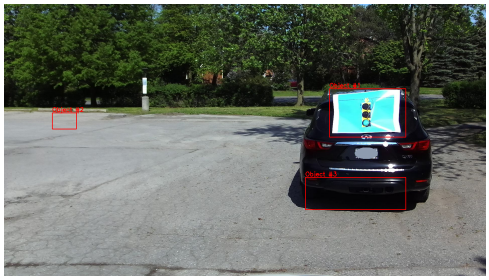

Fig. 1. PCA-OCSVM successful detection of adversarial object in a daytime scene.

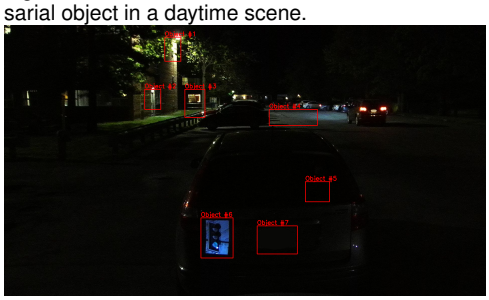

Fig. 2. PCA-OCSVM successful detection of adversarial object in nighttime scene.

TABLE 1
Table 1. Detection Accuracy (%) of PCA-OCSVM Model Across Lighting Conditions

Lighting Condition	Day(%)	Night(%)	Average(%)
Detection Accuracy	85.90%	24.28%	62.83%

The model performed well in detecting printed adversarial objects during daytime, achieving 85.90% accuracy, as shown in Table 1. This suggests that the pipeline effectively identified structural perturbations. The bounding boxes were accurately localized around the printed objects,

further supporting its detection capability. However, some false positives were observed, often caused by shadows, reflections, and dark textures, pointing to possible improvements in feature extraction or filtering.

In contrast, detection performance significantly declined during nighttime conditions, particularly for projected adversarial objects. Under these low-light scenarios, the model achieved an accuracy of 24.28%, highlighting challenges with adversarial projection detection when structural edges are weak or when projections blend into complex or textured backgrounds. Also, the model frequently misclassified other light sources such as vehicle headlights, signage, and reflective surfaces as adversarial. While it could detect clearly outlined projections in well-lit areas, the false positive rate increased in complex backgrounds.

Overall, the model achieved 62.83% average accuracy. It is most reliable in well-lit, high-contrast scenes but less effective in low-light or dynamic settings. Its reliance on intensity gradients and structural features made it susceptible to misclassifying license plates, censor bars, and shadows as adversarial. While it shows promise for certain controlled settings, further refinements are necessary to improve robustness and reduce false positives in real-world applications. Example outputs of successful detections in both lighting conditions are shown in Figures 1 and 2, further supporting the quantitative findings discussed above.

To provide a meaningful benchmark of the proposed PCA-OCSVM model, the pre-trained YOLOv7 object detection model was tested on the same dataset. YOLOv7 was evaluated under lighting variations and detection thresholds of 0.5 and 0.7.

In this context, an adversarial ghost attack was considered successful when YOLOv7 misclassified a fake object as a legitimate object. The primary metric used for this evaluation is the Attack Success Rate (ASR), the proportion of adversarial samples that were misclassified as legitimate objects.

TABLE 2
Table 2. Attack Success Rate (ASR) of YOLOv7 Under Varying Confidence Thresholds and Lighting Conditions

Lighting Condition	Confidence Rate	ASR
Day	0.5	33.12%
Day	0.7	12.39%
Night	0.5	36.07%
Night	0.7	26.49%

The results summarized in Table 2, reveal key trends. At a lower confidence threshold of 0.5, YOLOv7 was more susceptible to adversarial attacks in both lighting conditions. The ASR was 33.12% during daytime and increased to 36.07% at night, suggesting that lower detection thresholds make the model more permissive to adversarial attacks. When the confidence threshold was raised to 0.7, the ASR dropped substantially in both scenarios, reaching 12.39% during the day and 26.49% at night. These results demonstrate that higher confidence settings significantly improve model robustness by filtering out low-certainty detections that often include adversarial artifacts.

Furthermore, a consistent pattern emerges across both models: adversarial attacks tend to be more effective under nighttime conditions, where reduced lighting, contrast, and image clarity may obscure structural differences between real and fake objects. This reflects a general vulnerability in object detection systems, where environmental complexity and visual ambiguity can diminish detection accuracy and increase susceptibility to adversarial perturbations.

Fig. 3. YOLOv7 successful detection of adversarial object in a daytime scene.

Fig. 4. YOLOv7 successful detection of adversarial object in nighttime scene.

Comparing the ASR of the YOLOv7 model to the ASR (100-Detection Rate) of the PCA-OSVM model, it is quite clear that there is a difference in performance. Overall, the YOLOv7 model demonstrated a stronger resilience to adversarial ghost objects than the PCA-OCSVM model, particularly at higher confidence thresholds. However, the fact that it remains vulnerable, especially at night and under permissive detection settings, suggests that even high-performing deep learning models are not immune to adversarial manipulation. Example outputs of successful detections in both lighting conditions are shown in Figures 3 and 4, further supporting the quantitative findings discussed above.

5.2 RQ2: Why do different object detection models exhibit varying True Positive Rates and Attack Success Rates when processing images containing phantom objects under daylight conditions?

TABLE 3
Comparison of Faster R-CNN and DETR on Real and Phantom Object Detection

Model	CS	IoU	Transform	TPR (Real)	ASR
Faster R-CNN	0.3	0.1	`Resize(1000)`	99.88%	22.00%
DETR	0.3	0.1	`Resize(1000)`	100.00%	4.87%

The experiments were conducted using the IPRPAS dataset's Left_images1 folder (2D images taken in daylight conditions by left camera) for testing and both models were pre-trained on the COCO dataset.

Under consistent conditions (Confidence Score threshold = 0.3, IoU threshold = 0.1), and with all images resized to 1000 resolution, both models demonstrated strong abilities in detecting real objects. The DETR Model achieved a True Positive Rate (TPR) of 100% , outperforming Faster R-CNN, which had a TPR of 99.88%. These determined TPR values reflect the model's high accuracy in identifying real objects in standard, well-lit driving environments.

Faster R-CNN exhibited a significantly higher Attack Success Rate (ASR) of 22.00%, indicating that it is more prone to false positives. In contrast, DETR achieved a much lower ASR of 4.87%, demonstrating better robustness against adversarial attacks.

This suggests that architectural choices in 2D Vision-Based Object Detection Models (ODMs) play a critical role in determining their robustness to adversarial attacks. DETR with its transformer-based architecture and global attention mechanism demonstrates superior performance in daytime autonomous driving images by utilizing enhanced contextual reasoning and maintaining resilience against phantom object perturbations. In contrast, Faster R-CNN, despite achieving high detection accuracy, remains more vulnerable to adversarial interference, particularly in daylight scenarios. The fundamental difference, DETR's global context modeling versus Faster R-CNN's localized region proposals, signifies how the design of an ODM affects its resilience capabilities.

5.3 RQ3: What improved methods can be used to detect phantom traffic signs?

To address RQ3, we attempted to build upon the GhostBusters framework introduced in *Phantom of the ADAS* [1], which proposes a "Council of Experts" model combining multiple domain-specific neural networks (e.g., Context, Surface, Light, and Depth experts) to detect phantom traffic signs.

Due to the unavailability of the original dataset used in GhostBusters, we created an augmented version of the KITTI-STEP dataset. This involved inserting phantom traffic signs into real driving scenes while preserving spatial and brightness consistency through heuristic-based control over scale, position, and lighting. These augmented frames were processed using the lossless FFV1 codec to maintain optical flow fidelity required for training the GhostBusters model.

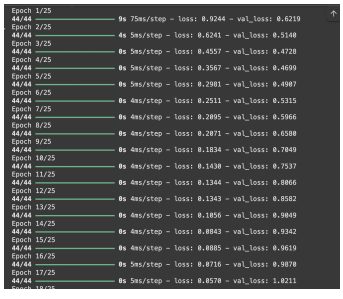

Fig. 5. Train/Validation Results for GhostBusters

Each "expert" model was trained individually using features extracted from a pretrained Faster R-CNN Inception-ResNet-v2 detector. Features were scaled and processed according to each expert's specifications:

- The **Context Model** masked out the central region of the feature map to detect contextual inconsistencies.
- The **Surface Model** trained on full RGB features to analyze textural consistency.
- The **Light Model** utilized pixel-level maximum RGB values to detect lighting anomalies.
- The **Depth Model** computed optical flow between consecutive frames using the Farneback algorithm and encoded the result in HSV space.

The final predictions were made by the **Combiner Model**, which aggregated outputs from all four experts. Despite these efforts, the model exhibited signs of extreme overfitting during training, especially when evaluated on unaugmented data. This was likely due to limited diversity in phantom sign placements and a lack of real-world adversarial samples, which constrained the generalization capabilities of the model.

Takeaways: While our dataset augmentation pipeline was successful in recreating the training flow for GhostBusters, the lack of sufficient and varied adversarial training samples led to overfitting. This underscores the importance of high-quality, diverse data when training multi-expert systems for adversarial detection. Future work will focus on expanding the augmented dataset and incorporating regularization techniques to improve generalization.

6 Group Member Contributios

Anneth was responsible for developing the PCA-OCSVM model and wrote the methodology, results, and takeaway sections for Research Question 1. Viren tested the YOLOv7 model on the dataset and contributed to the methodology, takeaway message, and citations for Research Question 1. He also managed the overall formatting of the report. Haani oversaw Research Question 2, implementing the relevant code and composing the corresponding methodology, results, and takeaway sections. Sidhardh led the work on Research Question 3, including code development and the writing of the methodology, results, and takeaway subsections. All group members collaboratively contributed to the Related Work section. The remaining sections were adapted from a previous report.

7 Replication Package

The experiment can be replicated by following the instructions in the replication package linked below. https://github.com/haanisyed/CISC351-Group-13

8 Conclusion and Future Work

This project explored the vulnerabilities of object detection models (ODMs) in autonomous vehicles to adversarial ghost and phantom attacks. We proposed a lightweight anomaly detection pipeline using Principal Component Analysis (PCA) and One-Class Support Vector Machines (OCSVM), which demonstrated strong performance in daytime settings but struggled under low-light conditions. This highlighted the difficulty of detecting adversarial perturbations in visually ambiguous environments.

A comparative analysis of Faster R-CNN and DETR models under adversarial settings revealed that architectural choices significantly impact model robustness. DETR, with its transformer-based global context modeling, outperformed Faster R-CNN in both detection accuracy and resistance to phantom object attacks, especially under daylight conditions.

For Research Question 3, while the GhostBusters framework for phantom traffic sign detection was partially implemented, the model encountered overfitting when trained on our custom-augmented KITTI-STEP dataset. This suggests a need for broader data diversity and further optimization of the training process.

Future work will focus on the following areas:

- Enhancing nighttime detection by incorporating multispectral data or spatiotemporal consistency.
- Reducing false positives through improved feature extraction and refinement of anomaly classification.
- Improving the GhostBusters pipeline to mitigate overfitting by exploring regularization strategies and dataset augmentation techniques.
- Conducting real-time inference tests to evaluate model performance under deployment conditions.

Together, these improvements aim to support the development of safer and more resilient object detection frameworks in autonomous driving systems.

9 References

[1] B. Nassi, Y. Mirsky, D. Nassi, R. Ben-Netanel, O. Drokin, and Y. Elovici, "Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, (New York, NY, USA), pp. 293–308, Association for Computing Machinery, 2020. event-place: Virtual Event, USA.

[2] M. Safarzadehvahed, M. Zulkernine, P. R. Marques de Araujo, and S. Givigi, "IPRPAS: A Dataset of Resembling Physical Adversarial Samples to Assess Object Detection by Intelligent Vehicles," 2025. Section: 2025-01-28 09:24:32.92.

[3] R. Muller, Y. Man, Z. B. Celik, R. Gerdes, and M. Li, "Physical Hijacking Attacks against Object Trackers," in *ACM Conference on Computer and Communications Security (CCS)*, pp. 1–13, 2022.

[4] Y. Man, R. Muller, M. Li, Z. B. Celik, and R. Gerdes, "That Person Moves Like A Car: Misclassification Attack Detection for Autonomous Systems Using Spatiotemporal Consistency," in *32nd USENIX Security Symposium (USENIX Security 23)*, (Anaheim, CA), pp. 6929–6946, USENIX Association, Aug. 2023.

[5] M. Weber, J. Xie, M. Collins, Y. Zhu, P. Voigtlaender, H. Adam, B. Green, A. Geiger, B. Leibe, D. Cremers, A. Osep, L. Leal-Taixe, and L.-C. Chen, "STEP: Segmenting and Tracking Every Pixel," in *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.