On the Assessment of Sensitivity of Autonomous Vehicle Perception

Apostol Vassilev, Munawar Hasan, Edward Griffor, Honglan Jin, Pavel Piliptchak, Mahima Arora
National Institute of Standards and Technology
Gaithersburg, MD
{apostol.vassilev, munawar.hasan, edward.griffor, honglan.jin, pavel.piliptchak, mahima.arora}@nist.gov

Thoshitha Gamage
Southern Illinois University Edwardsville
Edwardsville, IL
tgamage@siue.edu

Abstract

The viability of automated driving depends on perception systems that provide accurate and reliable real-time information for safe decision-making. However, these systems must operate not only under ideal conditions but also under challenging environments such as adverse weather, occluded roadway objects, and varying vehicle speeds, which can lead to detection errors and delayed responses. This work evaluates the robustness of autonomous vehicle perception using predictive sensitivity quantification based on an ensemble of models. We propose a notional architecture for perception evaluation that integrates multiple data sources through a ROS-based interface and an extensible AI pipeline. Experiments are conducted using CARLA-generated scenarios and real vehicle data, allowing assessment of simulation limitations. Results show that diminished lighting conditions, such as fog and low sun altitude, have the greatest impact on perception performance. Adversarial occlusions further increase model sensitivity. Greater distances to roadway objects and higher vehicle speeds also degrade perception performance, resulting in reduced perception robustness.

1. Introduction

As autonomous vehicles (AVs) approach real-world adoption, safe navigation in complex environments requires highly robust perception systems [7, 14]. Leading AV developers identify handling edge cases [2, 3] and rigorously evaluating overall system performance [3] as primary challenges. Validation must begin at the components level; specifically, AI-powered perception systems must accurately detect and classify objects to inform safe driving decisions [25, 26]. However, AI models often exhibit inconsistent predictions under minor input variations. These inconsistencies directly impact safety; false positives or negatives can cause "phantom" braking [5] or failure to avoid hazards, localization errors can lead to collisions [28], and object misclassifications, such as confusing a pedestrian with an inanimate object, can lead to catastrophic consequences. Novel or untrained situations further exacerbate these risks.

Mitigating these perception gaps remains an open research challenge. The standard approach involves training models on massive datasets, but collecting comprehensive real-world data is prohibitively expensive [35]. Consequently, industry leaders [6] and researchers [10, 41] increasingly utilize synthetic data. Yet, this introduces a "sim-to-real" gap, causing performance discrepancies between simulated and real-world environments [9]. As a result, midetections and misclassifications persist in real-world deployments, causing severe accidents [1, 17, 19, 23] and hindering public acceptance of current AV technology. Although some pilot deployments approach human-level performance, progress remains uneven, and even advanced systems demonstrate signs of brittleness under unusual road conditions [14].

To address these limitations, our research develops assessment criteria to evaluate the robustness of AV object detection and classification. We evaluate predictive sensitivity against real-world occlusions, vandalized signage, adverse weather, and time-of-day variations using a suite of CARLA scenarios [12]. Because CARLA cannot accurately simulate speed-induced perception effects, we complement these simulations with physical vehicle tests across varying speeds. Ultimately, by integrating simulated and physical camera data, we systematically assess perception robustness and analyze how these predictive outcomes directly impact safety-critical vehicle control responses.

2. Related Work and Background

AV perception systems rely on multiple sensors – such as Light Detection and Ranging (LiDAR), Radio Detection And Ranging (Radar), cameras, and ultrasonic sensors – to understand their environment [33]. *Sensor Fusion* [20] combines these data streams to leverage each modality's strengths. Despite these techniques, inconsistent perception performance remains a critical challenge, particularly under adverse conditions like heavy rain, fog, snow or low-light [44]. These conditions degrade sensor performance by introducing noise and reducing perception robustness [32, 38], issues that high-risk environments (e.g., dense urban traffic, vandalized traffic signs, and unstructured rural roads) further amplify.

To provide confidence measures for AV decision-making, perception systems utilize uncertainty estimation through Bayesian, non-Bayesian, and Gaussian methods. Bayesian techniques such as MC-Dropout [18], model predictive distributions rather than point estimates to improve outlier robustness [42]; however, they often prove too computationally expensive for real-time deployment. Conversely, non-Bayesian approaches like Deep Ensembles [22] effectively balance performance and efficiency [27], offering a practical pathway to incorporate uncertainty estimation and build fail-safe mechanisms [18]. By combining diverse models, ensemble methods capture multi-modal distributions and generally outperform Gaussian approaches in generalization [13, 34]. Building on this, our work employs an ensemble of five heterogeneous models. While we do not focus on ensemble design, we leverage this model diversity to analyze how perception variability affects AV performance.

3. Predictive Sensitivity Assessment

Our selected ensemble balances real-time capability, computational complexity, robustness, and scalability across two primary architectures:

- **YOLO Family:** The You Only Look Once (YOLO) architecture [30] provides highly accurate, real-time object detection. Specifically, we utilize the state-of-the-art **YOLOv8** [37] and **YOLOv9** [4] variants for their precision on established benchmarks.
- **DETR Family:** The DEtection TRansformer (DETR) [11] utilizes a transformer encoder-decoder network. We deploy **DETR** with ResNet50 and ResNet101 backends (DETR50/101), alongside the **Real-Time Detection Transformer (RT-DETR)** [45], which offers a highly accurate, real-time derivative.

To ensure consistent class definitions and enable fair comparisons across our CARLA generated and real-world datasets, each model is pre-trained on the MS COCO dataset [24].

3.1. Driving Scenario for Assessment

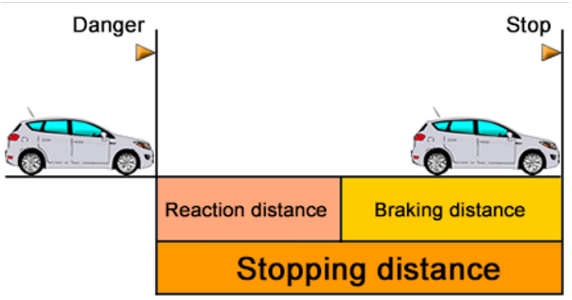


Figure 1. Simple Stopping Scenario

To assess perception robustness, our baseline scenario requires a vehicle to halt at a stop sign (Figure 1). Per Equation 1[1], safe total stopping distance (sd) combines reaction (rd) and braking (bd) distances, which depend on vehicle speed (s in km/h), reaction time (r in seconds), and road friction (f). Human reaction times range from 0.5 – 2.0s [21]; AVs average 0.8 – 0.9s (SD = 0.54s) [15], though computational delays or disengagements can extend this [31]. Friction (f) drops from 0.7 – 0.8 on dry asphalt to 0.1–0.6 on wet or icy surfaces [8].

$$rd = \frac{s \times r}{3.6}, \quad bd = \frac{s^2}{250 \times f}, \quad sd = rd + bd \tag{1}$$

Applying Equation 1 highlights how environments alter stopping requirements: at 48km/h (≈30mph) with a 1s AV reaction time, the required sd nearly doubles from 25.8m (dry, $f = 0.75$) to 50.7m (wet, $f = 0.25$).

3.2. Model Disagreement and Sensitivity Estimation

We adopt the aforementioned ensemble approach in Algorithm 1, focusing on a single class, i.e., a stop sign, denoted by $\mathcal{S}$. $\check{\theta}_{\mathcal{S}}$ denote the detection confidence threshold for $\mathcal{S}$ while $\theta_{\mathcal{S}}$ denotes the threshold for the ensemble. $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ denote the mean prediction probability and the standard deviation of the ensemble. Note, $\check{\theta}_{\mathcal{S}}$ and $\theta_{\mathcal{S}}$ are scenario-specific. Let $d_{\mathcal{S}}$ denote distance from object $\mathcal{S}$. If $sd \leq d_{\mathcal{S}}$ and $\mu_{\mathcal{S}} \geq \theta_{\mathcal{S}}$, then vehicle can comfortably stop before reaching $\mathcal{S}$. In contrast, $\mu_{\mathcal{S}} < \theta_{\mathcal{S}}$, at any point when $sd \geq d_{\mathcal{S}}$, results in unsafe driving conditions.

3.3. Safety Quadrant

Using Equation 1, we define a *safety quadrant* ($\mathcal{R}$) for our simple stopping scenario. Plotting the mean prediction probability ($\mu_{\mathcal{S}}$) against the distance to the target ($d_{\mathcal{S}}$), $\mathcal{R}$ occupies the region where $\mu_{\mathcal{S}} \geq \theta_{\mathcal{S}}$ and $d_{\mathcal{S}} > sd$. Bounded

[1] Adapted from https://tinyurl.com/44tadjsm for a "comfortable stop".

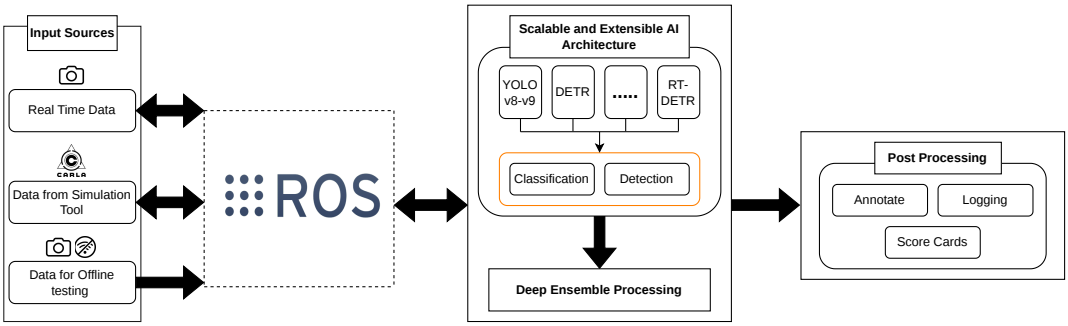


Figure 2. *Experiment Architecture:* Data from real-time, CARLA, or rosbag sources are transmitted via ROS to a scalable AI pipeline for object detection, ensemble processing, and post-processing.

Algorithm 1 Ensemble-Method

Require: $(M, \mathcal{D}, \mathcal{S}, \check{\theta}_{\mathcal{S}})$, where each model in M is trained on dataset $\mathcal{D}$ [24], $\mathcal{S}$ is the object class (stop sign) and $\check{\theta}_{\mathcal{S}}(= 0.20)$ is the detection confidence threshold for $\mathcal{S}$ for participating models.

Ensure: $|M| \geq 2$

$\hat{M} \leftarrow \phi$ ▷ Initialize ensemble

$\mu_{\mathcal{S}} \leftarrow \phi, \sigma_{\mathcal{S}} \leftarrow \phi$ ▷ Initialize mean and standard deviation of the ensemble

for $\check{d} \in \mathcal{D}$ **do**

  for $m \in M$ **do**

    if $Pr_{\mathcal{S}}[\check{d}][m] \geq \check{\theta}_{\mathcal{S}}$ **then** ▷ $Pr_{\mathcal{S}}[\check{d}][m]$ is the prediction probability by m on $\mathcal{S}$ for $\check{d}$

      $\hat{M}[\check{d}][m] \leftarrow Pr_{\mathcal{S}}[\check{d}][m]$

    else

      $\hat{M}[\check{d}][m] \leftarrow 0$

    end if

  end for

  $\mu_{\mathcal{S}}[\check{d}] \leftarrow$ mean for $\check{d}$ ▷ $\forall m \in M$

  $\sigma_{\mathcal{S}}[\check{d}] \leftarrow$ standard deviation for $\check{d}$ ▷ $\forall m \in M$

end for

Return $(\hat{M}, \mu_{\mathcal{S}}, \sigma_{\mathcal{S}})$

by $x = sd$ and $y = \theta_{\mathcal{S}}$, this quadrant signifies the ideal operational state of the AV's perception stack.

For safe, continous operation, once the system enters region $\mathcal{R}$, $\mu_{\mathcal{S}}$ must remain at or above the threshold $\theta_{\mathcal{S}}$ as the distance $d_{\mathcal{S}}$ decreases. A critical safety failure occurs under two conditions: either the system fails to enter $\mathcal{R}$ before reaching the stopping distance $d_{\mathcal{S}} \leq sd$, or $\mu_{\mathcal{S}}$ drops below $\theta_{\mathcal{S}}$ after initial entry.

For our dry road baseline, we set $sd = 25.55$ m and $\theta_{\mathcal{S}} = 0.75$. Setting $\theta_{\mathcal{S}} = 0.75$ prioritizes cross-model consensus and reduce false positives while allowing moderately confident predictions to contribute to $\mu_{\mathcal{S}}$. This aligns with commonly used confidence thresholds in safety-critical perception tasks [40, 43].

4. Methodology

Our experiment architecture[2] (Figure 2) integrates simulated and physical data streams via a unified ROS-based interface, enabling comprehensive benchmarking and controlled experimentation. To execute the CARLA [12] scenarios and physical tests outlined previously, sensor placement across both the simulation and the real-world vehicle mirrors the KITTI [16] data collection platform. Real-world data is captured using a Lucid camera[3] to specifically evaluate speed-induced perception degradation under rainy and nighttime conditions. While the current evaluation focuses on stop sign detection, this architecture readily generalizes to other objects within the AV domain.

4.1. The Control Experiment in Simulation

Figure 3. *Base Experiment:* View from the ego vehicle (*ego*) approaching the stop sign $\mathcal{S}$ in CARLA

In our baseline simulation (Figure 3), the *ego* vehicle approaches a stop sign ($\mathcal{S}$) at 48 km/h (≈30 mph). The environment is configured for optimal visibility: 0% cloudiness, precipitation, and fog density, a sun azimuth of 120°, a sun altitude of 56.67°, and no adversarial occlusions.

To isolate perception vulnerabilities, we systematically manipulate state variables governing weather and adversarial occlusions. Weather variables – cloudiness, precipita-

[2]Source Code:

[3]Lucid TRIO16S-CC: https://tinyurl.com/lucid-camera-specs

tion, and fog density – are evaluated at three discrete intervals ($0\%, 33.33\%, 66.67\%$) to simulate reduced brightness, sensor artifacts, and degraded visibility. Lighting conditions are altered via sun azimuth ($0°$, $120°$, $240°$) and altitude ($-10°$, $23.33°$, $56.67°$) to induce glare, shadows, and time-of-day variations. Adversarial occlusions encompass environmental obstacles (trees, bushes, poles, pedestrians), vehicular obstructions (ambulances, firetrucks), and targeted physical or digital alterations, such as graffiti [29], adversarial patches [29, 36], and low-resolution models.

These variables are evaluated across two experimental sets:

- **Experiment Set 1:** Applies Algorithm 1 for sensitivity analysis across discrete weather intervals and four progressive occlusion levels (none, low, moderate, high) to identify critical edge cases.
- **Experiment Set 2:** Expands weather parameters to continuous ranges (e.g., $0-100\%$ fog, $0°-360°$ sun rotation) and introduces complex adversarial combinations to isolate severe perception degradation.

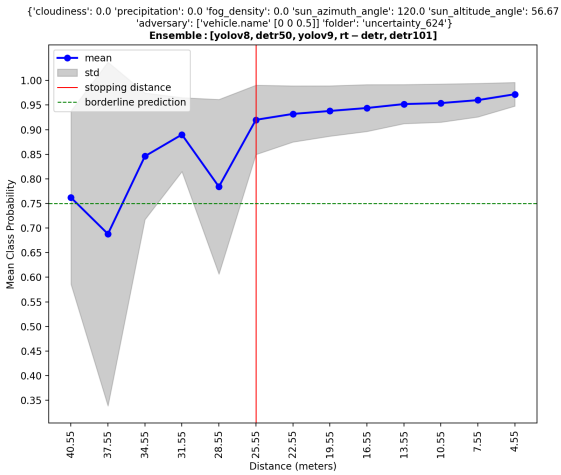


Figure 4. Baseline perception performance for $\mathcal{S}$ under optimal conditions, showing $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$. The safety quadrant ($\mathcal{R}$) is satisfied prior to the stopping distance.

Figure 4 illustrates ensemble performance under optimal baseline conditions, establishing a foundational control. The safety quadrant ($\mathcal{R}$) – bounded by $sd = 25.55$ m and $\theta_{\mathcal{S}} = 0.75$ – is satisfied. As the ego vehicle approaches $\mathcal{S}$, the ensemble mean ($\mu_{\mathcal{S}}$) monotonically increases, exceeding the confidence threshold well before the stopping distance. Concurrently, narrowing prediction variability ($\sigma_{\mathcal{S}}$, gray band) indicates high confidence and minimal predictive sensitivity. This control defines the ideal operational state, serving as a quantitative reference for assessing perception degradation in subsequent trials.

5. Results and Analysis

Our experimental evaluation encompasses 2,028 unique configurations of environmental and adversarial variables – 972 in Experiment Set 1[4] and 1,056 in Experiment Set 2[5]. Because our objective is to expose perception vulnerabilities, this analysis bypasses nominal outcomes and focuses strictly on edge cases that fail to satisfy the safety quadrant ($\mathcal{R}$), causing extreme perception degradation and critical safety failures (Sections 5.1 and 5.2).

5.1. Sensitivity Analysis (Experiment Set 1)

Analyzing the 972 scenarios in Experiment Set 1 reveals that moderate occlusion, fog density, and low sun altitude most severely degrade detection and localization. To illustrate these failure modes, two representative case studies demonstrating how these variables force the perception ensemble out of the safety quadrant ($\mathcal{R}$) are presented herewith.

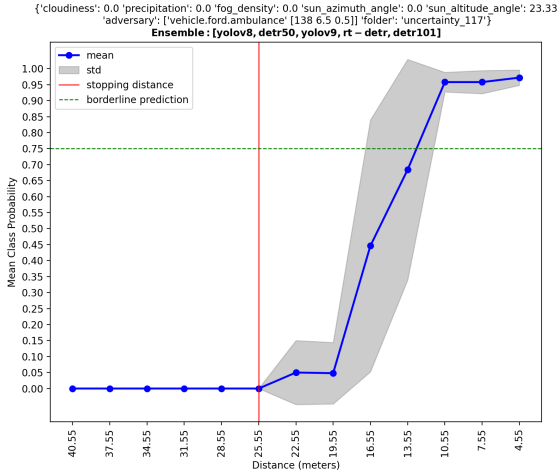
Case Study 1: Delayed Detection and High Predictive Sensitivity due to Moderate Occlusion Figure 5 shows Experiment 117 under optimal weather conditions (0% cloudiness, precipitation, and fog; sun altitude $23.33°$) but with moderate occlusion (see Figure 5a), thereby isolating the impact of adversarial object-induced occlusion on perception performance. $\mu_{\mathcal{S}}$ remains at or near zero for a substantial distance (see 5b), demonstrating that all participating models struggled significantly with detection and classification. Crucially, at the stopping distance ($sd = 25.55$m), $\mu_{\mathcal{S}}$ is effectively zero, indicating a complete failure to achieve the required confidence threshold ($\theta_{\mathcal{S}} = 0.75$) to enter region $\mathcal{R}$ before the vehicle reaches the critical braking point. The ensemble only achieves confidence above $\theta_{\mathcal{S}}$ at a much closer range, specifically below $d_{\mathcal{S}} = 13.55$m. Concurrent with this delayed detection, $\sigma_{\mathcal{S}}$ band exhibits high predictive sensitivity. While $\sigma_{\mathcal{S}}$ is narrow at close distances, it is significantly wider in the critical range between $d_{\mathcal{S}} \approx 25.55$ to 10.55m, indicating substantial disagreement among models. This poor performance is directly attributable to the partial occlusion of the stop sign by the parked adversarial vehicle. At 30mph, this scenario results in an unsafe operating condition, as delayed and uncertain perception may prevent the vehicle from stopping before $\mathcal{S}$.

Case Study 2: Compound Environmental Challenges Leading to Extreme Degradation of Perception Performance Figure 6 shows Experiment 689 under adverse conditions: 66.67% cloudiness, 33.33% precipitation,

[4]Calculated as 972 ($= 3^5 \times 4$) configurations: three discrete levels across five weather parameters, combined with four occlusion levels.

[5]Obtained by expanding the discretization of Set 1 parameters and introducing further adversaries.

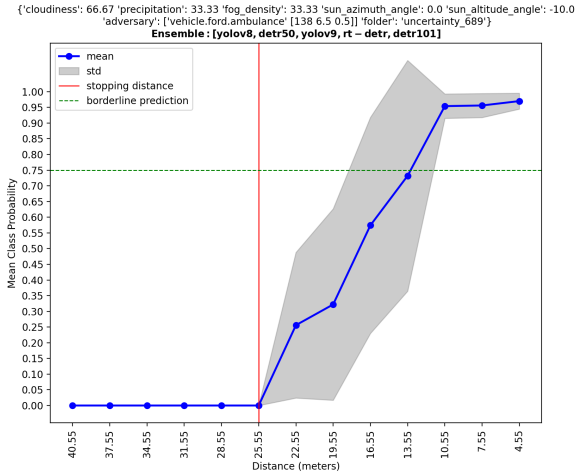
(a) CARLA simulation view of Experiment 117, illustrating the moderate occlusion of the stop sign by a parked ambulance.



(b) Mean prediction probability ($\mu_{\mathcal{S}}$) and standard deviation ($\sigma_{\mathcal{S}}$) for stop sign detection for Experiment 117.

Figure 5. Case Study 1: Perception performance and visual context under moderate occlusion. This scenario demonstrates delayed detection and high predictive sensitivity as the stop sign is partially obscured by an adversarial vehicle, leading to an unsafe condition.

(a) CARLA simulation view of Experiment 689, moderate occlusion, weather: high cloudiness, moderate precipitation/fog, low light.



(b) Mean prediction probability ($\mu_{\mathcal{S}}$) and standard deviation ($\sigma_{\mathcal{S}}$) for stop sign detection for Experiment 689.

Figure 6. Case Study 2: Perception performance and visual context under compound environmental challenges and moderate occlusion, demonstrates profound perception degradation leading to severely delayed detection.

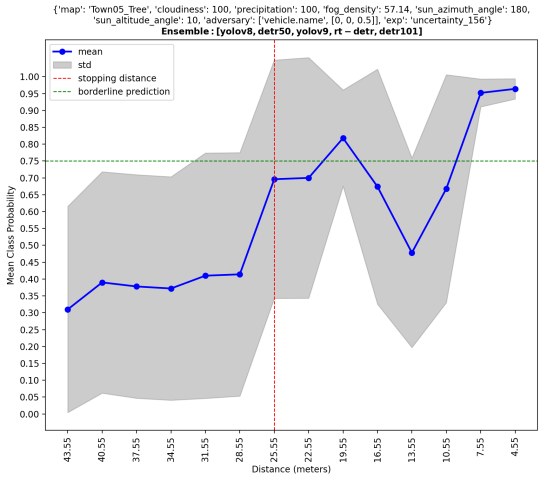
33.33% fog density, and low sun altitude ($-10.0°$), corresponding to nighttime illumination, with moderate occlusion (Figure 6a). Figure 6b reveals that $\mu_{\mathcal{S}}$ remains at or near zero for extended distances. At sd, $\mu_{\mathcal{S}}$ is negligible, indicating a complete failure to meet $\theta_{\mathcal{S}}$. Confidence above $\theta_{\mathcal{S}}$ is achieved at $d_{\mathcal{S}} \approx 11 - 12$m, a significant delay in reliable detection. Further, accompanied by persistent spread of $\sigma_{\mathcal{S}}$ across the critical range ($d_{\mathcal{S}} \approx 25.55$ to 10.55m). This reflects substantial ensemble disagreement during the critical detection phase. The combined effects significantly degrade the *ego* vehicle's ability to respond within required safety margins, yielding an unsafe operating condition.

Collectively, these cases demonstrate that moderate occlusion, precipitation, fog, and low illumination consistently induce delayed detection, depressed $\mu_{\mathcal{S}}$, and widened $\sigma_{\mathcal{S}}$ bands. When these factors combine (as in Case Study 2), perception failure is amplified, critically compromising the vehicle's ability to respond within required safety margins.

5.2. Extreme Sensitivity (Experiment Set 2)

Expanding to the 1,056 scenarios in Experiment Set 2, we evaluate how continuous, extreme environmental fac-

(a) CARLA simulation view of Experiment 156, adversarial object (a tree) partially covering the stop sign with extreme weather conditions.



(b) Mean prediction probability ($\mu_{\mathcal{S}}$) and standard deviation ($\sigma_{\mathcal{S}}$) for stop sign detection for Experiment 156.

Figure 7. Case Study 3: Perception performance and visual context under compound extreme conditions and occlusion by a tree, demonstrates catastrophic degradation, extreme predictive sensitivity, and loss of temporal consistency.

tors (e.g., 100% precipitation, direct sun glare) combined with complex adversarial objects completely shatter perception robustness. The following three case studies highlight catastrophic degradation, characterized by erratic temporal inconsistency and massive ensemble disagreement.

Case Study 3: Catastrophic Degradation from Compound Extreme Conditions Figure 7 illustrates a severe breakdown in perception robustness during Experiment 156. This scenario combines extreme environmental challenges – 100% cloudiness and precipitation, 57.14% fog density, and a 10° sun altitude indicative of dusk – with an adversarial tree partially occluding the stop sign ($\mathcal{S}$) from the ego vehicle's field of view (Figure 7a).

As shown in Figure 7b, the mean prediction probability ($\mu_{\mathcal{S}}$) exhibits highly erratic, non-monotonic behavior. At the required stopping distance (sd), $\mu_{\mathcal{S}}$ reaches only 0.70, failing to enter the safety quadrant ($\mathcal{R}$). Beyond sd, this instability persists: $\mu_{\mathcal{S}}$ briefly climbs above the confidence threshold ($\theta_{\mathcal{S}}$) at $d_{\mathcal{S}} = 19.55$m, but immediately drops back below it, fluctuating significantly before stabilizing only at a critically unsafe proximity ($d_{\mathcal{S}} \approx 8$m). This erratic trajectory represents a severe loss of temporal consistency. Furthermore, the exceptionally wide $\sigma_{\mathcal{S}}$ band across the entire approach indicates the perception system's inability to provide reliable safety margins.

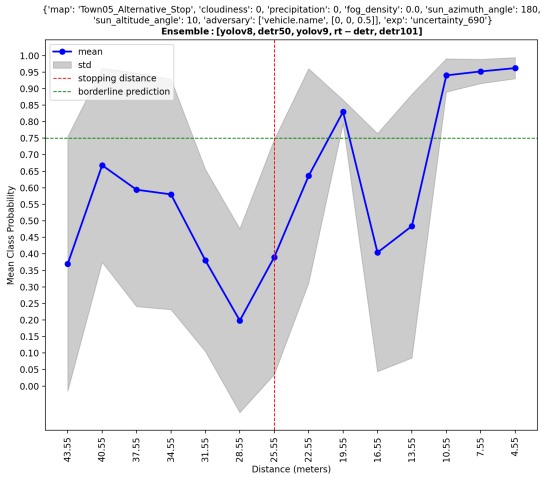
Case Study 4: Catastrophic Perception Performance Degradation from Adversarial Alteration and Sun Glare Experiment 690 (Figure 8) evaluates perception robustness against direct solar glare and targeted visual degradation. Under optimal weather conditions (0% cloudiness, precipitation, and fog), an adversarially altered stop sign is paired with a low sun altitude (-10.0°) and 180° azimuth. This geometry positions the sun directly in front of the ego vehicle, inducing intense glare and spectral deflection (Figure 8a).

As shown in Figure 8b, the ensemble mean ($\mu_{\mathcal{S}}$) exhibits significant instability. Starting at ≈ 0.65 at $d_{\mathcal{S}} = 40.55$ m, the probability drops to ≈ 0.59 at 37.55m, followed by a steep decline to ≈ 0.20 at 28.55m. At the stopping distance (sd), $\mu_{\mathcal{S}}$ is approximately 0.40, failing to satisfy the safety quadrant ($\mathcal{R}$). Beyond sd, $\mu_{\mathcal{S}}$ remains erratic: it briefly increases to ≈ 0.82 at 19.55m before suffering a severe drop to ≈ 0.40 at 16.55m. High confidence is sustained only at very close proximity ($d_{\mathcal{S}} \approx 8$m), confirming a profound loss of temporal consistency within the critical braking zone. The $\sigma_{\mathcal{S}}$ band remains wide until the final approach, signifying persistent model disagreement throughout the detection phase. Consistent with findings in Experiment Set 1, these results indicate that precise solar orientations can destabilize the perception system state even in the absence of adverse weather.

Case Study 5: Extreme Robustness Limits with Altered Sign and Severe Fog/Low Light Figure 9 illustrates the perception performance in Experiment 842, featuring an adversarially altered stop sign ($\mathcal{S}$), combined with a high fog density (71.43%) and a low sun altitude angle (10.0°), represent a challenging visual environment (Figure 9a).

As depicted in Figure 9b, $\mu_{\mathcal{S}}$ remains at extremely low values, fluctuating slightly, for an extensive range (e.g., ≈ 0.02 at $d_{\mathcal{S}} = 43.55$m, dropping to ≈ 0.02 at 28.55m).

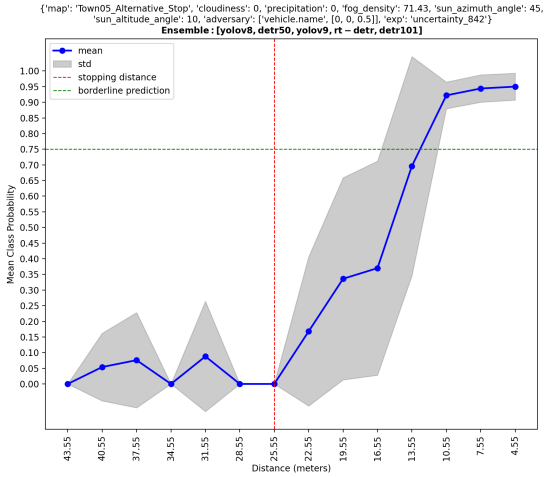
(a) CARLA simulation view of Experiment 690, illustrating the adversarially altered stop sign under severe sun glare conditions.



(b) Mean prediction probability ($\mu_{\mathcal{S}}$) and standard deviation ($\sigma_{\mathcal{S}}$) for stop sign detection for Experiment 690.

Figure 8. Case Study 4: Catastrophic predictive sensitivity leading to severe loss of temporal consistency.

(a) CARLA simulation view of Experiment 842, adversarially altered stop sign under severe fog and low light conditions.



(b) Mean prediction probability ($\mu_{\mathcal{S}}$) and standard deviation ($\sigma_{\mathcal{S}}$) for stop sign detection for Experiment 842.

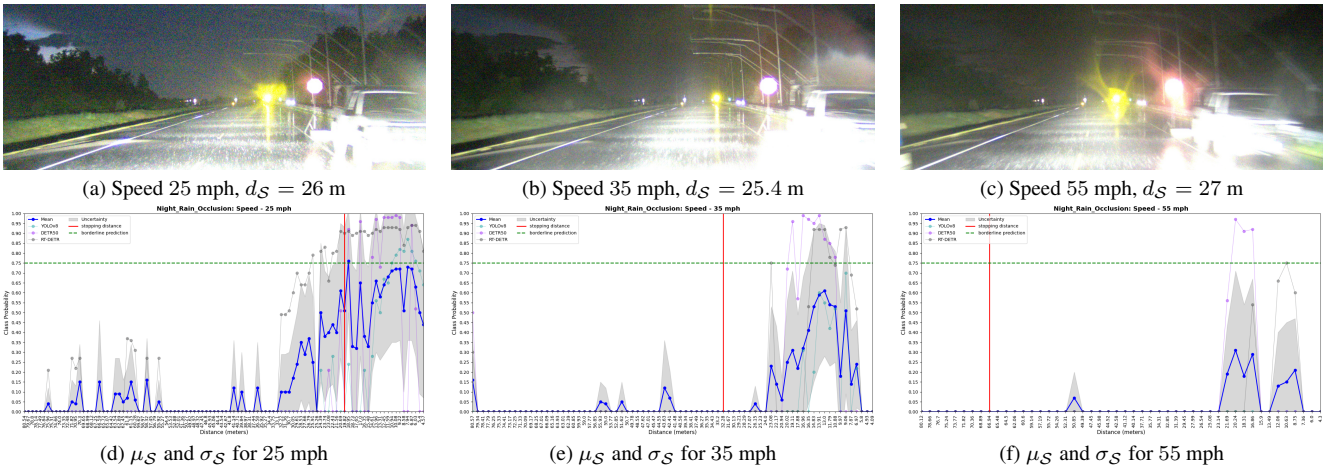
Figure 9. Case Study 5: Perception performance and visual context under adversarial stop sign alteration, high fog, and low light, showing profound predictive sensitivity leading to an unsafe condition.

Crucially, at sd, $\mu_{\mathcal{S}}$ is approximately 0.02, showing failure to enter $\mathcal{R}$. $\theta_{\mathcal{S}}$ is achieved around $d_{\mathcal{S}} \approx 11 - 12$m, which is far beyond any safe reaction distance. $\sigma_{\mathcal{S}}$ band is wide almost the entire detection range, indicating that models struggle not just with confidence, but also with basic agreement on object presence even at distances where high confidence is expected.

Collectively, these Set 2 experiments demonstrate that while discrete adverse conditions delay detection, compounding extreme variables – particularly intense sun glare, dense fog, and targeted adversarial patches – induce erratic temporal inconsistency and total predictive failure. Such catastrophic degradation underscores the critical vulnerability of current perception ensembles in unconstrained real-world environments.

5.3. Real-World Experiment

To evaluate speed-induced visual effects, we conducted physical vehicle tests at the VTTI [39] AV testing track. The *ego* vehicle, operating under manual control via a



(a) Speed 25 mph, $d_{\mathcal{S}} = 26$ m (b) Speed 35 mph, $d_{\mathcal{S}} = 25.4$ m (c) Speed 55 mph, $d_{\mathcal{S}} = 27$ m

(d) $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ for 25 mph (e) $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ for 35 mph (f) $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ for 55 mph

Figure 10. Real-World: Visual context under 25, 35 and 55 mph, and demonstration of effect of speed on $\mu_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$.

Dataspeed Inc.[6] drive-by-wire kit, utilized the image sensor architecture described in Figure 2. We profiled the vehicle at speeds of 40, 56, and 88km/h (25, 35, and 55mph). To simulate complex real-world edge cases, the target stop sign ($\mathcal{S}$) was partially occluded by a parked car, and we introduced artificial rain and nighttime settings (Figures 10a, 10b, and 10c).

As speed increases from 40 to 88km/h, the mean prediction probability ($\mu_{\mathcal{S}}$) progressively degrades (Figures 10d, 10e, and 10f). Higher speeds exacerbate motion blur and spectral deflection under these adverse conditions, significantly reducing image quality. Interestingly, the prediction variability ($\sigma_{\mathcal{S}}$) also decreases as speed increases. Rather than indicating high confidence, this shrinking standard deviation reveals that the participating models unanimously agree on a low detection confidence – a stark indicator of systemic perception failure. Consequently, across all three speed profiles, the ego vehicle fundamentally fails to enter the safety quadrant ($\mathcal{R}$), leading the perception stack into critically unsafe driving conditions.

6. Conclusion

This work systematically analyzed AV perception robustness using an ensemble of five state-of-the-art models across over 2,000 simulated and physical edge cases. Findings demonstrate that while isolated adverse conditions delay detection, the compounding effects of dense fog, heavy precipitation, low sun altitude, and adversarial occlusions lead to severe degradation. These variables significantly reduce ensemble consensus, induce temporal inconsistency, and consistently force the perception stack outside the defined safety quadrant ($\mathcal{R}$).

Physical evaluations indicate that high speeds under rainy and nighttime conditions amplify motion blur and spectral deflection. Crucially, these effects do not merely increase variability but result in unanimous model agreement on low detection confidence, indicating a systemic vulnerability. This research underscores that evaluating perception against isolated factors is insufficient; rigorous, multi-variable assessments are imperative to identify limits – such as those contributing to phantom braking or undetected hazards – before AVs can be reliably deployed in unconstrained environments.

[6] https://www.dataspeedinc.com/

References

[1] Self-Driving Uber Car Kills Pedestrian. https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html, 2018. Accessed: 2024-05-29. 1

[2] Cruise's Continuous Learning Machine Predicts the Unpredictable on San Francisco Roads. https://medium.com/cruise/cruise-continuous-learning-machine-30d60f4c691b, 2020. Accessed: 2025-08-20. 1

[3] Behind the Innovation: AI and ML at Waymo. https://www.youtube.com/watch?v=dtMMW0kTXIE, 2024. Accessed: 2025-08-20. 1

[4] YOLOv9: A Leap Forward in Object Detection Technology. https://docs.ultralytics.com/models/yolov9/, 2024. 2

[5] NHTSA Investigation Into Ford BlueCruise Fatal Crashes Reveals Shocking Detail. https://www.autoevolution.com/news/nhtsa-investigation-into-ford-bluecruise-fatal-crashes-reveals-shocking-detail-245894.html, 2025. 1

[6] Advancing Physical AI with NVIDIA Cosmos World Foundation Model Platform. https://developer.nvidia.com/blog/advancing-physical-ai-with-nvidia-cosmos-world-foundation-model-platform/, 2025. Accessed: 2025-05-29. 1

[7] Rasmus Adler, Patrik Feth, and Daniel Schneider. Safety Engineering for Autonomous Vehicles. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop (DSN-W)*, pages 200–205, 2016. 1

[8] Ulrik Andersson. What is Road Surface Friction Coefficient? https://intblog.onspot.com/road-surface-friction-coefficient-and-the-impact-of-automatic-snow-chains, 2024. 2

[9] Xiangyu Bai, Yedi Luo, Le Jiang, Aniket Gupta, Pushyami Kaveti, Hanumant Singh, and Sarah Ostadabbas. Bridging the Domain Gap between Synthetic and Real-World Data for Autonomous Driving. *ACM J. Auton. Transport. Syst.*, 1(2), 2024. 1

[10] Minh Cao and Ramin Ramezani. Data Generation Using Simulation Technology to Improve Perception Mechanism of Autonomous Vehicles. *Journal of Physics: Conference Series*, 2547(1):012006, 2023. 1

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Computer Vision –– ECCV 2020: 16th European Conference, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[12] CARLA Team. CARLA: Open-source Simulator for Autonomous Driving Research. https://carla.org/, 2024. 1, 3

[13] Ángela Casado-García and Jónathan Heras. Ensemble Methods for Object Detection. In *ECAI 2020*, pages 2688–2695. IOS Press, 2020. 2

[14] M.L. Cummings and Ben Bauchwitz. Identifying Research Gaps through Self-Driving Car Data Analysis. *IEEE Transactions on Intelligent Vehicles*, pages 1–10, 2024. 1

[15] Vinayak Dixit, Sai Chand, and Divya Nair. Autonomous Vehicles: Disengagements, Accidents and Reaction Times. *PLOS ONE*, 11:e0168054, 2016. 2

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision Meets Robotics: The KITTI Dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[17] Ben Goggins. A Tesla Owner Says His Car's 'Self-Driving' Technology Failed to Detect a Moving Train ahead of a Crash Caught on Camera. https://www.nbcnews.com/tech/tech-news/tesla-owner-says-cars-self-driving-mode-fsd-train-crash-video-rcna153345, 2024. Accessed: 2024-05-29. 1

[18] Ruben Grewal, Paolo Tonella, and Andrea Stocco. Predicting Safety Misbehaviours in Autonomous Driving Systems Using Uncertainty Quantification. In *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*, pages 70–81, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2

[19] The Intercept. Exclusive: Surveillance Footage of Tesla Crash on SF's Bay Bridge Hours After Elon Musk Announces "Self Driving Feature". https://theintercept.com/2023/01/10/tesla-crash-footage-autopilot/, 2024. Accessed: 2024-05-29. 1

[20] Jelena Kocić, Nenad Jovičić, and Vujo Drndarević. Sensors and Sensor Fusion in Autonomous Vehicles. In *2018 26th Telecommunications Forum (TELFOR)*, pages 420–425, 2018. 2

[21] Körkortonline. Stopping Distance = Reaction Distance + Braking Distance. https://korkortonline.se/en/theory/reaction-braking-stopping/. 2

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. 2

[23] Henry Lee. Feds Suspect Tesla that hit Fire Truck in Deadly I-680 Crash was on Autopilot. https://www.ktvu.com/news/tesla-that-hit-fire-truck-in-fatal-i-680-crash-was-on-autopilot-report, 2023. Accessed: 2024-05-29. 1

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014: 13th European Conference, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[25] John Molloy, Sepeedeh Shahbeigi, and John A. McDermid. Hazard and Safety Analysis of Machine-Learning-Based Perception Capabilities in Autonomous Vehicles. *Computer*, 57(11):60–70, 2024. 1

[26] Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C. de Albuquerque. Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4316–4336, 2021. 1

[27] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. 2

[28] Raj Haresh Patel, Jérôme Härri, and Christian Bonnet. Impact of Localization Errors on Automated Vehicle Control Strategies. In *2017 IEEE Vehicular Networking Conference (VNC)*, pages 61–68, 2017. 1

[29] Svetlana Pavlitska, Nico Lambing, and J Marius Zöllner. Adversarial Attacks on Traffic Sign Recognition: A Survey. In *2023 3rd International conference on electrical, computer, communications and mechatronics engineering (ICECCME)*, pages 1–6. IEEE, 2023. 4

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[31] Aleksander Rydzewski and Pawel Czarnul. Human Awareness versus Autonomous Vehicles view: comparison of reaction times during emergencies. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 732–739, 2021. 2

[32] Sensible4. The Problem with Weather: Rain and Fog Challenge the Sensors. https://sensible4.fi/rain-and-fog-challenge-the-sensors/. 2

[33] Varuna De Silva, Jamie Roche, and Ahmet M. Kondoz. Fusion of LiDAR and Camera Sensor Data for Environment Sensing in Driverless Vehicles. *ArXiv*, abs/1710.06230, 2017. 2

[34] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted Boxes Fusion: Ensembling Boxes from Different Object Detection Models. *Image and Vision Computing*, 107: 104117, 2021. 2

[35] Zhihang Song, Zimin He, Xingyu Li, Qiming Ma, Ruibo Ming, Zhiqi Mao, Huaxin Pei, Lihui Peng, Jianming Hu, Danya Yao, et al. Synthetic Datasets for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Vehicles*, 9 (1):1847–1864, 2023. 1

[36] Go Tsuruoka, Takami Sato, Qi Alfred Chen, Kazuki Nomoto, Yuna Tanaka, Ryunosuke Kobayashi, and Tatsuya Mori. WIP: Adversarial Retroreflective Patches: A Novel Stealthy Attack on Traffic Sign Recognition at Night. In *Proceedings of the Symposium on Vehicle Security and Privacy*, 2024. 4

[37] Ultralytics. Explore Ultralytics YOLOv8. https://docs.ultralytics.com/models/yolov8/, 2023. 2

[38] Jorge Vargas, Suleiman Alsweiss, Onur Toker, Rahul Razdan, and Joshua Santos. An Overview of Autonomous Vehicles Sensors and Their Vulnerability to Weather Conditions. *Sensors*, 21(16), 2021. 2

[39] VTTI. Virginia Tech Transportation Institute. 7

[40] Lei Yang, Xinyu Zhang, Jun Li, Li Wang, Minghan Zhu, Chuang Zhang, and Huaping Liu. Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11):6832–6844, 2023. 3

[41] Ze Yang, Sivabalan Manivasagam, Yun Chen, Jingkang Wang, Rui Hu, and Raquel Urtasun. Reconstructing Objects in-the-wild for Realistic Sensor Simulation. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11661–11668, 2023. 1

[42] Aayushi Zaveri. The Impact Of Bayesian Statistics On Analysis. https://mindthegraph.com/blog/bayesian-statistics/, 2024. 2

[43] Ruochen Zhang, Hyeung-Sik Choi, Dongwook Jung, Phan Huy Nam Anh, Sang-Ki Jeong, and Zihao Zhu. Auxdepthnet: Real-time monocular 3d object detection with depth-sensitive features. *arXiv preprint arXiv:2501.03700*, 2025. 3

[44] Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. Perception and Sensing for Autonomous Vehicles Under Adverse Weather Conditions: A Survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 146–177, 2023. 2

[45] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on Real-Time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024. 2