RACF: A Resilient Autonomous Car Framework with Object Distance Correction

Anonymous Author(s)

***Abstract*—Autonomous vehicles are increasingly deployed in safety-critical applications, where sensing failures or cyber-physical attacks can lead to unsafe operations, human injury, and severe physical damage. Reliable real-time perception is therefore critical for safe operation and practical deployment. In particular, vision-based distance estimation is vulnerable to environmental degradation and adversarial perturbations, while existing defenses are often reactive and may be too slow to promptly mitigate their impact on closed-loop control. We present a *Resilient Autonomous Car Framework (RACF)* that incorporates an *Object Distance Correction Algorithm (ODCA)* to maintain reliable obstacle distance estimation under sensing corruption and adversarial attacks. RACF improves perception-layer resilience through redundancy and diversity across a depth camera, LiDAR, and physics-based vehicle kinematics. When the obstacle distance estimated by the depth camera becomes inconsistent with LiDAR-based cross-sensor evidence, a residual-based gate activates ODCA to correct the corrupted measurement while preserving nominal observations, while vehicle kinematics provides a physics-aware constraint to stabilize the repair process. We implement the proposed RACF on the Quanser QCar 2 platform and evaluate its performance under multiple corruption levels and closed-loop stop-sign scenarios. Experimental results show that RACF achieves up to 35% RMSE reduction under strong corruption and improves stop compliance and braking latency while operating in real time. These results demonstrate that RACF provides an effective and lightweight framework for resilient perception in safety-critical autonomous driving systems.**

Note to Practitioners— This work addresses a practical challenge in autonomous driving: depth sensors may intermittently provide unreliable obstacle-distance measurements because of environmental degradation, sensor faults, or malicious interference. Such errors can delay or prevent timely braking even when independent sensing information remains available. The proposed framework checks whether the depth estimate is consistent with LiDAR evidence and activates distance correction only when the disagreement becomes significant. This conservative strategy avoids unnecessary modification of reliable measurements while improving distance recovery during corrupted intervals. Experiments on a Quanser QCar 2 demonstrate improved distance accuracy, stopping compliance, and braking latency under multiple corruption conditions. Because the method outputs a single corrected distance signal, it can be inserted before an existing planning or braking controller without redesigning the complete control architecture. The current validation is limited to an indoor, small-scale vehicle platform and controlled attack scenarios. Future work should evaluate the method under diverse weather, traffic, sensor-failure, and full-scale vehicle conditions.

***Index Terms*—Autonomous driving, perception resilience, time-series forecasting, adversarial attacks**

I. Introduction

Autonomous ground vehicles rely on accurate and timely perception to ensure safe closed-loop operation. Among perception components, obstacle distance estimation plays a critical role in braking and collision avoidance. When control-critical distance signals become corrupted due to adversarial attacks or physical malfunctioning, unsafe behavior such as delayed or missed braking may occur. Vision-based depth sensing, a primary source of obstacle distance estimation, is vulnerable to environmental degradation (e.g., glare and fog) [1] as well as physical and virtual adversarial perturbations [2]. Such corruption can propagate through the perception–planning–control pipeline and compromise safety [3]. Maintaining a reliable distance signal under intermittent sensing corruption remains an important practical challenge for real-world autonomous systems. This raises the central question addressed in this paper: *How can an autonomous system repair a corrupted control-critical distance signal without unnecessarily modifying measurements that remain reliable?*

Modern platforms typically employ multi-sensor redundancy, for example through camera–LiDAR fusion. However, redundancy alone does not guarantee stable behavior. LiDAR measurements can also suffer from sparsity and noise, and naïve modality switching or unconditional fusion may introduce brittle responses when one modality becomes unreliable [4]. Existing robustness approaches largely focus on input- or model-level defenses, such as adversarial patch mitigation and detector hardening [2], [5]–[7]. While these techniques improve detection robustness, they do not explicitly address the problem of repairing corrupted control-critical distance estimates while preserving measurements that remain reliable.

To address this problem, we introduce a cross-sensor trust-guided selective correction principle, in which the original distance measurement is preserved by default and correction is activated only when independent sensing evidence indicates potential corruption. Unlike unconditional fusion or full sensor replacement, this principle minimizes unnecessary modification of reliable observations while enabling targeted recovery during corrupted intervals.

We realize this principle through the *Resilient Autonomous Car Framework (RACF)*, a system-level resilience architecture designed to maintain reliable obstacle distance estimation under sensing corruption, adversarial attacks, and sensor anomalies. Rather than relying solely on robust perception models or unconditional sensor fusion, RACF combines cross-sensor

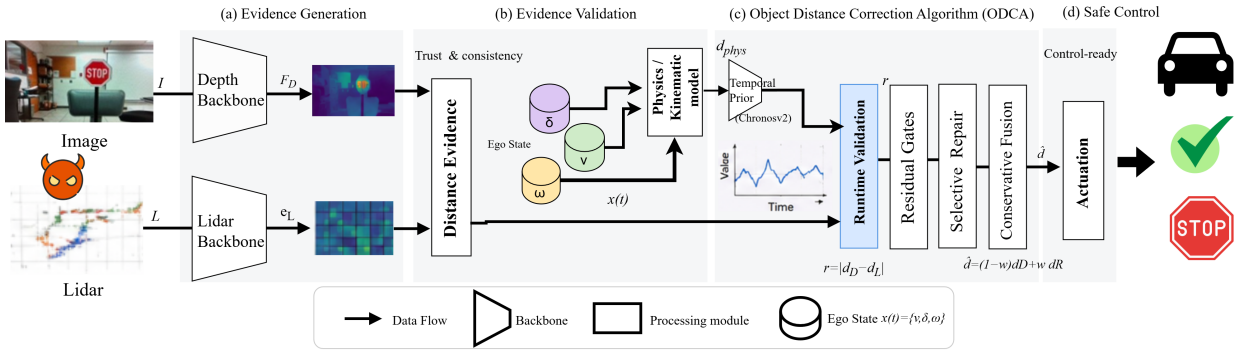


Fig. 1: **Resilient autonomous car framework.** Redundancy is achieved via depth and LiDAR sensing, while diversity is provided by a physics/kinematic distance evolution branch. The distance fusion module performs trust & consistency checks to produce a robust distance estimate for braking and control functions.

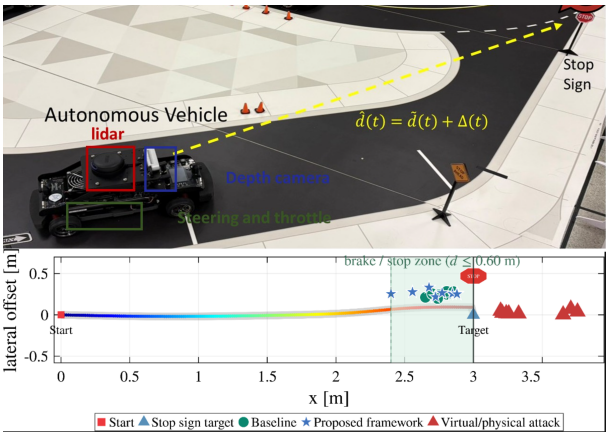


Fig. 2: **Selective distance correction for resilient braking.** Top: QCar testbed scenario under visual corruption, where attacked depth becomes inconsistent with LiDAR and can lead to unsafe distance estimation. Bottom: Closed-loop stop-sign approach outcomes. The shaded region denotes the safe brake/stop zone ($d \leq 0.60$ m). Baseline and the proposed framework stop within the safe range, while virtual and physical attacks lead to late or missed braking beyond the stop-sign target.

consistency monitoring, temporal prediction, and physics-aware validation to identify and mitigate corrupted distance measurements while preserving nominal system behavior.

Within RACF, depth resilience is formulated as a selective correction problem, where distance measurements are repaired only when sensor inconsistency indicates potential corruption. To achieve this objective, we introduce the *Object Distance Correction Algorithm (ODCA)*, a conservative correction module that uses cross-sensor trust to determine when a candidate repair should influence the control-ready distance estimate. A residual-based gating mechanism activates correction only when depth–LiDAR disagreement exceeds a predefined threshold. When the sensors agree, the original depth measurement is preserved; as the inconsistency increases, the contribution of the repaired estimate is progressively increased. In our implementation, ChronosV2 provides a frozen temporal prior, while LiDAR-grounded verification and vehicle kinematic constraints support trust evaluation and physically plausible repair. The proposed selective correction mechanism, however, is not restricted to a specific temporal forecasting backbone.

We implement RACF on the Quanser QCar 2 autonomous vehicle platform and evaluate its performance under weak, medium, and strong corruption attack scenarios. Experimental results demonstrate that RACF reduces distance-estimation errors by up to 35% under severe corruption without unnecessarily degrading reliable measurements under nominal operating conditions. Furthermore, closed-loop stop-sign experiments with physical adversarial patch attacks show improved stopping compliance and reduced braking latency, demonstrating the practical effectiveness of cross-sensor trust-guided selective correction for resilient autonomous driving.

The main contributions of this paper are summarized as follows:

- We introduce a **cross-sensor trust-guided selective correction principle** for control-critical distance estimation, in which nominal measurements are preserved by default and correction is activated only when independent sensing evidence indicates potential corruption.
- We realize this principle through the **Resilient Autonomous Car Framework (RACF)** and its **Object Distance Correction Algorithm (ODCA)**, which combine runtime cross-sensor trust evaluation, temporal priors, and physics-aware constraints to perform conservative residual-based recovery without unnecessarily modifying reliable observations.
- We validate RACF through extensive experiments on the Quanser QCar 2 platform under both controlled offline corruption and closed-loop physical adversarial patch attacks, demonstrating improved distance recovery, stopping compliance, and braking latency while preserving nominal operating performance.

II. Related Work

A. Autonomous Driving Architecture and Perception Dependencies

Autonomous driving systems operate through a perception–planning–control pipeline, where reliable sensing is essential for safe operation [8], [9]. Multi-sensor configurations, particularly camera and LiDAR combinations, are widely adopted to improve robustness through redundancy. Camera–LiDAR calibration and alignment methods reconcile cross-sensor bias and noise to enable consistent fusion under nominal conditions [10], while uncertainty-aware modeling further improves reliability under distribution shift [11]. However, many fusion pipelines implicitly assume reliable sensing and do not explicitly address intermittent corruption of a single modality that can be triggered by adversarial attacks or natural causes.

B. Adversarial and Environmental Attacks on Autonomous Perception

Vision-based autonomous perception components are vulnerable to both deliberate adversarial manipulation and natural sensing degradation. Physical adversarial patches can mislead traffic-sign and object detectors [5], [6], while adverse environmental conditions such as fog, low illumination, and reflective surfaces introduce systematic perception errors [1]. For example, perturbations can propagate to downstream planning and control modules, potentially resulting in unsafe operations.

C. Robustness and Sensor-Level Recovery Approaches

Most existing robustness techniques focus on image-level defenses or model hardening, aiming to improve visual robustness under attack [5], [6]. While effective at improving detection performance, these approaches do not directly correct the control-critical distance signal required for safe navigation planning and control. Multi-sensor fusion provides redundancy across modalities, but naive or unconditional fusion may degrade otherwise clean measurements when one sensor becomes unreliable [10]. Temporal forecasting offers an alternative perspective by exploiting temporal structure to recover corrupted signals. Recent multivariate forecasting models such as SOFTS [12] and foundation time-series models such as Chronos [13] demonstrate strong performance across domains. However, most forecasting-based approaches remain forecast-only and lack mechanisms for selectively correcting corrupted measurements while explicitly preserving uncompromised or normal measurements.

D. Research Gap

Existing robustness methods do not fully address the problem of maintaining a reliable obstacle-distance signal under intermittent sensing corruption. The key limitation is not only whether a perception model can detect or forecast corrupted measurements, but whether the system can selectively repair control-critical distance estimates without degrading nominal observations.

Specifically, three gaps remain. First, image-level and model-level defenses mainly improve visual detection robustness, but they do not directly stabilize the distance signal used by braking and control modules. Second, multi-sensor fusion improves nominal robustness, but unconditional fusion can propagate errors when one modality becomes unreliable and may also overwrite clean measurements. Third, temporal forecasting provides useful short-horizon priors, but forecast-only methods lack cross-sensor grounding and do not decide when correction should be activated.

These gaps motivate a selective, control-aware correction framework. RACF addresses this need by combining temporal priors, LiDAR-based consistency validation, and physically plausible constraints to repair corrupted depth measurements only when inconsistency is detected, while preserving nominal distance estimates under sensor agreement.

III. Resilient Autonomous Car Framework

The preceding discussion identifies a central limitation of existing perception-resilience approaches: corrupted control-critical distance measurements must be repaired without unnecessarily modifying observations that remain reliable. To address this limitation, we propose the *Resilient Autonomous Car Framework (RACF)*, which organizes resilient distance estimation around a cross-sensor trust-guided selective correction principle. Instead of continuously fusing or replacing sensor measurements, RACF preserves the primary distance estimate by default and activates recovery only when independent evidence indicates potential corruption.

This design is motivated by the fact that sensing redundancy does not necessarily imply that all available measurements should be treated as equally reliable or fused unconditionally. In safety-critical operation, an unreliable secondary sensor may propagate additional errors, while unnecessary replacement may distort a primary measurement that remains valid. RACF therefore separates resilient distance estimation into three functions: evidence generation, runtime trust evaluation, and selective recovery. This separation allows the system to first determine whether intervention is necessary before permitting a repaired estimate to affect the control-ready distance signal.

Within this architecture, the sensing sources serve distinct and complementary roles rather than acting as interchangeable distance measurements. The depth camera provides the primary control-critical distance estimate, LiDAR provides independent cross-sensor evidence for detecting inconsistency, and vehicle kinematics provides a diverse physical-consistency cue. LiDAR is therefore used primarily to assess trust rather than to directly replace the depth measurement. When the available evidence remains consistent, RACF retains the original depth estimate; when the evidence indicates potential corruption, the selective recovery mechanism is activated.

The temporal model serves a similarly constrained role. It generates a candidate repair prior rather than an always-active substitute for the observed measurement. Cross-sensor trust determines whether and to what extent this candidate repair

Algorithm 1 RACF Operational Workflow

Require: RGB image I, 2D LiDAR scan L, ego state $x(t) = \{v, \delta, \omega_z\}$
Ensure: Control-ready robust distance estimate $\hat{d}(t)$

(a) Evidence Generation
$F_D \leftarrow \text{DepthBackbone}(I)$
$F_L \leftarrow \text{LiDARBackbone}(L)$
$d_D \leftarrow \text{DistanceEvidence}(F_D)$
$d_L \leftarrow \text{DistanceEvidence}(F_L)$
(b) Evidence Validation
$d_{\text{phys}} \leftarrow f_{\text{kin}}(x(t))$
$d_{\text{prior}} \leftarrow \text{TemporalPrior}(d_{\text{phys}})$
$r \leftarrow |d_D - d_L|$
$r_{\text{phys}} \leftarrow \phi(d_D, d_L, d_{\text{prior}})$
$v_{\text{gate}} \leftarrow \text{RuntimeValidation}(r, r_{\text{phys}})$
(c) Object Distance Correction Algorithm
if $v_{\text{gate}} = \text{valid}$ **then**
    $\hat{d}(t) \leftarrow d_D(t)$
else
    $w \leftarrow \text{ResidualGate}(r, r_{\text{phys}})$
    $d_R \leftarrow \text{SelectiveRepair}(d_D, d_L, d_{\text{prior}})$
    $\hat{d}(t) \leftarrow (1 - w)d_D + wd_R$
end if
(d) Safe Control
$\text{Actuation}(\hat{d}(t))$
return $\hat{d}(t)$

is allowed to influence the final output, while temporal and physical priors constrain the magnitude and plausibility of the correction. By separating repair generation from repair activation, RACF mitigates corrupted observations while minimizing unnecessary changes to reliable measurements.

A. RACF Operational Workflow

Algorithm 1 summarizes the four-stage operational workflow of RACF shown in Fig. 1. First, multi-source evidence is generated from the depth camera, 2D LiDAR, and vehicle-state-driven physics priors. Second, the generated evidence is validated through cross-sensor and physics-based consistency analysis. When the evidence is considered reliable, the original depth estimate is retained. Otherwise, ODCA is activated to selectively repair the corrupted observation through residual gating and conservative fusion, producing a robust control-ready distance estimate for downstream planning and braking.

Resilient autonomous driving requires architectural mechanisms that preserve safety-critical functionality under sensing degradation and adversarial attacks. A robust design leverages both redundancy and diversity across multiple distance estimation sources to ensure correct measurement inputs for planning and control systems [8]. Within this architectural view, we consider three complementary methods for obstacle distance estimation: (i) depth camera measurements, (ii) LiDAR-based geometric range estimation, and (iii) physics-based distance estimation derived from vehicle kinematics.

Camera and LiDAR provide cross-modal redundancy, while the physics-based predictor introduces temporal consistency through vehicle dynamics constraints. Rather than blindly fusing these signals, the architecture evaluates cross-sensor consistency and applies conservative selection to correct unreliable measurements. Fig. 1 illustrates our Resilient Autonomous Car Framework (RACF) and highlights the perception-layer component addressed in this paper. In this work, we focus specifically on the perception layer by developing ODCA, a sensor-level correction mechanism that maintains depth camera reliability under intermittent corruption and attack, enabling robust distance estimation for control-critical tasks. In this approach, vehicle kinematics serves as a physics-based consistency constraint within ODCA, stabilizing the correction process without introducing additional control inputs. The control interface continues to operate on a single fused distance signal, preserving the original planning and braking architecture.

IV. Perception-Layer Realization in a Resilient Distance Architecture

A. Problem Formulation

Given a synchronized perception stream, ODCA receives a possibly corrupted depth estimate $\tilde{d}(t)$, confidence $\tilde{c}(t)$, LiDAR observation $\ell(t)$ when available, and vehicle-state measurements. The objective is to produce a control-ready fused distance estimate $\hat{d}_{\text{fused}}(t)$ that remains accurate and temporally stable under intermittent sensing corruption.

During online inference, the clean distance reference is not available. ODCA therefore relies only on the observed depth, confidence, LiDAR consistency, vehicle-state information, and temporal context to determine whether correction should be activated. Clean depth sequences collected under nominal sensing conditions are used only for offline evaluation as pseudo-ground-truth references.

B. Overview

Figure 1 illustrates the perception-layer realization of the Resilient Autonomous Car Framework (RACF), where safe braking depends on maintaining a trustworthy obstacle-distance signal under sensing degradation and adversarial interference. RACF combines redundant distance evidence from the depth camera and LiDAR with diverse physical-consistency information derived from vehicle kinematics. However, rather than continuously fusing all available measurements, RACF first evaluates whether intervention is necessary.

The key design principle is to separate candidate repair generation from repair activation. A temporal model may generate a plausible correction for a corrupted observation, but applying that correction continuously may unnecessarily alter measurements that remain reliable. RACF therefore preserves the original depth estimate by default and allows a candidate repair to influence the control-ready output only when independent depth–LiDAR inconsistency indicates potential corruption. When the sensors agree, the original measurement

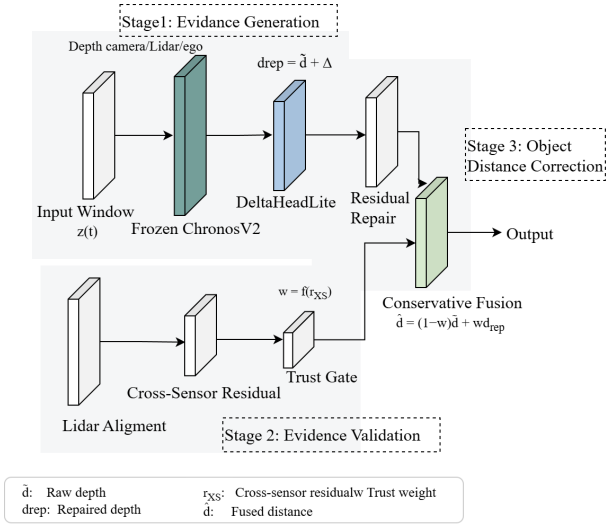


Fig. 3: **Perception-layer realization of ODCA within RACF.** A frozen temporal prior and lightweight repair head generate a candidate repaired distance. A depth–LiDAR residual is converted into a trust weight, which controls conservative fusion between the original and repaired distance estimates.

passes through unchanged; as the disagreement increases, the candidate repair is progressively introduced through a residual-based trust gate.

We instantiate this principle through ODCA, the perception-layer selective correction module shown in Fig. 3. The correction mechanism is independent of the specific temporal forecasting backbone. In our implementation, ChronosV2 serves as a frozen temporal prior, a lightweight additive delta head generates the candidate repair, and vehicle kinematics provides a physical-consistency constraint during training. The online gate relies on cross-sensor inconsistency to determine when and to what extent the candidate repair should affect the final control-ready distance estimate.

C. Online Inference

ODCA is deployed as a causal, per-step correction module. At each time t, it consumes the latest synchronized measurements $(\tilde{d}(t), \tilde{c}(t), \ell(t))$ (when LiDAR is available) and outputs a fused distance $\hat{d}_{\text{fused}}(t)$. Horizon notation is used only for offline training/evaluation.

D. Candidate Repair Generation

RACF does not require a specific forecasting backbone to generate the candidate repair. In this implementation, the repair generator is instantiated using ChronosV2 as a frozen probabilistic temporal prior. Given context length W, it produces S forecast samples whose mean and standard deviation form uncertainty-aware temporal features. We build the delta-head input by concatenating the latest observation features (e.g., $\tilde{d}(t), \tilde{c}(t)$) with $(\mu(t), \sigma(t))$ (and vehicle states when available). The sampling period Δt is estimated from

Algorithm 2 Object Distance Correction Algorithm (ODCA)

Require: Context length W, Chronos samples S
Require: Input window $\mathbf{x}_{t-W:t}$ with attacked depth $\tilde{d}$, LiDAR ℓ, ego states, and confidence $\tilde{c}$
Ensure: Control-ready fused distance $\hat{d}_{\text{fused}}(t)$

Stage 1: Evidence generation
Form context window $\mathbf{x}_{t-W:t}$
Extract $\tilde{d}(t)$, $\tilde{c}(t)$, $\ell(t)$, and ego states
Run frozen ChronosV2 with S samples
Compute forecast features $(\mu_i(t), \sigma_i(t))$
Construct feature tensor $\mathbf{z}(t) \in \mathbb{R}^{H\times 12}$
$\Delta(t) \leftarrow \text{DeltaHeadLite}(\mathbf{z}(t))$
$\hat{d}_{\text{rep}}(t) \leftarrow \tilde{d}(t) + \Delta(t)$
Stage 2: Evidence validation
if LiDAR $\ell(t)$ is available and valid **then**
  $\ell_{\to D}(t) \leftarrow \alpha\ell(t) + \beta$
  $r_{\text{XS}}(t) \leftarrow |\tilde{d}(t) - \ell_{\to D}(t)|$
  $w(t) \leftarrow \text{clip}\Big(\frac{r_{\text{XS}}(t)-\tau_{\text{low}}}{\tau_{\text{high}}-\tau_{\text{low}}}, 0, 1\Big)^{\gamma}$
else
  $r_{\text{XS}}(t) \leftarrow \text{N/A}$
  $w(t) \leftarrow 0$
end if
Stage 3: Object distance correction
$\hat{d}_{\text{fused}}(t) \leftarrow (1 - w(t))\tilde{d}(t) + w(t)\hat{d}_{\text{rep}}(t)$
Send $\hat{d}_{\text{fused}}(t)$ to the planning and braking controller
Output diagnostics $\{r_{\text{XS}}(t), w(t), \Delta(t)\}$
  return $\hat{d}_{\text{fused}}(t)$

timestamps (median inter-sample interval) or set to a fixed value (e.g., 0.02 s at 50 Hz).

E. Delta Repair Head

The delta head predicts an additive correction:

$$\hat{d}_{\text{rep}}(t) = \tilde{d}(t) + \Delta(t).$$

This residual formulation naturally supports *selective repair*: under nominal sensing, $\Delta(t)$ is driven toward zero and the final output remains close to the original measurement.

F. Cross-Sensor Alignment and Gate

To compare depth and LiDAR in a common domain, we use an affine alignment:

$$\ell_{\to D}(t) = \alpha\ell(t) + \beta,$$

where $\ell_{\to D}(t)$ denotes the LiDAR range aligned to the depth-distance domain. The alignment parameters (α, β) are estimated via robust least-squares over a short window using high-confidence depth samples (or an initial clean calibration run) and kept fixed within a run.

The gate is computed from depth–LiDAR inconsistency:

$$w(t) = \text{clip}\left(\frac{|\tilde{d}(t) - \ell_{\to D}(t)| - \tau_{\text{low}}}{\tau_{\text{high}} - \tau_{\text{low}}}, 0, 1\right)^{\gamma}.$$

When sensors agree, $w(t) \approx 0$ and $\hat{d}_{\text{fused}}(t) \approx \tilde{d}(t)$; when disagreement grows, $w(t)$ increases and the output relies more on the repaired signal.

In practice, we observed that direct regression between LiDAR range and camera-based depth estimates is highly unstable due to differences in sensing geometry and object localization noise. Empirically, naive regression produced large residual errors and unstable distance estimates in our dataset. Therefore, LiDAR is used only as a consistency trigger rather than as a direct substitute for the depth measurement.

G. Theoretical Analysis

We characterize three structural properties of ODCA: (i) exact preservation of the primary depth estimate under cross-sensor agreement, (ii) bounded intervention determined by the trust gate, and (iii) bounded sensitivity to perturbations in the LiDAR consistency reference. These properties concern the distance-correction mechanism and do not constitute a general closed-loop collision-avoidance guarantee.

Proposition 1 (Nominal preservation and selective error reduction). Let $d^*(t)$ denote the true distance, and define

$$\begin{aligned} e_D(t) &= \tilde{d}(t) - d^*(t), \\ e_R(t) &= \hat{d}_{\text{rep}}(t) - d^*(t), \\ e_F(t) &= \hat{d}_{\text{fused}}(t) - d^*(t). \end{aligned} \tag{1}$$

From the selective fusion rule,

$$\hat{d}_{\text{fused}}(t) = (1 - w(t))\tilde{d}(t) + w(t)\hat{d}_{\text{rep}}(t), \tag{2}$$

the fused error satisfies

$$e_F(t) = (1 - w(t))e_D(t) + w(t)e_R(t), \tag{3}$$

and therefore

$$|e_F(t)| \le (1 - w(t))|e_D(t)| + w(t)|e_R(t)|. \tag{4}$$

If

$$r_{\text{XS}}(t) \le \tau_{\text{low}}, \tag{5}$$

then $w(t) = 0$, and hence

$$\hat{d}_{\text{fused}}(t) = \tilde{d}(t). \tag{6}$$

Thus, ODCA exactly preserves the original depth measurement under cross-sensor agreement.

Moreover, if

$$e_D(t)e_R(t) \ge 0, \qquad |e_R(t)| < |e_D(t)|, \tag{7}$$

then, for any $w(t) \in (0, 1]$,

$$|e_F(t)| < |e_D(t)|. \tag{8}$$

Therefore, selective correction strictly reduces the absolute estimation error whenever the candidate repair lies on the same side of the true distance and is closer to it than the original depth estimate.

The condition $e_D(t)e_R(t) \ge 0$ requires the original and repaired estimates to lie on the same side of the true distance, while $|e_R(t)| < |e_D(t)|$ requires the candidate repair to be closer to the truth. These are sufficient, rather than necessary, conditions for strict error reduction.

Proposition 2 (Bounded intervention). Define the ODCA intervention as

$$\delta_F(t) = \hat{d}_{\text{fused}}(t) - \tilde{d}(t). \tag{9}$$

Then

$$\delta_F(t) = w(t)\left(\hat{d}_{\text{rep}}(t) - \tilde{d}(t)\right), \tag{10}$$

and therefore

$$|\delta_F(t)| = w(t)\left|\hat{d}_{\text{rep}}(t) - \tilde{d}(t)\right|. \tag{11}$$

If the residual repair is bounded as

$$\left|\hat{d}_{\text{rep}}(t) - \tilde{d}(t)\right| \le \Delta_{\max}, \tag{12}$$

then

$$|\delta_F(t)| \le w(t)\Delta_{\max} \le \Delta_{\max}. \tag{13}$$

Hence, the cross-sensor trust weight directly limits the magnitude of the modification applied to the primary distance measurement. In particular, $w(t) = 0$ gives zero intervention, whereas intermediate values of $w(t)$ permit only a proportional fraction of the candidate repair.

Proposition 3 (Bounded sensitivity to LiDAR-reference degradation). Let

$$r(t) = \left|\tilde{d}(t) - \ell_{\to D}(t)\right| \tag{14}$$

denote the nominal cross-sensor residual. Suppose that the aligned LiDAR reference is perturbed by $\eta_L(t)$, such that

$$\ell^{\text{deg}}_{\to D}(t) = \ell_{\to D}(t) + \eta_L(t), \qquad |\eta_L(t)| \le \bar{\eta}_L. \tag{15}$$

The degraded residual is

$$r_{\text{deg}}(t) = \left|\tilde{d}(t) - \ell^{\text{deg}}_{\to D}(t)\right|. \tag{16}$$

By the reverse triangle inequality,

$$|r_{\text{deg}}(t) - r(t)| \le |\eta_L(t)| \le \bar{\eta}_L. \tag{17}$$

Let the gate be written as

$$w(t) = g(r(t)), \tag{18}$$

where

$$g(r) = \text{clip}\left(\frac{r - \tau_{\text{low}}}{\tau_{\text{high}} - \tau_{\text{low}}}, 0, 1\right)^{\gamma}. \tag{19}$$

If g is Lipschitz continuous over the operating residual range with constant L_g, then

$$|w_{\text{deg}}(t) - w(t)| \le L_g\bar{\eta}_L. \tag{20}$$

Assuming that the candidate repair remains unchanged during this reference-only stress test, the difference between the degraded and nominal fused outputs satisfies

$$\begin{aligned} &\left|\hat{d}^{\text{deg}}_{\text{fused}}(t) - \hat{d}_{\text{fused}}(t)\right| \\ &\quad = |w_{\text{deg}}(t) - w(t)|\left|\hat{d}_{\text{rep}}(t) - \tilde{d}(t)\right|. \end{aligned} \tag{21}$$

If

$$\left|\hat{d}_{\text{rep}}(t) - \tilde{d}(t)\right| \leq \Delta_{\max}, \tag{22}$$

then

$$\left|\hat{d}^{\text{deg}}_{\text{fused}}(t) - \hat{d}_{\text{fused}}(t)\right| \leq L_g \bar{\eta}_L \Delta_{\max}. \tag{23}$$

Therefore, bounded degradation of the LiDAR consistency reference induces a bounded change in the ODCA output, provided that the repair magnitude is bounded and the gate operates over a Lipschitz-continuous region. This result is consistent with the LiDAR degradation stress test, in which performance deteriorates gradually rather than failing abruptly: LiDAR affects the final estimate through the bounded trust gate instead of being copied directly into the control-ready output.

In contrast, fixed-weight unconditional fusion modifies the primary measurement even when the two sensors agree. ODCA instead preserves the original depth estimate when $w(t) = 0$, bounds the intervention for $w(t) \in (0, 1)$, and limits the influence of bounded perturbations in the LiDAR consistency reference.

H. Training Objectives

Training follows a corruption-aware objective that preserves clean measurements while correcting corrupted segments: (i) **identity preservation** on clean/high-confidence steps, (ii) **minimal-change** regularization to discourage unnecessary correction, (iii) **cross-sensor consistency** to keep repaired depth aligned with $\ell_{\to D}$ when disagreement is high, and (iv) **kinematics regularization** applied on first differences of the repaired signal to encourage short-horizon physical plausibility (e.g., $\Delta d \approx -v\Delta t$). The total loss is:

$$\mathcal{L} = \lambda_{\text{ID}}\mathcal{L}_{\text{ID}} + \lambda_{\Delta 0}\mathcal{L}_{\Delta 0} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{kin}}\mathcal{L}_{\text{kin}},$$

with larger weights applied to low-confidence or attacked regions.

I. Residual-Based Diagnostic Signals

ODCA exposes interpretable residuals for diagnosis:

$$\begin{aligned} r_{\text{XS}}(t) &= \left|\tilde{d}(t) - \ell_{\to D}(t)\right|, \\ r_{\Delta}(t) &= \left|\Delta(t)\right|, \\ r_{\text{post}}(t) &= \left|\hat{d}_{\text{fused}}(t) - \ell_{\to D}(t)\right|. \end{aligned} \tag{24}$$

These residuals support attack diagnostic evaluation (AUROC/AUPRC) and complement recovery accuracy.

V. Experimental Setup

A. Overview

Vision-based depth estimation is vulnerable to environmental and adversarial corruption, potentially leading to unsafe braking. We therefore adopt a controlled attack-evaluation protocol that combines reproducible offline corruption with closed-loop physical patch attacks, enabling systematic assessment of both distance recovery and braking resilience.

B. QCar 2 Multi-Sensor Driving Benchmark

We collect synchronized multi-sensor data on a Quanser QCar 2 platform using the official indoor map at ≈50 Hz, totaling 67,821 samples across 13 sequences under speeds $\{1.0, 1.5, 2.0\}$ m/s and steering angles $\{0°, \pm 15°\}$.

Each timestamp includes depth-based distance and confidence, 2D LiDAR ranges, and vehicle state variables. The clean indoor depth signal is treated as pseudo-ground truth. All reported RMSE/MAE values are evaluated relative to this clean-depth baseline under the same sensing pipeline.

Stop signs are localized via YOLOv8-seg [14], and obstacle distance is computed as the median depth within the segmentation mask. LiDAR returns are clustered using DBSCAN [15], where the nearest cluster provides a cross-sensor reference.

C. Pseudo-Ground-Truth Construction

Obtaining precise ground-truth obstacle distances during closed-loop vehicle operation is challenging without external motion-capture systems or high-precision ranging infrastructure. Therefore, we construct a pseudo-ground-truth reference from clean measurements collected under nominal sensing conditions.

For each experimental trajectory, the vehicle follows the same driving protocol without attack injection or artificial corruption. The resulting clean depth sequence is used as the pseudo-ground-truth reference for the corresponding corrupted sequence generated under the same sensing pipeline.

Consequently, corrupted observations are compared against this clean reference to compute recovery metrics, including RMSE, MAE, and other resilience-related performance indicators. The clean reference is used only for offline evaluation and is not available to ODCA during online inference.

D. Hybrid Attack Evaluation

We evaluate resilience under both offline corruption and closed-loop physical patch attacks. Offline attacks introduce depth bias, confidence degradation, and blackout segments at three severity levels (weak/mid/strong), enabling controlled benchmarking against a clean reference (Fig. 5).

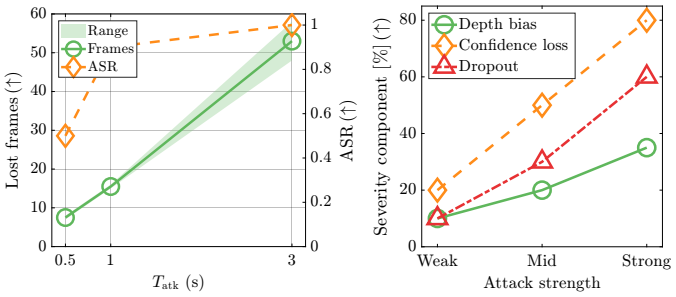
To emphasize safety-critical behavior, offline corruption is preferentially injected near the vehicle's braking region rather than uniformly over the trajectory. Specifically, attack segments are sampled around

$$I_{\text{imp}}(t) = \mathbf{1}\{d(t) \leq \min(d) + 0.2 \text{ m}\},$$

where $d(t)$ denotes the obstacle distance. This encourages attacks to occur near stop-sign approaches, where perception errors have the greatest influence on braking.

For real-world validation, we conduct closed-loop patch trials following [16]. Attack episodes are triggered after stable stop-sign detection and remain active during the critical braking window, with a short cooldown to prevent retriggering (Fig. 6). Per-frame detection status, estimated distance, brake commands, and timestamps are logged.

A trial is considered successful only if braking is triggered and the stopping distance satisfies a predefined safety



(a) Attack persistence (b) Attack configuration

Fig. 4: **Attack persistence and configuration.** (a) Longer patch duration expands the lost-detection window and increases attack success rate (ASR); the shaded band denotes the observed lost-frame range, while the green curve shows its midpoint. (b) Weak, mid, and strong attacks progressively increase depth bias, confidence loss, and dropout. Confidence loss is computed from the configured confidence value as $(1-c)\times 100\%$.

threshold; late braking is treated as failure. Since resilience depends on the relation between attack duration T_{atk} and system reaction time T_{react}, we report braking latency Δt_{brake} as a control-relevant metric and analyze suppression behavior under varying attack durations (Fig. 4(a)).

E. Attack Configuration

To systematically evaluate the resilience of RACF, we construct three levels of sensing corruption with progressively increasing severity. Fig. 4(b) summarizes the attack parameters used in our experiments. Each attack simultaneously affects depth accuracy, confidence reliability, and measurement availability through dropout. For visualization, confidence is reported as confidence loss, $(1-c)\times 100\%$, so that all severity components increase with attack strength. The weak setting represents mild sensing degradation, while the strong setting models severe corruption characterized by large distance bias, confidence collapse, and intermittent sensor blackout. These configurations provide a controlled benchmark for evaluating the ability of RACF to maintain reliable distance estimation under adverse sensing conditions.

VI. Experimental Results

A. Evaluation Metrics

We evaluate RACF from three perspectives: distance recovery, diagnostic utility, and closed-loop control performance.

Distance recovery. Recovery fidelity is measured using mean absolute error (MAE) and root mean square error (RMSE) between the fused estimate $\hat{d}_{\text{fused}}(t)$ and the clean pseudo-ground-truth reference. Lower MAE and RMSE indicate better distance recovery.

Diagnostic utility. We evaluate whether ODCA provides useful indicators of sensing corruption using AUROC and AUPRC. Two diagnostic scores are considered: the cross-sensor residual

$$r_{\text{XS}}(t) = \left|\tilde{d}(t) - \ell_{\to D}(t)\right|, \tag{25}$$

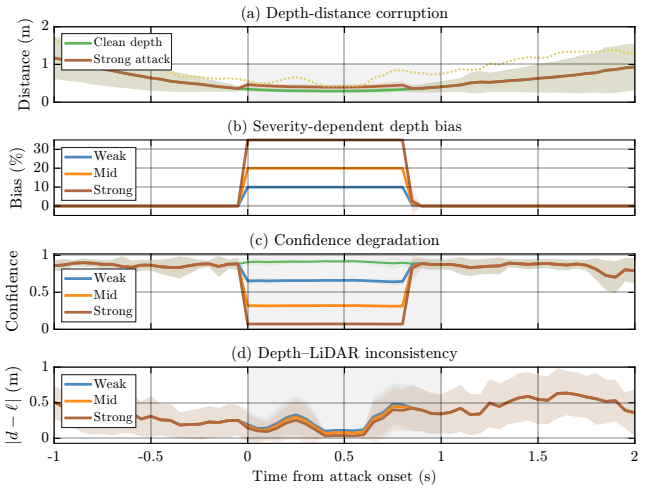
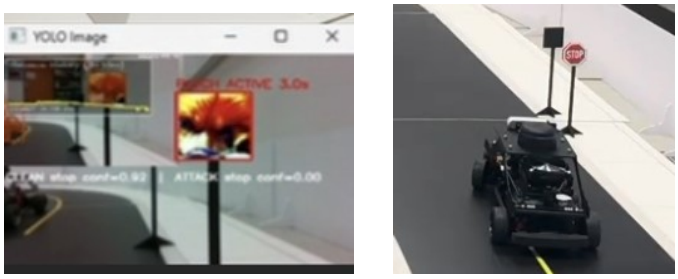


Fig. 5: **Offline depth-signal corruption dynamics.** The tiled dynamics are aligned to attack onset. Gray shading denotes the injected attack window, and colored bands denote mean $\pm$ standard deviation across aligned attack episodes. Weak, mid, and strong settings progressively increase depth bias, confidence degradation, and depth–LiDAR inconsistency.



(a) Detector view under patch attack. (b) Third-person QCar view.

Fig. 6: **Online physical adversarial patch attack.** (a) Detector view showing a patch attack that suppresses stop-sign detection and induces perception failure. (b) Third-person view of the QCar stopped in front of the patched stop sign during the closed-loop trial.

and the correction magnitude

$$r_{\text{chg}}(t) = \left|\hat{d}_{\text{fused}}(t) - \tilde{d}(t)\right|. \tag{26}$$

Here, r_{XS} captures depth–LiDAR inconsistency, while r_{chg} measures how strongly ODCA modifies the original depth estimate.

Robustness across attack strengths. To summarize robustness from weak to strong attacks, we report bounded degradation (BD) and resilience gain ratio (RGR). Bounded degradation measures how much RMSE increases from weak to strong attacks:

$$\text{BD} = \frac{\text{RMSE}_{\text{strong}} - \text{RMSE}_{\text{weak}}}{\text{RMSE}_{\text{weak}}}, \tag{27}$$

where lower values indicate more stable performance across attack strengths. The resilience gain ratio measures the relative

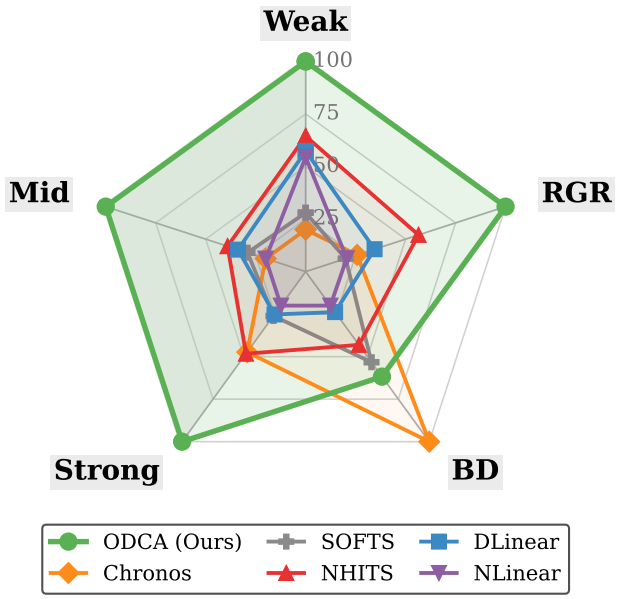


Fig. 7: **Resilience summary across attack strengths (higher is better).** Radar axes aggregate RMSE-based robustness at weak/mid/strong, bounded degradation (BD), and average gain vs. Chronos (RGR).

RMSE reduction of ODCA over the Chronos baseline:

$$\mathrm{RGR}(s) = 1 - \frac{\mathrm{RMSE}_{\mathrm{ours}}(s)}{\mathrm{RMSE}_{\mathrm{Chronos}}(s)}, \tag{28}$$

where $s \in \{\mathrm{weak}, \mathrm{mid}, \mathrm{strong}\}$. We report the average RGR across all attack strengths.

Closed-loop control performance. For physical QCar trials, we report stop compliance rate (SCR), attack success rate (ASR), and braking latency $\Delta t_{\mathrm{brake}}$. A stop is counted as successful only when braking is triggered and the final stopping distance satisfies $d_{\mathrm{brake}} \leq 0.60\,\mathrm{m}$. Braking latency is defined as the time from the first valid stop trigger to reaching the stopping threshold.

B. Experimental Comparison and Results

We compare ODCA with representative forecasting baselines, with citations embedded in Table I. A by-series split is adopted (70% train, 10% validation, 20% test). ChronosV2 remains frozen and only the 921-parameter delta head is trained. The prediction horizon is fixed at $H = 16$, corresponding to $\approx 0.32\,\mathrm{s}$ at 50 Hz, aligning with near-term braking decisions. Quantitative results are summarized in Table I, which reports recovery accuracy, model parameters, and strong-attack diagnostic quality. XS residuals exhibit strong separability across models, while ODCA demonstrates the highest Chg-AUPRC under strong attacks. In terms of complexity, ODCA maintains essentially the same total parameter scale as Chronos because the backbone is frozen and only a 921-parameter repair head is added; performance gains arise from cross-sensor validation and gated repair rather than increased model capacity. Fig. 7 summarizes robustness across weak/mid/strong attacks, bounded degradation (BD), and resilience gain relative to Chronos (RGR). Larger area indicates stronger robustness across regimes (BD is inverted so that higher is better).

a) Comparison with classical sensor baselines.: To contextualize the proposed framework, we compare ODCA with representative distance estimation approaches, including direct LiDAR measurements, EKF-based temporal filtering, and Chronos forecasting. These baselines provide reference points for traditional sensing, filtering, and forecasting pipelines rather than direct competitors to the proposed corruption-aware correction mechanism. For each baseline, the distance estimate is obtained directly from the corresponding sensor or filter output, and RMSE is computed with respect to the same clean-depth reference used as pseudo-ground truth in our evaluation protocol. This ensures a consistent comparison across methods under identical corrupted input sequences. While EKF fusion can reduce measurement noise through temporal filtering, it assumes approximately Gaussian noise and does not explicitly address intermittent or adversarial sensor corruption. The relatively large error of LiDAR measurements is mainly due to the sparse returns and limited angular resolution of the 2D LiDAR used on our platform.

b) Stress test with degraded LiDAR references.: Because ODCA uses LiDAR as the cross-sensor consistency reference, we further evaluate whether this reference assumption creates brittle behavior when LiDAR measurements are degraded. The experimental variable in this stress test is the LiDAR condition. Starting from the existing attacked-depth evaluation stream, we perturb the aligned LiDAR reference after synchronization as $\ell^{\mathrm{deg}}_{\to D}(t) = \ell_{\to D}(t) + b_L + \epsilon_L(t)$, where $\epsilon_L(t) \sim \mathcal{N}(0, \sigma_L^2)$. We additionally remove LiDAR observations with probability p_L to emulate intermittent missing returns. We then recompute the ODCA gate from $|\tilde{d}(t) - \ell^{\mathrm{deg}}_{\to D}(t)|$ and evaluate the final fused output against the same clean-depth pseudo-ground truth. This experiment isolates the consistency-reference assumption: it does not give ODCA access to clean depth at inference time and does not treat LiDAR as a replacement distance sensor. The LiDAR stress test uses the same held-out test sequences as Table I, but an independently generated strong-attack realization. Therefore, the clean-LiDAR baseline in Table III is not numerically identical to the strong-attack result reported in Table I. Within Table III, the attacked-depth stream is held fixed across all LiDAR conditions, and only the LiDAR reference is progressively degraded.

Table III shows that ODCA remains stable under mild LiDAR degradation, with RMSE increasing from 0.296 to 0.315. As degradation becomes moderate and severe, RMSE further increases to 0.343 and 0.373, respectively. These results indicate graceful rather than abrupt degradation as the LiDAR consistency reference becomes increasingly unreliable.

C. Closed-Loop Physical Implementation

We deploy QCar 2 on the official indoor map with multiple turns and intersection-like decision points (Fig. 8). A YOLO-based detector localizes the stop sign online, and braking is

TABLE I: **Distance recovery and diagnostic quality under attack.** Lower RMSE/MAE indicate better distance recovery (↓), while higher AUROC/AUPRC indicate better attack-related diagnostic quality (↑). Params reports total model parameters; for ODCA, this includes the frozen ChronosV2 backbone plus the 921-parameter repair head.

Model	Params (M)	Recovery Accuracy						Strong-Attack Diagnostics			
		Weak		Mid		Strong		XS		Chg	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	AUROC	AUPRC	AUROC	AUPRC
ODCA (Ours)	201.376	**0.111**	**0.093**	**0.193**	**0.170**	**0.323**	**0.283**	0.786	0.795	**0.581**	**0.610**
Chronos [13]	201.375	0.229	0.151	0.363	0.256	0.503	0.353	0.786	0.795	0.522	0.515
SOFTS [12]	6.903	0.207	0.171	0.330	0.232	0.649	0.487	0.894	**0.846**	0.461	0.336
NHITS [17]	3.036	0.144	0.104	**0.300**	**0.214**	0.498	0.365	**0.952**	0.817	0.321	0.142
DLinear [18]	0.016	0.153	0.117	0.315	0.223	0.657	0.530	**0.952**	0.817	0.381	0.156
NLinear [18]	0.008	0.157	0.116	0.363	0.261	0.708	0.577	**0.952**	0.817	0.452	0.172

Notes. Green cells indicate the best value in each column. Yellow cells indicate the second-best value when it helps compare strong baselines. XS denotes cross-sensor residual, and Chg denotes correction magnitude. Strong-attack diagnostics are retained from the full weak/mid/strong diagnostic analysis to keep the main table compact. SOFTS parameters are computed from the official default architecture under the same context and horizon setting used in our evaluation.

TABLE II: **Strong-attack method taxonomy and recovery comparison.** Lower RMSE indicates better distance recovery (↓).

Method	Range	Filter	Prior	Repair	RMSE ↓
LiDAR measurement	✓				0.600
EKF fusion [19]	✓	✓			0.450
Chronos [13]			✓		0.503
SOFTS [12]			✓		0.649
NHITS [17]			✓		0.498
DLinear [18]			✓		0.657
ODCA (Ours)	✓		✓	✓	**0.323**

Notes. All rows are evaluated on the same strong-attack split. ODCA uses LiDAR as a consistency trigger and ChronosV2 as a frozen temporal prior, then applies gated repair rather than directly substituting either signal.

TABLE III: **LiDAR degradation stress test.** Errors are computed between ODCA's final fused output and clean depth after recomputing the gate with degraded LiDAR. Lower is better.

LiDAR condition	LiDAR Cov.	Final Output Error		
		MAE	MSE	RMSE
Clean reference	1.00	0.221	0.088	0.296
Mild degradation	0.91	0.229	0.100	0.315
Moderate degradation	0.80	0.248	0.118	0.343
Severe degradation	0.66	0.262	0.139	0.373

Notes. The attacked-depth stream is held fixed across all four LiDAR conditions. Cov. denotes the fraction of valid LiDAR observations remaining after simulated dropout. Mild uses $(\sigma_L, b_L, p_L) = (0.05, 0, 0.10)$, moderate uses $(0.10, 0.05, 0.20)$, and severe uses $(0.20, 0.10, 0.35)$.

triggered based on accumulated detections and the estimated stop-sign distance.

a) Closed-loop metrics.: We report Stop Compliance Rate (SCR) and Attack Success Rate (ASR) for closed-loop trials, along with braking latency Δt_{brake}.

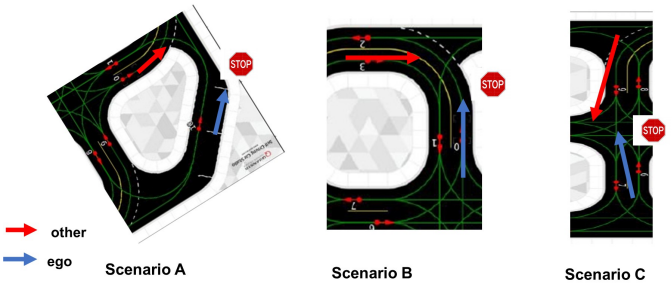


Fig. 8: **High-risk stop-sign scenarios on the indoor map.** The ego vehicle (blue arrow) encounters cross-traffic from other vehicles (red arrow) at conflict-prone intersections. The stop sign is placed in these regions to evaluate braking safety under attack. MAP-A/B/C emulate progressively more complex indoor driving layouts, linking the controlled QCar tests to real-world perception ambiguity and obstacle-interaction complexity.

b) Patch trials and integration.: Following the physical patch protocol in [16], we overlay an adversarial patch on the detected stop-sign region for a randomly sampled duration $T_{\text{atk}} \sim \mathcal{U}(0.5, 3.0)\,\text{s}$, aligned to the critical braking phase. To ensure consistent episode alignment across methods, an attack episode is triggered only after a stable stop-sign detection is observed, and a short cooldown prevents immediate retriggering. We log per-frame detection status and confidence, estimated stop-sign distance, brake command, and timestamps.

c) Patch coverage sensitivity.: In preliminary closed-loop sweeps, we found that attack success is dominated by the patch occlusion ratio $\rho \in [0, 1]$ over the stop-sign region. Here, ρ is an experimental attack-coverage parameter; values such as $\rho = 0.6$ are settings in our sensitivity sweep rather than fixed properties of ODCA. With partial coverage ($\rho \in \{0.3, 0.6\}$), detector confidence decreases but intermittent re-detections still occur, which remains sufficient for the controller to

TABLE IV: **Ablation under strong attacks.** Lower RMSE indicates better distance recovery (↓), while higher AUROC indicates stronger correction-magnitude diagnostic utility (↑).

$\mathcal{L}_{\text{ID}}$	$\mathcal{L}_{\Delta 0}$	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_{\text{kin}}$	**RMSE** ↓	**ΔRMSE**	**Chg-AUROC** ↑
✓				0.588	–	0.425
✓	✓			0.393	-0.195	0.609
✓	✓	✓		0.366	-0.222	**0.648**
✓	✓		✓	0.374	-0.214	0.593
✓	✓	✓	✓	**0.362**	-0.226	0.576

Notes. Checkmarks indicate active training losses. ΔRMSE is computed relative to $\mathcal{L}_{\text{ID}}$. $\mathcal{L}_{\Delta 0}$ yields the largest immediate recovery gain; $\mathcal{L}_{\text{cons}}$ improves diagnostic separability, while the full objective yields the best overall RMSE.

accumulate detections and trigger braking (ASR = 0). In contrast, near-full coverage ($\rho \in \{0.9, 1.0\}$) consistently suppresses detection throughout the braking window, resulting in missed or late braking beyond the safety margin (ASR = 1.0). Therefore, we treat $\rho \in [0.9, 1.0]$ as the effective strong physical attack regime in Table V.

d) Offboard deployment and evaluation.: Due to limited onboard compute, ODCA runs on an offboard server over a local network. Perception streams are generated at approximately 50 Hz, while the current prototype processes the latest available observation with an average end-to-end cycle time of approximately 49 ms, corresponding to an effective update rate of about 20 Hz. This update rate was sufficient for the low-speed QCar 2 stopping experiments considered in this study. The measured cycle includes perception transmission, ODCA inference, and control feedback. We conduct 10 closed-loop trials per setting (Clean, Attack, Attack+ODCA), with two stop encounters per run. A stop is counted as successful only if braking is triggered and $d_{\text{brake}} \leq 0.60$ m; late braking is treated as failure even if the vehicle eventually stops.

e) Statistical details: SCR is computed as SCR $= N_{\text{success}}/N$ with $N = 30$ stops per setting, and ASR is $1-$SCR under attack settings. Braking latency Δt_{brake} is reported as mean ± standard deviation over successful stops only, together with the sample size $N_{\Delta t}$.

D. Ablation on Training Losses

To validate the necessity of each loss component, we evaluate incremental training objectives under the **Strong** attack setting on the same held-out split as the main experiments. We report repair fidelity (RMSE) and diagnostic utility of correction magnitude (AUROC) as shown in Table IV.

E. Analysis and Discussion

ODCA achieves the lowest RMSE and MAE across all attack levels (Table I), demonstrating robust distance recovery under increasing corruption. While several baselines obtain high XS-AUROC/AUPRC, indicating that cross-sensor residual r_{XS} is an effective inconsistency cue, detection alone does not ensure accurate reconstruction. ODCA applies gated correction only under depth–LiDAR disagreement, improving both recovery accuracy and Change-AUROC/AUPRC.

TABLE V: **Closed-loop stop-sign trials on QCar 2,** $N = 30$ **per setting.** SCR and ASR are trial-level rates; Δt_{brake} is computed over successful stops only.

Setting	**SCR** ↑	**ASR** ↓	Δt_{brake} (s) ↓
Clean	**1.00** (30/30)	**0.00** (0/30)	0.1647 ± 0.016
Attack, no defense [16]	0.17 (5/30)	0.83 (25/30)	2.65 ± 0.21
Attack + ODCA [16]	0.80 (24/30)	0.20 (6/30)	0.1900 ± 0.030

Notes. Green cells indicate the best value in each column. Yellow cells indicate the defended attack setting, which restores most clean-condition stop compliance while substantially reducing attack success.

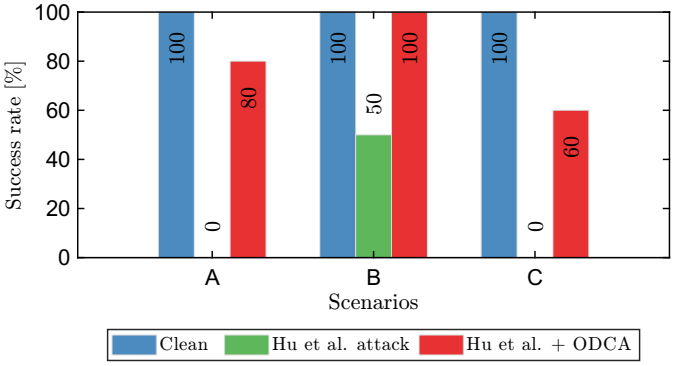


Fig. 9: **Scenario-level closed-loop stop success rates.** Success rates are reported for the three map configurations in Fig. 8 under the Hu *et al.* physical patch protocol [16]. The aggregate counts match Table V: Clean achieves 30/30 successful stops, the undefended patch attack succeeds in suppressing stops except for 5/10 trials in Scenario B, and ODCA recovers 8/10, 10/10, and 6/10 successful stops in Scenarios A–C, respectively.

LiDAR is not used as a direct substitute; instead, it acts as a consistency trigger, preserving nominal depth under agreement and activating repair only when necessary. Together with identity, consistency, and kinematic regularization ($\Delta d \approx -v\Delta t$), this yields stable and physically plausible recovery.

In closed-loop trials (Table V and Fig. 9), performance depends on visibility timing and reaction window. Residual failures (SCR $= 0.80$) occur mainly under the 3 s continuous attack regime, where attack persistence exceeds the effective causal recovery horizon. This reflects a reaction-time bound rather than instability of the repair mechanism. The current prototype operates offboard, and end-to-end latency may further reduce the effective repair window in short or turning approaches. Future work will integrate embedded deployment and explore complementary supervisory safety mechanisms to strengthen resilience under sustained attacks.

VII. Conclusion

We studied perception-layer resilience for autonomous driving under intermittent depth corruption and proposed **ODCA**, a conservative sensor substitution mechanism that preserves a control-critical distance signal for safe braking. Our approach leverages a pretrained time-series foundation model as a frozen temporal prior, capturing generic short-horizon dynamics without task-specific retraining. By learning only a

lightweight additive repair head on top of this foundation prior, ODCA enables correction-only-when-needed behavior through a depth–LiDAR inconsistency gate, while preserving nominal measurements under agreement. Experiments on the Quanser QCar 2 platform demonstrate consistent improvements over forecast-only baselines without degrading clean-condition performance. Closed-loop stop-sign trials with physical patch injection further validate improved stop compliance and reduced braking latency. Beyond ground vehicles, the proposed conservative cross-sensor substitution principle, grounded in foundation-model temporal priors and cross-sensor validation, provides a transferable blueprint for resilient perception in other safety-critical autonomous systems, such as unmanned aerial vehicle (UAV) collision avoidance.

REFERENCES

[1] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 2023.

[2] W. Zhu, X. Ji, Y. Cheng, S. Zhang, and W. Xu, "{TPatch}: A triggered physical adversarial patch," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 661–678.

[3] A. Wong, S. Cicek, and S. Soatto, "Targeted adversarial perturbations for monocular depth prediction," *Advances in neural information processing systems*, vol. 33, pp. 8486–8497, 2020.

[4] J. De Yeong, K. Panduru, and J. Walsh, "Exploring the unseen: A survey of multi-sensor fusion and the role of explainable ai (xai) in autonomous vehicles," 2025.

[5] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," in *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2017, arXiv:1712.09665.

[6] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] X. Liu, H. Yang, Z. Liu, L. Song, and H. Li, "Dpatch: An adversarial patch attack on object detectors," 2018, arXiv:1806.02299.

[8] C. Tsai, M. M. Abrar, and S. Hariri, "Security and resilience in autonomous vehicles: A proactive design approach," in *Connected, Cooperative, and Automated Mobility-Policy and Deployment Challenges, Threats, and Future Directions*. IntechOpen, 2026.

[9] C. Tsai and H. Salim, "Intelligent obstacle resilience in autonomous vehicles under security threats," in *2025 IEEE 12th International Conference on Cyber Security and Cloud Computing (CSCloud)*. IEEE, 2025, pp. 34–41.

[10] S. Verma, J. S. Berrio, S. Worrall, and E. Nebot, "Automatic extrinsic calibration between a camera and a 3d lidar using 3d point and plane correspondences," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 3906–3912.

[11] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[12] L. Han, X.-Y. Chen, H.-J. Ye, and D.-C. Zhan, "Softs: Efficient multivariate time series forecasting with series-core fusion," *Advances in Neural Information Processing Systems*, vol. 37, pp. 64 145–64 175, 2024.

[13] A. F. Ansari, O. Shchur, J. Küken, A. Auer, B. Han, P. Mercado, S. S. Rangapuram, H. Shen, L. Stella, X. Zhang *et al.*, "Chronos-2: From univariate to universal forecasting," *arXiv preprint arXiv:2510.15821*, 2025.

[14] M. Yaseen, "What is yolov8: An in-depth exploration of the next-generation object detector," *arXiv:2408.15857*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2408.15857

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[16] Y.-C.-T. Hu, B.-H. Kung, D. S. Tan, J.-C. Chen, K.-L. Hua, and W.-H. Cheng, "Naturalistic physical adversarial patch for object detectors," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7848–7857.

[17] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, and A. Dubrawski, "Nhits: Neural hierarchical interpolation for time series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 6, 2023, pp. 6989–6997.

[18] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 121–11 128.

[19] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.