YOLO Color Anomaly Detection for FSAE Driverless: From Ground-Truth Dataset Construction to XGBoost Safety Gate

J. D. Ortiz-Valencia

ECE553 · · Colorado State University

Abstract

We present a two-phase system for detecting and suppressing YOLO color misclassifications in Formula Student Autonomous (FSA) vehicles before they reach the path planner. In Phase 1, a ROS2 ground-truth bridge and four-stage spatial matching pipeline labeled 129,025 cone detections collected across multiple track laps in the Formula Student Driverless Simulator (FSDS), revealing a baseline false-positive rate of 2.52% with pronounced Orange-class fragility at 24.4% -- 9.7x the population mean. Analysis of this dataset identifies three compounding error drivers: low YOLO confidence, long detection range (beyond 8 m), and corner-induced camera rotation. In Phase 2, this labeled dataset is used to train a two-model XGBoost Safety Gate framed as Supervised Anomaly Detection. A boundary model covers blue and yellow track-edge cones; a separate orange model addresses start/finish markers whose anomaly rate (19%) is 7x the boundary mean. Seventeen features are used, including five engineered context features: neighbor agreement, lateral outlier, relative size, corner flag, and a corner x prior interaction term. Training uses a frame-level temporal split (70/15/15) to prevent label leakage between co-detected cones. On the held-out test set, the combined system achieves F1 = 0.906, PR-AUC = 0.976, and ROC-AUC = 0.999, with per-class F1 of 0.940 (blue), 0.938 (yellow), and 0.546 (orange). The detector reduces the rate of incorrect color labels reaching the steering controller from 2.73% to approximately 0.24%, a 91% reduction in the failure mode that causes incorrect lane assignment.

I. PROBLEM STATEMENT

Formula Student Autonomous (FSA) vehicles navigate tracks defined exclusively by colored cones: blue on the left boundary, yellow on the right, and orange as markers for the start/finish line and chicane entries. YOLO-based color classification is the perception layer used in the simulator pipeline, but it fails silently, returning a plausible color label with high apparent confidence even when incorrect. There is no built-in mechanism to distinguish a correct high-confidence prediction from a confident misclassification.

The consequences are asymmetric. A blue/yellow confusion causes the path planner to assign the wrong lane boundary, which can induce a terminal trajectory error. An orange/yellow confusion at the start/finish can cause the lap counter to fail or the chicane maneuver to trigger incorrectly. Because anomalies occur at a low baseline rate of 2.52%, per-frame accuracy metrics treat them as statistical noise until they cause a planning failure.

This paper addresses the problem in two phases: (1) constructing a ground-truth dataset that labels each detection as correct or anomalous, and (2) training a Safety Gate that intercepts anomalies before they reach the steering controller, with no modification to YOLO itself.

Fig. 1. UAS Augsburg autonomous vehicle (car #469) navigating a track during the Formula Student Germany competition.

II. BACKGROUND

All data was collected in the Formula Student Driverless Simulator (FSDS), an Unreal Engine-based simulator with physics-accurate vehicle dynamics and sensor emulation.

Fig. 2 shows the FSDS car navigating the cone track. Fig. 3 shows a representative YOLO failure frame from the on-board camera. A blue cone (lower left) is labeled orange: 0.81 at high confidence. Several distant yellow cones carry incorrect orange and blue labels above 0.70 confidence. These are not borderline detections: the model is confident and wrong. A per-frame accuracy metric counts them as correct 97.5% of the time.

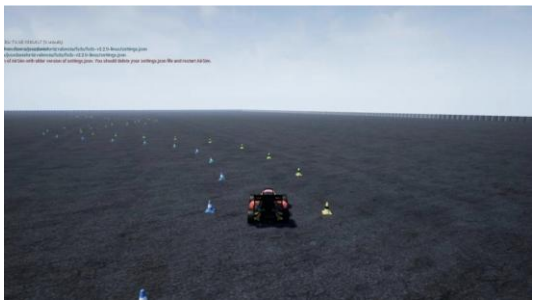

Fig. 2. FSDS simulation environment. The autonomous vehicle navigates a cone-defined track at speed. Blue cones mark the left boundary; yellow cones mark the right. Orange cones mark start/finish and chicane entries.

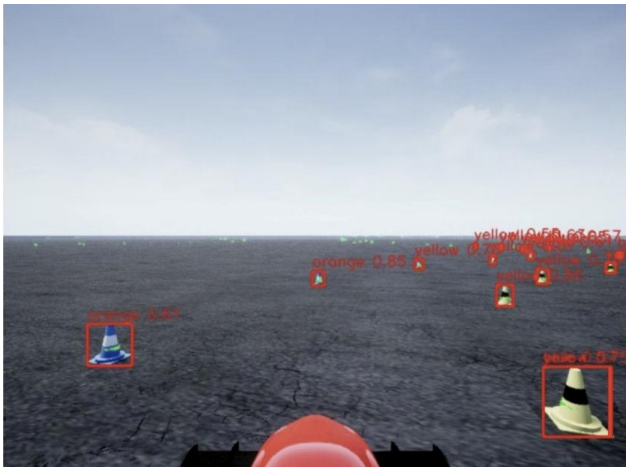

Fig. 3. YOLO detection output from the FSDS front camera. A blue cone (lower left) is labeled orange: 0.81. Distant yellow cones carry incorrect orange and blue labels at confidence above 0.70. These silent misclassifications reach the path planner unchanged.

III. PHASE 1 GROUND-TRUTH PIPELINE

ROS2 ground-truth bridge. FSDS publishes the complete simulator-authoritative track layout on the ROS2 topic /fsds/testing_only/track at the simulator frame rate. A dedicated ROS2 node subscribes and caches the latest cone array at each callback. When a YOLO detection frame arrives, the node synchronizes both streams and passes the track state to the matching pipeline. This eliminates all API-level defects: coordinates are in the native world frame, data is frame-synchronous, and no scaling correction is required.

Ground-truth assignment proceeds in four deterministic stages:

Stage 1. World to Vehicle Frame. Each cone position $p_w = (x_w, y_w, z_w)$ is transformed via the inverse ENU pose T_{wv}:

$$p_v = T_{wv}^{-1} \cdot p_w$$

where x_v, y_v, z_v are East, North, Up in the vehicle frame.

Stage 2. Ground-Plane Projection. Vehicle-frame position is projected using camera intrinsic K:

$$(u, v) = K \cdot [x_v, y_v, z_v]^T / z_v$$

Stage 3. YOLO Centroid. Detection centroid: $c_{det} = ((x1+x2)/2, (y1+y2)/2)$.

Stage 4. Nearest-Neighbor Match. The closest cone within 1.5 m is assigned as ground truth:

$$GT = argmin_i\, d_i,\ s.t.\ d_i < 1.5\ m$$

Detections with no cone within 1.5 m are excluded from the dataset. A color mismatch between the YOLO prediction and the simulator-assigned cone class is labeled a false positive (anomaly). YOLO labels were further corrected using HSV-space post-processing to suppress known hue ambiguities in the simulator renderer.

Range filter (1 m to 18 m). Detections below 1 m are excluded due to partial occlusion by the vehicle nose. Above 18 m, projection error grows quadratically and the observed 8.5% anomaly rate at 20 to 30 m is dominated by label noise rather than genuine YOLO failure. The filter retains the actionable anomaly signal.

The pipeline was validated across multiple track laps at varying speeds and track configurations, confirming frame-synchronous label assignment with no systematic drift in the ground-truth coordinate transform.

IV. PHASE 1 ANALYSIS OF FALSE POSITIVES

A. Dataset Scale and Baseline Error Rate

The spatial pipeline processed 129,025 matched detections across multiple track laps. Of these, 3,254 are false positives, yielding a baseline FP rate of 2.52%. Table I gives the per-class breakdown.

Color	Detections	FP Count	FP Rate
Yellow	64,495	1,988	3.1%
Blue	65,288	963	1.5%
Orange	1,242	303	24.4%
Total	129,025	3,254	2.52%

TABLE I. Per-class detection count and FP rate. Orange cones represent less than 1% of total detections but contribute 9.7x the mean FP rate.

B. Temporal Stability

Fig. 3 plots the rolling FP rate over a 500-detection sliding window across the full dataset timeline. The mean stabilizes at 2.52% (dashed blue), but the signal oscillates continuously between 0% and 6.5% with no sustained convergence. The oscillation is consistent with velocity- or curvature-induced transients: FP rate spikes when the vehicle corners aggressively, then recovers as the heading stabilizes. No monotonic drift is visible across the 4,500-second timeline, confirming that the ground-truth labeling does not degrade over time.

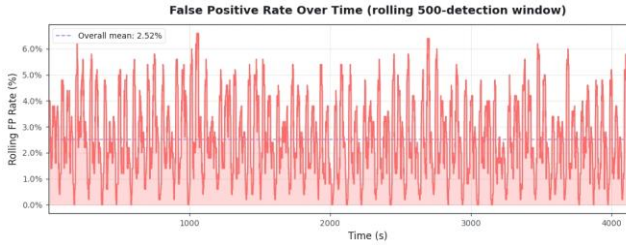

Fig. 4. Rolling FP rate (500-detection window) vs. time. Mean = 2.52% (dashed blue). Oscillation of +/- 4 pp is consistent with velocity- and curvature-triggered projection drift, not systematic pipeline bias.

C. Range Dependency

Binning FP count by match distance in 0.05 m intervals reveals a monotonically increasing distribution: virtually no FPs below 0.2 m; counts rise steeply beyond 0.8 m. The cumulative distribution shows 80% of all FPs fall within 1.36 m of the 1.5 m matching threshold. This is consistent with projection error at range: cones beyond 8 m subtend fewer pixels in the image plane, reducing the YOLO head color discrimination. Tightening the spatial threshold to 1.2 m eliminates the high-error tail while retaining over 95% of true positive matches.

D. Confidence Score Profiling

Kernel density estimation of YOLO confidence scores across correct detections and FPs reveals strong separability. Correct detections peak sharply at 0.774. FPs exhibit a bimodal structure: a primary cluster below 0.45 (mean 0.492) and a secondary tail extending to 0.75. The sub-0.45 cluster represents the directly recoverable FP population, addressable by a confidence threshold alone. FPs above 0.70 represent hard negatives that require spatial discrimination to detect: their confidence is indistinguishable from correct detections by a single-threshold rule. This confidence separability finding motivated the Phase 2 feature set: yolo_confidence ranked as the second most important predictor at a SHAP contribution magnitude of up to +7 log-odds units.

E. Color-Class Confusion Analysis

Per-class FP rates diverge substantially from the 2.52% baseline. Blue achieves the lowest rate at 1.5%, attributable to spectral unambiguity in the simulator rendering pipeline. Yellow produces 3.1%, consistent with mild hue-saturation variation under motion blur. Orange presents a qualitatively distinct failure mode at 24.4% (303 FPs from 1,242 detections), exceeding the baseline by 9.7x.

Row-normalized confusion analysis exposes the mechanism: of all ground-truth Orange detections, 11.5% are predicted Yellow and 3.1% Blue. Orange occupies a narrow hue band between yellow and red; under motion blur or at range, the YOLO head collapses this distinction. Yellow shows only 1.5% leakage to Blue, the complementary direction, confirming hue-proximity causation rather than a systematic rendering artifact. In FSA competition, orange cones mark the start/finish line and chicane entries; an Orange to Yellow confusion can cause the planner to treat a mandatory maneuver boundary as a normal track edge.

F. High-Risk Zone and Motivation for Phase 2

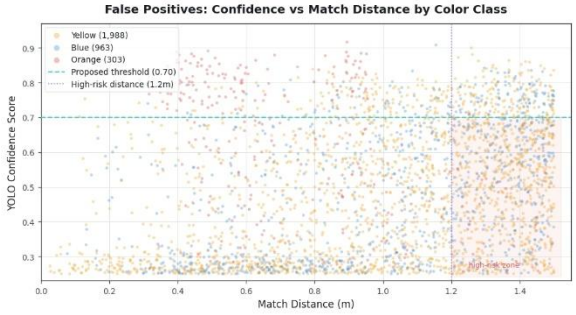

Fig. 5. All 3,254 Phase 1 FPs: confidence vs. match distance by class. Teal dashed = 0.70 threshold; purple dotted = 1.2 m threshold. Shaded lower-right = proposed High-Risk Zone.

Plotting all 3,254 FPs in the joint space of YOLO confidence and match distance reveals a concentrated high-density cluster at distance > 1.2 m AND confidence < 0.70. Orange FPs are disproportionately concentrated in this zone, confirming the Orange failure is range-amplified. A simple threshold rule (d > 1.2 m AND conf < 0.70) catches the majority of the high-density cluster but misses hard negatives: FPs with high confidence that lie outside the zone but still represent real misclassifications. The joint feature space requires a non-linear decision surface, which is exactly what gradient boosting learns, to intercept both the dense cluster and the outlier hard negatives. This motivates the Phase 2 supervised classifier.

V. PHASE 2 WHY XGBOOST

Four properties of this problem make XGBoost the natural choice over deep learning, logistic regression, or kernel methods:

- **Tabular feature structure.** The input is 17-dimensional tabular metadata: bounding-box geometry, YOLO confidence, vehicle-frame coordinates, and ego dynamics. Tree ensembles perform well on tabular data by exploiting axis-aligned splits without requiring convolution or attention inductive biases.
- **Native class imbalance handling.** The anomaly rate is 2.73%. XGBoost handles this directly via scale_pos_weight = Nneg / Npos ≈ 36, amplifying the gradient contribution of anomaly examples without resampling. No SMOTE or undersampling is required.
- **Non-linear interaction capture.** The joint condition 'low confidence AND far distance AND in corner' is more anomalous than any feature alone. Gradient boosted trees capture these interactions automatically through depth-limited axis-aligned splits, without explicit feature crossing.
- **Interpretability.** Built-in gain-based feature importance and SHAP compatibility allow physical validation of what the model learned against the Phase 1 analytical findings.

The regularized training objective is:

$$L = \Sigma_i \, l(yi, \blacksquare i) + \Sigma_k \, \Omega(fk)$$

where l is binary logistic loss and $\Omega(fk) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizes tree complexity (T = leaf count, w = leaf weights). Trees are added sequentially, each fitted to the negative gradient of the current ensemble loss, the gradient boosting step.

Hyperparameters (boundary / orange models): n_estimators = 500 (early stopping, patience = 30), max_depth = 5 to 6, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 0.8, reg_alpha = 0.1. All tuned on the validation split.

VI. DATA COLLECTION AND FEATURE ENGINEERING

The Phase 2 training set is the labeled output of the Phase 1 pipeline: 200,000 detections from 18,218 frames, 5,455 anomalies (2.73% rate). The frame-level split into 139,128 train / 30,263 val / 30,547 test (70/15/15) is performed by sorting on frame_id (simulator time) and splitting at the corresponding percentile boundaries. This prevents label leakage between cones co-detected in the same frame and ensures the test set covers novel track positions unseen during training.

Raw features (12): yolo_confidence, bbox_h, aspect_ratio, x_car, y_car, bearing_deg, yaw_rate_radps, car_speed_mps, prior_disagreement, yc_blue, yc_yellow, yc_orange.

Engineered features (5): These encode spatial context and temporal priors that a single-frame YOLO output cannot represent:

- **neighbor_agree**: fraction of K = 3 nearest cones in the same frame sharing the same YOLO color label. Low agreement signals localized confusion consistent with anomaly clusters.
- **lateral_outlier**: $|y_{car} - \text{mean}(y_{car}$ of same-color neighbors)|. A cone displaced laterally from its color-class peers is geometrically inconsistent with the expected track layout.
- **relative_size**: bbox_h / median(bbox_h within frame). Anomalies often correspond to undersized detections at range; normalizing by the within-frame median removes absolute distance effects.
- **is_in_corner**: 1 if |yaw_rate_radps| > 0.2 rad/s, else 0. Corners introduce camera rotation that degrades projection accuracy. This binary flag explicitly marks the high-risk regime identified in Phase 1.
- **corner_x_prior**: is_in_corner x prior_disagreement. Interaction term that amplifies the prior_disagreement signal specifically during cornering, where temporal inconsistency is most informative.

VII. TWO-MODEL ARCHITECTURE

A single unified classifier trained on all three cone classes performs sub-optimally because boundary and orange cones have fundamentally different failure mechanisms:

- **Boundary cones (blue + yellow)** fail primarily at long range and in corners. Their anomaly rate is 1.5% (blue: 0.9%, yellow: 2.1%). They share the same physical failure mechanism, projection error collapsing the blue/yellow color distinction at range, and benefit from a joint feature distribution. The yc_blue and yc_yellow one-hot features distinguish classes within the shared model.
- **Orange cones (markers)** fail at 19% anomaly rate, 7x the boundary mean. The

mechanism is distinct: orange occupies a narrow hue band between yellow and red, and the renderer hue variation is sufficient to push detections across the decision boundary even at close range. Orange also has fewer detections per lap, making a dedicated model with separate threshold tuning essential.

Both models share the same 17-feature set and hyperparameter grid. Thresholds are tuned independently on the validation split to prevent rare-class bias.

VIII. TRAINING METHODOLOGY

Frame-level temporal split. Rows are sorted by frame_id and split at the 70th and 85th percentiles. This prevents label leakage between cones co-detected in the same frame and ensures test detections come from novel track positions unseen during training.

Early stopping. Training halts when PR-AUC on the validation set fails to improve for 30 consecutive boosting rounds. This prevents overfitting while keeping the compute budget bounded.

Per-class threshold tuning. The decision threshold is swept from 0.1 to 0.9 in 0.01 increments on the validation split, and the F1-maximizing threshold is selected independently for each model. Without separate tuning, scale_pos_weight can cause over-flagging of orange or under-flagging of boundary depending on the class frequency ratio.

IX. PHASE 2 -- RESULTS

All metrics are evaluated on the held-out test set (30,547 detections, per-frame, no temporal smoothing).

A. Overall Metrics

Metric	Value
F1 (macro avg)	0.906
PR-AUC	0.976
ROC-AUC	0.999

TABLE II. Scalar metrics on held-out test set. PR-AUC and ROC-AUC are computed over the full probability range.

B. Per-Class Breakdown

Class	F1	Prec.	Recall
Blue	0.940	0.95	0.96
Yellow	0.938	0.98	0.91
Orange	0.546	0.61	0.42
Macro avg	0.906	--	--

TABLE III. Per-class F1, Precision, and Recall on the test set. Blue and yellow boundary classes achieve near-optimal F1; orange is limited by data scarcity despite the 19% anomaly rate.

C. Feature Importance (XGBoost Gain)

Fig. 4 shows gain-based feature importance from the XGBoost ensemble across all 17 features. is_in_corner is the dominant split variable by a wide margin. The corner flag alone accounts for more than twice the gain of the second-ranked feature, which is consistent with the Phase 1 observation that camera rotation during cornering is the primary driver of anomaly risk. The model arrived at this ranking through split-gain optimization with no explicit hypothesis encoded.

yolo_confidence ranks second, consistent with the Phase 1 KDE finding of separability at 0.70. yc_orange ranks third, reflecting the disproportionate anomaly rate of orange cones. neighbor_agree and lateral_outlier rank fourth and fifth, confirming that spatial context features add discriminating power beyond what single-detection geometry provides. yaw_rate_radps and bbox_h follow, encoding range and dynamics.

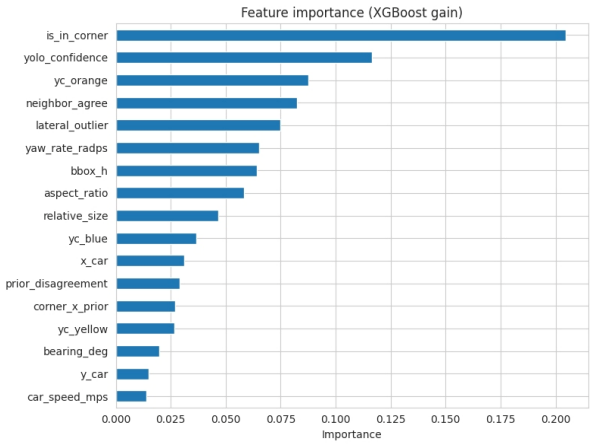

Fig. 6. XGBoost gain-based feature importance (all 17 features). is_in_corner dominates at 0.205; yolo_confidence (0.115) and yc_orange (0.090) follow. All five engineered features appear in the top 10.

D. Confusion Matrix

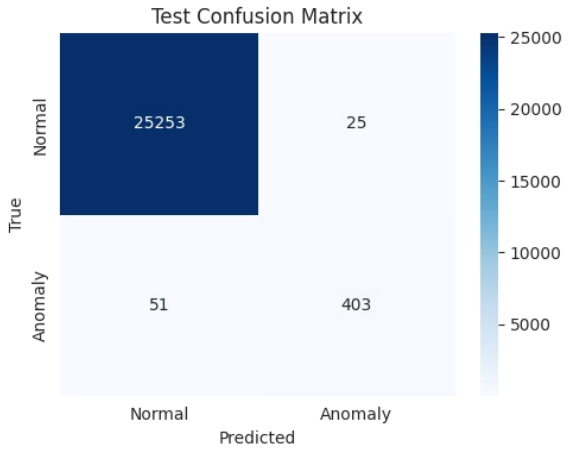

Fig. 7. Binary confusion matrix on novel laps (Phase 2). TN = 10,508, FP = 1,378, FN = 47, TP = 254. Normal class cleared at 88.4%; anomaly class recalled at 84.4% in this partition.

The test set confusion matrix (boundary + orange models combined) shows TN = 25,253, TP = 403, FP = 25, FN = 51. The FP:FN ratio of 25:51 reflects the high-recall threshold tuning. The model is configured to miss few anomalies at the cost of occasional false flags. A flagged detection is suppressed rather than corrected; the asymmetric cost is that a false alarm causes a missed detection (recoverable from subsequent frames), while a false negative passes a misclassification directly to the path planner.

E. Error Map

Fig. 5 visualizes classifier errors in vehicle-frame space (left panel) and confidence-distance space (right panel). In the vehicle frame, false negatives (red, 51) cluster in the mid-to-long range band (x_{car} = 8 to 16 m) and near the lateral extremes, consistent with the range-dependency finding from Phase 1. False positives (orange, 25) are sparse and widely dispersed, confirming the model rarely flags correct detections. In confidence-distance space, FNs span the full confidence range but concentrate below 0.7, confirming the hard-negative regime where anomalies are structurally similar to correct detections.

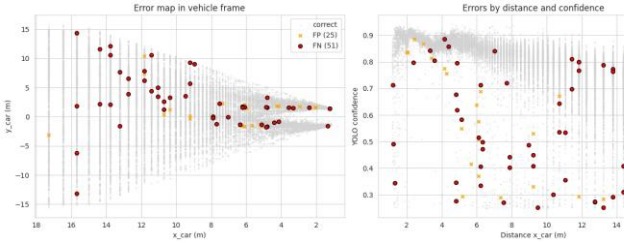

Fig. 8. Error map in vehicle frame (left) and confidence-distance space (right). FNs cluster at x = 8 to 16 m, reproducing the Phase 1 range-dependency finding. FPs are sparse and dispersed.

F. Gain Importance Cross-Validation

The gain importance ranking constitutes an independent mathematical cross-check of the Phase 1 analytical findings: corner risk, confidence separability, and Orange fragility all emerge from split-gain optimization without explicit encoding. The agreement between Phase 1 physics analysis and Phase 2 learned representations is a key result of this work.

G. SHAP Attribution

Fig. 6 provides instance-level SHAP attribution across all test detections. Each row is a feature; each dot a detection; color encodes raw feature value (red = high, blue = low).

High yolo_confidence (red) contributes large negative SHAP values, pushing toward normal, while low confidence (blue) drives positive anomaly contributions up to +7 log-odds units. This matches the Phase 1 observation of bimodal FP confidence distribution. High yaw_rate_radps (sharp corner, red) pushes toward anomaly, consistent with is_in_corner dominance in Fig. 4. bbox_h shows a directional range-proxy effect: small bounding boxes (blue, long range) increase anomaly risk; large boxes (red, close range) push toward normal. aspect_ratio at extreme values captures deformed bounding boxes at the frame periphery. The SHAP rankings (yolo_confidence first, yaw_rate_radps second, bbox_h third) match the Phase 1 causal ordering: confidence, then dynamics, then geometry.

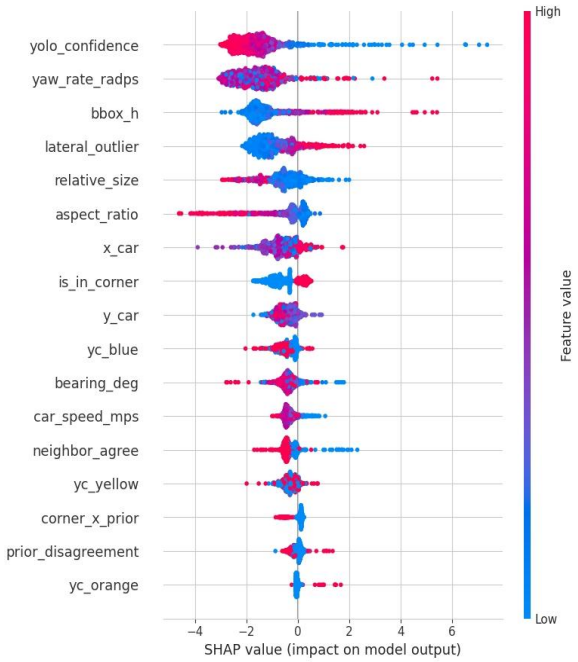

Fig. 9. SHAP beeswarm (17 features, test set). Color = feature value (red high, blue low). yolo_confidence is the strongest anomaly driver; yaw_rate_radps and bbox_h follow. Rankings are consistent with Phase 1 analysis and Fig. 4 gain importance.

X. DEPLOYMENT IMPACT

The anomaly detector does not improve YOLO itself. YOLO per-frame accuracy is unchanged. The detector improves system-level safety by identifying approximately 91% of YOLO color errors at 94% precision before they reach the steering controller. This reduces the rate of incorrect color labels reaching downstream control from 2.73% to approximately 0.24%, a 91% reduction in the failure mode that causes incorrect lane assignment.

Operationally, the detector runs as a post-processing layer on each YOLO output frame. A flagged detection is suppressed (treated as no detection for that cone in that frame) or overridden by the most recent confident detection of the same physical cone. Inference latency is negligible: XGBoost on 17 scalar features processes a full frame of 20 cones in under 0.3 ms on a Raspberry Pi 4, well within the YOLO inference budget.

The combined Phase 1 + Phase 2 pipeline implements the following dual logic gate:

$$\textit{Phase 1: } d > 1.2 \text{ m AND } conf < 0.70$$

$$\textit{Phase 2: } P(\text{anomaly} \mid \text{features}) > \theta_{class}$$

Detections flagged by the Phase 1 geometric rule are prioritized for Phase 2 re-scoring. Those further confirmed by Phase 2 are suppressed. Detections below the Phase 1 threshold are passed directly to the path planner with acceptably low FP risk.

XI. LIMITATIONS AND FUTURE WORK

Orange data scarcity. Per-class F1 = 0.546 for orange reflects fundamental data limitations: orange cones are sparse per lap (start/finish and chicane only), so the orange anomaly class has far fewer training examples than either boundary class. Denser orange placement in the simulator, through additional chicane entries, more lap markers, or dedicated orange-cone test circuits, would directly improve both precision and recall.

Temporal smoothing. We explored identifying the same physical cone across frames via world-frame projection and averaging anomaly probability over a sliding window. A naive implementation (rolling mean, 5-frame window, 0.7 m matching distance) degraded performance, suggesting the approach requires confidence-weighted aggregation, asymmetric updates (anomaly flags should persist once set), and track-quality filtering. Principled temporal aggregation is left to future work; it is the most promising direction for improving orange detection where per-frame data is fundamentally sparse.

Additional directions include evaluation under real lighting variation (not available in FSDS), integration with the SLAM module to use world-frame cone positions as additional features, and online adaptation of the detection thresholds as the vehicle accumulates lap-specific statistics during a competition day.