# CAESAR++: Uncertainty-Driven Contextual Reasoning for Trustworthy and Explainable Road Object Detection

Anh-Thu Mai[1,2], Marina Nicolas[1], Patricia Ladret[2], Alice Caplier[2]

[1] STMicroelectronics, Grenoble, France
[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, France

anh-thu.mai@st.com

## Résumé

*Cet article présente CAESAR++, une approche pour la détection d'objets routiers qui combine la prédiction conforme, un raisonnement contextuel adaptatif et des cartes de saillance à double couleur. CAESAR++ calibre d'abord les incertitudes de classification et de localisation au moyen d'une procédure conforme en deux étapes, puis agrandit dynamiquement la fenêtre de contexte autour de chaque détection en fonction de son niveau d'incertitude, et produit enfin des explications au niveau de l'objet qui distinguent les indices sensoriels locaux des indices contextuels. Les expériences montrent des gains constants en précision, en calibration de l'incertitude et en stabilité des explications, sans réentraîner les modèles de base.*

## Mots-clés

*Intelligence artificielle explicable, détection d'objets routiers, prédiction conforme, raisonnement contextuel.*

## Abstract

*This paper introduces CAESAR++, a framework for road object detection that combines conformal prediction, adaptive contextual reasoning, and dual-color saliency maps. CAESAR++ first calibrates classification and localization uncertainty using a two-step conformal procedure, then enlarges the context window around each detection in proportion to its uncertainty, and finally produces object-wise explanations that disentangle bottom-up sensory evidence from top-down contextual cues. Experiments indicate consistent improvements in detection accuracy, uncertainty calibration, and explanation stability without retraining the base models.*

## Keywords

*Explainable artificial intelligence, road object detection, conformal prediction, contextual reasoning.*

## 1 Introduction

Reliable perception is a central requirement for autonomous vehicles and advanced driver-assistance systems. In real-world urban traffic, detectors must recognize and localize a variety of objects despite occlusions, adverse weather, cluttered backgrounds and small apparent sizes. Recent deep learning detectors have made considerable progress [1], yet they often remain overconfident and brittle under challenging conditions in real-world scenarios [2].

Two aspects are particularly critical for trustworthy perception. First, the system should provide well-calibrated uncertainty on both class labels and bounding boxes so that downstream modules can take risk-aware actions. Second, the system should expose human-understandable explanations that reveal which visual cues drive each detection, in order to support debugging and user trust [3].

Existing approaches usually address these aspects separately. Uncertainty estimation methods, including Bayesian approximations and ensembles [4, 5], improve calibration but rarely show how to act on the estimated uncertainty. Conformal prediction provides distribution-free coverage guarantees [6, 7], and recent work has adapted it to detection [8, 9]. However, these methods are often limited to reporting coverage statistics, without using uncertainty to guide refinement or explanation.

Conversely, explanation techniques for object detectors [10] usually rely on local visual features and fixed context. Current systems often either waste computation on easy cases or leave false detections unresolved.

This leads to a persistent gap between reliability and explainability in safety-critical settings, where these two capabilities are not fully unified within a single framework.

This work proposes CAESAR++, which extends our previous context-aware explanation framework CAESAR [11] into an integrated system that:

— produces statistically valid uncertainty estimates on labels and bounding boxes;

— uses these estimates to adapt the amount of contextual information considered for each detection;

— generates dual-color saliency maps that clearly separate local features from contextual reasoning;

Although the proposed methodology is mainly designed for urban road scenes with complex interactions and challenging conditions, we also evaluate its generalization on diverse datasets under controlled out-of-context settings to demonstrate its robustness beyond the target domain.

This paper is a conference-length version of a manuscript submitted to *Neurocomputing* (Elsevier), currently under review after revision.

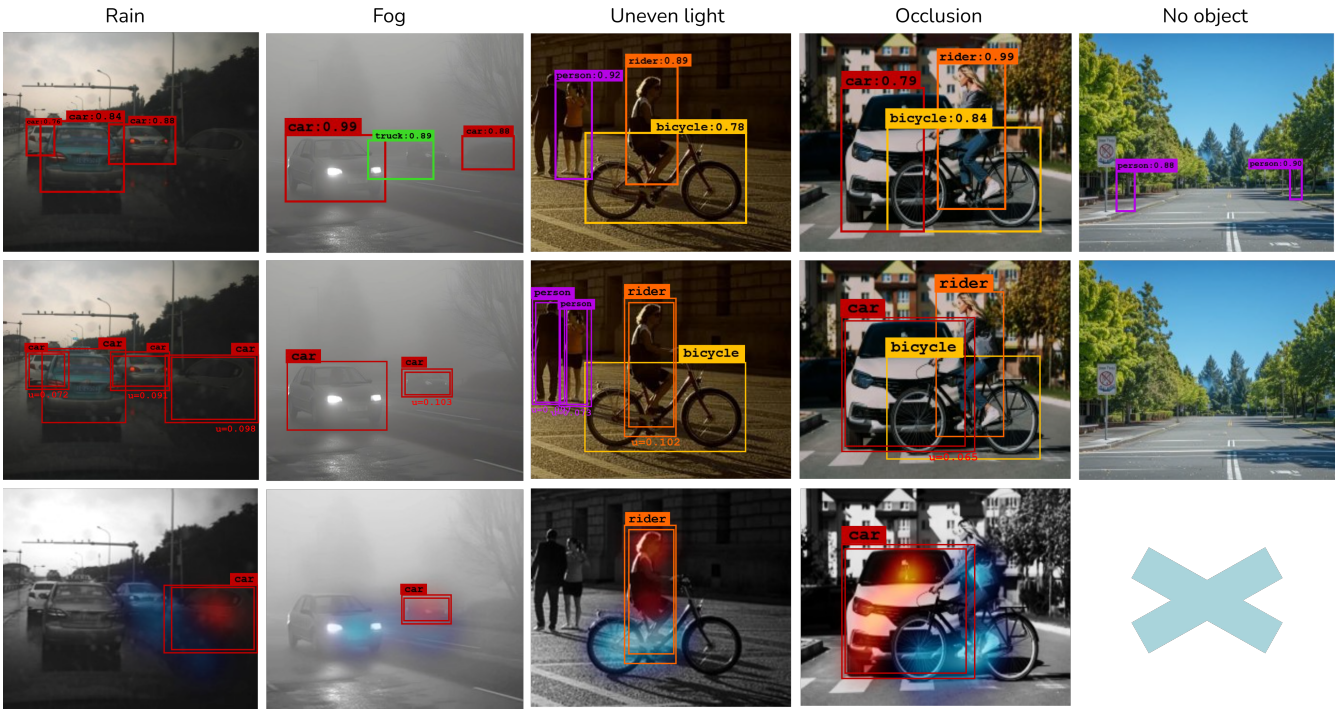


Figure 1: Overview of CAESAR++ main contributions. Top row: baseline detections in challenging urban conditions, with missed or mislocalized objects. Middle row: corresponding refined detections produced by adaptive contextual reasoning. Bottom row: dual-color saliency maps where red highlights sensory evidence and blue highlights contextual cues.

**Contributions.** The main contributions are as follows:

— A detector-agnostic pipeline that combines a two-step conformal prediction procedure [9] with context-driven refinement for road object detection.

— An end-to-end contextual reasoning mechanism that adaptively expands the spatial window around each detection proportionally to its calibrated uncertainty.

— A dual-color saliency scheme based on Grad-CAM++ [12, 13] that disentangles bottom-up cues from top-down context, providing instance-wise explanations with improved visual clarity.

Figure 1 illustrates consistent improvements obtained by CAESAR++ over a baseline detector in challenging scenes. Overall, the framework enhances detection performance while also making the decision process more transparent.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the CAESAR++ framework. Section 4 reports the main experimental results. Section 5 provides additional insights and implications. Section 6 concludes the paper and outlines future work.

## 2 Related work

**Uncertainty in object detection.** Uncertainty estimation is essential to detect overconfident errors in safety-critical applications. Bayesian formulations and dropout-based approximations [4] offer principled tools to estimate epistemic and aleatoric uncertainty, while deep ensembles provide strong empirical calibration [5]. Conformal prediction offers an alternative that yields finite-sample coverage guarantees under mild assumptions [6, 7]. Recent works adapt conformal prediction to object detection [8, 9], with separate calibration of classification and localization. These studies mainly focus on coverage metrics and interval width, and do not exploit uncertainty to steer context usage or explanation.

**Explainability for detectors.** Gradient-based explanation methods such as Grad-CAM and Grad-CAM++ [12, 13], and perturbation-based approaches such as RISE [14], have become standard for visualizing the internal reasoning of convolutional networks. Several extensions build object-specific saliency maps for detectors [15, 16, 17], and human attention has been proposed as an additional signal [18]. These works considerably improve instance-level interpretability, but largely treat all detections in the same manner and use fixed context ranges. As a result, explanations can become noisy on ambiguous cases and waste computation on very confident detections.

**Contextual reasoning and computation strategy.** Human observers strongly rely on scene context when local evidence is insufficient [19]. Modern detectors incorporate context via attention mechanisms and relation modules [20, 21], which improves robustness in cluttered scenes. In parallel, dynamic computation strategies adjust the amount of processing to the difficulty of the input [22, 23]. However, most context modules are integrated into the detector architecture and require retraining. CAESAR++ instead provides a detector-agnostic, post-hoc mechanism that triggers context expansion and refinement.

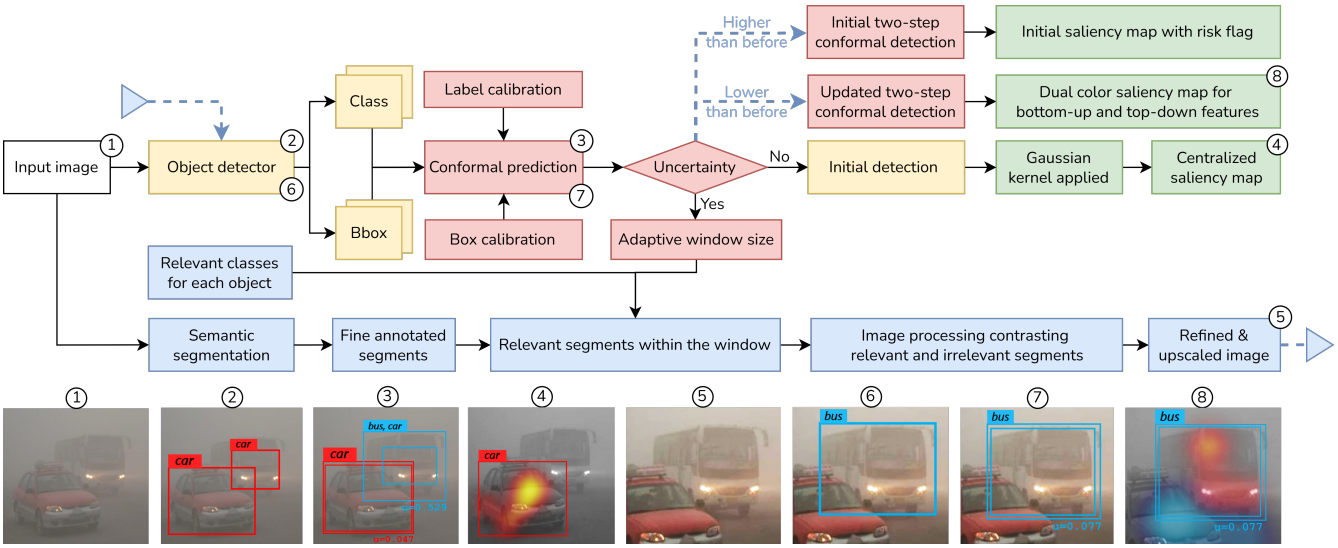


Figure 2: Overall CAESAR++ pipeline showing the successive stages from initial detection (yellow) to uncertainty estimation (red), contextual refinement (blue), and post-hoc explainability (green). The detector first generates predictions; conformal calibration then produces uncertainty scores that guide adaptive context expansion using semantic segmentation; the refined patch is reprocessed, and dual-color saliency maps are finally generated. Numbered images illustrate corresponding steps.

# 3 CAESAR++ framework

Figure 2 summarizes the end-to-end workflow of our proposed method. The following subsections provide a detailed description of each main component in turn.

## 3.1 Context modeling from segmentation

CAESAR++ builds on the context computation strategy introduced in our prior work CAESAR [11]. To model the typical contextual environment associated with each road object category, we employ a lightweight semantic segmentation network tailored to our use case. Specifically, we use DeepLabV3+ [24] with a MobileNetV2 backbone [25], pretrained on MS COCO and fine-tuned on Cityscapes [26]. The model is evaluated on selected images from TJU-DHD-Traffic [27], BDD100K [28], and Pascal VOC [29]. For each instance of a road object class, we define a local window centered at its centroid and count the distinct semantic classes intersecting this window. The resulting counts are normalized over the dataset, and for each road object category we retain the most frequent context classes as well as co-occurring road objects. Table 1 reports the resulting context sets. Road objects themselves are also considered as potential context for other road objects, reflecting the structured nature of traffic scenes [19]. This design deliberately focuses on meaningful in-domain context modeling, rather than on out-of-distribution scenarios beyond the scope of this work.

## 3.2 Two-step conformal calibration

For each image in a calibration set, the base detector outputs candidate bounding boxes and class probabilities. Based on the procedure adopted by [9], CAESAR++ derives nonconformity scores for labels and bounding boxes and computes empirical quantiles that will be used at test time.

Table 1: Relevant context classes for each road object category. Context categories are indicated in italics and co-occurring road objects are underlined.

| Road Object | Relevant Classes |
|---|---|
| Person | *road*, *sidewalk*, *pole*, person, rider, bicycle |
| Rider | *road*, *sidewalk*, *wall*, rider, bicycle, motorcycle |
| Car | *road*, *traffic sign*, *pole*, person, rider, car |
| Truck | *road*, *traffic light*, *building*, car, truck, bus |
| Bus | *road*, *traffic sign*, *building*, car, truck, bus |
| Tram | *road*, *fence*, *vegetation*, person, car, bicycle |
| Motorcycle | *road*, *wall*, *building*, person, rider, motorcycle |
| Bicycle | *road*, *sidewalk*, *wall*, person, rider, bicycle |

For labels, the nonconformity score is defined as one minus the predicted probability of the true class. The empirical quantile at level $(1-\alpha)$ defines a threshold; at test time, the conformal label set for a given detection contains all classes whose complement probability does not exceed this threshold, to yield marginal coverage guarantees on labels.

For bounding boxes, we use conformalized quantile regression [30], where each coordinate is predicted with lower and upper quantiles, and the nonconformity score is defined as the largest deviation outside this predictive interval. The corresponding quantile then determines a symmetric expansion of the predicted box, yielding a conformal prediction set for localization.

To drive context adaptation, CAESAR++ introduces a normalized uncertainty score $u$ in $[0,1]$ by aggregating label and box nonconformity, which is then used to derive the adaptive threshold for context expansion. Low values indicate confident detections, whereas high values correspond to ambiguous cases.


Figure 3: From left to right: baseline detections, refined detections with CAESAR++, and dual-color explanations for individual objects. Red indicates sensory evidence, blue indicates contextual support. Uncertainty scores omitted for clarity.

### 3.3 Adaptive context selection

Given the conformal outer box and the normalized uncertainty score of a detection, CAESAR++ determines how much surrounding context should be incorporated. Confident detections are left unchanged, whereas uncertain ones are reprocessed with a larger contextual region.

The context window is expanded using a simple linear rule that progressively enlarges the region from the outer box toward the full image as uncertainty increases. This choice provides a direct and monotonic relationship between uncertainty and context expansion, while remaining easy to interpret and free of additional hyperparameters. We also evaluated nonlinear mappings, but we retained the linear formulation for its simplicity and stable behavior.

To avoid unnecessary reprocessing, CAESAR++ further relies on a data-driven certainty criterion derived from the recent distribution of uncertainty scores [31]. Detections whose uncertainty falls below a low adaptive threshold are processed directly with the original box, while the remaining detections undergo contextual enhancement. Within the selected region, semantic masks preserve the relevant classes listed in Table 1 and suppress irrelevant areas before the refined patch is fed back to the detector.

### 3.4 Dual-color saliency maps

To explain individual detections, CAESAR++ constructs two complementary saliency maps per object using Grad-CAM++ [12, 13] as the baseline approach.

A bottom-up map is obtained by applying Grad-CAM++ to the original detection without context expansion. It captures the direct sensory evidence for the prediction. A second map is computed for the refined detection after context refinement. The difference between the refined and original maps highlights top-down contextual cues.

Bottom-up saliency is visualized in red and top-down contribution in blue. For detections classified as certain, only the bottom-up component is displayed, smoothed by a Gaussian kernel [32] to enhance spatial coherence. For uncertain detections, both components are shown, providing a clear view of how context modifies the decision. As illustrated in Fig. 3, this representation improves interpretability over the initial detector output. It offers a compact and informative view of the interplay between appearance and context, helping to understand both the detection update and the underlying decision mechanism.

## 4 Results

### 4.1 Experimental setup

We evaluate CAESAR++ on three datasets: TJU-DHD-Traffic [27], BDD100K [28] and Pascal VOC 2012 [29]. Each contributes 5000 validation images. TJU-DHD-Traffic contains diverse driving scenes with strong environmental variations. BDD100K covers dense urban environments. Pascal VOC includes a broader set of scenes and objects, which is less specific to driving but useful to test generalization across various scenarios.

We consider six detectors representing one-stage, two-stage and transformer-based paradigms, including YOLOv8m [33], IA-YOLO [34], EfficientDet-D4 [35], L-SSD [36], RT-DETR-R50 [37], and Faster R-CNN [38]. CAESAR++ is applied as a post-processing module without retraining.

Detection performance is measured with mAP@0.5:0.95 and mAP@0.5, false negative rate (FNR), false discovery rate (FDR), F1-score and frames per second (FPS). Uncertainty quality is assessed through label and box coverage and mean prediction interval width [7]. Explanations are evaluated with Deletion and Insertion metrics [14], the Energy-Based Pointing Game (EBPG) [39], and Average Stability (AvgS) for robustness [40].

Table 2: Detection results comparison on TJU-DHD-Traffic validation set with and without CAESAR++ (95% CI). Metrics reported: mAP@0.50:0.95 (↑), mAP@0.5 (↑), False Negative Rate (FNR, ↓), False Discovery Rate (FDR, ↓), F1-Score (↑), and FPS (↑). Better results are **bolded**.

| Detector brief description | Method | mAP@0.50:0.95 | mAP@0.5 | FNR ↓ | FDR ↓ | F1-Score ↑ | FPS ↑ |
|---|---|---|---|---|---|---|---|
| One-stage, anchor-based | (1) YOLOv8m | $50.8 \pm 1.4$ | $73.4 \pm 1.5$ | $8.2 \pm 0.8$ | $7.8 \pm 0.5$ | $78.2 \pm 1.7$ | $58.1 \pm 0.6$ |
| | **(1) + CAESAR++** | $\mathbf{53.7 \pm 0.7}$ | $\mathbf{79.1 \pm 0.8}$ | $\mathbf{4.8 \pm 0.4}$ | $\mathbf{4.2 \pm 0.3}$ | $\mathbf{82.9 \pm 0.7}$ | $52.8 \pm 1.8$ |
| One-stage, anchor-based, image adaptive processing | (2) IA-YOLO | $52.2 \pm 1.2$ | $75.1 \pm 1.4$ | $5.1 \pm 0.8$ | $5.4 \pm 0.7$ | $80.5 \pm 1.7$ | $79.7 \pm 1.2$ |
| | **(2) + CAESAR++** | $\mathbf{54.8 \pm 0.8}$ | $\mathbf{80.3 \pm 0.7}$ | $\mathbf{2.7 \pm 0.3}$ | $\mathbf{2.2 \pm 0.2}$ | $\mathbf{84.7 \pm 0.8}$ | $75.8 \pm 2.1$ |
| One-stage, EfficientNet architecture | (3) EfficientDet-D4 | $52.5 \pm 1.1$ | $75.8 \pm 0.9$ | $7.2 \pm 0.7$ | $4.9 \pm 0.5$ | $80.8 \pm 1.2$ | $35.8 \pm 0.5$ |
| | **(3) + CAESAR++** | $\mathbf{55.4 \pm 0.7}$ | $\mathbf{81.2 \pm 0.5}$ | $\mathbf{4.3 \pm 0.4}$ | $\mathbf{2.1 \pm 0.3}$ | $\mathbf{84.1 \pm 0.5}$ | $32.1 \pm 1.1$ |
| One-stage, anchor-based, lightweight | (4) L-SSD | $46.9 \pm 1.0$ | $71.4 \pm 0.8$ | $8.7 \pm 0.6$ | $4.9 \pm 0.5$ | $76.9 \pm 0.8$ | $99.9 \pm 0.7$ |
| | **(4) + CAESAR++** | $\mathbf{50.4 \pm 0.6}$ | $\mathbf{77.1 \pm 0.4}$ | $\mathbf{5.4 \pm 0.4}$ | $\mathbf{2.8 \pm 0.2}$ | $\mathbf{80.3 \pm 0.4}$ | $94.3 \pm 1.6$ |
| Transformer-based, real-time end-to-end | (5) RT-DETR-R50 | $51.2 \pm 0.9$ | $73.8 \pm 0.8$ | $6.9 \pm 0.6$ | $4.8 \pm 0.4$ | $81.4 \pm 1.0$ | $68.3 \pm 1.6$ |
| | **(5) + CAESAR++** | $\mathbf{53.5 \pm 0.5}$ | $\mathbf{77.0 \pm 0.4}$ | $\mathbf{4.3 \pm 0.4}$ | $\mathbf{2.2 \pm 0.3}$ | $\mathbf{84.2 \pm 0.6}$ | $66.1 \pm 2.1$ |
| Two-stage, region proposal networks | (6) Faster R-CNN | $53.4 \pm 0.4$ | $75.4 \pm 0.5$ | $4.7 \pm 0.4$ | $4.2 \pm 0.3$ | $82.3 \pm 0.7$ | $9.8 \pm 0.2$ |
| | **(6) + CAESAR++** | $\mathbf{56.6 \pm 0.2}$ | $\mathbf{80.9 \pm 0.3}$ | $\mathbf{2.1 \pm 0.2}$ | $\mathbf{1.3 \pm 0.2}$ | $\mathbf{86.2 \pm 0.4}$ | $7.4 \pm 0.8$ |

Table 3: Accuracy gains in mAP@0.50:0.95 after applying CAESAR++ on YOLOv8m across object sizes and eight road object classes in cross-dataset validation. Object sizes are stratified following the COCO challenge. Values are mean ± standard deviation in percentage points.

| Train | Test | Small | Medium | Large | Person | Rider | Car | Truck | Bus | Tram | M.cycle | Bicycle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TJU-DHD | PascalVOC | 5.1±0.6 | 3.0±0.4 | 1.1±0.3 | 4.3±0.5 | 3.9±0.6 | 3.4±0.4 | 1.8±0.4 | 1.7±0.3 | 0.7±0.3 | 5.2±0.7 | 5.4±0.8 |
| TJU-DHD | BDD100K | 5.9±0.5 | 3.8±0.4 | 1.4±0.2 | 5.2±0.4 | 4.5±0.5 | 3.8±0.3 | 2.0±0.3 | 1.9±0.2 | 0.8±0.2 | 6.3±0.6 | 6.4±0.6 |
| PascalVOC | TJU-DHD | 3.6±0.7 | 2.3±0.5 | 0.9±0.3 | 3.6±0.5 | 3.5±0.6 | 3.1±0.4 | 1.5±0.5 | 1.4±0.4 | 0.3±0.1 | 4.7±0.7 | 4.6±0.8 |
| PascalVOC | BDD100K | 4.0±0.7 | 2.5±0.5 | 0.9±0.3 | 3.8±0.6 | 3.7±0.5 | 3.3±0.4 | 1.6±0.5 | 1.4±0.4 | 0.4±0.1 | 4.8±0.7 | 4.8±0.7 |
| BDD100K | TJU-DHD | 5.5±0.6 | 4.2±0.4 | 1.3±0.2 | 5.0±0.4 | 4.4±0.5 | 3.8±0.3 | 1.9±0.3 | 1.8±0.3 | 0.8±0.2 | 6.1±0.6 | 6.2±0.6 |
| BDD100K | PascalVOC | 4.7±0.6 | 2.8±0.4 | 1.1±0.2 | 3.9±0.5 | 3.9±0.6 | 3.4±0.4 | 1.8±0.5 | 1.7±0.3 | 0.5±0.3 | 5.1±0.7 | 5.3±0.7 |

## 4.2 Detection performance

Table 2 presents results on TJU-DHD-Traffic for a subset of detectors, with and without CAESAR++. In all cases, CAESAR++ improves mAP and reduces both FNR and FDR. The gain in mAP@0.5:0.95 reaches about three percentage points for some detectors, and the reduction in FNR and FDR often exceeds fifty percent relative.

An important trend is that FDR often decreases nearly as much as FNR. This suggests that CAESAR++ is effective at suppressing uncertain false alarms by incorporating semantically relevant context, while still reducing missed detections. In safety-critical road perception, this is a desirable trade-off: fewer false positives improve reliability, and fewer false negatives better support safe navigation.

From a deployment perspective, the computational cost remains moderate because only uncertain detections are reprocessed. As a result, the FPS drop is limited for the one-stage and transformer-based detectors, whereas Faster R-CNN remains the computational bottleneck in absolute terms due to its inherently heavier design. The additional computation depends on the proportion of uncertain detections. Detectors with stronger baselines generate fewer uncertain cases and therefore incur smaller slowdowns. In most configurations CAESAR++ maintains real-time performance of the base models, for example more than 70 FPS with IA-YOLO.

Cross-dataset experiments with YOLOv8m (Table 3) show that the largest gains are achieved for small objects and for classes such as pedestrians, riders, bicycles and motorcycles, which are both critical for safety and particularly difficult to detect. This is consistent with the fact that small instances provide limited appearance evidence inside the original bounding box, so neighboring structure becomes more valuable for disambiguation.

## 4.3 Visual explanations

Figure 3 shows qualitative examples with YOLOv8m. The baseline detector misses several objects and sometimes misplaces bounding boxes. After CAESAR++ refinement, detections are more accurate, and the dual-color visual explanations reveal which parts of the image support each decision. Red regions capture the object appearance itself, whereas blue regions highlight contextual structures such as road, sidewalks or nearby vehicles. This separation is particularly useful when local appearance is weak, because it makes explicit how contextual information helps resolve ambiguity caused by occlusion, clutter, or low contrast.

Table 4 quantitatively compares CAESAR++ with three object-specific peer explainers [15, 16, 17] across detectors. The reported metrics capture complementary aspects of explanation quality. Insertion measures how much the detector score increases as salient pixels are progressively revealed, Deletion measures how quickly the score decreases as they are removed, EBPG evaluates spatial alignment with the ground-truth bounding box, and AvgS assesses the stability of explanations under perturbations.

Table 4: Comparison of object-specific explainers on TJU-DHD-Traffic with YOLOv8m, Faster R-CNN and RT-DETR-R50. Metrics: Deletion (Del., ↓), Insertion (Ins., ↑), EBPG (↑), AvgS (↓). The best result for each metric is in **bold**.

| Method | YOLOv8m | | | | Faster R-CNN | | | | RT-DETR-R50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Del. | Ins. | EBPG | AvgS | Del. | Ins. | EBPG | AvgS | Del. | Ins. | EBPG | AvgS |
| D-RISE | 0.22 | 0.54 | 0.51 | 0.13 | 0.21 | 0.57 | 0.51 | 0.13 | 0.24 | 0.54 | 0.47 | 0.15 |
| D-CLOSE | **0.19** | 0.55 | 0.55 | 0.11 | 0.21 | 0.66 | **0.78** | 0.16 | 0.22 | 0.56 | 0.55 | 0.12 |
| D-MFPP | 0.22 | 0.58 | **0.64** | 0.09 | **0.18** | 0.63 | 0.69 | 0.09 | **0.18** | 0.57 | **0.59** | 0.10 |
| CAESAR++ | 0.21 | **0.65** | 0.62 | **0.08** | 0.17 | **0.69** | 0.71 | **0.07** | 0.19 | **0.64** | 0.56 | **0.09** |

CAESAR++ achieves the best Insertion and AvgS results, showing that its explanations consistently identify decision-relevant regions and remain reliable under small input changes. Its Deletion scores are also competitive, which indicates that the highlighted areas are not only visually meaningful but also influential for the detector response. The slightly lower EBPG values are consistent with the design of CAESAR++. Unlike other methods that restrict explanations to the object interior, CAESAR++ is explicitly context-aware and may attribute part of the decision to surrounding structures when they help disambiguate the detection. This is not a limitation of the method, but a reflection of the fact that road-object recognition often depends on scene context, especially for uncertain, partially occluded, or small instances.

As a result, CAESAR++ provides explanations that are less box-constrained but more informative and interpretable, since they expose both what the detector directly sees and how context changes the final decision.

# 5 Discussion

CAESAR++ is particularly effective in urban driving scenes, where the semantic relationships learned during context modeling remain representative of the scene at test time. In this setting, the uncertainty score serves as a meaningful indicator of ambiguity, the adaptive window retrieves relevant surrounding cues, and the explanation branch makes the contribution of context explicit. This behavior matches the intended design of the framework, which is to exploit regular scene structure in order to improve both detection reliability and interpretability, while remaining fully compatible with the underlying detector.

Its behavior is less favorable when the scene departs from the semantic prior encoded during context construction. In out-of-distribution or unusual situations, uncertainty may still increase, but the contextual evidence available in the enlarged window may be incomplete or poorly aligned with the expected traffic pattern. In such cases, the refinement stage is triggered on weaker contextual support, and the improvement in detection or explanation quality may be limited. This is an inherent consequence of the method's design, which targets structured road scenes rather than open-world conditions with unsupported semantics.

A second potential limitation arises in sparse scenes or in configurations where informative cues lie outside the adaptive crop. In such cases, the method may still detect ambiguity correctly, but the expanded region does not necessarily provide enough additional semantic evidence to justify the extra computation. This leads to a diminishing return effect, where refinement is correctly triggered but the contextual gain remains limited. For this reason, in latency-sensitive deployment scenarios, contextual reprocessing should be applied selectively, for instance by capping the number of refined detections or by restricting the mechanism to the object classes for which context is most informative and operationally relevant.

Regarding the balance between error types, the framework lowers both false negatives and false discoveries, but not always to the same extent. In real-world perception, this trade-off should be interpreted in light of the downstream application: missed detections are generally more critical than occasional false alarms, especially for vulnerable road users. As a result, the operating point should be chosen according to the target system requirements, with priority given to recall when safety is the primary concern. The latency overhead introduced by CAESAR++ is therefore acceptable when a moderate increase in computation is compatible with the application, but it should be managed carefully in real-time deployments.

# 6 Conclusion and perspectives

This paper presented CAESAR++, a detector-agnostic framework for road object detection that jointly addresses uncertainty quantification, adaptive contextual reasoning and visual explainability. By coupling a two-step conformal prediction with context-aware refinement and dual-color saliency maps, CAESAR++ improves detection accuracy, calibrates uncertainty and yields robust, plausible, and interpretable explanations. Experiments across multiple detectors and datasets show consistent gains, especially for small and difficult objects, with reasonable cost.

Future work will focus on reducing computational overhead through lighter segmentation backbones, single-pass refinement without re-detection, and more selective contextual expansion. We also plan to explore online recalibration and adaptive preprocessing to reduce reliance on segmentation in atypical scenes, while extending the framework to stronger domain shifts and related tasks such as instance segmentation and tracking.

# Acknowledgments

Most of the computations presented in this paper were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr), which is supported by the Grenoble research community.

# References


[1] Narges Saeedizadeh, Seyed Mohammad Jafar Jalali, Burhan Khan, and Shady Mohamed. Cutting-edge deep learning methods for image-based object detection in autonomous driving: In-depth survey. *Expert Systems*, 42(4), 2025.

[2] Tirupathamma Mudavath and Anooja Mamidi. Object detection challenges: Navigating through varied weather conditions—a comprehensive survey. *Journal of Ambient Intelligence and Humanized Computing*, 16(2):443–457, 2025.

[3] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[5] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30:6405–6416, 2017.

[6] Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 388–395, 2007.

[7] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.

[8] Leo Andeol, Thomas Fel, Florence de Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: An application to railway signaling. In *Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204, pages 36–55. PMLR, 2023.

[9] Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction. In *European Conference on Computer Vision*, pages 363–398, 2025.

[10] Ruoxi Qi, Guoyang Liu, Jindi Zhang, and Janet Hui-Wen Hsiao. Do saliency-based explainable ai methods help us understand ai's decisions? the case of object detection ai. In *Annual Meeting of the Cognitive Science Society*, volume 46, Rotterdam, the Netherlands, 2024.

[11] Anh-Thu Mai, Marina Nicolas, Patricia Ladret, and Alice Caplier. Robust road object detection with caesar: Context-aware explanations via semantic attribution and refinement. In *IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2025.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[13] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018.

[14] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

[15] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11438–11447, 2021.

[16] Van Binh Truong, Truong Thanh Hung Nguyen, Vo Thanh Khang Nguyen, Quoc Khanh Nguyen, and Quoc Hung Cao. Towards better explanations for object detection. In *Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pages 1385–1400, 2024.

[17] Alain Andres, Aitor Martinez-Seras, Ibai Laña, and Javier Del Ser. On the black-box explainability of object detection models for safe and trustworthy industrial applications. *Results in Engineering*, 24:103498, 2024.

[18] Guoyang Liu, Jindi Zhang, Antoni B. Chan, and Janet H. Hsiao. Human attention guided explainable artificial intelligence for computer vision models. *Neural Networks*, 177:106392, 2024.

[19] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

[20] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[21] Xinfang Zhong, Wenlan Kuang, and Zhixin Li. Adaptive graph reasoning network for object detection. *Image and Vision Computing*, 151, 2024.

[22] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *European Conference on Computer Vision*, pages 420–436, 2018.

[23] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S. Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8817–8826, 2018.

[24] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 833–851, 2018.

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[26] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[27] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing*, 30:207–219, 2021.

[28] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2020.

[29] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. Pascal visual object classes 2012, 2025.

[30] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3543–3553, 2019.

[31] Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1322–1330, 2020.

[32] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2016.

[33] Rejin Varghese and M. Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024.

[34] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *AAAI Conference on Artificial Intelligence*, pages 1792–1800, 2022.

[35] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10778–10787, 2020.

[36] Huilin Wang, Huaming Qian, Shuai Feng, and Wenna Wang. L-ssd: Lightweight ssd target detection based on depth-separable convolution. *Journal of Real-Time Image Processing*, 21(2), 2024.

[37] 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, 2023.

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[39] Jianming Zhang, Saeed Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *International Journal of Computer Vision*, pages 1084–1102, 2018.

[40] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods, 2018.