Indonesian Journal of Electrical Engineering and Computer Science
Vol. 44, No. 1, October 2026, pp. 209~220
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v44.i1.pp209-220

Evaluating the reliability of computer-vision-based traffic data under diverse weather conditions

Shafiqullah Ahmadzai[1,2,3], Emre Demi[4]
[1]Department of Civil Engineering, Institute of Graduate Studies, Antalya Bilim University, Antalya, Türkiye
[2]Department of Civil Engineering, Faculty of Engineering, Paktia University, Gardez, Afghanistan
[3]Department of Civil Engineering, Mustaqbal University, Kabul, Afghanistan
[4]Department of Civil Engineering, Faculty of Engineering and Natural Sciences, Antalya Bilim University, Antalya, Türkiye

Article Info

Article history:

Received Mar 30, 2026
Revised Aug 21, 2026
Accepted Sep 9, 2026

Keywords:

Adverse weather
Intelligent transportation systems
Object detection
Reliability
Traffic data quality

ABSTRACT

Intelligent transportation systems (ITS) increasingly derive traffic data from camera-based perception. Yet the reliability of that data under adverse weather conditions is rarely quantified in terms that traffic engineers can act on. This study evaluates detection reliability on real adverse-weather imagery drawn from a public benchmark dataset, partitioned using the dataset's own recorded weather and time-of-day attributes into balanced subsets of 736 images each. Three one-stage detectors and one two-stage detector were evaluated under identical conditions, and a bootstrap confidence interval accompanies each reported value. Rain and snow produced no statistically detectable change in overall mean average precision, whereas night-time operation reduced it by up to 38%. Small-object average precision, however, fell by 38–41% under precipitation and by as much as 67% at night, a degradation that aggregate accuracy conceals entirely. Objective measures of image signal quality were associated with per-image detection performance under precipitation but not in favorable conditions. The results indicate that aggregate accuracy is an inadequate governance metric for vision-derived traffic data, and that reliability monitoring should be stratified by object scale and by time of day.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Shafiqullah Ahmadzai
Department of Civil Engineering, Institute of Graduate Studies
Antalya Bilim University
Antalya, Türkiye
Email: 2423201@std.antalya.edu.tr

1. INTRODUCTION

Camera sensors are among the most widely deployed sources of automated traffic data. They are inexpensive relative to lidar; they can be mounted on existing roadside infrastructure, and a single unit simultaneously yields vehicle counts, classification, speed estimates, and trajectory information. As intelligent transportation systems (ITS) and advanced driver assistance systems (ADAS) expand, an increasing share of the data used for traffic operations, incident detection, and automated decision-making originates from computer vision rather than from in-pavement sensors. Vision is central to these systems, which depend on visual sensors for guidance and decision-making. The camera is also, however, the modality most sensitive to weather and lighting [1].

This dependence creates an exposure that is not present in loop detectors or radar: optical sensing degrades when the atmosphere between the camera and the scene changes. Rain produces motion streaks and lens occlusion; snow introduces high-contrast moving particles; fog reduces contrast through atmospheric

scattering; and night operation reduces the available signal while introducing glare and sensor noise. Each of these alters the image statistics that a detector was trained to expect, and detection performance falls accordingly [1]–[3].

A substantial body of work has documented this degradation. Reviews of detection under adverse weather catalog both classical and deep-learning mitigation strategies [1], [2]. Empirical evaluations report large losses in mean average precision under distortion, and note that two-stage architectures do not necessarily degrade in the same manner as one-stage architectures [3], [4]. Other work has pursued mitigation directly, through weather-informed image enhancement [5], generative image translation [6], domain adaptation [7], and the construction of dedicated adverse-condition benchmarks [8]. Edge deployment has been examined as a means of reducing the latency between perception and response [9].

Three limitations recur across this literature. First, much of the available robustness evidence comes from synthetically corrupted imagery, in which fog, rain, or noise is applied computationally to clear-weather photographs [10]. The physical models behind such data are simplified, and the resulting synthetic-to-real domain gap is repeatedly associated with reduced generalization in real weather [2]; simulated environments may not fully reproduce real-world driving conditions, including unpredictable driver behavior and dynamic traffic patterns [1]. Second, reported values are presented as point estimates without confidence intervals or significance tests [1], [2], [9], so it cannot be determined whether a difference between two architectures or two conditions exceeds sampling variation; in [10], for example, weather is judged to have little effect on detection performance on the basis of point estimates alone. Third, and most importantly for transportation practice, results are reported in the vocabulary of computer vision rather than in the vocabulary of traffic data quality (TDQ). A traffic engineer responsible for a data feed must reason about dimensions such as accuracy, completeness, validity, and timeliness [11], whereas mean average precision averages precision across recall levels, intersection-over-union thresholds, and object classes, and therefore maps onto none of these directly.

The consequence is a gap between the two communities. Computer-vision research establishes that detection performance is sensitive to adverse conditions, but the magnitude of that sensitivity varies sharply between conditions and is rarely disaggregated; TDQ frameworks, in turn, specify what a reliable data feed must satisfy. To the best of our knowledge, the correspondence between the two has not been made explicit or measured. This study addresses that gap. The contributions of this work are as follows:

- A mapping between object-detection performance metrics and the TDQ dimensions used in ITS practice, formulated so that each dimension is associated with a measurable quantity rather than a qualitative judgment.
- A replication and extension of the real-imagery evaluation protocol of [10], in which a public benchmark dataset is partitioned using its own recorded weather and time-of-day attributes rather than by applying synthetic corruption. The protocol is extended here from a single detector to four architectures spanning two detector families and three model scales, and the resulting condition-wise degradation is reported under uncertainty quantification.
- A statistical treatment in which every reported value carries a bootstrap confidence interval, and in which comparisons between architectures are paired while comparisons between weather conditions are not, reflecting the fact that different conditions comprise different images.
- An analysis relating objective measures of image signal quality, computed per image, to the detection performance obtained on that same image, quantifying the extent to which degradation is associated with measurable loss of signal rather than asserting the relationship qualitatively.

The remainder of the paper is organized as follows. Section 2 presents the proposed reliability assessment framework. Section 3 describes the datasets, detection models, evaluation metrics, and statistical procedures. Section 4 reports and discusses the results. Section 5 concludes.

2. PROPOSED RELIABILITY ASSESSMENT FRAMEWORK

TDQ is conventionally characterized through a small set of dimensions: accuracy, completeness, validity, timeliness, coverage, and accessibility [11]. These dimensions were formulated for traffic data systems in general and are independent of the sensing modality used to produce the data. When the source of the data is a camera and a detection model, each dimension has a natural counterpart among the quantities that computer-vision evaluation already produces, but that correspondence is not usually stated.

The framework proposed here makes the correspondence explicit and, where possible, measurable. Table 1 sets out the mapping. Its purpose is not to redefine either field but to allow a detection result to be read as a statement about data quality: a fall in recall under snow is a loss of completeness; a fall in average precision at a strict intersection-over-union threshold, without a corresponding fall at a permissive threshold, is a loss of validity rather than of accuracy; an increase in inference latency is a loss of timeliness even when detection quality is unchanged.

Multi-criteria assessment of this kind is established practice in transportation evaluation, where composite indicator frameworks are routinely used to summarize system performance across dimensions that are not directly commensurable [12]. One dimension resists this treatment. Accessibility concerns whether data can be obtained when required, which depends on system architecture, network availability, and institutional arrangements rather than on the detector itself. It is retained in the framework for completeness and discussed qualitatively, but no detection metric is claimed to measure it.

Table 1. Mapping between TDQ dimensions and computer-vision metrics

TDQ measure	Definition in the ITS context	Corresponding CV metric	What this study measures
Accuracy	Degree of agreement between a reported value and a value assumed correct [11].	mAP@0.5, mAP@0.5:0.95 [13]	Per-condition mean average precision with 95% confidence intervals.
Completeness	Degree to which required data values are present [11].	Recall; missed-detection rate [13]	Recall behavior and precision-recall curves per condition.
Validity	Degree to which data values satisfy acceptance requirements [11].	AP at strict IoU thresholds [13]	AP@0.75 relative to AP@0.5, as a localisation-quality measure.
Timeliness	Degree to which values are available when required [11].	Inference latency; frames per second [9]	Latency measured on the study hardware for every architecture.
Coverage	Degree to which the sensing system observes the intended domain [11].	AP by object scale (small/medium/large)	Loss of small-object performance, reported per condition.
Accessibility	Degree to which data can be obtained when needed [11].	Not directly measurable from detection output	Discussed qualitatively; identified as outside the scope of this study.

3. METHOD

3.1. Data source

All imagery used in this study is drawn from a public benchmark dataset. The authors collected no data, and no camera was operated for this work; consequently, no camera specifications, mounting height, or calibration parameters apply to the study, and the acquisition characteristics are those documented by the dataset authors. The dataset is BDD100K [14], which provides 100,000 driving images at 1280×720 resolution with two-dimensional bounding-box annotations for ten object categories.

The property of BDD100K that makes it suitable here is that each image includes an attribute block recording the weather condition, scene type, and time of day at which it was captured. Adverse condition subsets can therefore be obtained by selecting genuinely adverse photographs, rather than by applying synthetic corruption to clear-weather images. The evaluation is performed on the validation split, whose annotations are public.

3.2. Construction of condition subsets

Images were partitioned by the recorded weather attribute into clear, rainy, snowy, foggy, and overcast subsets, and by the time-of-day attribute into a night subset. The clear-daytime subset serves as the reference condition against which all others are compared. The validation split contains 5,346 clear, 1,239 overcast, 769 snowy, 738 rainy, and 13 foggy images, and 5,258 daytime, 3,929 night, and 778 dawn or dusk images. The foggy subset is far too small to support a balanced comparison and was therefore excluded; this exclusion is revisited in Section 4.8.

Average precision is sensitive to the number of images over which it is computed, so subsets of unequal size are not directly comparable. Every subset was therefore reduced by random sampling, under a fixed seed of 42, to the size of the smallest qualifying subset, following the balanced-subset approach adopted in prior robustness benchmarking [10]. A minimum subset size of 150 images was imposed: a condition with fewer images than this threshold was excluded from the balanced comparison rather than allowing it to reduce every other condition to its size. This yielded five conditions, each containing 736 images, the compositions of which are given in Table 2. The training split was not pooled.

3.3. Annotation provenance

Bounding-box annotations were produced and quality-controlled by the authors of the source dataset, not by the present authors, using the annotation protocol described in the dataset publication [14]. No re-annotation, correction, or verification of the released labels was performed in this study. Objects annotated as crowded regions were excluded from evaluation in accordance with standard practice.

3.4. Detection models and category mapping

Both one-stage and two-stage architectures were evaluated to test the widely reported claim that one-stage detectors are comparatively robust under degraded imaging [4], rather than assuming it. The one-

stage models are YOLOv8n, YOLOv8s, and YOLOv8m [15]; the two-stage model is Faster R-CNN with a ResNet-50 feature pyramid backbone [16], [17]. All models use weights learned on the COCO dataset [18]; No fine-tuning on the evaluation dataset was performed. The study therefore measures zero-shot robustness, that is, the behavior of an off-the-shelf detector deployed in conditions to which it has not been adapted. This corresponds to a realistic ITS deployment scenario and avoids confounding weather robustness with the quality of dataset-specific adaptation.

Because the label spaces of the two datasets differ, evaluation is restricted to the categories present in both. The categories pedestrian and rider in the source dataset are merged into a single person category, consistent with the semantics of the pretrained models. The traffic sign category has no counterpart in the pretrained label space. It is excluded, since evaluating a model on a category it was never trained to predict would yield an artificially null result. Eight categories remain: person, car, truck, bus, train, motorcycle, bicycle, and traffic light.

Table 2. Composition of the balanced evaluation subsets

Condition	Images	Annotations	Boxes/image	Small objects
Clear daytime	736	11,420	15.5	5,952
Rainy	736	10,655	14.5	5,113
Snowy	736	11,107	15.1	5,468
Overcast	736	12,251	16.6	6,488
Night	736	9,573	13.0	4,836

3.5. Evaluation metrics

Detection quality is reported using the standard COCO protocol [18]. Average precision (AP) is the area under the interpolated precision-recall curve for a given category at a given intersection-over-union (IoU) threshold, and mean average precision (mAP) is its average over categories. Two thresholds are reported: mAP@0.5, which accepts a detection if its overlap with the ground-truth box exceeds 0.5, and mAP@0.5:0.95, which averages over 10 thresholds from 0.5 to 0.95 in steps of 0.05 and is therefore sensitive to both localization precision and detection success. AP@0.75 is reported separately as a strict-localization measure. The choice of evaluation metric is not neutral under class imbalance, which is pronounced in road scenes where cars dominate the annotation count; metric selection under such imbalance has been examined systematically [19].

Average precision is additionally reported by object scale, using the COCO area conventions of small (area below 32^2 pixels), medium, and large, since loss of small-object performance is a principal failure mode under degraded imaging and corresponds to the coverage dimension of Table 1. Precision-recall curves are reported per condition and per category, so that the trade-off between false positives and missed detections can be inspected directly rather than inferred from a single scalar. Inference latency is measured on the study hardware and reported as milliseconds per image and frames per second, corresponding to the timeliness dimension. A low confidence threshold of 0.001 is applied at inference. A higher threshold truncates the precision-recall curve and inflates average precision, making it inappropriate for a study whose subject is the shape of that curve.

3.6. Statistical procedure

Mean average precision is computed from a pooled, rank-ordered list of detections rather than as a mean of independent per-image values and consequently has no closed-form standard error. Uncertainty is therefore estimated by bootstrap resampling over images [20]. For each of 300 replicates, images are drawn with replacement from the subset and the metric is recomputed; the reported 95% interval is the percentile interval of the resulting distribution. With 300 replicates, the smallest resolvable p-value is 0.0033, and values at that floor are reported as $p < 0.004$.

Two distinct comparisons are made, each requiring different treatment. Comparisons between architectures are performed under identical conditions, where both models are evaluated on the same images; the same resample indices are therefore applied to both models, and the difference is taken within each replicate, thereby removing variance due to image difficulty and yielding a paired test. Comparisons between weather conditions involve different images, and pairing would be invalid; a two-sample bootstrap is used instead. Every reported p-value is accompanied by the corresponding effect size and confidence interval, since a difference may be statistically distinguishable from zero while being operationally negligible.

3.7. Image signal quality measures

To examine how far detection loss is associated with measurable degradation of the image signal, seven objective measures were computed for every evaluated image: signal-to-noise ratio expressed in

decibels, computed as the ratio of mean to standard deviation of the luminance channel; an estimated peak signal-to-noise ratio obtained against a median-filtered reference, the residual being treated as noise; root-mean-square contrast; the variance of the Laplacian, as a measure of sharpness; the mean of the dark channel, which increases with atmospheric scattering and therefore serves as a haze density proxy [21]; the Shannon entropy of the luminance histogram; and the proportion of pixels exceeding a gradient-magnitude threshold, as a measure of edge density.

Average precision was then computed for each image individually, using the same matching used for the aggregate metrics, and the Spearman rank correlation coefficient quantified the association between each quality measure and per-image detection performance. A rank correlation is used in preference to a linear one because the relationships are monotonic rather than linear and because average precision is bounded, which violates the assumptions of the linear coefficient at the extremes of its range.

3.8. Implementation

Detection and evaluation were implemented in Python 3.12.13 with PyTorch 2.11.0, NumPy 2.0.2, Ultralytics YOLO 8.4.117, pycocotools, OpenCV, and SciPy. Latency measurements were taken on an NVIDIA Tesla T4 GPU (16 GB, driver 580.82.07, CUDA 13.0); the detection sweep reported in Section 4 was executed on CPU, which does not affect detection outputs. The bootstrap implementation was verified against the reference pycocotools accumulation before use, reproducing mAP@0.5:0.95, mAP@0.5, mAP@0.75 and AP at all three object scales to a tolerance below 10^{-9}. The evaluation code and derived result files are available from the corresponding author on reasonable request

4. RESULTS AND DISCUSSION

4.1. Detection performance across conditions

Table 3 reports detection performance for every model and condition. Each mean average precision value is accompanied by a 95% percentile bootstrap interval obtained from 300 resamples of the images. Two observations follow immediately. First, the ordering of conditions is consistent across all three YOLO variants: overcast and clear daytime are indistinguishable, rain and snow produce modest reductions, and night is markedly the worst condition. Second, the confidence intervals are wide relative to the differences between conditions, which has direct consequences for what can and cannot be claimed from these measurements.

Table 3. Detection performance by model and condition (N = 736 images per condition)

Model	Condition	mAP@0.5:0.95 [95% CI]	mAP@0.5	mAP@0.75	AP-S	AP-M	AP-L
YOLOv8n	Clear daytime	0.139 [0.129, 0.176]	0.235	0.132	0.021	0.158	0.555
YOLOv8n	Overcast	0.138 [0.128, 0.171]	0.238	0.129	0.023	0.182	0.399
YOLOv8n	Rainy	0.132 [0.121, 0.165]	0.230	0.128	0.017	0.142	0.393
YOLOv8n	Snowy	0.132 [0.113, 0.163]	0.222	0.136	0.011	0.142	0.496
YOLOv8n	Night	0.086 [0.073, 0.114]	0.171	0.074	0.007	0.092	0.208
YOLOv8s	Clear daytime	0.181 [0.170, 0.224]	0.300	0.177	0.042	0.215	0.595
YOLOv8s	Overcast	0.214 [0.190, 0.291]	0.356	0.209	0.038	0.248	0.472
YOLOv8s	Rainy	0.180 [0.165, 0.224]	0.307	0.173	0.024	0.199	0.491
YOLOv8s	Snowy	0.169 [0.151, 0.203]	0.282	0.166	0.025	0.184	0.521
YOLOv8s	Night	0.120 [0.105, 0.170]	0.224	0.110	0.015	0.169	0.260
YOLOv8m	Clear daytime	0.226 [0.209, 0.280]	0.372	0.222	0.067	0.272	0.670
YOLOv8m	Overcast	0.227 [0.217, 0.268]	0.397	0.215	0.066	0.305	0.461
YOLOv8m	Rainy	0.216 [0.199, 0.269]	0.365	0.212	0.041	0.243	0.540
YOLOv8m	Snowy	0.205 [0.173, 0.252]	0.344	0.209	0.042	0.221	0.592
YOLOv8m	Night	0.172 [0.146, 0.235]	0.308	0.164	0.050	0.209	0.332
Faster R-CNN	Clear daytime	0.204 [0.195, 0.250]	0.391	0.184	0.094	0.251	0.569

4.2. Statistical comparison against the clear-daytime reference

Table 4 reports the change in mAP@0.5:0.95 relative to the clear-daytime reference, along with the 95% confidence interval for the difference and the two-sided bootstrap p-value. Because the conditions comprise different images, an unpaired two-sample bootstrap is used.

Table 4. Change in mAP@0.5:0.95 relative to clear daytime

Model	Condition	Change	Relative	95% CI of difference	p
YOLOv8n	Overcast	−0.001	−0.6%	[−0.039, +0.029]	0.853
YOLOv8n	Rainy	−0.007	−5.1%	[−0.043, +0.023]	0.547
YOLOv8n	Snowy	−0.007	−5.3%	[−0.052, +0.023]	0.380
YOLOv8n	Night	−0.053	−38.0%	[−0.092, −0.026]	< 0.004
YOLOv8s	Rainy	−0.001	−0.6%	[−0.044, +0.033]	0.853
YOLOv8s	Snowy	−0.012	−6.6%	[−0.059, +0.023]	0.300
YOLOv8s	Overcast	+0.033	+18.3%	[−0.019, +0.094]	0.353
YOLOv8s	Night	−0.061	−33.6%	[−0.109, −0.018]	0.013
YOLOv8m	Overcast	+0.001	+0.4%	[−0.055, +0.045]	0.947
YOLOv8m	Rainy	−0.010	−4.3%	[−0.062, +0.030]	0.627
YOLOv8m	Snowy	−0.021	−9.1%	[−0.085, +0.022]	0.280
YOLOv8m	Night	−0.054	−23.7%	[−0.113, +0.009]	0.100

The result that carries the most weight is negative. Rain and snow do not produce a statistically detectable change in overall mean average precision for any of the three models tested. The point estimates are consistently in the expected direction, and their magnitudes (roughly 4–9% relative for the medium model) are plausible, but every confidence interval includes zero. At the sample size used here, these effects cannot be distinguished from sampling variation.

Night is different. For the two smaller models, the degradation is large and statistically distinguishable from zero: −38.0% for YOLOv8n ($p < 0.004$) and −33.6% for YOLOv8s ($p = 0.013$). For YOLOv8m, the point estimate is a comparable −23.7%, but the interval crosses zero ($p = 0.100$), so night degradation is not established for that model at this sample size.

This finding should be stated carefully. A non-significant result is not evidence that an effect is absent; it is evidence that the study did not detect one. The appropriate conclusion is that, on real adverse-weather imagery, night-time operation is the dominant degradation mode for aggregate detection accuracy. At the same time, precipitation effects are smaller than this study can resolve.

4.3. Object scale and the coverage dimension

Aggregate mean average precision conceals a much larger effect. Table 5 reports average precision for small objects, using the COCO area convention, relative to clear daytime. The contrast with Table 4 is the central result of this study. Under rain, YOLOv8m loses 4.3% of its overall mean average precision but 39% of its small-object average precision. Under snow, the corresponding figures are 9.1% and 38%. At night, the two smaller models lose roughly two-thirds of their small-object performance.

Overall accuracy is therefore a misleading summary of detector behavior under adverse conditions. A traffic data feed reporting a mean average precision that has fallen only slightly may simultaneously be failing to detect a large proportion of distant or physically small road users. In terms of the framework in Section 2, this is a loss of coverage that is invisible in the accuracy dimension. In road scenes, small objects are disproportionately pedestrians, cyclists, and distant vehicles, which is precisely the population for which detection failure carries the greatest safety consequences [22], [23]. The intrinsic difficulty of small-object detection, limited spatial and contextual information, low effective resolution, and susceptibility to background interference is well documented [24], but its interaction with adverse conditions is rarely quantified separately from aggregate accuracy. The per-class results support this reading. For YOLOv8m, traffic light average precision falls from 0.161 in clear daytime to 0.076 in rain, 0.084 in snow, and 0.044 at night, a 73% loss at night for a class that is both small and critical to signal-state inference. Figure 1 shows the effect of object size and weather conditions on detection performance.

Table 5. Small-object average precision (AP-S) relative to clear daytime

Model	Clear	Rainy	Snowy	Overcast	Night
YOLOv8n	0.021	0.017 (−19%)	0.011 (−50%)	0.023 (+9%)	0.007 (−67%)
YOLOv8s	0.042	0.024 (−41%)	0.025 (−40%)	0.038 (−7%)	0.015 (−65%)
YOLOv8m	0.067	0.041 (−39%)	0.042 (−38%)	0.066 (−2%)	0.050 (−25%)

4.4. Precision-recall behavior

Precision-recall curves for each condition are shown in Figure 2, and per-class curves within conditions are shown in Figure 3. The curves are computed from the same evaluation that produces the values in Table 3, so figures and tables are mutually consistent by construction. They show that the loss under adverse conditions is concentrated at the high-recall end of the curve: precision at low recall is largely

preserved, while the recall attainable at any given precision falls. The practical reading is that adverse conditions cause missed detections rather than spurious ones.

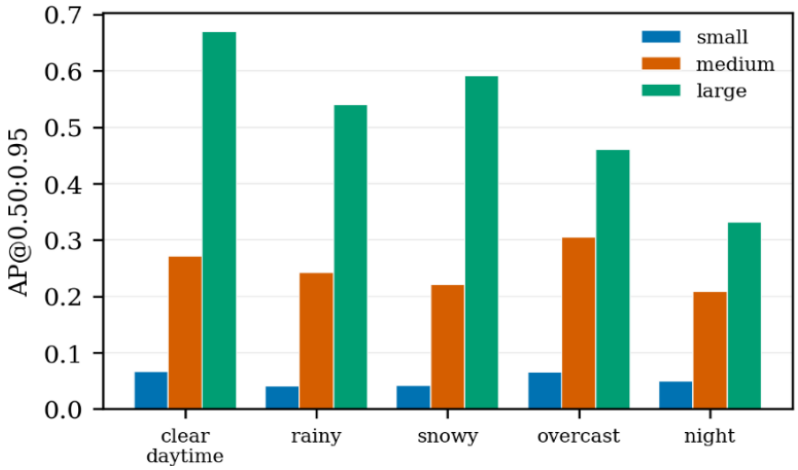

Figure 1. Average precision by object scale across conditions (YOLOv8m)

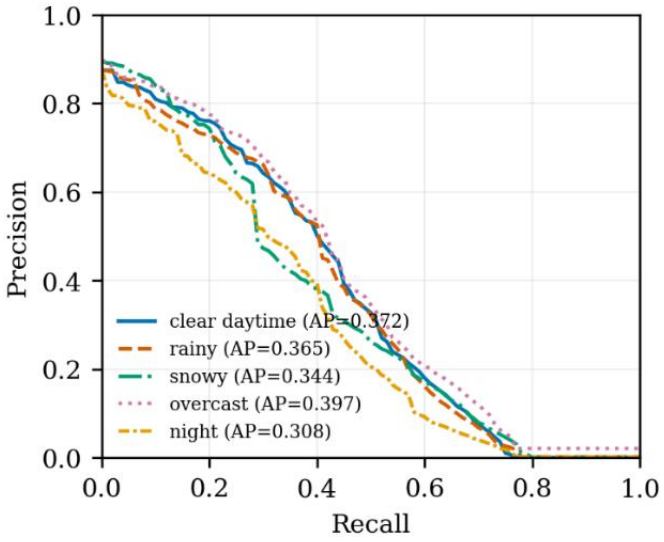

Figure 2. Precision–recall curves by condition (YOLOv8m)

4.5. Architecture comparison

Within clear daytime, where both architecture families were evaluated on identical images, the comparison is paired: the same bootstrap resamples were applied to both models and the difference taken within each replicate. Table 6 shows that YOLOv8m performs best among the YOLO models.

The widely reported advantage of one-stage detectors is not reproduced as a property of the architecture family. YOLOv8m outperforms Faster R-CNN by a statistically distinguishable margin, but YOLOv8s and YOLOv8n are both significantly worse than the same two-stage baseline. What separates the models here is capacity, not the one-stage or two-stage design. Claims of one-stage superiority that do not control for model size should therefore be treated with caution.

Faster R-CNN also shows a distinctive profile: it achieves the highest mAP@0.5 of any model (0.391) and by far the best small-object performance (AP-S = 0.094, against 0.067 for YOLOv8m), while recording the lowest AP@0.75 (0.184). It finds objects, including small ones, more reliably than the one-stage models, but localizes them less precisely. In the framework of Section 2, Faster R-CNN offers better completeness and coverage at the cost of validity.

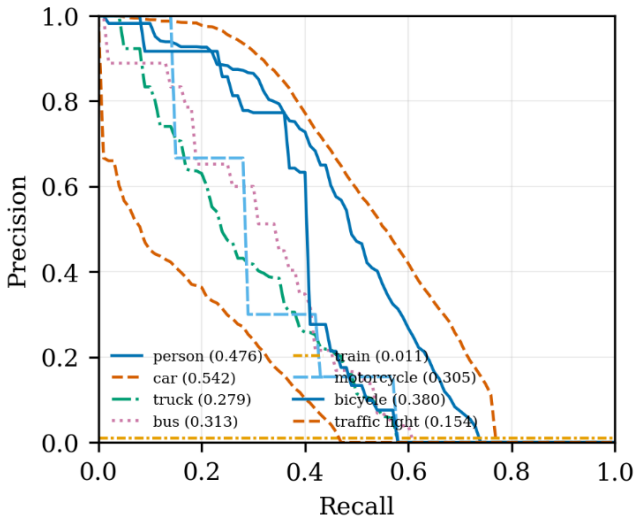

Figure 3. Precision–recall curves by object category at night (YOLOv8m)

Table 6. One-stage against two-stage detection, clear daytime (paired bootstrap)

Comparison	Difference in mAP@0.5:0.95	95% CI	p
YOLOv8m − Faster R-CNN	+0.021	[+0.010, +0.040]	< 0.004
YOLOv8s − Faster R-CNN	−0.024	[−0.036, −0.016]	< 0.004
YOLOv8n − Faster R-CNN	−0.065	[−0.083, −0.057]	< 0.004

4.6. Timeliness

Inference latency measured on the study hardware separates the architectures by an order of magnitude: 13.2 ms per image for YOLOv8n (75.9 fps), 13.1 ms for YOLOv8s (76.1 fps), 20.0 ms for YOLOv8m (50.1 fps), and 126.9 ms for Faster R-CNN (7.9 fps). Read against the framework, this is the timeliness dimension, where the architecture families genuinely diverge. Faster R-CNN's advantages in completeness and coverage are achieved at roughly one-tenth the throughput. For a roadside unit required to process a live video stream, this is decisive in a way that a two-point difference in mean average precision is not [9], [17].

4.7. Image signal quality and detection performance

For each evaluated image, seven objective signal-quality measures were computed and correlated with the average precision obtained on that image. Table 7 reports condition-level means. Signal quality and detection performance decline together across conditions. Night imagery records a negative signal-to-noise ratio in decibels, meaning that the standard deviation of the luminance channel exceeds its mean; the noise floor is comparable to the signal itself. Pooled across all 3,680 evaluated images, every quality measure correlates positively with per-image average precision: root-mean-square contrast (Spearman $\rho = 0.317$), edge density ($\rho = 0.310$), entropy ($\rho = 0.303$), dark channel ($\rho = 0.285$), Laplacian variance ($\rho = 0.283$) and signal-to-noise ratio ($\rho = 0.255$), all with $p < 10^{-54}$.

Table 7. Mean image signal quality and per-image detection performance (YOLOv8m)

Condition	SNR (dB)	RMS contrast	Dark channel	Mean image AP
Clear daytime	4.80	0.234	0.240	0.387
Overcast	3.97	0.257	0.276	0.388
Snowy	2.76	0.198	0.169	0.345
Rainy	2.41	0.209	0.199	0.332
Night	−0.82	0.128	0.062	0.267

The within-condition analysis is more informative than the pooled figure and qualifies the pooled figure. Within rain ($\rho = 0.299$, $p = 1.1 \times 10^{-16}$) and snow ($\rho = 0.229$, $p = 3.5 \times 10^{-10}$) the association between signal-to-noise ratio and detection performance is clear. Within clear daytime ($\rho = -0.032$, $p = 0.39$), overcast ($\rho = -0.005$, $p = 0.89$) and night ($\rho = 0.060$, $p = 0.10$) it is absent. The interpretation is that signal quality

predicts detection performance only once it has degraded. In favorable conditions, variation in image quality is not the limiting factor on detection; scene composition and object scale dominate. Under precipitation, image degradation becomes the binding constraint and individual image quality becomes predictive. At night, the association again disappears, which suggests that night-time failure is not principally a signal-to-noise phenomenon but is driven by other factors: reduced object visibility, glare, and the mismatch between night imagery and the daylight-dominated distribution on which the detectors were trained [8], [25].

This is an association, not a demonstration of causation. Weather conditions differ in scene composition, object density, object distance, and signal quality, and these factors cannot be separated in observational data of this kind. The annotation counts in Table 2 differ across subsets, confirming that the conditions are not matched on content. The correct reading is that measurable signal degradation is associated with reduced detection performance under precipitation, and that this association does not hold in conditions where signal quality is adequate.

4.8. Interpretation within the TDQ framework

When read against Table 1, the results provide a differentiated picture of how camera-derived traffic data degrades. Accuracy, measured by aggregate mean average precision, is more robust than the literature would suggest; rain and snow do not measurably reduce it at this sample size. Coverage, measured by small-object average precision, degrades severely under conditions where accuracy appears stable, with losses of 38–41% in precipitation and up to 67% at night. A data feed governed by accuracy alone would not detect this.

Validity, measured by the ratio of strict-threshold to permissive-threshold performance, distinguishes the architectures: Faster R-CNN detects more but localizes less precisely. Timeliness varies by an order of magnitude across architectural families and is independent of weather conditions. Completeness follows coverage, with the missed detections concentrated among small objects. The practical implication for ITS is that a single accuracy figure is not an adequate governance metric for a vision-derived traffic data feed. Monitoring should be stratified by object scale and by time of day, because these are the axes along which the data actually fails.

4.9. Limitations

Statistical power. With 736 images per condition and 300 bootstrap replicates, this study can resolve differences of roughly five percentage points in mean average precision. Smaller effects, including the 4–9% precipitation effects observed here, cannot be distinguished from zero. The non-significant results in Table 4 should be read as undetected, not absent.

Fog excluded. The validation split contains only 13 foggy images, far below the threshold required for a balanced comparison. Fog is therefore absent from this study, despite being among the most frequently discussed adverse conditions in the literature [1], [7]. This is a consequence of the source dataset's composition and is a genuine gap. Incomplete architecture comparison. Faster R-CNN was evaluated only on the clear-daytime subset. The architecture comparison in Section 4.5 therefore establishes how the families differ under favorable conditions but not how that difference behaves as conditions degrade.

Zero-shot evaluation. All models use COCO-pretrained weights without fine-tuning on the evaluation dataset. The results describe off-the-shelf deployment in unadapted conditions, which is realistic for many installations but understates what a dataset-adapted detector would achieve [7], [26]. Confounded conditions. As noted in Section 4.7, weather conditions differ in scene content as well as in image quality, so the signal-quality analysis supports association rather than causation.

Geographic and sensor scope. The source dataset was collected in the United States with a particular camera configuration. Generalization to other road environments, vehicle fleets, and sensor types is not established here. Category coverage. Evaluation is restricted to the eight categories common to both label spaces. Traffic signs, a category of direct relevance to ITS, are excluded because they have no counterpart in the pretrained label space.

5. CONCLUSION

This study evaluated the reliability of camera-derived traffic data under real adverse-weather conditions, using imagery selected based on the recorded weather and time-of-day attributes of a public benchmark dataset rather than on synthetic corruption. Three one-stage detectors and one two-stage detector were assessed on balanced subsets of 736 images per condition, with bootstrap confidence intervals reported for every measurement and a mapping established between detection metrics and the TDQ dimensions used in ITS practice.

Four findings emerge. First, precipitation did not produce a statistically detectable change in aggregate mean average precision for any detector tested: the point estimates ranged from −0.6% to −9.1% relative to clear daytime, but every confidence interval included zero at the sample size used. Second, night-time operation was the dominant degradation mode for aggregate accuracy, reducing mean average precision by 38.0% for YOLOv8n and 33.6% for YOLOv8s. Third, and most consequentially, small-object average precision degraded severely under exactly the conditions in which aggregate accuracy appeared stable, falling by 38–41% under precipitation and by up to 67% at night; traffic light detection lost 73% of its average precision at night. Fourth, the frequently reported advantage of one-stage over two-stage architectures was not reproduced as a property of the architecture family: the medium one-stage model outperformed the two-stage baseline. In contrast, both smaller one-stage models performed significantly worse than the two-stage baseline, indicating that model capacity, rather than the detection paradigm, accounts for the difference.

The practical implication is that a single accuracy figure is not sufficient to govern a vision-derived traffic data feed. A feed whose reported mean average precision has fallen only marginally under rain may nevertheless be failing to detect a large proportion of small road users in a population consisting disproportionately of pedestrians, cyclists and distant vehicles. Reliability monitoring for such systems should therefore be stratified by object scale and by time of day, and should treat coverage and timeliness as quantities distinct from accuracy rather than as aspects of it.

These conclusions are bounded by the limitations set out in Section 4.9. In particular, the non-significant precipitation results indicate effects that this study could not resolve rather than effects shown to be absent; fog was excluded because the source dataset contains too few foggy images to support a balanced comparison; and the architecture comparison was completed only for the clear-daytime condition. Extending the evaluation to a larger sample, to a dedicated adverse-condition dataset with adequate fog coverage, and to the full architecture comparison across all conditions would allow the precipitation effects reported here as undetected to be either established or excluded.

FUNDING INFORMATION

No funding was received for this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Shafiqullah Ahmadzai	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Emre Demi	✓	✓		✓			✓			✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

This study uses publicly available benchmark data. The authors collected no new data. The BDD100K dataset [14] is available from its distributors, subject to acceptance of its license terms. The derived per-condition evaluation subsets, detection outputs, and result files generated in this study are available from the corresponding author upon reasonable request.

Declaration of generative AI use. Generative AI assistance was used to develop the evaluation scripts that compute the reported metrics and to edit the English of this manuscript. No AI tool was used to generate or manipulate research data, figures, or tables. All imagery is from the public BDD100K benchmark, all reported values are outputs of detector inference and statistical evaluation performed by the

authors, and all figures were rendered directly from those outputs. The authors reviewed and verified all code, results, and text, and take full responsibility for the content of this manuscript.

REFERENCES

[1] N. U. A. Tahir, Z. Zhang, M. Asim, J. Chen, and M. ELAffendi, "Object detection in autonomous vehicles under adverse weather: a review of traditional and deep learning approaches," *Algorithms*, vol. 17, no. 3, p. 103, Feb. 2024, doi: 10.3390/a17030103.

[2] Y. Song and Y. Lu, "A review of unmanned visual target detection in adverse weather," *Electronics (Switzerland)*, vol. 14, no. 13, p. 2582, Jun. 2025, doi: 10.3390/electronics14132582.

[3] V. S. Patel, K. Agrawal, and T. V. Nguyen, "A comprehensive analysis of object detectors in adverse weather conditions," in *2024 58th Annual Conference on Information Sciences and Systems, CISS 2024*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/CISS59072.2024.10480197.

[4] A. Munir, A. J. Siddiqui, S. Anwar, A. El-Maleh, A. H. Khan, and A. Rehman, "Impact of adverse weather and image distortions on vision-based UAV detection: a performance evaluation of deep learning models," *Drones*, vol. 8, no. 11, p. 638, Nov. 2024, doi: 10.3390/drones8110638.

[5] E. Bayramov and Z. Istenes, "Weather-informed vision enhancement for autonomous vehicles in adverse conditions," *Pollack Periodica*, vol. 20, no. 3, pp. 88–95, Oct. 2025, doi: 10.1556/606.2025.01240.

[6] U. Gurbindo, A. Brando, J. Abella, and C. König, "Object detection in adverse weather conditions for autonomous vehicles using Instruct Pix2Pix," in *Proceedings of the International Joint Conference on Neural Networks*, 2025. doi: 10.1109/IJCNN64981.2025.11228623.

[7] J. Li, R. Xu, J. Ma, Q. Zou, J. Ma, and H. Yu, "Domain adaptive object detection for autonomous driving under foggy weather," in *Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, IEEE, Jan. 2023, pp. 612–622. doi: 10.1109/WACV56688.2023.00068.

[8] C. Sakaridis, D. Dai, and L. Van Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Oct. 2021, pp. 10745–10755. doi: 10.1109/ICCV48922.2021.01059.

[9] M. Rahmati, "Edge AI-powered real-time decision-making for autonomous vehicles in adverse weather conditions," 2025, [Online]. Available: http://arxiv.org/abs/2503.09638

[10] C. Michaelis *et al.*, "Benchmarking robustness in object detection: autonomous driving when winter is coming," 2020, [Online]. Available: http://arxiv.org/abs/1907.07484

[11] Battelle, Cambridge_Systematics, and Texas_Transportation_Institute, "Traffic data quality measurement," Federal highway administration. Accessed: Feb. 05, 2026. [Online]. Available: http://ntl.bts.gov/lib/jpodocs/repts_te/14058.htm

[12] Y. Bai, J. Wang, J. Su, Q. Zhou, and S. He, "Assessment of urban rail transit development using DPSIR-Entropy-TOPSIS and obstacle degree analysis: A case study of 27 Chinese cities," *Physica A: Statistical Mechanics and its Applications*, vol. 663, p. 130439, Apr. 2025, doi: 10.1016/j.physa.2025.130439.

[13] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A survey on performance metrics for object-detection algorithms," in *International Conference on Systems, Signals, and Image Processing*, IEEE, Jul. 2020, pp. 237–242. doi: 10.1109/IWSSIP48289.2020.9145130.

[14] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2020, pp. 2633–2642. doi: 10.1109/CVPR42600.2020.00271.

[15] Ultralytics, "Ultralytics YOLO, version 8.4.117." Accessed: Mar. 01, 2026. [Online]. Available: https://github.com/ultralytics/ultralytics

[16] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[18] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS, no. PART 5, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[19] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices," *Remote Sensing*, vol. 16, no. 3, p. 533, Jan. 2024, doi: 10.3390/rs16030533.

[20] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman and Hall, 1993.

[21] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011, doi: 10.1109/TPAMI.2010.168.

[22] S. Ahmed, M. N. Huda, S. Rajbhandari, C. Saha, M. Elshaw, and S. Kanarachos, "Pedestrian and cyclist detection and intent estimation for autonomous vehicles: A survey," *Applied Sciences (Switzerland)*, vol. 9, no. 11, p. 2335, Jun. 2019, doi: 10.3390/app9112335.

[23] S. Chadwick, W. Maddetn, and P. Newman, "Distant vehicle detection using radar and vision," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2019, pp. 8311–8317. doi: 10.1109/ICRA.2019.8794312.

[24] M. Nikouei *et al.*, "Small object detection: A comprehensive survey on challenges, techniques and real-world applications," *Intelligent Systems with Applications*, vol. 27, 2025, doi: 10.1016/j.iswa.2025.200561.

[25] H. Lin *et al.*, "A study on data selection for object detection in various lighting conditions for autonomous vehicles," *Journal of Imaging*, vol. 10, no. 7, p. 153, Jun. 2024, doi: 10.3390/jimaging10070153.

[26] S. Burdorf, K. Plum, and D. Hasenklever, "Reducing the amount of real world data for object detector training with synthetic data," 2022, [Online]. Available: http://arxiv.org/abs/2202.00632.

BIOGRAPHIES OF AUTHORS



Shafiqullah Ahmadzai received the B.Eng. degree in Civil Engineering from Paktia University, Gardez, Afghanistan, in 2023, and is currently pursuing the M.Sc. degree in Civil Engineering at Antalya Bilim University, Antalya, Türkiye, where he holds High Honour standing. He served as a Lecturer in Civil Engineering at Paktia University from 2023 to 2024 and currently lectures in transportation, foundation, and hydraulic engineering at Mustaqbal University, Kabul, Afghanistan. His research interests include intelligent transportation systems, traffic flow theory and intersection optimization, computer vision for automated traffic data collection, and multi-criteria decision-making methods such as AHP and TOPSIS applied to urban mobility planning. His master's research assesses urban traffic management and prioritizes signalized intersections in Kabul using a ten-criterion AHP/TOPSIS evaluation framework, supported by field data collected across fifteen high-congestion intersections. He can be contacted by email: 2423201@std.antalya.edu.tr.



Emre Demir received the B.Sc. degree in Civil Engineering from Boğaziçi University, Istanbul, Türkiye, in 2007, the M.Sc. degree in Transport Engineering and Operations from Newcastle University, Newcastle upon Tyne, U.K., in 2008, and the Ph.D. degree in Civil Engineering (Transportation) from the University of Missouri–Columbia, MO, USA, in 2013. He joined Antalya Bilim University, Antalya, Türkiye, in 2014, where he is currently an Associate Professor in the Department of Civil Engineering, Faculty of Engineering and Natural Sciences. His research interests include transportation engineering and traffic systems, network and facility-location optimisation, intelligent transportation systems, fuzzy and multi-criteria decision-making methods, and geospatial analysis for sustainable urban mobility. He has published in journals including the International Journal of Fuzzy Systems, the Journal of Cleaner Production, Neural Computing and Applications, and the International Journal of Geographical Information Science, and serves on the editorial and advisory board of the Journal of Intelligent Transportation Systems and Applications. He can be contacted at email: emre.demir@antalya.edu.tr.