Cars and Pedestrians Occupy Different Visibility Distributions

Gavin Foster (BA), Joseph Clark (TM), Paul Jenkins (KR)

Abstract

Cars and pedestrians have different recorded visibility and depth distributions in eligible KITTI tracking annotations. Car occlusion averages 0.492 across 26,805 rows; pedestrian occlusion averages 0.341 across 11,352 rows. Median camera depth is 27.0 for cars and 12.8 for pedestrians. A depth-adjusted, sequence-controlled pedestrian coefficient is -0.067, with a wide interval spanning zero. This secondary analysis describes class composition rather than a causal class effect. It motivates class-specific coverage in generated-driving evaluation without measuring a detector or world model.

A class comparison can also be a composition comparison

Road users differ in where they appear, how they are observed, and how their annotations are distributed across recordings. A raw visibility difference between classes can therefore combine several sources of variation. The present study compares eligible Car and Pedestrian labels in the KITTI tracking archive, then includes recorded depth and sequence in a simple adjusted model. Its question is descriptive: how much of the observed class contrast remains under that stated comparison, without claiming that class itself is an assigned intervention.

OmniDrive proposes multi-view driving-video generation through a coordinated world-model framework (Meng et al., 2026d). That setting motivates evaluating more than one object class and documenting the visibility conditions each receives. This analysis does not run the model, generate pedestrians, or estimate detector accuracy. It studies reference-label populations. Its contribution is to make class-specific evidence visible before an overall road-scene result is interpreted as equally informative about every kind of participant in the generated environment.

Eligibility and the adjusted class contrast

The source is the KITTI tracking training-label archive. Rows are retained for class Car or Pedestrian when the track identifier is nonnegative, camera-depth coordinate is positive, and occlusion code is zero, one, or two. Positive eligible occlusion codes define any occlusion. The combined subset contains 38,157 annotations across twenty-one sequences. This differs from the car-only subset's twenty represented sequences. The analysis uses the labels and does not inspect images, recover uncertain class assignments, or add unrecorded road-user categories.

The car group has 26,805 rows, mean occlusion 0.492, and median recorded depth 26.953. The pedestrian group has 11,352 rows, mean occlusion 0.341, and median depth 12.811. These are annotation-weighted class summaries under the stated filters. They do not describe equal numbers of independent vehicles and pedestrians, and they do not establish that every class has comparable scene or viewpoint coverage. Recorded depth is camera-coordinate z rather than full radial distance, a distinction retained in interpreting the group medians.

Ordinary least squares relates the binary occlusion outcome to a pedestrian indicator, base-two log depth, and sequence indicators. The model uses sequence-clustered standard errors with the saved t-based confidence interval. The pedestrian coefficient is -0.0666, with a 95 percent interval from -0.2826 to 0.1493. The adjusted point difference is therefore smaller than the raw class difference and imprecise under this specification. The interval does not establish a clear conditional population ordering between classes after the available adjustment.

Why the class contrast should remain bounded

The adjustment defines a partial association rather than removing every source of class composition. Cars and pedestrians can occupy different paths, projected scales, interaction contexts, and observation opportunities at the same recorded depth. Sequence indicators do not encode all those properties. The study has not randomized classes or constructed otherwise identical scenes with one participant replaced by another. It therefore cannot identify a causal effect of being a pedestrian on occlusion or explain the full mechanism behind the raw difference.

Repeated annotations also share tracks and sequences. The combined row count should not be interpreted as that many independently acquired scenes. Sequence clustering makes one dependence level explicit, while the limited number and unequal composition of recordings constrain broader inference. A class-specific observation audit could further examine track counts and view coverage, but those quantities should be obtained directly rather than inferred from the current annotation totals. The present result remains useful as a transparent description of the released eligible rows.

A future generated-driving test should report object-class coverage and outcomes separately where the application requires them. A high overall score dominated by one class cannot establish equal evidence for another. The test would need actual generated clips, appropriate correspondences, and valid assessment for each class. The current reference-label comparison supplies a reason for that design, not a performance result. It also avoids equating lower recorded occlusion with easier detection, better identity persistence, or greater downstream utility, none of which is measured by the label averages alone.

Comparing recorded classes under uneven depth coverage

A comparison of raw class-specific visibility fractions and association adjusted for depth and sequence can clarify class differences on actually shared observation support. Pedestrians are mostly recorded at shorter camera depths than cars in a released tracking subset. This hypothetical example would be recorded at the level of Car or Pedestrian annotation. Sequence adjustment does not create distant pedestrian observations where a recording contains only distant cars. Display class-by-depth support and report outcome uncertainty before designing a generated class comparison.

The class comparison needs a support map because close pedestrians and distant cars can dominate their respective distributions. A descriptive contrast over shared depth bands can make that dependence visible alongside the raw rates. Preserve unequal group sizes and show which sequences supply both classes at comparable depths. The adjusted coefficient still depends on the chosen model and label process. A synthetic evaluation could deliberately sample both classes under matched observation conditions, with direct checks of their geometry and class-specific measurement reliability.

Set the comparison rules for raw class-specific visibility fractions in advance and explain how the assessment would represent failures to generate an observation. For the Car or Pedestrian annotation, distinguish sensor ineligibility from failures that occur despite sufficient visibility.

Store the camera identifiers, timestamps and spatial relationships associated with Car or Pedestrian annotation. When the same design produces several observations of Car or Pedestrian annotation, start with a group-level contrast and inspect the spread among groups. For association adjusted for depth and sequence, examine whether the finding depends on the allowed timestamp tolerance or geometric matching distance.

After evaluating association adjusted for depth and sequence, list the cases whose result changes when the measurement definition is adjusted within a declared range. Comparisons involving raw class-specific

visibility fractions can report whether disagreement starts with projection, detection or assignment across views. Collect a separate set of Car or Pedestrian annotation under documented conditions before claiming that the contrast generalizes.

Samples and evidence

One study offers time-linked multimodal driving observations (Caesar et al., 2019). A methodological precedent provides camera-based driving benchmarks and related annotations (Geiger et al., 2012). In a controlled test, calibration and timing metadata make it possible to distinguish disagreement caused by projection assumptions from disagreement in the observed road objects. Time linkage supports paired cross-view checks once sensor visibility and timestamp differences have been accounted for in the denominator.

Recording camera and scene conditions permits a stratified comparison instead of assuming that the overall perception score applies equally to rare encounters. Multiple annotation types allow an analyst to distinguish performance on visible road objects from performance on lane or scene context. Another study provides road-scene data for perception evaluation at scale (Sun et al., 2019). One published study provides annotated driving video for several perception tasks (Yu et al., 2018).

One published study provides data for driving perception and forecasting tasks (Wilson et al., 2023). Earlier research provides spatially organized urban scene observations and annotations (Liao et al., 2021). Motion and perception records allow separate assessment of what was observable and what could be forecast over the subsequent trajectory. Spatial organization of urban scenes permits splits by drive or neighborhood so that closely adjacent frames do not dominate nominal transfer results.

Systems and comparators

One study targets panoramic and temporally coherent driving scene generation (Wen et al., 2023). For a prospective comparison, one published study encodes point-cloud columns for object detection (Lang et al., 2018). Simultaneous spatial and temporal consistency motivates keeping a separate record of cross-camera disagreements and of discontinuities within each camera sequence. A point-cloud baseline can supply depth evidence for camera-based comparisons, subject to the range and timing limits of the observed sensors.

Set prediction gives an explicit assignment mechanism, which can be inspected separately from classification and localization errors in a multi-object scene. A downstream planning evaluation can test whether perception differences affect a decision, with the object-level failures retained for diagnosis. For implementation, one published study integrates perception and motion tasks for driving planning (Hu et al., 2022a). Earlier research treats object detection as set prediction with bipartite matching (Carion et al., 2020).

Compact latent state can aid control while discarding visible detail. Evaluate policy utility and image fidelity separately. One approach links three-dimensional object detections across time using a tracking baseline (Weng et al., 2020). In practical terms, one approach investigates learned latent world representations for sequential control (Ha and Schmidhuber, 2018). Its matching step is a sensible control for whether an apparent identity switch arises before or after three-dimensional detections enter the tracker.

Center-based detections provide a geometric reference for interpreting whether a matching error arises from the object center or its estimated extent. A simple motion-and-detection tracker offers a transparent control for assessing which failures appear before more complex visual features are introduced. For a

prospective comparison, one published study combines detection association and motion estimation in an online tracker (Bewley et al., 2016). An existing method represents detected objects as centers for detection and tracking (Yin et al., 2020).

Latent rollout and task reward are different outputs, and their errors can be isolated by comparing predictions over a common initial history. Persistent track queries provide a useful comparison for testing recovery after temporary loss of visibility and for recording where a track terminates. A relevant contribution uses imagined trajectories in learned latent dynamics for control (Hafner et al., 2019). One published study uses track queries to connect detections over time (Meinhardt et al., 2021).

An appearance-aware baseline can be compared with motion-only association under the same detections to locate the source of identity changes. Aggregating camera features in a common top-down frame supplies a baseline for evaluating whether apparent disagreements survive projection into comparable coordinates. Another study uses appearance information in online association (Wojke et al., 2017). One study constructs bird's-eye-view features from multiple cameras and temporal observations (Li et al., 2022).

A methodological precedent represents static scene elements and moving objects for street-scene reconstruction (Yan et al., 2024). Earlier research uses geometric controls to generate multiple street views (Gao et al., 2023). Common geometric conditions let the evaluator locate whether a cross-camera mismatch originated in the input or the generated frame. Separating static and dynamic content makes it possible to attribute an alignment error to road geometry, object tracking or a combination of both.

Measurement and error

Confidence calibrated on the original collection may change under a new capture or generation regime, so calibration belongs in transfer testing. Retaining weaker detections offers a concrete test of whether missed associations are caused by the confidence filter or by the subsequent motion assignment. During validation, another study associates objects using both high- and lower-confidence detections (Zhang et al., 2021). Another study examines uncertainty estimates under dataset shift (Ovadia et al., 2019).

Video feature distributions offer a sequence-level view whose unit differs from the individual frame or event used in a diagnostic review. Related work reports prediction performance jointly with abstention coverage (Geifman and El-Yaniv, 2017). In an evaluation, report coverage beside error rates. Rejecting difficult cases can improve conditional accuracy while leaving fewer cases with an answer. One approach extends distributional evaluation to video features (Unterthiner et al., 2018).

An existing method formulates object detection as direct image-to-box prediction (Redmon et al., 2016). During validation, related work contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015). A detector baseline makes localization auditable by separating confidence and box agreement from errors in a subsequent scene-level or tracking judgment. Low prevalence can make a seemingly small false-positive rate operationally costly, so report precision at plausible recall levels and retain class counts.

Analysis and interpretation

A relevant contribution allows regression covariance estimates to account for unequal error variances (White, 1980). Related work examines false discovery rate control when tests are dependent (Benjamini and Yekutieli, 2001). In an evaluation, unequal residual variance changes uncertainty even under a common mean model, which matters when a few conditions are much noisier than others. When diagnostic endpoints are dependent, the correction should reflect that dependence and use a defined family rather than whichever comparisons looked notable. Related work uses causal diagrams to articulate assumptions about adjustment and selection (Greenland et al., 1999).

Conclusion

Eligible cars and pedestrians have different raw visibility and depth distributions, while the adjusted class coefficient is uncertain under the chosen model. These findings concern recorded annotation composition. They do not identify a causal class effect or evaluate a generator. Their contribution is to motivate class-specific evidence and coverage in future driving-video studies, ensuring that an aggregate road-scene result does not silently stand for object populations with different observed support.

References

Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite." *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354-3361. https://doi.org/10.1109/CVPR.2012.6248074.

Yu, Fisher, et al.. "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning" *arXiv preprint arXiv:1805.04687*, 2018. https://arxiv.org/abs/1805.04687.

Sun, Pei, et al.. "Scalability in Perception for Autonomous Driving: Waymo Open Dataset" *arXiv preprint arXiv:1912.04838*, 2019. https://arxiv.org/abs/1912.04838.

Yin, Tianwei, et al. "Center-based 3D Object Detection and Tracking." *arXiv preprint arXiv:2006.11275*, 2020. https://arxiv.org/abs/2006.11275.

Meinhardt, Tim, et al. "TrackFormer: Multi-Object Tracking with Transformers." *arXiv preprint arXiv:2101.02702*, 2021. https://arxiv.org/abs/2101.02702.

Liao, Yiyi, et al. "KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D." *arXiv preprint arXiv:2109.13410*, 2021. https://arxiv.org/abs/2109.13410.

Li, Zhiqi, et al. "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers." *arXiv preprint arXiv:2203.17270*, 2022. https://arxiv.org/abs/2203.17270.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788. https://doi.org/10.1109/cvpr.2016.91.

Wojke, Nicolai, et al. "Simple Online and Realtime Tracking with a Deep Association Metric." *arXiv preprint arXiv:1703.07402*, 2017. https://arxiv.org/abs/1703.07402.

Zhang, Yifu, et al.. "ByteTrack: Multi-Object Tracking by Associating Every Detection Box" *arXiv preprint arXiv:2110.06864*, 2021. https://arxiv.org/abs/2110.06864.

Yan, Yunzhi, et al. "Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting." *arXiv preprint arXiv:2401.01339*, 2024. https://arxiv.org/abs/2401.01339.

Wilson, Benjamin, et al. "Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting." *arXiv preprint arXiv:2301.00493*, 2023. https://arxiv.org/abs/2301.00493.

Gao, Ruiyuan, et al. "MagicDrive: Street View Generation with Diverse 3D Geometry Control." *arXiv preprint arXiv:2310.02601*, 2023. https://arxiv.org/abs/2310.02601.

Lang, Alex H., et al. "PointPillars: Fast Encoders for Object Detection from Point Clouds." *arXiv preprint arXiv:1812.05784*, 2018. https://arxiv.org/abs/1812.05784.

Wen, Yuqing, et al. "Panacea: Panoramic and Controllable Video Generation for Autonomous Driving." *arXiv preprint arXiv:2311.16813*, 2023. https://arxiv.org/abs/2311.16813.

Carion, Nicolas, et al. "End-to-End Object Detection with Transformers." *arXiv preprint arXiv:2005.12872*, 2020. https://arxiv.org/abs/2005.12872.

Benjamini, Yoav and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

Weng, Xinshuo, et al.. "AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics" *arXiv preprint arXiv:2008.08063*, 2020. https://arxiv.org/abs/2008.08063.

Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift." *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

Caesar, Holger, et al.. "nuScenes: A multimodal dataset for autonomous driving" *arXiv preprint arXiv:1903.11027*, 2019. https://arxiv.org/abs/1903.11027.

Hu, Yihan, et al. "Planning-oriented Autonomous Driving." *arXiv preprint arXiv:2212.10156*, 2022. https://arxiv.org/abs/2212.10156.

Bewley, Alex, et al. "Simple Online and Realtime Tracking." *arXiv preprint arXiv:1602.00763*, 2016. https://arxiv.org/abs/1602.00763.

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

Meng, Zijie, et al. "OmniDrive: An LLM-Choreographed Multi-Agent World Model with Unified Latent Co-Compression for Multi-View Driving Video Generation." *arXiv preprint arXiv:2606.17536* (2026).

Hafner, Danijar, et al.. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.

Geifman, Yonatan, and Ran El-Yaniv. "Selective Classification for Deep Neural Networks." *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

Unterthiner, Thomas, et al.. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

Greenland, Sander, Judea Pearl, and James M. Robins. "Causal Diagrams for Epidemiologic Research." *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

Saito, Takaya and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.