Occlusion Transitions within Observed Vehicle Tracks

Ryan Lewis (PE), Simon Brooks (TO), Arthur Hill (BD)

Abstract

Most adjacent eligible KITTI car annotations retain their previous binary visibility state. Among 26,203 consecutive-frame pairs, the visible-to-occluded rate is 1.80 percent, while the occluded-to-visible rate is 3.62 percent. These are conditional annotation transitions, not tracker errors or a complete model of physical visibility dynamics. This secondary analysis preserves track identity and frame adjacency. It motivates event-based evaluation of generated driving video while making clear that a label archive alone cannot establish temporal consistency in a synthesized scene.

A transition needs a linked observation history

An overall occlusion fraction says how often a label appears, but not how that state changes through a track. Two collections can have the same average with very different temporal patterns. Studying transitions requires explicit vehicle linkage and a rule for time adjacency. The present analysis examines consecutive recorded frame numbers within eligible KITTI car tracks. It asks how frequently the binary occlusion label changes, without interpreting every gap in the filtered archive as an observed transition.

OmniDrive proposes coordinated multi-view driving-video generation (Meng et al., 2026d). Its setting makes persistence through visibility changes a relevant assessment target. The current study does not inspect generated trajectories or evaluate the focal model. It uses public reference labels to define a temporal condition that a future test could investigate. An observed transition frequency in real annotations can guide sampling and endpoint design, but it is not evidence that a generator reproduces those dynamics or preserves compatible objects through them.

Constructing adjacent annotation pairs

Eligible rows are Car annotations with nonnegative track identifiers, positive camera-depth coordinate, and occlusion codes zero, one, or two. Any positive eligible occlusion code becomes the binary occluded state. Sequence identifiers are combined with track identifiers before sorting by frame. The analysis pairs an annotation with its predecessor only when their frame numbers differ by exactly one. This excludes transitions across gaps in the eligible record rather than silently treating them as the same elapsed observation interval.

The eligible archive has 26,805 rows in 574 sequence-specific tracks across twenty sequences. It yields 26,203 adjacent pairs. Of those, 13,021 remain visible, 12,475 remain occluded, 238 move from visible to occluded, and 469 move from occluded to visible. The visible-to-occluded conditional rate is therefore 238 divided by the 13,259 pairs starting visible, or 1.80 percent. The reverse rate is 469 divided by the 12,944 pairs starting occluded, or 3.62 percent.

The denominators differ because each rate conditions on its starting state. Comparing only the two transition counts would mix this difference in exposure with the transition behavior. Most observed adjacent pairs retain their previous binary state. That persistence is a feature of the released eligible labels, not an estimate of a universal road process. The analysis does not convert frame transitions into continuous-time hazards, infer unobserved intervals, or assume that every recorded vehicle has a complete history from physical arrival to departure.

Temporal persistence and the evaluation boundary

Annotation transitions can reflect changing visibility, coding conventions, observation boundaries, and the selection imposed by eligibility filters. The binary construction also combines different positive occlusion levels. A stable binary state can contain changes within that combined category. A more detailed transition study could retain the original codes, but it would answer another question. The present summary intentionally reports a simple defined state and preserves the information needed to understand what that simplification omits.

The paired observations are not independent. Many come from the same vehicle and recording sequence. This study reports descriptive counts and conditional rates rather than attaching a row-independent uncertainty interval that would exaggerate replication. A broader inferential study would need an explicit sampling target and an appropriate dependence structure. Likewise, these labels are not tracker predictions: a transition from visible to occluded is not a false negative or identity switch unless an actual algorithm's outputs are assessed under a separate matching protocol.

A future generated-video study could use declared occlusion events and assess object identity before, during, and after them. It would need synchronized evidence, a correspondence rule, and an uncertainty policy for genuinely unobservable intervals. A favorable average visual score would not establish success through those events. The present analysis supplies a motivation for event-level assessment while leaving its model outputs and performance outcomes unmeasured. That separation allows the reference archive to inform a test without pretending that its labels already validate a synthetic world's temporal behavior.

Transitions over a defined observation interval

A comparison of consecutive-frame state change and state comparison across an intervening gap can clarify conditional visibility transitions at the declared frame spacing. Two records separated by missing frames can appear adjacent when a filtered track is flattened. This hypothetical example would be recorded at the level of consecutive eligible car-frame pair. Pooling gaps with adjacent rows can imply observations during periods never recorded. Preserve frame number differences and count the starting-state denominator for each transition direction.

Create the transition table from actual frame-number differences. For each current state, count the adjacent eligible successors and calculate its conditional rate. Label gaps separately and keep their lengths available. A future measurement can inspect whether gaps correspond to severe occlusion or another acquisition feature. Combining them with consecutive frames would change the observation interval. The current transition table does not assess tracker predictions, which would require association outputs linked to the original annotations and a separate error rule.

A result about conditional visibility transitions at the declared frame spacing needs a denominator that includes all eligible attempts, including ones without a usable score. For the consecutive eligible car-frame pair, distinguish sensor ineligibility from failures that occur despite sufficient visibility.

Store the camera identifiers, timestamps and spatial relationships associated with consecutive eligible car-frame pair. A useful second view of state comparison across an intervening gap gives each independent group equal weight and retains the row-weighted result for comparison. For state comparison across an intervening gap, examine whether the finding depends on the allowed timestamp tolerance or geometric matching distance.

Use the comparison with consecutive-frame state change to identify which parts of the finding survive a reasonable change in the decision threshold. Comparisons involving consecutive-frame state change can

report whether disagreement starts with projection, detection or assignment across views. Collect a separate set of consecutive eligible car-frame pair under documented conditions before claiming that the contrast generalizes.

Samples and evidence

Another study provides road-scene data for perception evaluation at scale (Sun et al., 2019). Related work provides camera-based driving benchmarks and related annotations (Geiger et al., 2012). In an evaluation, recording camera and scene conditions permits a stratified comparison instead of assuming that the overall perception score applies equally to rare encounters. 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. In a controlled test, one study provides data for driving perception and forecasting tasks (Wilson et al., 2023). Motion and perception records allow separate assessment of what was observable and what could be forecast over the subsequent trajectory. One published study offers time-linked multimodal driving observations (Caesar et al., 2019).

Multiple annotation types allow an analyst to distinguish performance on visible road objects from performance on lane or scene context. Closed-loop assessment can reveal compounding consequences of small initial errors; open-loop frame evaluation answers a different question about recorded scenes. For implementation, one published study develops a closed-loop benchmark for vehicle planning (Caesar et al., 2021). An existing method provides annotated driving video for several perception tasks (Yu et al., 2018).

Systems and comparators

Aggregating camera features in a common top-down frame supplies a baseline for evaluating whether apparent disagreements survive projection into comparable coordinates. Structured inputs permit a comparison of controlled motion while keeping the initial scene and available layout observations explicit in the experimental record. For implementation, one approach constructs bird's-eye-view features from multiple cameras and temporal observations (Li et al., 2022). Another study uses structured conditions from driving scenes to guide video generation (Wang et al., 2023a).

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

A simple motion-and-detection tracker offers a transparent control for assessing which failures appear before more complex visual features are introduced. Another study addresses consistency of multi-view driving video synthesis (Li et al., 2023b). Use camera calibration to assess cross-view consistency. Two plausible images can disagree about the same object's location. One study combines detection association and motion estimation in an online tracker (Bewley et al., 2016).

Common geometric conditions let the evaluator locate whether a cross-camera mismatch originated in the input or the generated frame. Related work integrates perception and motion tasks for driving planning (Hu et al., 2022a). An existing method uses geometric controls to generate multiple street views (Gao et

al., 2023). For a prospective comparison, a downstream planning evaluation can test whether perception differences affect a decision, with the object-level failures retained for diagnosis.

A relevant contribution targets panoramic and temporally coherent driving scene generation (Wen et al., 2023). Earlier research represents static scene elements and moving objects for street-scene reconstruction (Yan et al., 2024). Simultaneous spatial and temporal consistency motivates keeping a separate record of cross-camera disagreements and of discontinuities within each camera sequence. For a prospective comparison, separating static and dynamic content makes it possible to attribute an alignment error to road geometry, object tracking or a combination of both.

Persistent track queries provide a useful comparison for testing recovery after temporary loss of visibility and for recording where a track terminates. One approach represents detected objects as centers for detection and tracking (Yin et al., 2020). In practical terms, related work uses track queries to connect detections over time (Meinhardt et al., 2021). Center-based detections provide a geometric reference for interpreting whether a matching error arises from the object center or its estimated extent.

Measurement and error

A detector baseline makes localization auditable by separating confidence and box agreement from errors in a subsequent scene-level or tracking judgment. In practical terms, confidence calibrated on the original collection may change under a new capture or generation regime, so calibration belongs in transfer testing. One published study formulates object detection as direct image-to-box prediction (Redmon et al., 2016). One approach examines uncertainty estimates under dataset shift (Ovadia et al., 2019).

Related work provides object annotations and standardized visual recognition tasks (Lin et al., 2014). A relevant contribution formulates classification decisions with explicit costs for different errors (Elkan, 2001). Object annotations support localized checks, with the class mapping and annotation policy recorded before evaluating cases absent from the detector vocabulary. An application with asymmetric errors calls for costs to be stated separately from the metric so that operating thresholds can be recalculated.

Retaining weaker detections offers a concrete test of whether missed associations are caused by the confidence filter or by the subsequent motion assignment. An existing method combines detection and association considerations in tracking assessment (Luiten et al., 2020). In practical terms, related work associates objects using both high- and lower-confidence detections (Zhang et al., 2021). Report the detection and association components as well as the aggregate so that tracking continuity can be interpreted alongside object availability.

Analysis and interpretation

A missingness analysis distinguishes failures from observations that were never eligible and makes the assumptions behind complete-case summaries visible to readers. A simulator can vary visibility, traffic and camera placement independently while leaving the underlying events observable for verification against the rendered frames. For a prospective comparison, another study separates missing-data assumptions from observed-data summaries (Little and Rubin, 2002). One published study provides a configurable driving simulation environment (Dosovitskiy et al., 2017). Earlier research contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015).

Unequal residual variance changes uncertainty even under a common mean model, which matters when a few conditions are much noisier than others. During validation, a longitudinal formulation can separate

baseline differences between units from changes within the same unit as the sequence progresses. A relevant contribution allows regression covariance estimates to account for unequal error variances (White, 1980). Earlier research models repeated longitudinal measurements with random effects (Laird and Ware, 1982).

Conclusion

Adjacent eligible car labels are highly persistent, with different conditional rates in the two directions of binary visibility change. The result is a descriptive transition table under a strict consecutive-frame rule. It does not measure a tracker, a continuous-time process, or OmniDrive's outputs. Its contribution is a clearly defined temporal condition for future generated-scene evaluation, preserving track linkage and missing intervals rather than replacing them with an undifferentiated frame average.

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.

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

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).

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.

Luiten, Jonathon, et al.. "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking" *arXiv preprint arXiv:2009.07736*, 2020. https://arxiv.org/abs/2009.07736.

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

Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.

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.

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.

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.

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.

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.

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

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.

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.

Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

Wang, Xiaofeng, et al. "DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving." *arXiv preprint arXiv:2309.09777*, 2023. https://arxiv.org/abs/2309.09777.

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.

Dosovitskiy, Alexey, et al. "CARLA: An Open Urban Driving Simulator." *arXiv preprint arXiv:1711.03938*, 2017. https://arxiv.org/abs/1711.03938.

Lin, Tsung-Yi, et al. “Microsoft COCO: Common Objects in Context.” *Lecture Notes in Computer Science*, 2014, pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48.

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.

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.

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.

Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

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

Caesar, Holger, et al. "NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles." *arXiv preprint arXiv:2106.11810*, 2021. https://arxiv.org/abs/2106.11810.

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.

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.

Li, Xiaofan, et al. "DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model." *arXiv preprint arXiv:2310.07771*, 2023. https://arxiv.org/abs/2310.07771.