Journal of Measurement and Evidence
Volume 1 · Number 1 · 2026

When a Stationary Vehicle Moves in the Image: Camera and Subject Contributions in Virtual KITTI 2

Jordan Allen[1] · Adrian Porter[2] · Ryan King[3]

[1] Greece · [2] Estonia · [3] Latvia

Original publication: October 2, 2026 | Article 672

Abstract

A vehicle can be stationary in a scene while its projected location changes substantially. In five Virtual KITTI 2 clone scenes, 77.85 percent of 4,633 source-labeled stationary vehicle pairs exceeded a declared normalized image-plane displacement threshold. Among 6,697 moving pairs, a camera-first geometric decomposition assigned the larger component to camera change in 61.46 percent, and the two component vectors opposed one another in 79.12 percent. These are measurements from released synthetic geometry, not new real-driving observations. The decomposition is path-dependent and additive by construction; the empirical findings concern its component magnitudes and directions. Apparent image motion alone does not identify the movement of the vehicle reference point.

Geometric reanalysis of a public synthetic benchmark. No optical-flow network or subject-motion model was run.

When a Stationary Vehicle Moves in the Image: Camera and Subject Contributions in Virtual KITTI 2

Geometric reanalysis of a public synthetic benchmark. No optical-flow network or subject-motion model was run.

Abstract

A vehicle can be stationary in a scene while its projected location changes substantially. In five Virtual KITTI 2 clone scenes, 77.85 percent of 4,633 source-labeled stationary vehicle pairs exceeded a declared normalized image-plane displacement threshold. Among 6,697 moving pairs, a camera-first geometric decomposition assigned the larger component to camera change in 61.46 percent, and the two component vectors opposed one another in 79.12 percent. These are measurements from released synthetic geometry, not new real-driving observations. The decomposition is path-dependent and additive by construction; the empirical findings concern its component magnitudes and directions. Apparent image motion alone does not identify the movement of the vehicle reference point.

1. The coordinate of a moving image point has two histories

An image location changes when an object moves relative to a camera. That relation leaves several possibilities: the object may move, the camera may move, or both may move in partially canceling directions. A scalar image-displacement magnitude cannot by itself distinguish those possibilities. This becomes particularly relevant when a system is expected to preserve or separately control camera movement and subject movement.

The research question is how often camera change supplies a large contribution to apparent motion of vehicle reference points in a public scene collection. We examine source-labeled stationary and moving objects and calculate a declared geometric decomposition from the supplied world and camera coordinates. The study uses released synthetic ground truth. It does not estimate motion from pixels or assess whether a learned model correctly disentangles those sources.

The disentangling objective in Meng, 2026 motivates this measurement distinction. Our analysis does not implement its attention mechanism or evaluate its generated outputs. It asks what an elementary apparent-motion observation can support before such a system is introduced. A benchmark fact about camera and object coordinates can clarify an evaluation prerequisite without being evidence for or against a particular generation architecture.

2. A public virtual world supplies information absent from an ordinary video

We use the metadata release of Yohann et al., 2020, an adaptation of the virtual-world resource introduced in Gaidon et al., 2016. The archive supplies camera parameters, object reference-point locations, bounding-box information, and track identities. Its geometry allows a counterfactual projection of an object point held fixed while the camera changes. An ordinary video without those annotations would require additional estimation assumptions to carry out the same calculation.

We retain five clone scenes and the left camera only. Weather variants, yaw variants, and the second stereo camera are not treated as independent replications of the world dynamics. This avoids inflating scene count by reusing related renderings of the same underlying configuration. The observed population is five supplied virtual trajectories, rather than fifty independent driving environments or a random sample of real road traffic.

The geometric framework in Hartley and Zisserman, 2004 provides the relevant distinction between world coordinates, camera coordinates, and projection. Reconstruction systems such as Schonberger and Frahm, 2016 and Mur-Artal and Tardos, 2016 estimate related quantities from images, but are not run here. Depth methods in Ranftl et al., 2019 and Godard et al., 2018 also address information that we obtain directly from the source geometry. Their performance is neither assumed nor used as an uncertainty estimate for the released annotations.

The driving collections in Geiger et al., 2012, Caesar et al., 2019, Sun et al., 2019, and Yu et al., 2018 provide other acquisition and annotation settings. The simulator in Dosovitskiy et al., 2017 adds a separate synthetic setting rather than a real-world replication. No observations from any of these sources enter the fitted regression. Their comparison clarifies why the present Virtual KITTI metadata should retain its source identity: relation to driving benchmarks does not itself validate transfer of a geometric association to other sensors or scenes.

3. Define a visible reference point, not a dense motion field

The source's object position is the center of the bottom face of its three-dimensional bounding box. We use that same point in world and camera coordinates. It is not the center of a two-dimensional image rectangle or a tracked image texture. Its projected displacement therefore measures one reference-point trajectory, rather than dense optical flow, every visible surface point, or a perceptual judgment about how the entire car appears to move.

For each scene and track in the left camera, we join consecutive frame observations. Both endpoints must have supplied depth between five and eighty metres, occupancy ratio at least 0.25, and truncation ratio no greater than 0.1. Those restrictions define a relatively visible, positive-depth sample. They exclude heavily occluded, strongly truncated, very distant, and very near cases. The conclusions should not be generalized to every object present in the rendered scenes.

The source occupancy ratio describes non-occluded object pixels independently of truncation, while truncation describes image-boundary loss. We preserve those distinct roles rather than combining them into a single generic visibility flag. Track identities join the same annotated vehicle across frames. That use of supplied identities does not evaluate a tracking method. Tracking and association questions addressed by Weng et al., 2020 and Luiten et al., 2020 require an algorithm to recover relationships that are given to us here.

The retained sample contains 11,330 object-frame pairs. Scene contributions are 2,916, 1,038, 643, 1,168, and 5,565 pairs for scenes 01, 02, 06, 18, and 20, respectively. The largest scene therefore contributes nearly half the object-frame observations. We retain object-frame weighting in the regression and disclose those counts; we do not describe the sample as an equal-scene experiment simply because it has five scene labels.

4. Validate the coordinate transformation before interpreting it

Camera extrinsic matrices transform the released world point into a camera-space point. We apply that transformation and compare it to the independently supplied camera-coordinate fields for the same annotated point. The largest absolute coordinate discrepancy across retained current observations is about 0.000163 metres. This numerical agreement supports the chosen matrix direction and field parsing within the precision of the released text metadata.

We obtain normalized image-plane coordinates by dividing camera x and y by positive camera depth. These dimensionless coordinates do not include the focal-length conversion to pixels. A displacement

threshold of 0.002 therefore belongs to this declared normalized plane. It is not a universal perceptual-motion threshold, a claim about a screen's pixel resolution, or an estimate of how easily a human observer detects motion in a rendered video.

Measurement-error considerations in Carroll et al., 2006 remain relevant because the matrices and points are stored with finite textual precision. The coordinate-consistency check does not certify perfect ground truth or an error-free renderer. It verifies that our calculation matches the supplied representation closely enough for the stated descriptive exercise. We retain that distinction rather than treating a successful parser check as independent validation of the physical reality of a synthetic scene.

5. A camera-first decomposition with a declared order

Let the current point be observed through the current camera and the next point through the next camera. The total apparent change is the difference between those two projected positions. To define a camera component, we additionally project the current world point through the next camera. The change from its original projection to that intermediate projection is assigned to camera change. Moving the point from its current to next world position under the next camera supplies the remaining subject component.

The two vectors add exactly to the total apparent change. This is an algebraic construction, not an empirical discovery. Reversing the order of the counterfactual operations could assign different component values because perspective projection is nonlinear. The components are consequently interpretable only under the stated camera-first path. We do not label them uniquely identified causal contributions to a perceptual outcome.

We calculate the Euclidean magnitude of each vector and whether their dot product is negative. A negative dot product indicates opposing components in the normalized image plane. Their magnitudes need not add to the total magnitude, since vectors can cancel. This distinction matters when a small apparent displacement is mistakenly read as evidence that neither camera nor subject moved much.

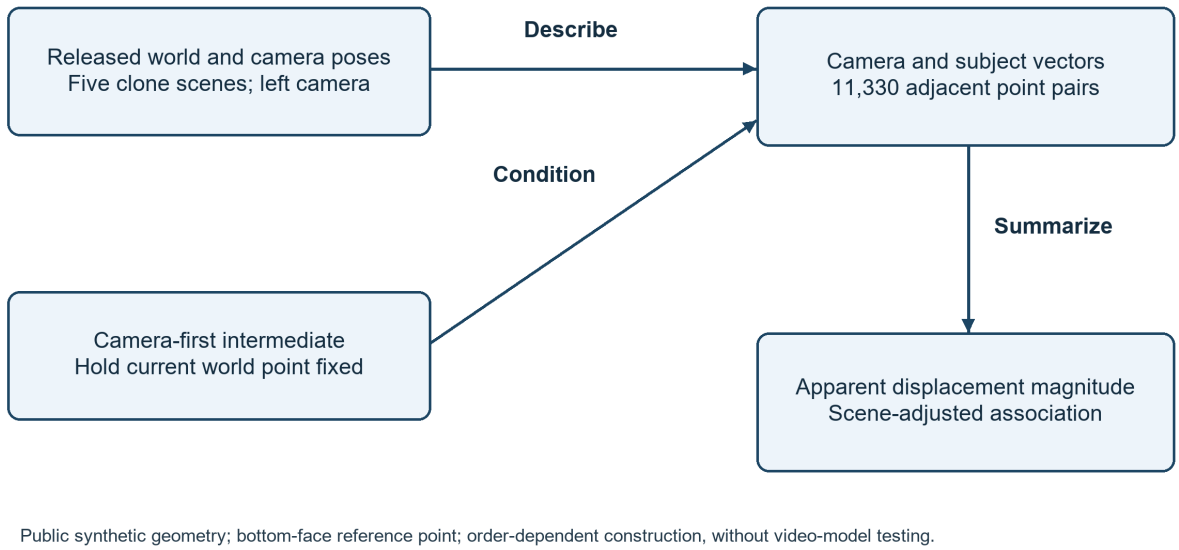

Figure 1. Released camera and world coordinates define a camera-first intermediate projection and two additive displacement vectors.

6. Apparent movement is common among source-labeled stationary pairs

The sample contains 4,633 pairs labeled stationary by the source's between-frame moving flag. Of these, 77.85 percent have total normalized-plane displacement greater than 0.002. That is an exact descriptive fraction for the retained source-labeled subset and threshold. It shows that a simple threshold on apparent point movement can flag many objects whose supplied moving label is false. It is not a measured false-positive rate of an actual deployed motion detector.

Among the 6,697 pairs labeled moving, the camera component has larger magnitude than the subject component in 61.46 percent. The component vectors oppose one another in 79.12 percent. These fractions show why the magnitude of total motion is not a direct proxy for the magnitude of world-object displacement. They remain conditional on the camera-first order, visibility restrictions, and public synthetic trajectories used here.

Across all retained pairs, mean total displacement magnitude is 0.01054, mean camera-component magnitude 0.00968, and mean subject-component magnitude 0.00365. Their means do not satisfy a scalar addition identity and should not be expected to do so. The valid identity concerns the two-dimensional vectors before magnitudes are taken. Reporting only scalar means without that explanation would invite an incorrect decomposition interpretation.

7. A descriptive regression retains separate geometric predictors

We regress apparent bearing-displacement magnitude on the object's world-displacement magnitude, camera-center translation magnitude, camera rotation angle, inverse current depth, and scene indicators. Distances refer to adjacent frames, not a speed inferred from an assumed frame rate. Rotation is measured in radians between the supplied camera orientations. The predictors summarize different parts of the geometry while leaving their directional relationships compressed into scalar magnitudes.

Covariance clusters by scene following the dependence considerations in Cameron and Miller, 2015, with the recorded correction and a t reference using four degrees of freedom. There are only five clusters, so inference is fragile. Heteroskedasticity correction in White, 1980 alone would not address shared scene trajectories. Repeated-observation frameworks such as Laird and Ware, 1982 are possible alternatives, but are not additional models fitted in this report.

Table 1. Geometric associations with normalized apparent point-displacement magnitude, including scene fixed effects.

Model	Term	Estimate (SE)	95% CI	N
M1	Object world displacement	0.0038 (0.0067)	[-0.0149, 0.0225]	11330
M1	Camera translation	0.0076 (0.0023)	[0.0014, 0.0139]	11330
M1	Camera rotation, radians	0.7667 (0.1162)	[0.4442, 1.0892]	11330
M1	Inverse current depth	0.3487 (0.0713)	[0.1507, 0.5468]	11330
M1: fit	R-squared = 0.4881	-	-	11330

Model labels: M1 = Apparent bearing change and motion components.

The model R-squared is 0.4881. Camera translation has coefficient 0.00764, with interval 0.00139 to 0.01389. Camera rotation has coefficient 0.76670 per radian, with interval 0.44421 to 1.08919. Inverse depth has coefficient 0.34872, with interval 0.15069 to 0.54675. These are conditional descriptive

associations among quantities derived from the same geometry, not independently identified effects of changing a camera setting.

The world-displacement coefficient is 0.00380, with interval -0.01492 to 0.02253. Its uncertainty and scalar specification prevent a confident directional conclusion from that coefficient alone. It does not imply that object movement has no geometric consequence. Directional cancellation, depth, selection, and shared scene structure can make a simple magnitude regression an incomplete summary of the underlying projection relationships.

8. The measurement result does not validate a generation or tracking system

Agreement concepts in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 concern measurement properties that require their own appropriate units and repetitions. The coordinate-consistency check here is not a new inter-observer reliability study. Nor do the component fractions evaluate identity preservation, correct texture motion, or visual quality. No rendered frames, predicted tracks, optical-flow fields, or learned subject representations enter the outcome calculation.

Validation concerns in Roberts et al., 2017 and selection-bias reasoning in Varma and Simon, 2006 are relevant if the sample restrictions or thresholds are later tuned to support a method comparison. The present report keeps the declared depth and visibility rules and the one normalized threshold. Influence analysis under Cook, 1977 could identify dominant scenes or tracks, but no scene is removed because its geometry weakens a desired association. Five virtual scenes remain a small basis for generalization.

Missing-data analysis in Little and Rubin, 2002 also applies to objects absent from one endpoint or excluded by visibility criteria. Such objects can have systematically different motion. The retained consecutive pairs do not estimate behavior during full occlusion or reappearance, and the source identities do not supply the missing image evidence. A broader evaluation would need to retain those cases under a separately defined target rather than treating them as ordinary observed motion pairs.

9. A useful distinction for evaluating motion claims

The interpretation principles in Wasserstein and Lazar, 2016 and concerns in Ioannidis, 2005 support keeping the exact descriptive fractions separate from the uncertain regression coefficient and from broader methodological claims. The clear empirical point is about a representation: apparent reference-point movement often contains a large camera contribution in the retained geometry. It does not establish an optimal architecture for separating motion or the external validity of a benchmark score on real roads.

A future evaluation could apply specified models to actual frames and compare their inferred camera and subject components against a declared geometric target. It would need to decide whether the target uses this camera-first path, another decomposition, or a different motion quantity entirely. Keeping that definition explicit would prevent an apparently favorable score from mixing bounding-box motion, optical flow, camera translation, and subject movement as if they were one interchangeable endpoint.

Documentation under Gebru et al., 2021 records scene and camera selection, the bottom-face reference point, projection validation, pair eligibility, decomposition order, and the threshold. Wilkinson et al., 2016 motivates preserving the public metadata and processed vectors while retaining Naver's source attribution and data terms. The study supplies an auditable example of why camera movement and subject movement require distinct measurements. Its empirical scope remains the five released synthetic trajectories and the declared geometric construction.

References

1. Hartley, Richard; Zisserman, Andrew. (2004). Multiple View Geometry in Computer Vision. Cambridge University Press. https://doi.org/10.1017/cbo9780511811685.
2. 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.
3. 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.
4. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.
5. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
6. Caesar, Holger, et al. "nuScenes: A multimodal dataset for autonomous driving" *arXiv preprint arXiv:1903.11027*, 2019. https://arxiv.org/abs/1903.11027.
7. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
8. 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.
9. 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.
10. Godard, Clément; Mac Aodha, Oisin; Firman, Michael; Brostow, Gabriel. (2018). Digging Into Self-Supervised Monocular Depth Estimation. https://arxiv.org/abs/1806.01260.
11. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
12. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
13. Mur-Artal, Raul; Tardos, Juan D.. (2016). ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. https://arxiv.org/abs/1610.06475.
14. 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.
15. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
16. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
17. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.
18. Gaidon, Adrien; Wang, Qiao; Cabon, Yohann; Vig, Eleonora. (2016). Virtual Worlds as Proxy for Multi-Object Tracking Analysis. CVPR 2016, 4340-4349. https://openaccess.thecvf.com/content_cvpr_2016/html/Gaidon_Virtual_Worlds_as_CVPR_2016_paper.html.
19. 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.
20. Ranftl, René; Lasinger, Katrin; Hafner, David; Schindler, Konrad; Koltun, Vladlen. (2019). Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. https://arxiv.org/abs/1907.01341.
21. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

22. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

23. Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

24. Dosovitskiy, Alexey; Ros, German; Codevilla, Felipe; Lopez, Antonio; Koltun, Vladlen. (2017). CARLA: An Open Urban Driving Simulator. arXiv:1711.03938. https://arxiv.org/abs/1711.03938.

25. Meng, Zijie. (2026). OrthoMotion:Disentangling Camera and Subject Motion via Geometry Semantics Orthogonal Attention. arXiv:2606.22835. https://arxiv.org/abs/2606.22835.

26. Schonberger, Johannes L.; Frahm, Jan-Michael. (2016). Structure-from-Motion Revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.445.

27. Varma, Sudhir and Richard Simon. "Bias in error estimation when using cross-validation for model selection." *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

28. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

29. Geiger, A.; Lenz, P.; Urtasun, R.. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2012.6248074.

30. Yohann, Cabon,; Naila, Murray,; Martin, Humenberger,. (2020). Virtual KITTI 2. https://arxiv.org/abs/2001.10773.