Uncertainty and Assessment
Volume 1 · Number 1 · 2026

Moving in the World or Moving in the Image? Agreement between Two Explicit Motion Definitions

Cameron Evans[1] · Andrew Stone[2] · James Harrison[3]

[1] Greece · [2] Croatia · [3] Serbia

Original publication: October 9, 2026 | Article 681

Abstract

A world-motion label and an image-motion threshold need not describe the same event. On 11,330 tracked Virtual KITTI 2 point pairs, a normalized-image displacement gate disagrees with the source moving label in 6,727 cases. It is positive for 1,796 source-stationary pairs and negative for 4,931 source-moving pairs. These are disagreements between constructs, not measured detector errors. The analysis identifies the consequences of substituting apparent motion for a world-motion label, using actual geometry and five-scene working uncertainty while leaving the appropriate definition to the intended task.

Moving in the World or Moving in the Image? Agreement between Two Explicit Motion Definitions

Abstract

A world-motion label and an image-motion threshold need not describe the same event. On 11,330 tracked Virtual KITTI 2 point pairs, a normalized-image displacement gate disagrees with the source moving label in 6,727 cases. It is positive for 1,796 source-stationary pairs and negative for 4,931 source-moving pairs. These are disagreements between constructs, not measured detector errors. The analysis identifies the consequences of substituting apparent motion for a world-motion label, using actual geometry and five-scene working uncertainty while leaving the appropriate definition to the intended task.

1. The word moving leaves an observation frame unstated

A vehicle can remain fixed in the world while moving across a camera's image. Another vehicle can change its world position while showing only a small projected displacement because of depth, direction, or camera motion. Calling both observations moving without stating the frame can create an apparent disagreement that is really a change of construct. The research question is how much a declared apparent-motion gate agrees with the source world-motion label on the same tracked reference points, and where the disagreement occurs.

The distinction is relevant to the camera and subject separation addressed by Meng, 2026. Our comparison does not implement that model or treat the released motion label as a universal correctness standard for its outputs. We calculate an apparent-motion indicator from geometric projections and compare it with the source isMoving metadata. Each represents a different operational question. The purpose is to identify what is lost when one is used as a stand-in for the other, rather than to rank two motion-detection algorithms that were never run.

The public synthetic resource in Yohann et al., 2020, following the earlier setting in Gaidon et al., 2016, provides both tracked world information and calibrated camera geometry. This allows the apparent indicator to be computed without image-based tracking or optical-flow estimation. That advantage should not be overstated. The observed point is a tracked bottom-face reference, not an entire object silhouette, and the sample comprises five selected clone scenes. The comparison therefore concerns explicit geometric labels in a bounded source setting, with no claim about human motion perception or all road environments.

2. Definitions before agreement coefficients

The framework in Hartley and Zisserman, 2004 explains why world displacement and image displacement are different quantities. Projection combines the point's world location with camera pose and depth. The pose treatments in Barfoot, 2017, and estimation formulations represented by Triggs et al., 2000, Lourakis and Argyros, 2009, and Lepetit et al., 2009, likewise require a declared reference frame and observation model. We estimate none of those models here. The supplied geometry is sufficient to show why two legitimate motion definitions can disagree even when there is no measurement error or failed detector.

The measurement perspective in Stevens, 1946 therefore precedes a numerical agreement score. A binary source label and a thresholded continuous projection can be compared, but they should not be assumed to be alternate noisy measurements of a single latent event. Carroll et al., 2006 distinguishes uncertainty in measurements from the relation between a proxy and its target. Our apparent indicator becomes a proxy

for world movement only if a particular task chooses to use it that way. The resulting mismatch must then be interpreted relative to that choice.

Agreement literature makes additional distinctions. Bland and Altman, 1986 and Lin, 1989 address comparisons among measurement procedures, while Fleiss, 1971 concerns agreement across categorical ratings. Reliability treatments in Koo and Li, 2016 and Shrout and Fleiss, 1979 emphasize the design behind an agreement coefficient. None is calculated as a substitute for the present confusion counts. A single agreement number would not reveal whether disagreement comes from image motion in source-stationary cases or from subthreshold image motion in source-moving cases. We keep those directions explicit and avoid treating either label as the other's definitive ground truth.

The label-noise discussion in Natarajan et al., 2013 further cautions against declaring disagreement a corrupted annotation by default. A source-stationary object with a positive apparent displacement can be entirely consistent with a moving camera. A source-moving object with a small apparent displacement can likewise be geometrically coherent. Calling those records false positives and false negatives would imply a target definition that the comparison has not established. We use source-stationary apparent positive and source-moving apparent negative instead, preserving the construct distinction in the result labels.

3. One cohort, two operational labels

The retained cohort has 11,330 adjacent point pairs from five clone scenes and the left camera. Track and frame keys identify the same source point at consecutive frames. Both endpoint camera depths must fall between five and eighty meters, occupancy must be at least 0.25, and truncation ratio no greater than 0.1. An inherited intermediate projection must remain in front of the camera. These visibility and depth restrictions are fixed before the present label comparison, and no pair is removed because the labels disagree.

The source moving indicator is read from the current record's released isMoving field. We do not rederive or repair it using the magnitude of the projected motion. There are 6,697 source-moving and 4,633 source-stationary pairs. The apparent indicator is one when the norm of the change in normalized image coordinates is at least 0.01. This cutoff is an illustrative geometric gate chosen for the present comparison. It is not an exact angular threshold in radians, a learned decision boundary, or a source-validated definition of physical motion.

Normalized coordinates remove the direct numerical pixel focal scale, but they do not remove depth, camera rotation, or the direction of point motion. The apparent value is not dense optical flow and is not a bounding-box center estimated from images. It follows a single three-dimensional reference point through two camera projections. A large object could have internal or edge motion that this reference does not capture. Conversely, a projected point can move substantially without the object changing its world position, which is precisely why the two indicators deserve separate names.

The source scenes are unevenly represented. Scene20 supplies 5,565 pairs and Scene01 supplies 2,916, while Scene02, Scene18, and Scene06 supply 1,038, 1,168, and 643. Their moving-label compositions also differ sharply. Scene18 has only two source-stationary pairs, whereas Scene01 has many. This imbalance limits any intuitive description of an average stationary object. We retain the actual counts and use a scene-clustered working summary, rather than balancing labels retrospectively in a way that would alter the finite population being described.

4. The research model and what the table estimates

Figure 1 presents both definitions as outputs from different parts of the observation description. The world-motion label is not an outcome caused by the apparent gate, and the gate does not evaluate a learned detector. The first statistical outcome is an indicator that the two labels differ. The second is the apparent-positive indicator restricted to source-stationary pairs. The latter quantifies an especially interpretable disagreement direction while leaving open whether apparent motion is undesirable or useful for a particular application.

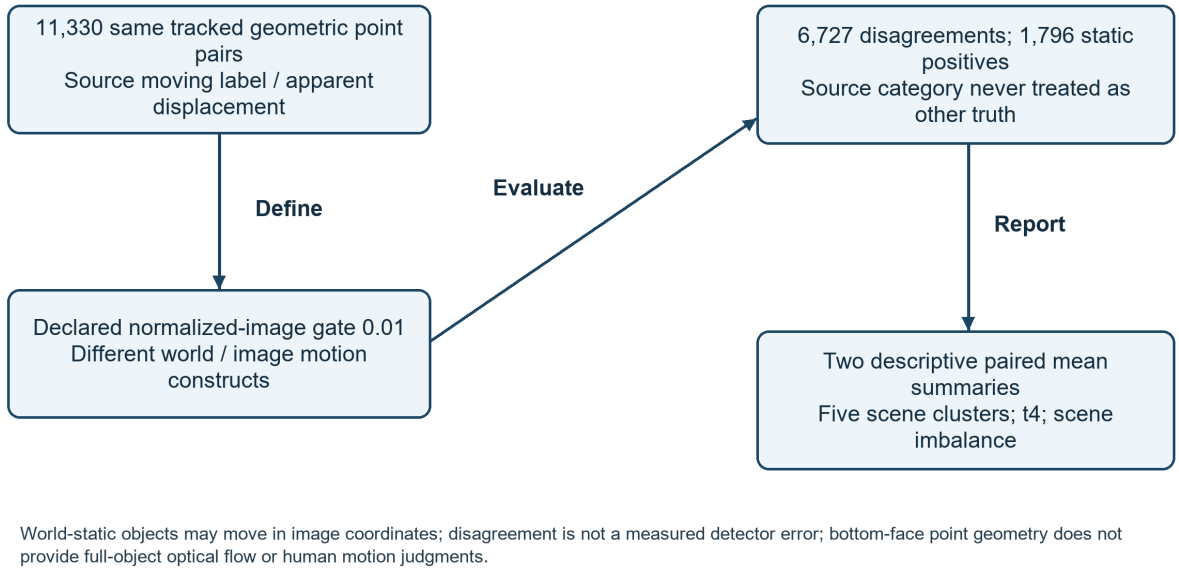

Figure 1. *A source world-motion label and a declared normalized-image gate classify the same tracked point pairs. Camera projection and thresholding can change the classification meaning without creating an annotation or detector failure.*

Both outcomes are summarized by intercept-only OLS with covariance clustered by the five scenes, finite-sample correction, and a Student reference with four degrees of freedom. White, 1980 motivates heteroskedasticity-aware covariance, and Cameron and Miller, 2015 highlights the limited number of independent working groups. The first fit contains all 11,330 pairs; the second contains the 4,633 source-stationary pairs. There are no controls, learned calibration parameters, or fixed effects. Repeated tracks and shared source trajectories remain inside the scene groups rather than being treated as independent replications.

The table's intervals describe uncertainty under that working model, not the probability that a motion label is physically correct. The finite confusion counts are exact for the stored table and chosen threshold. Different scenes, another gate, or another reference point could change them. We do not tune the cutoff to maximize agreement and then report the tuned result as validation. The reporting perspective in Wasserstein and Lazar, 2016 favors showing these definitional and sampling limits together with the effect magnitudes rather than reducing the analysis to a significance label.

Table 1. *Intercept-only OLS means for label disagreement and apparent-positive status among source-stationary pairs. N counts eligible point pairs in each fit. Covariance uses five scene clusters with finite-sample correction and a t reference with four degrees of freedom. There are no controls. Neither model estimates detector sensitivity against an agreed physical ground truth.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.593733 (0.070584)	[0.397762, 0.789705]	11330
M1: fit	R-squared = 0.0000	-	-	11330
M2	Mean outcome/contrast	0.387654 (0.080737)	[0.163491, 0.611817]	4633
M2: fit	R-squared = 0.0000	-	-	4633

Model labels: M1 = Mean disagreement between source and apparent-motion definitions; M2 = Apparent-positive fraction among source-stationary pairs.

5. Disagreement is common and runs in both directions

Among the 4,633 source-stationary pairs, 2,837 are below the apparent gate and 1,796 meet it. Among the 6,697 source-moving pairs, 4,931 are below the gate and 1,766 meet it. The total disagreement count is consequently 6,727. Its mean is 0.593733, with standard error 0.070584 and a 95% interval from 0.397762 to 0.789705. These values state how different the operational labels are in this cohort; they do not show that either label is poorly measured relative to its own definition.

The apparent-positive fraction among source-stationary pairs is 0.387654, with an interval from 0.163491 to 0.611817. The corresponding fraction among source-moving pairs is 0.263700, reported descriptively. The ordering may seem surprising if the gate is assumed to measure world motion directly. It is compatible with the different camera paths, depths, directions, and source-category composition. We do not identify a single mechanism from these group rates, and we do not interpret their difference as a causal camera effect.

Scene-level disagreement counts range from 216 in Scene02 to 3,678 in Scene20, reflecting both scene size and label behavior. Scene18 has 1,021 disagreements among 1,168 retained pairs, nearly all of which carry the source-moving label. These counts make the scene-cluster limitation concrete. A large number of adjacent records from that scene does not create a large number of independent environments. The broad interval is therefore part of the result, rather than an inconvenience to be removed by using an unjustified independent-pair covariance.

The threshold interpretation in Fawcett, 2006 is relevant only after the intended event is specified. If a workflow truly seeks image displacement above the declared gate, a source-stationary apparent positive is a positive observation rather than an error. If it seeks world movement, the same record illustrates a proxy mismatch. Gneiting and Raftery, 2007 similarly motivates stating what a prediction or decision is rewarded for. The task definition determines whether disagreement has an operational cost; the confusion table alone cannot settle that normative choice.

6. Tracking and benchmark implications

Tracking metrics such as Luiten et al., 2020 and the evaluation setting in Dendorfer et al., 2021 concern richer outputs than a binary motion label. Bewley et al., 2016, Wojke et al., 2017, Zhang et al., 2021, and Weng et al., 2020 address association or tracking with different observations and representations. We execute none of them, and the current geometric disagreement cannot be interpreted as their accuracy. A detector might track a stationary object's image trajectory perfectly while a world-motion label remains stationary. Conflating those outputs would penalize a valid observation for failing to answer a different question.

The real-world benchmark in Geiger et al., 2012 also illustrates why source and task definitions should travel with annotations. The present clone geometry uses a synthetic setting and one source reference point. It does not establish how the same gate would agree with labels produced under a different annotation pipeline, sensor configuration, or traffic composition. Nor does it assess a human judge's uncertainty about motion. More complicated measurements may introduce genuine label or estimation error, but those mechanisms are not required to explain the construct disagreement demonstrated here.

The operational implication is to record whether moving means displacement in the world, displacement relative to the camera, image-plane change, or a task-specific hazard. A threshold should additionally state its units and time interval. Those definitions can then be evaluated against the intended consequence rather than against a convenient surrogate. A gate that is unsuitable for identifying independent world motion might still be useful for image stabilization, rendering analysis, or allocating visual attention. The current data measure none of those downstream benefits, so they remain applications to evaluate rather than conclusions.

Documentation follows Gebru et al., 2021 and Wilkinson et al., 2016 by retaining scene, frame and track keys, source labels, normalized displacements, gate flags, and all confusion counts. The resulting record distinguishes this construct-agreement question from an analysis of continuous apparent-motion associations or decomposition order. Human review must still judge whether related studies using the same source offer independent scientific value. The present claim remains modest and checkable: the two explicitly defined labels disagree often on the same stored geometric points, without requiring a faulty label or a failed motion estimator.

Raw agreement also depends on the prevalence of the two labels. A source scene containing almost entirely moving objects has a different opportunity for disagreement than one containing mostly stationary objects, even before the apparent-displacement distribution is considered. That is why the saved scene counts include both category totals and apparent-positive totals. We do not adjust the sample to equalize them after viewing results or interpret a single pooled proportion as an intrinsic reliability property of the source label.

Threshold choice introduces another dependence. Lowering the apparent gate would classify more pairs as image-moving and could reduce one disagreement direction while increasing the other. There is no task-independent optimum because the two definitions do not share an agreed target event. We report the declared gate rather than an unexecuted sensitivity curve or a tuned maximum-agreement value. A follow-up application should specify its consequence and costs before deciding whether the apparent gate is a useful proxy for the world-motion question it actually needs to answer.

7. Conclusion

The declared apparent-motion gate disagrees with the source moving label for about fifty-nine percent of retained point pairs, in both directions. That finding concerns a substitution between world and image motion definitions. It does not measure detector error or demonstrate corrupted annotations. Motion evaluations should identify the frame, threshold, and intended event before reporting agreement, because a coherent observation in one construct can be a disagreement only when silently judged against another.

References

1. Shrout, Patrick E.; Fleiss, Joseph L.. (1979). Intraclass correlations: Uses in assessing rater reliability.. Psychological Bulletin. https://doi.org/10.1037/0033-2909.86.2.420.
2. 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.
3. Natarajan, Nagarajan; Dhillon, Inderjit S.; Ravikumar, Pradeep; Tewari, Ambuj. (2013). Learning with Noisy Labels. Advances in Neural Information Processing Systems 26. https://papers.nips.cc/paper_files/paper/2013/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html.
4. 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.
5. Triggs, Bill; McLauchlan, Philip F.; Hartley, Richard I.; Fitzgibbon, Andrew W.. (2000). Bundle Adjustment - A Modern Synthesis. Lecture Notes in Computer Science. https://doi.org/10.1007/3-540-44480-7_21.
6. 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.
7. 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.
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. Lepetit, Vincent; Moreno-Noguer, Francesc; Fua, Pascal. (2009). EPnP: An Accurate O(n) Solution to the PnP Problem. International Journal of Computer Vision. https://doi.org/10.1007/s11263-008-0152-6.
10. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
11. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
12. Lourakis, Manolis I. A.; Argyros, Antonis A.. (2009). SBA: A software package for generic sparse bundle adjustment. ACM Transactions on Mathematical Software. https://doi.org/10.1145/1486525.1486527.
13. 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.
14. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
15. 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.
16. 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.
17. Meng, Zijie. (2026). OrthoMotion:Disentangling Camera and Subject Motion via Geometry Semantics Orthogonal Attention. arXiv:2606.22835. https://arxiv.org/abs/2606.22835.

18. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

19. Bewley, Alex; Ge, Zongyuan; Ott, Lionel; Ramos, Fabio; Upcroft, Ben. (2016). Simple Online and Realtime Tracking. arXiv:1602.00763. https://arxiv.org/abs/1602.00763.

20. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

21. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

22. Barfoot, Timothy D.. (2017). State Estimation for Robotics. Cambridge University Press, first edition. https://doi.org/10.1017/9781316671528.

23. 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.

24. Hartley, Richard; Zisserman, Andrew. (2004). Multiple View Geometry in Computer Vision. Cambridge University Press. https://doi.org/10.1017/cbo9780511811685.

25. Fleiss, Joseph L.. (1971). Measuring nominal scale agreement among many raters.. Psychological Bulletin. https://doi.org/10.1037/h0031619.

26. 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.

27. Wojke, Nicolai; Bewley, Alex; Paulus, Dietrich. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. arXiv:1703.07402. https://arxiv.org/abs/1703.07402.

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

29. 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.

30. Dendorfer, Patrick; Ošep, Aljoša; Milan, Anton; Schindler, Konrad; Cremers, Daniel; Reid, Ian; Roth, Stefan; Leal-Taixé, Laura. (2021). MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. International Journal of Computer Vision 129, 845-881. https://doi.org/10.1007/s11263-020-01393-0.