Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

A Single Depth Proxy across Stereo Baselines: Two-Scene Photometric Evidence

Gregory Reynolds[1] · Noah Morgan[2] · Evan Anderson[3]

[1] Iceland · [2] New Zealand · [3] Canada

Original publication: October 6, 2026 | Article 594

Abstract

One global depth proxy cannot describe every pixel's parallax, but the size of its practical penalty depends on the scene and comparison. Using nine equally spaced views of two Middlebury scenes, we compared a median-disparity plane with a warp based on the supplied disparity map. Across 477 retained tile-view cells, the single-plane photometric penalty increased with baseline separation after adjustment. The expected additional interaction with local disparity spread was uncertain, with an interval spanning both signs. This supports the usefulness of depth-varying geometry in these images without establishing a simple local-spread rule. It is a two-scene reprojection audit, not a camera-transfer generator benchmark.

Exploratory geometry study using public multi-view images and reference disparities. Human scientific and authorship review is outstanding.

A Single Depth Proxy across Stereo Baselines: Two-Scene Photometric Evidence

Exploratory geometry study using public multi-view images and reference disparities. Human scientific and authorship review is outstanding.

Abstract

One global depth proxy cannot describe every pixel's parallax, but the size of its practical penalty depends on the scene and comparison. Using nine equally spaced views of two Middlebury scenes, we compared a median-disparity plane with a warp based on the supplied disparity map. Across 477 retained tile-view cells, the single-plane photometric penalty increased with baseline separation after adjustment. The expected additional interaction with local disparity spread was uncertain, with an interval spanning both signs. This supports the usefulness of depth-varying geometry in these images without establishing a simple local-spread rule. It is a two-scene reprojection audit, not a camera-transfer generator benchmark.

1. A global motion description meets local geometry

A camera translation can produce different image displacements at different depths. A global description of the motion may therefore be meaningful without specifying every pixel's correspondence. We examine what is lost when a depth-varying scene is replaced by one median-disparity plane in a real multi-view sequence. The research question concerns the resulting photometric mismatch and whether its penalty changes with camera separation and local disparity heterogeneity.

The scale-calibrated motion-transfer argument in Meng, 2026 motivates distinguishing the magnitude of a camera move from the depth scale against which it is realized. Our comparison does not implement that system, its consistency metric, or a video generator. Nor do we attribute a single-plane assumption to it. We instead study a simpler approximation to expose a boundary of global summaries: correctly describing one typical parallax magnitude does not specify the complete spatial displacement field.

Middlebury's structured-light work provides accurate reference disparities for challenging scenes (Scharstein and Szeliski, 2003). Its associated multi-view release contains equally spaced, rectified camera views. This makes the relation between view separation and disparity particularly transparent. We use the images as observed, retaining the difference between a supplied correspondence reference and a newly estimated depth map. The depth-varying condition is therefore an oracle geometric reference, not a learned model whose prediction accuracy is being assessed.

2. Where this comparison fits in vision evaluation

The stereo taxonomy in Scharstein and Szeliski, 2002 establishes the importance of evaluating correspondence against suitable ground truth. Structure-from-motion reconstruction in Schonberger and Frahm, 2016 and pose estimation in Mur-Artal and Tardos, 2016 address related geometry problems with different observations and objectives. None is run here. We do not recover an absolute translation, perform bundle adjustment, or evaluate pose drift. The released rectification and disparity are inputs to a controlled image-warp comparison.

Monocular depth estimation in Godard et al., 2018 and cross-dataset depth work in Ranftl et al., 2019 highlight the roles of image formation, occlusion, and scale. Their relevance is to interpretation rather than a performance contest. Photometric error can remain nonzero even with accurate geometry because of visibility changes, illumination, and interpolation. A lower reprojection error therefore does not by

itself establish that a reconstruction method has recovered every physically meaningful property of the scene.

Driving resources such as Geiger et al., 2012, Liao et al., 2021, Caesar et al., 2019, Sun et al., 2019, and Yu et al., 2018 cover wider environments and sensor configurations. Driving-generation work in Gao et al., 2023 and Wang et al., 2023 has additional visual and temporal objectives. The two tabletop-style stereo scenes used here do not stand in for those settings. Their advantage is a small, inspectable geometric experiment; their disadvantage is limited diversity and no evidence about moving subjects or generated multi-view consistency.

3. Two propositions and a common comparison region

The first proposition is that the single-plane penalty changes with the absolute number of view steps from the reference. Larger separation magnifies the displacement difference between a global plane and a spatially varying disparity field, but can also create more occlusion and interpolation difficulty. The second proposition asks whether that baseline association is stronger in tiles with greater reference disparity dispersion. It is a testable descriptive interaction, not an assumption imposed on the result.

The outcome is the plane warp's mean absolute RGB error minus the depth-varying warp's error. Positive values favor the depth-varying condition. Both errors are evaluated on the same pixels for each view comparison. Pixels must have a supplied positive reference disparity and remain within the target image under both warps. This common-region rule avoids allowing one policy to improve its score merely by dropping locations where the other policy is evaluated.

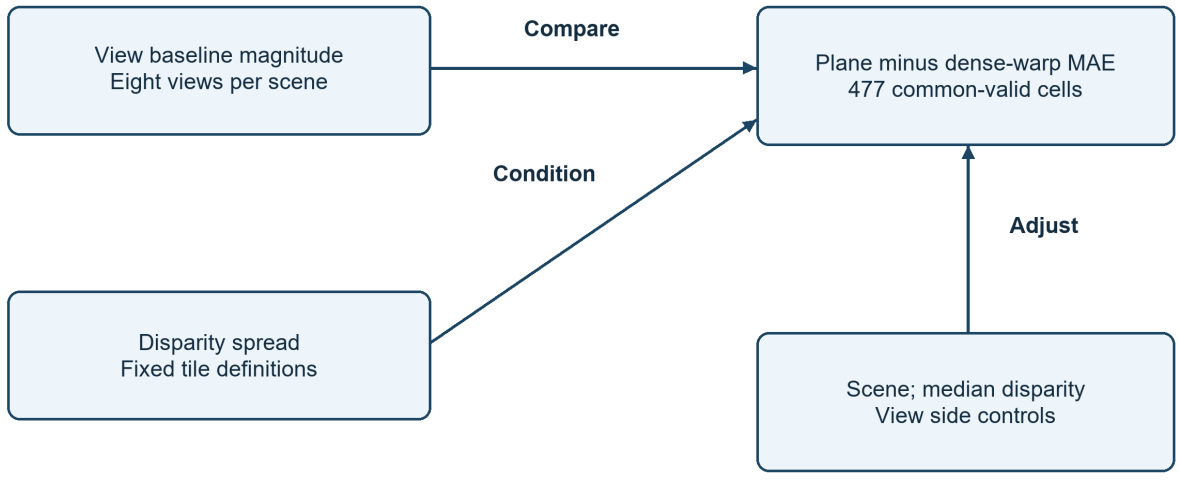

Figure 1. Baseline separation and local disparity dispersion describe the paired photometric penalty of a global plane. Scene, tile median disparity, and view direction are adjustment terms; sixty spatial tiles define covariance clusters.

4. Images, disparities, and view separation

We use the Cones and Teddy quarter-resolution multi-view archives. Each contains nine color views and reference disparity maps for views two and six. The images are 450 by 375 pixels, rectified so the intended correspondence displacement is horizontal. The source describes the views as equally spaced along the camera x-axis. Disparities are encoded at four gray levels per pixel, with zero marking unknown disparity. We divide positive encoded values by four.

View two is the fixed reference. For target view k, its relative baseline factor is k minus two divided by four, because the supplied disparity corresponds to the separation between views two and six. The depth-varying horizontal coordinate is the reference x coordinate minus that factor times its supplied disparity. The plane condition replaces each pixel disparity with the global median over valid reference pixels. The median is 32.25 pixels in Cones and 30.75 in Teddy.

Target RGB values are sampled by bilinear interpolation. We then average absolute channel differences from the reference image. These are differences in encoded eight-bit RGB values, not a linear-light radiometric measurement. No color normalization or exposure correction is applied. Structural comparison in Wang et al., 2004 and perceptual features in Zhang et al., 2018 could define other losses, but neither is computed here or assumed equivalent to the selected photometric criterion.

5. Spatial aggregation without claiming hundreds of scenes

Each image is divided into thirty nonoverlapping tiles of 75 by 75 pixels. Eight nonreference views for each of two scenes produce 480 candidate tile-view cells. A cell is retained if at least one thousand common valid pixels remain under the two warp conditions. Three cells fail this rule, leaving 477. The analysis therefore contains repeated views of sixty spatial tiles, not 477 independent images or hundreds of independently sampled camera scenes.

Within each tile, reference disparity dispersion is its standard deviation over pixels with valid reference disparity. The tile median disparity is retained separately. Both quantities are computed before considering either photometric loss and remain fixed across target views. Absolute baseline separation ranges from one to six view steps. A direction indicator distinguishes targets to the right of the reference from the two views on its other side.

The common validity mask addresses image boundaries and missing reference disparity. It does not certify that a reference point is visible in every target view. We do not infer an occlusion mask from low photometric error or remove pixels because they weaken the desired result. Missing-data principles in Little and Rubin, 2002 help distinguish the explicit geometric exclusion rule from a claim that the retained pixels are representative of every surface in either scene.

6. Estimation and uncertainty

Ordinary least squares regresses the paired error penalty on baseline steps, disparity dispersion, their interaction, tile median disparity, view direction, and a scene indicator. This is a compact association model over the observed comparisons. It does not identify a causal effect of changing scene depth independently of texture, object arrangement, or visibility. Those properties vary together within the two real scenes.

Covariance clusters by spatial tile, allowing its repeated target-view outcomes to co-vary. We use a finite-sample correction and t-based intervals across sixty clusters. Cameron and Miller, 2015 motivates that repeated-unit treatment, while White, 1980 supplies the heteroskedastic covariance foundation. Neighboring tiles can still share surfaces and illumination, and only two scenes are available. Tile clustering cannot produce an estimate of across-scene generalization uncertainty from those two sources.

The distinction parallels structured-validation arguments in Roberts et al., 2017 and selection concerns in Varma and Simon, 2006. A random partition of tile-view rows would not create a new environment. We fit and describe the full declared comparison rather than calling an internal row split an independent camera-transfer test. The model R-squared describes these observations only. A future predictive rule would need separate scenes and a declared evaluation design.

Table 1. Baseline and disparity-dispersion associations with the plane-minus-depth-varying photometric penalty. Scene, tile median disparity, and view direction remain in the fit. Standard errors cluster by tile.

Model	Term	Estimate (SE)	95% CI	N
M1	Absolute view baseline	2.5756 (0.8601)	[0.8546, 4.2966]	477
M1	Tile disparity SD	-0.0586 (0.4505)	[-0.9602, 0.8429]	477
M1	Baseline x disparity SD	-0.1809 (0.1920)	[-0.5651, 0.2033]	477
M1	Tile median disparity	-0.2585 (0.0859)	[-0.4303, -0.0867]	477
M1: fit	R-squared = 0.2949	-	-	477

Model labels: M1 = Single-plane minus depth-varying warp RGB MAE.

7. Depth variation helps, but the local interaction is unclear

In Cones, equal-cell mean absolute RGB error is 25.47 for the plane condition and 10.06 for depth-varying warping, a penalty of 15.41. In Teddy, the corresponding means are 18.70 and 10.68, a penalty of 8.02. These are averages of retained tile-view cell means. They are not pixel-count-weighted whole-image errors, and the two scene summaries are not independent replications of a learned model comparison.

The adjusted baseline coefficient is 2.5756 encoded-channel error units per view step, with standard error 0.8601 and interval 0.8546 to 4.2966. The baseline-by-dispersion interaction is -0.1809, with standard error 0.1920 and interval -0.5651 to 0.2033. The local dispersion main term is also uncertain. We therefore retain evidence for a baseline association while rejecting a simple narrative that the expected positive local-spread interaction was clearly demonstrated.

Tile median disparity has coefficient -0.2585, with interval -0.4303 to -0.0867, and the view-direction term is uncertain. The full model R-squared is 0.2949. These adjusted signs can reflect texture, object position, and overlap with the common-region mask as well as geometry. A negative median-disparity coefficient should not be advertised as a universal physical relationship between object distance and the value of depth-aware warping.

8. Measurement and mechanism remain distinct

Measurement-error considerations in Carroll et al., 2006 are relevant even with a high-quality reference map. Disparity quantization, boundary ambiguity, interpolation, and source acquisition affect what the loss means. Agreement concepts in Bland and Altman, 1986, Lin, 1989, and Koo and Li, 2016 also distinguish a lower discrepancy from full equivalence to a physical target. We estimate neither a reliability coefficient nor an absolute depth error; the measured response is a paired encoded-color reprojection difference.

Influence diagnostics following Cook, 1977 could identify spatial regions that dominate the baseline slope. We do not delete tiles according to their outcome or change the common-pixel threshold after viewing the regression. Guidance in Wasserstein and Lazar, 2016 and concerns in Ioannidis, 2005 support retaining the uncertain interaction. Its failure to exclude zero is not proof that local geometry is irrelevant; it means this model and these scenes do not clearly establish the proposed additional association.

A deeper investigation would separate texture from geometry, use explicit visibility labels, and evaluate more independent scenes. It could compare several global depth summaries without selecting a favorable one on the final evaluation set. It could also examine a genuine camera-transfer generator with both

geometric and visual outcomes. Those would be new experiments. The present oracle warp does not measure the quality of an estimated depth field or establish how a generative backbone responds to a calibrated pose.

9. A bounded geometric lesson

Local dispersion is only one aspect of departure from the global plane. A tile can have almost constant disparity while lying far from the scene-wide median, producing substantial displacement mismatch despite very little internal spread. Conversely, a tile containing several depths can include many pixels near that median. The regression includes tile median disparity as a linear adjustment, but does not fully represent every nonlinear distance from the selected plane. This helps explain why the uncertain interaction should not be read as a general rejection of the role of scene geometry.

Boundary retention also changes with view separation. At larger baselines, common in-bounds pixels may omit a different part of a tile than at smaller baselines. We preserve a fixed minimum count and evaluate both warps on the same retained pixels within each cell, which makes the paired comparison fair at that cell. It does not make the evaluated surface region identical across all target views. A study requiring that stronger comparability would need an intersection mask across views and would answer a somewhat narrower visibility question.

The chosen reference disparity has quarter-pixel encoding precision, and the target image is sampled between pixels. These are distinct numerical approximations. The first concerns the supplied geometric label; the second concerns how encoded color values are interpolated at a projected location. Neither operation recreates a physically captured intermediate viewpoint exactly. Recognizing them prevents a residual photometric error from being attributed entirely to a failed geometric assumption or entirely to an incorrect depth label without further evidence.

Finally, the two scenes differ in texture and depth layout as well as their median disparity. Their average penalties therefore should not be interpreted as an experiment that changes only scene scale. The multi-view capture controls camera separation within each scene, while the across-scene contrast remains descriptive. This is precisely why the manuscript treats the result as a small geometric audit and preserves the full tile-view record instead of presenting two favorable averages as a universal camera-motion transfer benchmark.

The materials follow Gebru et al., 2021 by preserving source scale factors, view indices, tile definitions, masks, and interpolation choices. Wilkinson et al., 2016 motivates keeping the original small archives and processed cell losses with stable local hashes. The analysis needs no personal attributes, new scene collection, or paid rendering service. The model figure represents the actual fitted variables, and the result table is generated from the saved regression object rather than an illustrative numerical example.

The empirical conclusion is specific. Spatially varying reference disparity gives substantially better photometric agreement than one median-disparity plane in these two sequences, and the plane penalty grows with baseline after adjustment. The additional local-dispersion interaction remains uncertain. That combination supports careful treatment of local geometry while keeping a clear boundary between a reproducible two-scene approximation audit and validated scale-calibrated video generation.

References

1. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
2. Liao, Yiyi; Xie, Jun; Geiger, Andreas. (2021). KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. arXiv:2109.13410. https://arxiv.org/abs/2109.13410.

3. 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.

4. Wang, Xiaofeng; Zhu, Zheng; Huang, Guan; Chen, Xinze; Zhu, Jiagang; Lu, Jiwen. (2023). DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving. arXiv:2309.09777. https://arxiv.org/abs/2309.09777.

5. Scharstein, Daniel; Szeliski, Richard. (2002). A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. International Journal of Computer Vision. https://doi.org/10.1023/a:1014573219977.

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

7. 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.

8. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

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

10. Scharstein, Daniel; Szeliski, Richard. (2003). High-accuracy stereo depth maps using structured light. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.. https://doi.org/10.1109/cvpr.2003.1211354.

11. 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.

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. Gao, Ruiyuan; Chen, Kai; Xie, Enze; Hong, Lanqing; Li, Zhenguo; Yeung, Dit-Yan; Xu, Qiang. (2023). MagicDrive: Street View Generation with Diverse 3D Geometry Control. arXiv:2310.02601. https://arxiv.org/abs/2310.02601.

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. 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.

16. 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.

17. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

18. 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.

19. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

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

21. 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.

22. Meng, Zijie. (2026). ParaScale: Scale-Calibrated Camera-Motion Transfer via a Gauge-Invariant Parallax Number. https://arxiv.org/abs/2606.19805.

23. 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.

24. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

25. Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

26. Godard, Clément; Mac Aodha, Oisin; Firman, Michael; Brostow, Gabriel. (2018). Digging Into Self-Supervised Monocular Depth Estimation. https://arxiv.org/abs/1806.01260.

27. 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.

28. 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.

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. 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.