Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

A Geometry Gate Meets an Average Goal, Not a Pixelwise Guarantee: A Two-Scene Parallax Study

Michael Bennett[1] · Cameron Walker[2] · Evan Anderson[3]

[1] Cyprus · [2] Canada · [3] United States

Original publication: November 5, 2026 | Article 664

Abstract

A threshold can satisfy its source-scene objective without becoming an error guarantee. We chose a disparity-dispersion gate on Middlebury cones and froze it before evaluating teddy. It retained 86 of 240 held-scene tile-view comparisons; their mean plane-versus-depth-aware RGB error increment was 1.47 channel levels, below the declared goal of two. Nevertheless, 25 retained comparisons exceeded that goal, and the mean's interval extended to 2.11. The experiment illustrates the distinction between an observed average selection target and reliable control of individual reconstruction errors.

Exploratory research draft using public stereo images and supplied disparity. Human scientific review and final document checks remain pending.

A Geometry Gate Meets an Average Goal, Not a Pixelwise Guarantee: A Two-Scene Parallax Study

> Exploratory research draft using public stereo images and supplied disparity. Human scientific review and final document checks remain pending.

Abstract

A threshold can satisfy its source-scene objective without becoming an error guarantee. We chose a disparity-dispersion gate on Middlebury cones and froze it before evaluating teddy. It retained 86 of 240 held-scene tile-view comparisons; their mean plane-versus-depth-aware RGB error increment was 1.47 channel levels, below the declared goal of two. Nevertheless, 25 retained comparisons exceeded that goal, and the mean's interval extended to 2.11. The experiment illustrates the distinction between an observed average selection target and reliable control of individual reconstruction errors.

1. What does passing a geometric threshold promise?

Camera-motion applications often use a geometric summary to decide when a simplified image transformation is acceptable. A small disparity spread or parallax score may suggest that a single plane will approximate a more detailed correspondence field. Turning that suggestion into a gate creates a decision problem: which comparisons should be retained, and what loss is the retained set expected to satisfy? Without a stated objective, a threshold can appear more protective than the evidence supporting it.

Our research question is whether a geometry gate selected to meet a source-scene average photometric-penalty goal retains that goal on another captured scene. We use the cones scene for cutoff selection and teddy for held evaluation. The source work on structured-light stereo Scharstein and Szeliski, 2003 and the taxonomy of stereo correspondence evaluation Scharstein and Szeliski, 2002 establish the provenance of the supplied images and disparity. The experiment uses those released measurements, rather than newly generated footage or an estimated depth map whose accuracy would introduce another evaluation problem.

Meng, 2026 proposes scale-calibrated camera-motion transfer through a gauge-invariant parallax number. Its focus on geometry and motion scale motivates examining how geometric summaries enter a threshold rule. We do not implement that number, its transfer procedure, or its reported experiments. Our risk score is a separately declared product of view separation and supplied disparity dispersion. It is intentionally simple enough to expose the difference between a geometric cutoff, a source average loss target, and the distribution of losses after selection.

The analytical proposition is that source-selected average control may transport approximately while still failing at individual comparisons. This is not a contradiction. An average objective allows some units to exceed the target, and a change of scene can alter texture, occlusion, and the relationship between geometry and encoded RGB error. The experiment therefore reports coverage, retained-set mean loss, uncertainty, and exceedance counts. It does not treat a favorable average as a universal safety statement.

2. Geometry, measurement, and the simplified warp

Multiple-view geometry Hartley and Zisserman, 2004 provides the conceptual basis for relating correspondence to camera separation. The released nine-view scenes are rectified and equally spaced, allowing the view index to define a relative separation. Reference view two supplies the image and disparity. For target view k, the disparity displacement uses the fraction obtained by dividing k minus two by four. The supplied disparity file is divided by four to recover its documented pixel units. No absolute translation in meters is reconstructed.

The detailed reference warp varies horizontal displacement pixel by pixel according to the supplied disparity. The simplified warp replaces that disparity with the median among valid reference pixels in a seventy-five-by-seventy-five tile. This is a local plane, not a global scene plane. Both use bilinear interpolation of the same target image and are scored against the same reference pixels. The simplified policy is therefore more spatially adaptive than a whole-scene median, while remaining much less detailed than the supplied correspondence field.

This computation differs from estimating a complete camera system. COLMAP Schonberger and Frahm, 2016, bundle adjustment Triggs et al., 2000, sparse bundle adjustment software Lourakis and Argyros, 2009, and EPnP Lepetit et al., 2009 address reconstruction or pose-estimation components that are not fitted here. We do not optimize camera parameters, recover a point cloud, or estimate uncertainty in a calibrated pose. Naming those boundaries matters because a gate based on supplied disparity has privileged geometric information that an operational image-only system may lack.

Likewise, learned depth methods such as MiDaS Ranftl et al., 2019 and Monodepth2 Godard et al., 2018 would supply predicted geometry with its own errors. ORB-SLAM2 Mur-Artal and Tardos, 2016 would introduce a tracking and mapping process. None is run in this experiment. The observed gate performance is conditional on the released reference disparity, with no evidence that a noisy estimated substitute would preserve the same retained cells or loss relationship. It should be treated as a diagnostic under known geometry before a more demanding deployment study.

3. Reference-only risk and source-only selection

Each scene is divided into five rows and six columns of tiles, producing thirty reference tiles. Every nonreference view gives one candidate comparison per tile, for 240 candidates per scene. A comparison remains eligible only when at least one thousand pixels have positive reference disparity and both detailed and plane-based target coordinates are in bounds. Cones retains 237 candidate cells after this rule; teddy retains all 240. The same common pixels score both warps within an eligible comparison.

The geometric risk score is the absolute view-separation fraction multiplied by the ninetieth percentile of absolute disparity deviations from the tile median. The percentile uses all valid reference disparity pixels in that tile, not target RGB residuals. The resulting risk has units of horizontal pixels. It summarizes dispersion in the displacement approximation, rather than estimating a probability of a bad rendering. Its ninetieth-percentile definition also leaves a tail outside that summary and does not certify occlusion-free correspondence.

Seven candidate cutoffs are fixed at 0.5, 1, 2, 4, 8, 16, and 100 risk pixels. For each cutoff on cones, the source sweep records coverage and the mean signed difference between plane-warp RGB MAE and detailed-warp RGB MAE. A cutoff is eligible if it retains at least five comparisons and that mean difference is no greater than two encoded channel levels. Among eligible cutoffs, the rule selects the greatest coverage, breaking equal-coverage ties toward the larger cutoff.

The two numerical values of two belong to different units. The selected risk cutoff happens to be two pixels, while the objective permits a mean increment of two RGB channel levels. Their equality as numbers is not a conversion law or an error bound. Pixel displacement interacts with image content, boundaries, occlusion, and interpolation. Measurement-error reasoning Carroll et al., 2006 reinforces why a geometric proxy should not be silently equated with the downstream photometric outcome it helps screen.

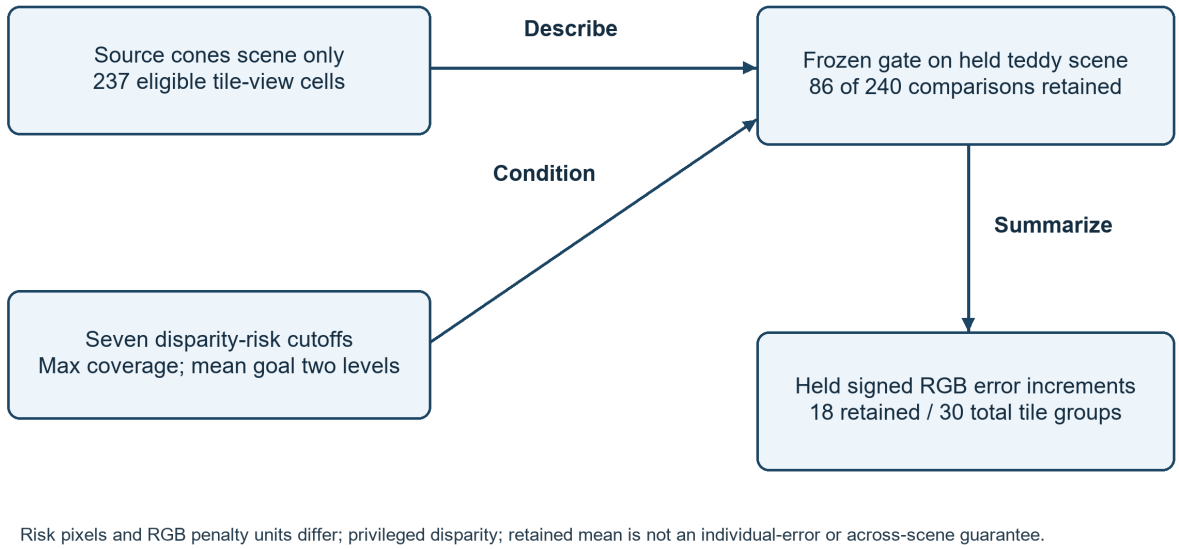


Figure 1. Source-only gate selection and a frozen held-scene loss evaluation.

Figure 1 separates the source objective from the held-scene result. Teddy RGB losses never enter cutoff selection. Once the source rule chooses the cutoff, it is applied to held geometric risk without adjustment. The selection framework resembles the coverage-error question in selective prediction Geifman and El-Yaniv, 2017, but the present outcome is signed image-warp error rather than a classifier's mistake indicator. Calibration research Guo et al., 2017 concerns another object again: the score here is neither a forecast probability nor a calibrated confidence value.

4. Loss, inference, and the relevant denominators

For an eligible tile-view comparison, RGB MAE averages absolute channel differences over the common reference pixels and three encoded eight-bit channels. We subtract the detailed-warp MAE from the plane-warp MAE. The resulting increment can be negative; the supplied geometric warp is not guaranteed to have smaller photometric error when occlusions, interpolation, or appearance effects remain. Calling the outcome a penalty is a descriptive shorthand for this signed contrast, not an imposed positivity constraint.

Two intercept-only OLS models summarize teddy. M1 uses the eighty-six comparisons accepted by the frozen gate; M2 uses all 240 eligible comparisons. The respective covariance groups are the eighteen contributing retained tiles and all thirty held tiles. Finite-sample cluster corrections and t references with seventeen and twenty-nine degrees of freedom accompany the means. Robust covariance principles White, 1980 and cluster-inference guidance Cameron and Miller, 2015 motivate reporting these group counts, which differ from the number of repeated tile-view rows.

The models do not estimate a causal treatment effect of accepting a comparison. Accepted and all-comparison samples overlap, and geometry determines selection. Their two intervals are not a direct test of a between-group difference. Structured validation Roberts et al., 2017 helps define the stronger

separation actually used: one scene selects the threshold and another supplies the reported test outcomes. The model-selection concerns in Varma and Simon, 2006 are addressed by freezing the source-selected cutoff, without claiming that two scenes represent a broad sampled population.

The source search itself is conditional on one scene and a coarse grid. Selecting the largest feasible observed coverage does not account for uncertainty in the source mean, and a different grid could select another operating point. We do not revise the two-level objective or insert an intermediate cutoff after viewing teddy. A study designed to certify a mean bound would need an explicit procedure for uncertainty after selection, a representative scene population, and an adequate number of independent scenes. The present held evaluation is useful precisely because it exposes what the simple source rule actually does without adding those unperformed guarantees.

Repeated views also explain why row counts and cluster counts cannot be exchanged. A retained tile may contribute several target views, each sharing the same reference image content and disparity field. Treating all eighty-six retained comparisons as unrelated images would overstate the available independent structure. Grouping by reference tile acknowledges one source of dependence, but neighboring tiles still belong to the same physical scene and can share objects or appearance. The resulting conditional intervals summarize this recorded image arrangement; they cannot serve as uncertainty over a population of new camera environments.

We also count accepted comparisons whose signed increment exceeds the two-level mean goal. This is a descriptive diagnostic, not a newly optimized threshold. ROC methodology Fawcett, 2006 could support a separate threshold-classification study with declared bad-outcome labels, but no ROC curve is reported here. Cost-sensitive decision theory Elkan, 2001 would require explicit consequences of retaining and rejecting comparisons. Those costs are absent, so maximum source coverage subject to the stated mean goal is the complete decision objective in this experiment.

5. What transferred, and what did not become guaranteed

On cones, cutoffs 0.5, 1, and 2 meet the declared mean constraint, retaining 38, 58, and 92 comparisons. Their mean increments are 0.2656165, 0.6643161, and 1.2833082 channel levels. The next cutoff, four pixels, retains 145 comparisons but has mean increment 2.1741013 and is ineligible. Larger candidates also fail the mean condition. The selected two-pixel cutoff retains 92 of 237 source comparisons, or about 38.8%, before any teddy losses are examined.

The frozen gate retains eighty-six of 240 teddy comparisons, giving coverage 0.3583333. Their mean signed increment is 1.4707520 channel levels. All held comparisons have mean increment 3.5176503. The accepted mean is below the two-level source objective in the realized held scene. This is the finite-sample transport result. It is useful evidence about the declared gate, but it remains narrower than a statement that future scenes or every accepted comparison will meet the objective.

Table 1. Held-scene signed RGB error increments under the frozen source gate

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	1.4708 (0.3036)	[0.8302, 2.1113]	86
M1: fit	R-squared = 0.0000	-	-	86
M2	Mean outcome/contrast	3.5177 (0.7160)	[2.0533, 4.9820]	240
M2: fit	R-squared = 0.0000	-	-	240

Model labels: M1 = Accepted held-scene plane-minus-dense RGB MAE; M2 = All held-scene plane-minus-dense RGB MAE.

Table 1 reports mean-only OLS models without controls or fixed effects. M1 has N86 and eighteen tile clusters, using t17; M2 has N240 and thirty tile clusters, using t29. Standard errors use corrected cluster covariance. The outcome is plane minus detailed warp RGB MAE in encoded channel levels. Intercept-only R-squared is zero. These models summarize overlapping held samples and do not test a causal effect of gating or supply independent across-scene uncertainty.

The accepted mean has standard error 0.3036231 and interval from 0.8301633 to 2.1113408. The all-comparison mean has standard error 0.7159826 and interval from 2.0533014 to 4.9819992. In particular, the accepted interval extends above the two-level objective. We therefore report an observed held average below the goal, rather than claiming statistically established mean control. Twenty-five accepted comparisons exceed two levels individually, which is entirely compatible with the source rule having constrained only an average.

6. Interpreting a geometry-assisted decision

The gate works locally in a sensible direction: it retains a subset with a smaller observed average increment than the complete held set. However, that comparison does not identify whether the gain comes from view separation, disparity coherence, texture, or their joint distribution. The risk product combines the first two factors, and the retained images differ in additional ways. An influence analysis in the tradition of Cook, 1977 could examine individual tiles, but no unreported diagnostic is used to remove difficult accepted comparisons from the results.

The choice of loss is equally important. Encoded RGB MAE weights all channel deviations linearly and does not measure physical radiance, perceived motion quality, or downstream driving performance. The scoring perspective of Gneiting and Raftery, 2007 encourages aligning the reported criterion with the actual decision object. Here the source objective and held result use precisely the same signed contrast. A claim about another quality criterion would need its own measurements. The contextual guidance of Wasserstein and Lazar, 2016 also discourages replacing this evidence with a significance label that ignores the chosen goal and its units.

Moving toward practical motion systems would require broader data. KITTI Geiger et al., 2012 and KITTI-360 Liao et al., 2021 provide driving-related evidence settings, while Waymo Sun et al., 2019, nuScenes Caesar et al., 2019, and BDD100K Yu et al., 2018 organize other large-scale driving observations. These sources are methodological context, not extra test scenes in the present table. None justifies transporting a static two-scene cutoff into dynamic traffic, independently moving subjects, or a generated multi-view video without further evaluation and appropriate source-specific measurements.

7. A reviewable gate rather than a universal bound

The saved record contains the source threshold sweep, eligibility counts, the selected cutoff, held acceptance indicators, common-pixel counts, tile identifiers, and structured estimates. Data attribution follows the original Middlebury release, with source-specific reuse conditions retained. Dataset documentation Gebru et al., 2021 makes the privileged disparity and source-selection rule visible, while FAIR stewardship Wilkinson et al., 2016 supports preserving identifiers and transformations. The numerical gate can therefore be inspected as an actual decision rule rather than an unspecified favorable threshold.

The experiment answers its question conditionally. A gate chosen on cones retained a teddy subset whose observed mean error increment remained below the source goal. It did not guarantee individual error

control, and its interval did not establish a universally bounded mean. Coverage, the realized average, the interval, and the exceedance count each describe a different aspect of that result. Keeping those quantities together is the main lesson for reporting geometry-based acceptance rules from small public-data studies.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. Hartley, Richard; Zisserman, Andrew. (2004). Multiple View Geometry in Computer Vision. Cambridge University Press. https://doi.org/10.1017/cbo9780511811685.
6. 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.
7. 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.
8. 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.
9. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.
10. 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.
11. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
12. 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.
13. Meng, Zijie. (2026). ParaScale: Scale-Calibrated Camera-Motion Transfer via a Gauge-Invariant Parallax Number. https://arxiv.org/abs/2606.19805.
14. 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.
15. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
16. 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.
17. Godard, Clément; Mac Aodha, Oisin; Firman, Michael; Brostow, Gabriel. (2018). Digging Into Self-Supervised Monocular Depth Estimation. https://arxiv.org/abs/1806.01260.
18. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
19. 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.
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. 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.

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. 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.

24. 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.

25. 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.

26. 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.

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

28. 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.

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

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