Task-Relevant Depth Quality Metrics for Suction Grasping

Shivansh Inamdar
University of Pennsylvania
shivansh@sas.upenn.edu

Abstract

Depth foundation models are increasingly used as a perception backbone for embodied manipulation, and are typically evaluated with metrics that measure global accuracy (RMSE, MAE, AbsRel) but fail to capture the local geometric properties that determine suction grasp success. These properties include surface planarity within the contact patch, surface normal accuracy at grasp points, and contact patch completeness near object boundaries. We propose four task-relevant depth quality metrics grounded in suction contact mechanics and evaluate two depth foundation models (Depth Anything V2, Marigold) against raw structured-light sensor depth on 1,200 images from the GraspNet-1Billion dataset, using synthetic ground-truth depth rendered from object meshes. Our results reveal a consistent rank reversal: the raw depth sensor achieves two to three times better RMSE than the foundation models, yet scores worse than at least one foundation model on every task-relevant metric. The foundation models produce geometrically coherent surfaces (smooth, complete, with consistent normals) despite worse metric accuracy, and suction grasping rewards coherence over accuracy. This suggests that standard metrics can mislead practitioners selecting depth backbones for embodied manipulation, and that hybrid pipelines using sensor depth for positioning and foundation-model depth for grasp evaluation and final approach may be beneficial. An earlier version was accepted to the ICRA 2026 ASAB and Rigorous Perception workshops.

1. Introduction

Depth foundation models such as Depth Anything V2 [9] and Marigold [4] are increasingly used as perception backbones for embodied manipulation. These models, like other depth estimation methods, are typically evaluated using pixel-wise metrics such as RMSE, MAE, and AbsRel. Recent work has shown these metrics are insufficient, being insensitive to local curvature perturbations, alignment-biased, and task-dependent, leading to proposed task-based alternatives [6, 8]. However, these alternatives focus on computer vision tasks (SLAM, stereo, 3D reconstruction). None address robotic manipulation, where depth quality at suction cup contact patches directly determines grasp success. This gap is most pronounced exactly where depth perception needs to be robust: reflective surfaces, object boundaries, and regions with missing depth, where structured-light sensors degrade.

For suction grasping, grasp success depends on *local* geometric properties: surface planarity within the contact patch (for seal formation [7]), surface normal accuracy (for approach direction and wrench resistance [1]), and depth completeness (for full contact). Global metrics like RMSE do not capture these properties. Prior work evaluates depth for grasping indirectly, by measuring downstream grasp success rates [1, 7]. This conflates depth quality with grasp planner quality, since a poor planner on good depth is indistinguishable from a good planner on poor depth.

We propose four metrics that evaluate depth quality directly at grasp-relevant locations, independent of any specific planner: Contact Patch Planarity Error (CPPE), Surface Normal Angular Error (SNAE), Grasp Wrench Resistance Error (GWRE), and Contact Patch Completeness (CPC). Evaluating on GraspNet-1Billion [2] with three depth methods, we find a consistent rank reversal: methods with worse RMSE produce better geometry for suction grasping, demonstrating that task-specific depth evaluation is necessary for manipulation, as selecting depth methods based on standard metrics alone could lead to worse grasp performance.

2. Related Work

Depth evaluation metrics. Wu et al. [8] systematically tested the sensitivity of existing depth metrics to controlled perturbations and found that all widely used metrics have near-zero sensitivity to curvature perturbations, precisely the kind of error that affects suction seal formation. They proposed RelNormal, a metric based on relative surface normals, to address this gap. BenchDepth [6] took a different approach, arguing for evaluating depth via downstream task performance rather than pixel-wise accuracy. They showed that method rankings shift across tasks (SLAM, stereo, 3D reconstruction), with no single depth model performing best on all tasks. However, neither work addresses robotic ma-

nipulation, where depth quality requirements are determined by contact physics rather than visual fidelity.

Suction contact mechanics. Dex-Net 3.0 [7] formalized suction grasp success as two conditions: seal formation (modeled via a quasi-static spring system on the contact ring) and wrench resistance (the ability to resist gravity after pickup). SuctionNet-1Billion [1] simplified this into continuous scores (S_{seal}, S_{wrench}) and introduced a Normal STD baseline that predicts grasp quality from surface normal variance alone. Sim-Suction [5] showed that evaluating seal formation across the full contact area (960 vertices) rather than only at the rim (as in Dex-Net) significantly improves prediction accuracy. Jiang et al. [3] demonstrated that simple geometric features, such as normal variance and plane-fit residuals, predict grasp quality competitively with full physics-based models. Together, these works establish that local geometric properties within the contact patch determine suction grasp success, motivating our use of these properties as depth quality metrics.

Depth estimation methods. We evaluate representative methods from two paradigms: Depth Anything V2 [9], a monocular depth model trained on large-scale data, and Marigold [4], which repurposes a diffusion-based image generator for depth estimation. Both predict depth from RGB only and represent recent state-of-the-art approaches in monocular depth estimation. We compare these against raw structured-light sensor depth (Intel RealSense D435) to understand how learned geometric priors differ from direct measurement in terms of task-relevant quality.

3. Task-Relevant Metrics

Given a ground-truth depth map D_{gt} (rendered from object meshes) and an estimated depth map D_{est}, we evaluate at candidate grasp points randomly sampled on object surfaces. For each point $\mathbf{p}$ with suction cup radius r, the contact patch $\mathcal{P}(\mathbf{p}, r)$ is the set of pixels within the projected circle.

CPPE (Contact Patch Planarity Error) measures the mean squared distance of back-projected 3D points within the contact patch to their best-fit plane, following SuctionNet-1Billion's S_{fit} formulation [1], which, along with similar geometric features [3], has been shown to predict grasp quality:

$$\text{CPPE}(\mathbf{p}, r) = \frac{1}{|\mathcal{P}|} \sum_{q_i \in \mathcal{P}} d(q_i, \pi)^2 \tag{1}$$

where π is the least-squares plane and $d(q_i, \pi)$ is the signed point-to-plane distance. Lower CPPE indicates a flatter contact surface, better for seal formation.

SNAE (Surface Normal Angular Error) measures the angular error between estimated and ground-truth surface normals at grasp-relevant pixels:

$$\text{SNAE}(\mathbf{p}, r) = \frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} \arccos(|\hat{n}_{\text{est}}(q) \cdot \hat{n}_{\text{gt}}(q)|) \tag{2}$$

Unlike SuctionNet's Normal STD baseline [1], which measures normal consistency within a single depth map, and Wu et al.'s RelNormal [8], which measures relative normal angles between patches, SNAE measures *absolute accuracy* of each normal against ground truth. A depth method can produce smooth, consistent normals that are systematically wrong. Lower SNAE indicates more accurate normals, critical for correct grasp approach direction.

CPC (Contact Patch Completeness) measures the fraction of pixels in the contact patch with valid depth that belongs to the same object as the grasp center:

$$\text{CPC}(\mathbf{p}, r) = \frac{|\{q \in \mathcal{P} : D(q) > 0 \wedge L(q) = L(\mathbf{p})\}|}{|\mathcal{P}|} \tag{3}$$

where L is the segmentation label. This captures the "no holes in the contact ring" requirement from Dex-Net 3.0 [7]. Higher CPC indicates a more complete contact patch, better for seal formation. Note that CPC is the only metric where higher is better.

GWRE (Grasp Wrench Resistance Error) captures the task-relevant consequence of normal errors. While SNAE treats all normal errors equally, GWRE weights them by their impact on wrench resistance, which depends on surface tilt relative to gravity. It measures the change in wrench resistance score [1] caused by normal estimation error:

$$\text{GWRE}(\mathbf{p}) = |S_w(\hat{n}_{\text{est}}) - S_w(\hat{n}_{\text{gt}})| \tag{4}$$

where $S_w = 1 - \min(1, |\tau_e|/\tau_{\text{thre}})$ is SuctionNet's elastic torque-based wrench score, τ_e is the gravity-induced elastic torque, and τ_{thre} is the material-dependent torque threshold. Lower GWRE indicates that depth estimation errors have less impact on wrench resistance prediction.

4. Experiments

Dataset. We use GraspNet-1Billion [2]: 60 scenes (30 with previously seen objects + 30 with novel objects) with 88 objects, Intel RealSense D435 RGB-D images, accurate 6D poses, and 3D mesh models. We evaluate 20 views per scene (1,200 images total), randomly sampling 50 grasp points per image on object surfaces with a 15mm cup diameter. Ground-truth depth and normals are rendered from object meshes using known 6D poses.

Methods. We evaluate three architecturally distinct depth sources: (1) **RealSense (raw)**: unprocessed sensor depth. (2) **Depth Anything V2** [9]: a monocular depth model trained on large-scale data (Base, metric indoor) that predicts depth from RGB only. (3) **Marigold** [4]: a diffusion-based depth model (4 denoising steps) that also predicts from RGB only. Both learned methods are affine-aligned to sensor depth via least-squares.

Results. Fig. 1 provides a qualitative example of the differences our metrics capture. Tables 1 and 2 show results on

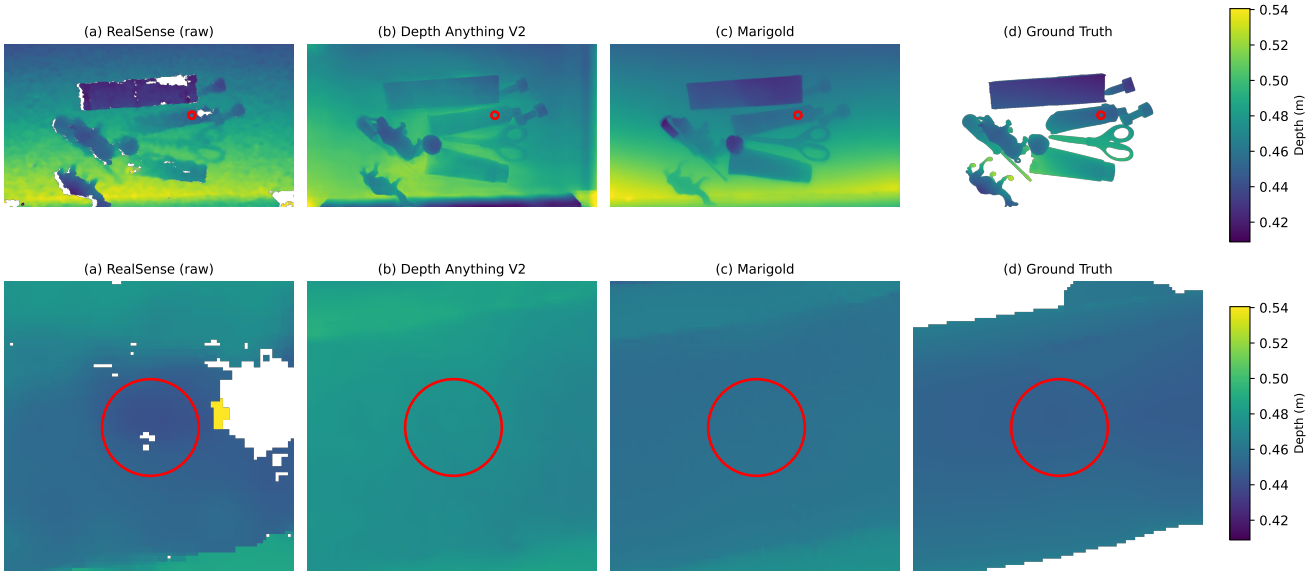

Figure 1. Depth comparison across methods. Top: full scene view. Bottom: zoomed into the contact patch (red circle). The RealSense sensor has missing depth (holes) within the patch, while both learned methods produce complete, smooth surfaces. However, the learned methods blur object boundaries compared to the ground truth. Ground truth is rendered from object meshes provided by GraspNet-1Billion (table surface is not included in the dataset's mesh models).

test-seen and test-novel splits respectively. The core finding is consistent across both: RealSense achieves two to three times better RMSE but scores worse than at least one learned method on every task-relevant metric. Marigold produces the flattest contact patches (lowest CPPE), while Depth Anything V2 produces the most accurate normals (lowest SNAE). Both learned methods achieve higher contact patch completeness (CPC). GWRE values are small across all methods, likely because objects on the tabletop present mostly horizontal graspable surfaces where gravity torque variation is minimal.

Table 1. Test-seen results (600 images, 30 scenes with objects seen during training). Bold = best per column.

Method	RMSE↓	AbsRel↓	CPPE↓	SNAE↓	GWRE↓	CPC↑
RealSense	**0.013**	**0.016**	1.97e-6	37.7°	1.07e-4	0.878
DA V2 [9]	0.038	0.081	1.69e-6	**33.0°**	**6.3e-5**	0.914
Marigold [4]	0.036	0.073	**1.37e-6**	37.2°	7.1e-5	**0.914**

Table 2. Test-novel results (600 images, 30 scenes with previously unseen objects). Novel objects amplify depth quality differences: CPPE values are higher across all methods, and the gap between methods widens. Bold = best per column.

Method	RMSE↓	AbsRel↓	CPPE↓	SNAE↓	GWRE↓	CPC↑
RealSense	**0.015**	**0.021**	2.64e-6	39.7°	1.07e-4	0.859
DA V2 [9]	0.036	0.073	2.88e-6	**39.1°**	**7.6e-5**	0.906
Marigold [4]	0.035	0.071	**1.63e-6**	42.5°	8.4e-5	**0.906**

Results are consistent across test-seen and test-novel splits. Novel objects show higher CPPE across all methods (e.g., 34% higher for the raw sensor), and the gap between sensor and Marigold CPPE widens from 1.4 times on seen objects to 1.6 times on novel objects, suggesting that complex geometry amplifies depth quality differences. Notably, Depth Anything V2's CPPE degrades on novel objects (2.88e-6, worse than the sensor's 2.64e-6), while Marigold's remains strong (1.63e-6). However, Marigold's SNAE shows the reverse pattern, degrading from 37.2° to 42.5° on novel objects, suggesting that each method's generalization weakness manifests on different geometric properties. Rankings are also robust at a larger 25mm cup diameter: Marigold retains the lowest pooled CPPE (5.54e-6 vs. RealSense's 8.83e-6), and both learned methods retain higher CPC (0.85 vs. 0.80).

Cross-metric correlations. To verify that our four metrics capture distinct failure modes rather than redundant information, we compute pairwise Pearson correlations pooled across all three methods and 1,200 images ($n = 3{,}600$). CPPE and CPC are moderately negatively correlated ($r = -0.358$): patches with worse planarity tend to have better completeness (interior points have full coverage but sensor noise) while patches near edges have missing depth but flatter remaining surfaces. This confirms they capture complementary failure modes. SNAE and GWRE are moderately positively correlated ($r = 0.356$), as expected since GWRE depends on normal accuracy, but the correlation is far from unity, confirming that GWRE adds information by weight-

ing normal errors by their task-relevant consequence. CPPE and GWRE are essentially uncorrelated ($r \approx 0$), indicating that contact patch planarity and wrench resistance capture independent aspects of depth quality.

5. Discussion

Our results demonstrate that a depth method can have two to three times worse RMSE yet produce geometry better suited for suction grasping. The learned methods are not more *accurate* but more *geometrically coherent*: they produce smooth, complete surfaces with consistent normals, which is what seal formation requires, even though their absolute depth values are less precise. This happens because a structured-light sensor like the RealSense measures real distances but with pixel-level noise, edge artifacts, and missing depth. Models like Depth Anything V2 and Marigold predict what surfaces *should* look like based on training on large image corpora. They get absolute distances wrong but produce smooth, continuous surfaces with consistent normals, because that is what real-world surfaces look like in photographs. A suction cup does not care if depth is off by 2cm globally; it cares if the local surface is flat and complete at the contact point. Wu et al. [8] showed that standard metrics are insensitive to curvature perturbations, which is precisely the kind of depth error that degrades seal formation. Our task-relevant metrics detect these errors where standard metrics cannot, extending BenchDepth's task-dependent evaluation argument [6] into robotic manipulation. Our argument is orthogonal to improvements in depth estimation itself: a more accurate estimator by standard metrics is not necessarily a better one for suction grasping if those metrics do not correlate with seal-formation geometry.

The per-split analysis reveals an additional nuance: different methods degrade on different geometric properties when faced with novel objects. Depth Anything V2's CPPE degrades on novel objects (worse than the raw sensor), while Marigold's SNAE degrades significantly. This suggests that no single learned method universally preserves all task-relevant geometric properties, and that the choice of depth method may need to be informed by which geometric property is most critical for a given grasp scenario. Multiple task-relevant metrics are necessary to capture these distinct failure modes.

The method that minimizes global pixel error is not the method that maximizes seal-formation geometry; hybrid pipelines are a practical response to this mismatch. These findings have direct implications for depth perception in manipulation. First, these metrics can inform sensing strategy: e.g., if CPPE is high at a candidate grasp point, a different viewpoint may reduce planarity error. Second, they suggest that manipulation systems could benefit from *hybrid depth pipelines* that use different depth representations at different stages of the grasp. Sensor depth provides accurate absolute distances needed for collision-free approach planning, while model-predicted depth provides the geometric coherence needed for evaluating seal formation at the contact patch. Indeed, our evaluation already relies on this complementarity: the learned models are affine-aligned to sensor depth, inheriting the sensor's metric accuracy while providing superior local geometry. A more sophisticated pipeline could dynamically select or blend depth representations based on the task-relevant metric values at candidate grasp points.

6. Limitations and Future Work

Limitations. Ground truth depth is rendered from object meshes rather than captured by a reference sensor, which means our GT represents ideal geometry but may not capture all real-world surface properties (e.g., texture, material reflectance). The evaluation is limited to rigid objects on a tabletop; deformable objects, transparent/reflective materials, and bin-picking scenarios would present different challenges for both sensors and learned methods. Grasp points are randomly sampled on object surfaces; stratified sampling by surface difficulty (object interiors, near edges, tilted surfaces) would better reveal where each metric adds the most value and could show stronger differentiation for GWRE on tilted surfaces. CPC currently measures only depth completeness (whether valid depth exists) but not boundary accuracy (whether the depth values near edges are geometrically correct). However, depth artifacts at object boundaries typically also increase CPPE, which can help indirectly catch boundary inaccuracy. Finally, GWRE shows limited differentiation on this dataset, likely because tabletop objects present mostly horizontal graspable surfaces where gravity torque variation is minimal.

Future work. Several directions could strengthen and extend this work. First, validating our metrics against SuctionNet's per-point grasp quality annotations (S_{seal}, S_{wrench}) would confirm that depth regions with high CPPE or low CPC correspond to reduced graspability. Second, a controlled sensitivity study that synthetically degrades GT depth and measures how grasp planner performance degrades as a function of our metrics vs. RMSE would directly demonstrate the predictive value of task-relevant metrics. Third, evaluating on a larger variety of objects and scenes, including bin-picking with non-horizontal surfaces, would test whether GWRE becomes a stronger differentiator in those settings. Finally, integrating these metrics into a perception pipeline that dynamically selects viewpoints or depth representations to minimize task-relevant error at candidate grasp points is a natural extension of the hybrid pipeline concept discussed above.

References

[1] Hanwen Cao, Hao-Shu Fang, Wenhai Liu, and Cewu Lu. SuctionNet-1billion: A large-scale benchmark for suction grasping. *IEEE Robotics and Automation Letters*, 6(4), 2021. 1, 2

[2] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. GraspNet-1billion: A large-scale benchmark for general object grasping. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[3] Ping Jiang, Junji Oaki, Yoshiyuki Ishihara, Junichiro Ooga, Haifeng Han, Atsushi Sugahara, et al. Learning suction graspability considering grasp quality and robot reachability for bin-picking. *Frontiers in Neurorobotics*, 2022. 2

[4] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[5] Juncheng Li and David J. Cappelleri. Sim-Suction: Learning a suction grasp policy for cluttered environments using a synthetic benchmark. *IEEE Transactions on Robotics*, 2023. 2

[6] Zhenyu Li, Haotong Lin, Jiashi Feng, Peter Wonka, and Bingyi Kang. BenchDepth: Are we on the right way to evaluate depth foundation models? *arXiv preprint arXiv:2507.15321*, 2025. 1, 4

[7] Jeffrey Mahler, Matthew Matl, Xinyu Liu, Albert Li, David Gealy, and Ken Goldberg. Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 1, 2

[8] Shuo Wu, Jacob Nugent, Wenhao Yang, and Jia Deng. Toward a better understanding of monocular depth evaluation. *arXiv preprint arXiv:2510.19814*, 2025. 1, 2, 4

[9] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 3