Task-Relevant Depth Quality Metrics for Suction Grasping

Shivansh Inamdar

***Abstract*— Standard depth evaluation metrics (RMSE, MAE) measure global accuracy 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 three depth estimation methods on 1,200 images from the GraspNet-1Billion dataset. Our results reveal a consistent rank reversal: the raw depth sensor achieves two to three times better RMSE than learned methods, yet scores worse than at least one learned method on every task-relevant metric. Learned 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 methods for manipulation, and that hybrid pipelines using sensor depth for positioning and model depth for grasp evaluation and final approach may be beneficial.**

I. Introduction

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 [5], [4]. 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.

For suction grasping, grasp success depends on two primary conditions from the Dex-Net 3.0 contact model [1]: (1) *seal formation*: the surface within the suction cup contact patch must be sufficiently planar for an airtight seal, and (2) *wrench resistance*: the contact normal must allow the grasp to resist gravity. Related properties such as depth completeness within the contact patch also affect success. These conditions depend on *local* geometric properties (planarity, normal direction) that global metrics like RMSE do not capture. Sim-Suction [8] demonstrated this concretely, showing that higher-resolution geometric evaluation within the contact patch significantly improves grasp prediction accuracy over coarser approximations.

Prior work evaluates depth for grasping indirectly, by measuring downstream grasp success rates [1], [2]. 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 instead 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 [3] 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.

S. Inamdar is with the University of Pennsylvania, Philadelphia, PA, USA. shivanshi@sas.upenn.edu

II. 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 [2], which, along with similar geometric features [9], 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 [2], which measures normal consistency within a single depth map, and Wu et al.'s RelNormal [5], 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 [1]. 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 [2] caused by normal estimation error: $\text{GWRE}(\mathbf{p}) = |S_w(\hat{n}_{\text{est}}) - S_w(\hat{n}_{\text{gt}})|$, 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.

TABLE I

DEPTH QUALITY METRICS ACROSS 1,200 IMAGES (60 SCENES). BOLD INDICATES BEST PER COLUMN. STANDARD METRICS (RMSE, ABSREL) CONSISTENTLY RANK THE SENSOR BEST, WHILE TASK-RELEVANT METRICS OFTEN FAVOR LEARNED METHODS.

Method	RMSE↓	AbsRel↓	CPPE↓	SNAE↓	GWRE↓	CPC↑
RealSense	**0.014**	**0.018**	2.30e-6	38.7°	1.07e-4	0.8681
DA V2 [6]	0.037	0.077	2.28e-6	**36.1°**	**7.0e-5**	0.9099
Marigold [7]	0.036	0.072	**1.50e-6**	39.9°	7.8e-5	**0.9101**

III. EXPERIMENTS

Dataset. We use GraspNet-1Billion [3]: 60 scenes (30 test-seen + 30 test-novel) 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** [6]: a monocular depth foundation model (Base, metric indoor) that predicts depth from RGB only. (3) **Marigold** [7]: 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. Table I shows the main finding: 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.

Results are consistent across test-seen and test-novel splits. Novel objects show 34% higher CPPE overall, and the gap between sensor and Marigold CPPE widens from 1.4× on seen objects to 1.6× 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 consistent across suction cup diameters (15mm and 25mm).

IV. 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. Wu et al. [5] 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 [4] into robotic manipulation.

This finding has direct implications for active perception in manipulation. First, these metrics inform *what* to re-sense: e.g., if CPPE is high at a candidate grasp point, a different viewpoint may reduce planarity error. Second, they suggest that manipulation systems should *actively switch* between depth representations depending on the task stage: sensor depth for coarse positioning (where metric accuracy matters) and model-predicted depth for final grasp evaluation (where geometric coherence determines seal formation). 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.

Limitations and future work. Our evaluation uses ground truth rendered from meshes and is limited to rigid objects on a tabletop. CPC currently measures only depth completeness, not boundary accuracy, and GWRE shows limited differentiation on this dataset. Extending to a larger variety of objects and scenes, along with stratified grasp point sampling by surface difficulty, would address these limitations and reveal where each metric adds the most value. Additional future directions include validating metrics against SuctionNet's grasp quality annotations, measuring how grasp planner performance degrades as a function of our metrics vs. RMSE, and integrating these metrics into an active perception loop that selects viewpoints or depth representations to minimize task-relevant error.

REFERENCES

[1] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, & K. 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 ICRA*, 2018.

[2] H. Cao, H.-S. Fang, W. Liu, & C. Lu, "SuctionNet-1Billion: A large-scale benchmark for suction grasping," *IEEE RA-L*, vol. 6, no. 4, 2021.

[3] H.-S. Fang, C. Wang, M. Gou, & C. Lu, "GraspNet-1Billion: A large-scale benchmark for general object grasping," in *Proc. IEEE/CVF CVPR*, 2020.

[4] Z. Li, H. Lin, J. Feng, P. Wonka, & B. Kang, "BenchDepth: Are we on the right way to evaluate depth foundation models?" *arXiv preprint arXiv:2507.15321*, 2025.

[5] S. Wu, J. Nugent, W. Yang, & J. Deng, "Toward a better understanding of monocular depth evaluation," *arXiv preprint arXiv:2510.19814*, 2025.

[6] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, & H. Zhao, "Depth Anything V2," in *Proc. NeurIPS*, 2024.

[7] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, & K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proc. IEEE/CVF CVPR*, 2024.

[8] J. Li & D. J. Cappelleri, "Sim-Suction: Learning a suction grasp policy for cluttered environments using a synthetic benchmark," *IEEE T-RO*, 2023.

[9] P. Jiang, J. Oaki, Y. Ishihara, J. Ooga, H. Han, & A. Sugahara *et al.*, "Learning suction graspability considering grasp quality and robot reachability for bin-picking," *Frontiers in Neurorobotics*, 2022.