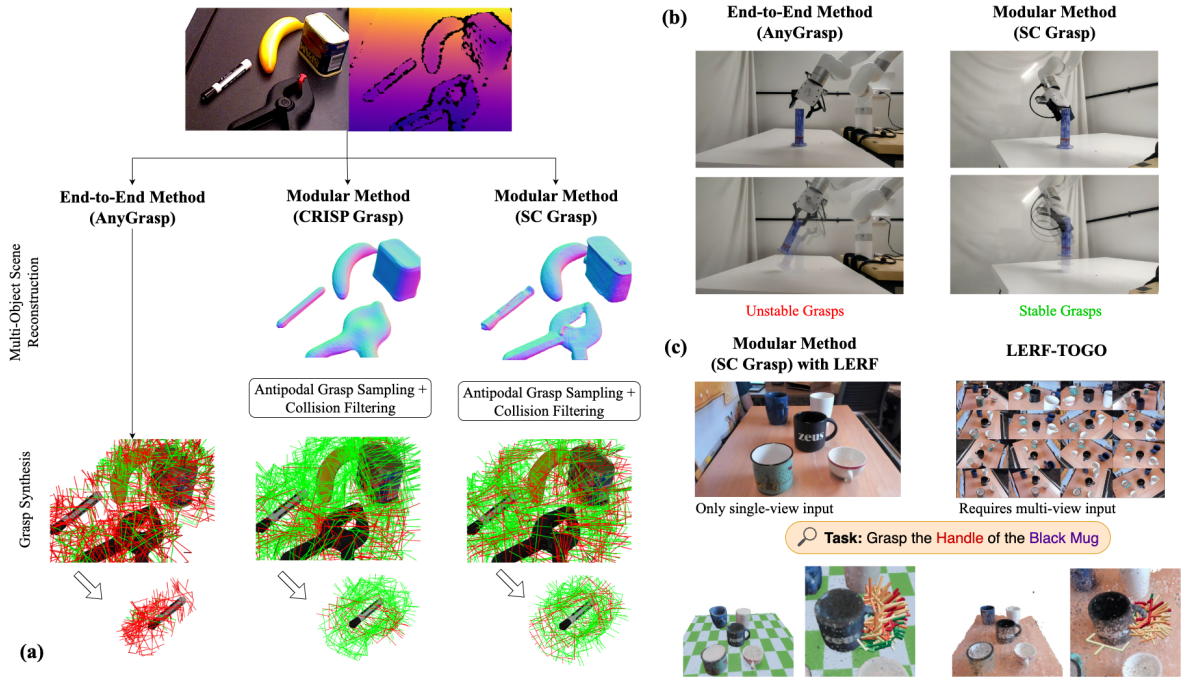
Object Pose and Shape Estimation for Grasping: Does it Work?

Pavan Karke[1*], Kushal Shah[1*], Gaurav Singh[1], Md Faizal Karim[1], K Madhava Krishna[1], and Rajat Talak[2,3]

[*]Equal Contributions, [1]Robotics Research Center, IIIT Hyderabad, [2]Massachusetts Institute of Technology, [3]National University of Singapore



(a) Illustrates the workings of three baseline methods for grasp synthesis: a state-of-the-art, end-to-end method (AnyGrasp), and two modular methods (CRISP Grasp and SC Grasp), which first estimate the object pose and shape for all objects in the scene and then synthesize grasps using antipodal sampling.We observe that modular methods synthesize more successful grasps (green) than the end-to-end method; unsuccessful grasps are shown in red. (b) The end-to-end methods synthesize more unstable grasps compared to modular methods in our real-world experiments. (c) The modular methods are augmented with vision-language models to yield task-oriented grasps from just single-view RGB-D input. We observe comparable performance compared to the state-of-the-art baseline LERF-TOGO that uses multi-view input.

Extended Abstract — The problem of object pose and shape estimation has seen key advancements lately. Encoder-decoder models have shown category-agnostic shape encoding capacity, whereas diffusion models, utilizing the recent progress in novel-view synthesis, have demonstrated open-set generalizability. In this work, we ask the question:

Q: Are the object pose and shape estimation methods mature enough, such that when used with antipodal grasp sampling, can outperform the end-to-end grasp synthesis methods?

We explore this question in detail by scoping our study to parallel grippers, 6 DoF grasps, and single-view RGB(-D) image as input. We implement and compare a state-of-the-art, end-to-end grasp synthesis method and two modular methods, which first estimate the object pose and shape for all objects in the scene, and generates grasps using antipodal sampling. We observe that the modular methods outperform the end-to-end method in all our experiments. The modular methods are able to synthesize plenty of grasps, even for small objects, where the end-to-end methods fail. The effectiveness of the modular methods is contingent on the accuracy of the pose and shape estimation, and suffers partial degradation in cluttered scenes — a limitation of the existing pose and shape estimation methods. We also investigate into the failure modes for our two modular methods, which use two different ways of object pose and shape estimation; one based on an encoder-decoder model, while another a diffusion model. Finally, we demonstrate that the modular methods can be augmented with vision-language models to yield task-oriented grasps from just single-view RGB-D image as input. We notice comparable performance to the state-of-the-art LERF-TOFO baseline.

A. Background

The problem of estimating the complete geometry of an object from a single- or partial-view is very important in robotics. It can help robots navigate a scene as well as grasp and manipulate objects with very few observations. It is a capability that mobile manipulators aught to have. Classical works on grasping, for instance, relied on the assumption that such a complete object geometry is available for grasp synthesis [1], [2]. Knowing the complete geometry, material, and physical properties, the grasp synthesis problem was framed as an optimization problem to seek the grasp that is most likely to succeed [2], [3]. Recent works have used these physics constraints to develop grasp sampling approaches given a 3D mesh of the object [4]–[6]. However, estimating the complete geometry, from single- or partial-view, is an underspecified problem and cannot be solved without sufficient induced priors.

The problem of estimating the complete geometry of objects given a single-view RGB(-D) image has seen key advancements lately [7]–[23]. First, encoder-decoder models have been able to estimate the pose and shape of an object, across object categories [13]–[17]. Previously, researchers were working on category-level object pose and shape estimation, where the template object shape would be known [8], [10]–[12], [24]–[28]. It was believed that such a category-agnostic generalization would not be possible. Second, diffusion models have been able to show open-set generalization [18]–[23]. This leveraged the progress on novel-view synthesis [29], which once framed as a pose conditioned image-to-image generation problem and trained using a large diffusion model shows zero-shot generalizability.

Today's grasp synthesis approaches are dominated by end-to-end regression methods which regress the grasp pose directly from a single- or partial-view of a scene [4], [30]–[35]. These methods were initially inspired by the success of the end-to-end deep learning methods in computer vision [3], [35]. They have now shown remarkable capability in terms of increased accuracy, generalizability to unseen objects, and robustness compared to traditional machine learning or analytical methods. End-to-end methods have also enabled faster run times for real-time operations, unlike sampling-based methods [35]. The end-to-end methods are now being extended to task-oriented grasping [36], [37], language-driven grasping [37], [38], dynamic object grasping [34], and cross embodiment grasping [39].

B. Contributions

In this work, we inquire if this recent progress in object pose and shape estimation can be leveraged for grasping (see $\mathcal{Q}$ on page 1). We explore this question via experimental analysis. Scoping our study to parallel grippers, 6 DoF grasps, and single-view RGB-D image as input we analyze and compare a state-of-the-art, end-to-end grasp synthesis method and two modular method, which first estimate object pose and shape of all the objects in the scene, and generate grasps using antipodal sampling [4], [5]. We evaluate the grasp synthesis methods with collisions, force closure, and grasp stability. We conduct our experiments in a physics simulator, real-world datasets, and real-world grasping experiments with Xarm7 manipulator arm affixed with a parallel-jaw gripper and an Intel RealSense D455 camera.

Our main contributions in this paper are our observations from the experimental analysis.

First, we make the following key observations in our experiments:

1) The modular methods outperform the end-to-end methods in all our experiments.
2) The modular methods are able to generate plenty of grasps for each object, whereas for the end-to-end methods it depends on the object size.
3) Performance of modular methods depreciates as we add clutter to the scene.

Our studies indicate that the single-view object pose and shape estimation methods are mature enough to enable successful grasp synthesis and execution with antipodal grasp sampling.

Second, we answer whether we can trivially combine the grasps synthesized by the end-to-end model with the estimated object pose and shape for improved performance. We observe that simple strategies such as using the estimated shape and pose to filter colliding grasps does not yield significant gains. Therefore, if one aims to combine the best of both worlds (*i.e.,* enhance grasp synthesis using the estimated object pose and shape), a non-trivial solution is needed.

Third, we analyze the run-time requirements and failure modes of the two modular pipelines. Comparing the failure modes for the two modular approaches, we observe that most of the grasp synthesis errors are due to scale and pose errors for the first method that uses the encoder-decoder model. In contrast, a significant 29.6% of the grasping errors in the second method are caused due to the shape estimation errors.

Fourth, we show that the modular approaches can be extended with vision-language models (*e.g.,* CLIP) to enable language- and task-oriented grasps synthesis. We demonstrate our modular approaches on the LERF-TOGO dataset [37], and qualitatively compare with LERF-TOGO as our baseline [37]. Our result shows the promise of robust language-conditioned grasps from single-view RGBD.

C. Concluding Remark

Our work investigates the question whether this progress in object pose and shape estimation can be leveraged for the task of grasp synthesis. In the year 2021, the answer to this question was a strong negative as Sundermeyer *et al.* [32] wrote:

> "A complete 3D reconstruction enables traditional grasp planning. However, learned single-view reconstructions are often ambiguous, coarse and require class conditioning. Multiple views for 3D scanning are beneficial but not always obtainable, take additional time and typically assume a static scene."

Our results show that we are at an inflection point, and that the advances in object pose and shape estimation have began to show advantages to the downstream tasks like grasping.

REFERENCES

[1] K. Shimoga, "Robot grasp synthesis algorithms: A survey," *Intl. J. of Robotics Research*, vol. 15, no. 3, pp. 230–266, 1996.

[2] A. Bicchi and V. Kumar, "Robotic grasping and contact: a review," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2000.

[3] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesisa survey," *IEEE Trans. Robotics*, vol. 30, no. 2, pp. 289–309, 2014.

[4] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017.

[5] G. Zhai, Y. Zheng, Z. Xu, X. Kong, Y. Liu, B. Busam, Y. Ren, N. Navab, and Z. Zhang, "Da^2 dataset: Toward dexterity-aware dual-arm grasping," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 4, pp. 8941–8948, 2022.

[6] C. Eppner, A. Mousavian, and D. Fox, "A billion ways to grasps - an evaluation of grasp sampling schemes on a dense, physics-based grasp data set," in *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, 2019.

[7] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 16611–16621, 2021.

[8] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651, 2019.

[9] X. Liu, R. Zhang, C. Zhang, G. Wang, J. Tang, Z. Li, and X. Ji, "GDRNPP: A geometry-guided and fully learning-based object pose estimator," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 47, no. 7, pp. 5742–5759, 2025.

[10] Y. Fu and X. Wang, "Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset," *Conf. on Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 27469–27483, 2022.

[11] K. Chen and Q. Dou, "Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation," in *Intl. Conf. on Computer Vision (ICCV)*, pp. 2773–2782, 2021.

[12] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency," in *Intl. Conf. on Computer Vision (ICCV)*, pp. 3560–3569, 2021.

[13] C.-Y. Wu, J. Johnson, J. Malik, C. Feichtenhofer, and G. Gkioxari, "Multiview compressive coding for 3d reconstruction," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[14] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, "LRM: Large reconstruction model for single image to 3d," 2024.

[15] J. Shi, R. Talak, H. Zhang, D. Jin, and L. Carlone, "CRISP: Object pose and shape estimation with test-time adaptation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025. .

[16] K. A. Vasudev, A. Gupta, and S. Tulsiani, "Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Z. Huang, S. Stojanov, A. Thai, V. Jampani, and J. M. Rehg, "Zeroshape: Regression-based zero-shot shape reconstruction," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," *arXiv preprint arXiv:2404.07191*, 2024.

[19] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Intl. Conf. on Computer Vision (ICCV)*, 2023.

[20] H. Jun and A. Nichol, "Shap-e: Generating conditional 3d implicit functions," *arXiv preprint arXiv:2305.02463*, 2023.

[21] D. Xu, Y. Jiang, P. Wang, Z. Fan, Y. Wang, and Z. Wang, "NeuralLift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 views," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[22] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "RealFusion: 360 reconstruction of any object from a single image," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[23] A. Agarwal, G. Singh, B. Sen, T. Lozano-Prez, and L. P. Kaelbling, "SceneComplete: Open-World 3D Scene Completion in Complex Real World Environments for Robot Manipulation," *arXiv preprint arxiv:2410.23643*, Oct. 2024.

[24] G. Wang, F. Manhardt, J. Shao, X. Ji, N. Navab, and F. Tombari, "Self6D: Self-supervised monocular 6D object pose estimation," in *European Conf. on Computer Vision (ECCV)* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), pp. 108–125, Nov. 2020.

[25] Y. Zhang and J. Leonard, "ShapeICP: Iterative category-level object pose and shape estimation from depth," 2024.

[26] D. Chen, J. Li, Z. Wang, and K. Xu, "Learning canonical shape space for category-level 6d object pose and size estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 11973–11982, 2020.

[27] M. Tian, M. H. Ang, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation," in *European Conf. on Computer Vision (ECCV)*, pp. 530–546, Springer, 2020.

[28] M. Lunayach, S. Zakharov, D. Chen, R. Ambrus, Z. Kira, and M. Z. Irshad, "Fsd: Fast self-supervised single rgb-d to categorical 3d objects," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pp. 14630–14637, IEEE, 2024.

[29] D. Watson, W. Chan, R. Martin-Brualla, J. Ho, A. Tagliasacchi, and M. Norouzi, "Novel view synthesis with diffusion models," *arxiv preprints arxiv:2210.04628*, 2022.

[30] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "GraspNet-1Billion: A large-scale benchmark for general object grasping," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6DoF closed-loop grasping from low-cost demonstrations," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 3, pp. 4978–4985, 2020.

[32] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," 2021.

[33] M. Gou, H.-S. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, "RGB Matters: Learning 7-DoF grasp poses on monocular rgbd images," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021.

[34] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Trans. Robotics*, vol. 39, pp. 3929–3945, Jun. 2023.

[35] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, and A. Cosgun, "Deep learning approaches to grasp synthesis: A review," *IEEE Trans. Robotics*, vol. 39, pp. 3994–4015, Jun. 2023.

[36] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *Intl. J. of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

[37] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," 2023.

[38] A. D. Vuong, M. N. Vu, B. Huang, N. Nguyen, H. Le, T. Vo, and A. Nguyen, "Language-driven grasp detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[39] Z. Wei, Z. Xu, J. Guo, Y. Hou, C. Gao, Z. Cai, J. Luo, and L. Shao, "D(R,O) Grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2025.