CaTGrasp: Learning Category-Level Task-Relevant Grasping from Simulation

Bowen Wen[1,2], Wenzhao Lian[1], Kostas Bekris[2] and Stefan Schaal[1]

***Abstract*—Task-relevant grasping is critical for industrial assembly, where downstream manipulation tasks constrain the set of valid grasps. Learning how to perform this task, however, is challenging, since task-relevant grasp labels are hard to define and annotate. There is also yet no consensus on proper representations for modeling or off-the-shelf tools for performing task-relevant grasps. This work proposes a framework to learn task-relevant grasping for industrial objects without the need of time-consuming real-world data collection or manual annotation. To achieve this, the entire framework is trained solely in simulation, including supervised training with synthetic label generation and self-supervised, hand-object interaction. In the context of this framework, this paper proposes a novel, object-centric canonical representation at the category level, which allows establishing dense correspondence across object instances and transferring task-relevant grasps to novel instances. Extensive experiments on task-relevant grasping of densely-cluttered industrial objects are conducted in both simulation and real-world setups, demonstrating the effectiveness of the proposed framework. Code and data are available at https://github.com/wenbowen123/catgrasp**

I. Introduction

Robot manipulation often requires identifying a suitable grasp that is aligned with a downstream task. An important application domain is industrial assembly, where the robot needs to perform constrained placement after grasping an object [2], [3]. In such cases, a suitable grasp requires stability during object grasping and transporting while avoiding obstructing the placement process. To tackle this problem, this work aims to learn category-level, task-relevant grasping solely in simulation, circumventing the requirement of manual data collection or annotation efforts. In addition, during the test stage, the trained model can be directly applied to novel object instances with previously unseen dimensions and shape variations, saving the effort of acquiring 3D models or re-training for each individual instance. In summary, the contributions of this work are the following: (a) A novel framework for learning category-level, task-relevant grasping of densely cluttered industrial objects and targeted placement; (b) This work models dense, point-wise task relevance on 3D shapes by representing it as hand-object contact heatmaps generated in a self-supervised manner in simulation; (c) We propose "Non-Uniform Normalized Object Coordinate Space" (NUNOCS) representation for learning category-level object 6D poses and 3D scaling, which establishes more reliable dense correspondence across object instances; (d) Direct sim-to-real transfer is achieved by leveraging domain randomization [4], bi-directional alignment [5], and domain-invariant, hand-object contact heatmaps modeled in a category-level canonical space.

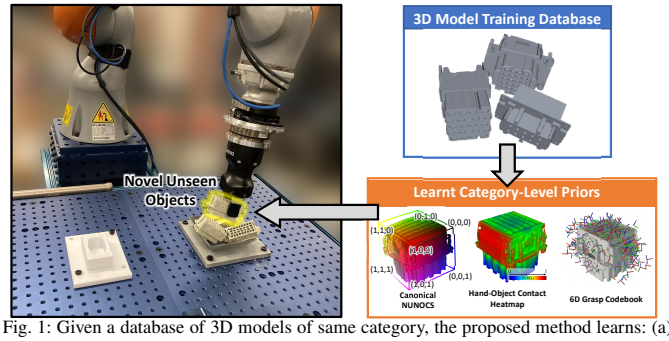


Fig. 1: Given a database of 3D models of same category, the proposed method learns: (a) an object-centric NUNOCS representation that is canonical for the object category, (b) a heatmap that indicates the task achievement success likelihood dependent on the hand-object contact region during the grasp, and (c) a codebook of stable 6D grasp poses. The heatmap and the grasp poses are transferred to real-world, novel unseen object instances during testing for solving task-relevant grasping.

II. Problem Statement

Given a dense clutter of the same type of novel unseen objects, the objective is to compute task-relevant 6D grasp poses that enables direct transportation for the downstream task. The inputs to the framework are listed below:

- A collection of 3D models $\mathcal{M}_C$ belonging to category C for training (e.g., Fig. 1 top-right). They do not include any testing instances of the same category, i.e., $M_C^{\text{test}} \notin \mathcal{M}_C$.
- A downstream placement task T_C corresponding to the category (e.g., Fig. 1), including a matching receptacle and the criteria of placement success.
- A depth image I_D of the scene for grasp planning during the test stage.

III. Approach

Fig. 2 summarizes the proposed framework. Offline, given a collection of models $\mathcal{M}_C$ of the same category, synthetic data are generated in simulation for training the *NUNOCS Net*, *Grasping Q Net* and *3D U-Net*. Then, self-interaction in simulation provides hand-object contact experience, which is summarized in task-relevant heatmaps for grasping. The canonical NUNOCS representation allows the aggregation of category-level, task-relevant knowledge across instances. Online, the category-level knowledge is transferred from the canonical NUNOCS model to the segmented target object via dense correspondence and 9D pose estimation, guiding the grasp candidate generation and selection.

[1]Intrinsic Innovation LLC in CA, USA. {wenzhaol, sschaal}@intrinsic.ai. This research was conducted during Bowen's internship at Intrinsic. This work has been previously accepted to appear at ICRA 2022 [1].

[2]Rutgers University in NJ, USA. {bw344, kostas.bekris}@cs.rutgers.edu.

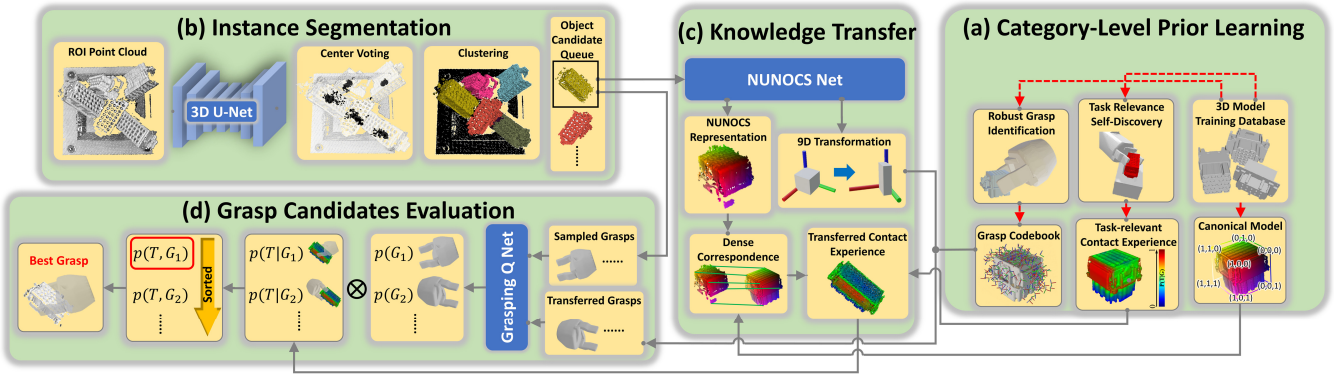


Fig. 2: Overview of the proposed framework. **Right: (a)** Given a collection of CAD models for objects of the same category, the NUNOCS representation is aggregated to generate a canonical model for the category. The CAD models are further utilized in simulation to generate synthetic point cloud data for training all the networks (*3D U-Net, NUNOCS Net* and *Grasping Q Net*). Meanwhile, the category-level grasp codebook and hand-object contact heatmap are identified via self-interaction in simulation. **Top-left: (b)** A *3D U-Net* is leveraged to predict point-wise centers of objects in dense clutter, based on which the instance segmentation is computed by clustering. **Center: (c)** The *NUNOCS Net* operates over an object's segmented point cloud and predicts its NUNOCS representation to establish dense correspondence with the canonical model and compute its 9D pose $\xi_o \in \{SE(3) \times R^3\}$ (6D pose and 3D scaling). This allows to transfer the precomputed category-level knowledge to the observed scene. **Bottom-left: (d)** Grasp proposals are generated both by transferring them from a canonical grasp codebook and directly by sampling over the observed point cloud. IK-infeasible or in-collision (using FCL [6]) grasps are rejected. Then, the *Grasping Q Net* evaluates the stability of the accepted grasp proposals. This information is combined with a task-relevance score computed from the grasp's contact region. The entire process can be repeated for multiple object segments to find the currently best grasp to execute according to $P(T,G) = P(T|G)P(G)$. Red dashed arrows occur in the offline training stage only.

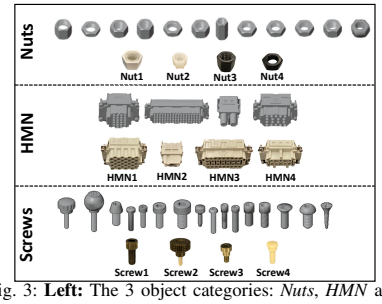
IV. Experiments

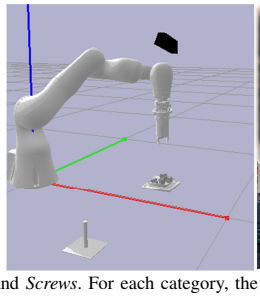
Evaluations are performed in similar setups in simulation and the real-world. The hardware is composed of a Kuka IIWA14 arm, a Robotiq Hand-E gripper, and a Photoneo 3D camera, as in the wrapped figure. Simulation experiments are conducted in PyBullet, with the corresponding hardware components modeled and gravity applied to manipulated objects. At the start of the bin-picking process, a random number of object instances (between 4 to 6) of the same type are randomly placed inside the bin to form a cluttered pile. Experiments for each of the 12 object instances have been repeated 10 times in simulation and 3 times in real-world, with different arbitrarily formed initial pile configurations. This results in approximately 600 and 180 grasp evaluations in simulation and real-world respectively for each evaluated approach. For each bin-clearing scenario, its initial pile configuration is recorded and set similarly across all evaluated methods for fair comparison. After each grasp, its stability is evaluated by a lifting action. If the object drops, the grasp is marked as failure. For stable grasps, additional downstream category-specific placement tasks are performed to further assess the task-relevance. A stable grasp is further examined and marked as a task-relevant grasp, if the placement also succeeds. Otherwise, it is marked as a task-irrelevant grasp, though being stable. The placement receptacles are CAD designed and 3D printed for each object instance with tight placement tolerances ($< 3mm$). For evaluation purposes, the placement planning is performed based on manually annotated 6D in-hand object pose post-grasping. This effort is beyond the scope of this work.

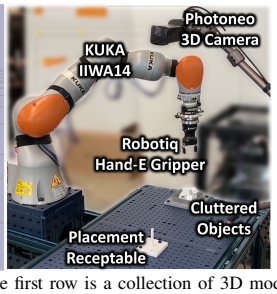
Our proposed method is compared against:

- `PointNetGPD` [7]: A state-of-the-art method on robust grasping. For fair comparison, the network is retrained using the same synthetic training data of industrial objects as our method. At test time, it directly samples grasp proposals over the raw point cloud without performing instance segmentation [7].
- `Ours-NA`: A variant of our method that does not consider task-relevant affordance but still transfers category-level grasp knowledge. Only $P(G)$ is used for ranking grasp candidates.
- `Ours-NOCS`: A variant of our method by replacing the NUNOCS representation with NOCS [8] for solving the category-level pose, while the remainings are the same as our framework.

The quantitative results in simulation and real-world are shown in Fig. 3. The success rate excludes the task-irrelevant or failed grasps. As demonstrated in the two tables, `Ours` significantly surpasses all baselines measured by the success rate on task-relevant grasping in both simulation and real-world.







	Nuts	HMN	Screws	Total
PointnetGPD	53.3%	49.2%	45.0%	49.2%
Ours-NA	51.1%	58.3%	50.0%	53.1%
Ours-NOCS	75.6%	71.7%	80.0%	75.7%
Ours	**97.8%**	**88.3%**	**93.3%**	**93.1%**

	Nuts	HMN	Screws	Total
PointnetGPD	33.3%	35.0%	38.0%	35.4%
Ours-NA	40.0%	43.3%	42.9%	42.1%
Ours-NOCS	70.0%	58.3%	52.5%	60.3%
Ours	**93.3%**	**83.3%**	**86.7%**	**87.8%**

Fig. 3: **Left:** The 3 object categories: *Nuts*, *HMN* and *Screws*. For each category, the first row is a collection of 3D models used for learning in simulation. The second row is the novel unseen instances used during testing. **Middle:** Hardware setup in simulation and real-world. **Right:** Percentage of task-relevant grasping in simulation (top) and real-world (bottom) respectively.

References

[1] B. Wen, W. Lian, K. Bekris, and S. Schaal, "CaTGrasp: Learning Category-Level Task-Relevant Grasping in Clutter from Simulation," *ICRA*, 2022.

[2] A. S. Morgan, B. Wen, J. Liang, A. Boularias, A. M. Dollar, and K. Bekris, "Vision-driven compliant manipulation for reliable, high-precision assembly tasks," *RSS*, 2021.

[3] J. Luo, O. Sushkov, R. Pevceviciute, W. Lian, C. Su, M. Vecerik, N. Ye, S. Schaal, and J. Scholz, "Robust multi-modal policies for industrial assembly via reinforcement learning and demonstrations: A large-scale study," *RSS*, 2021.

[4] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS 2017*.

[5] B. Wen and et al, "se (3)-tracknet: Data-driven 6D pose tracking by calibrating image residuals in synthetic domains," in *IROS*, 2020.

[6] J. Pan, S. Chitta, and D. Manocha, "Fcl: A general purpose library for collision and proximity queries," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 3859–3866.

[7] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "PointNetGPD: Detecting grasp configurations from point sets," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

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