Accepted in IEEE International Conference on Robotics and Automation (ICRA) 2022

CaTGrasp: Learning Category-Level Task-Relevant Grasping in Clutter 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://sites.google.com/view/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 [1], [2]. In such cases, a suitable grasp requires stability during object grasping and transporting while avoiding obstructing the placement process. For instance, a grasp where the gripper fingers cover the thread portion of a screw can impede its placement through a hole. Grasping a screw in this manner is not a task-relevant grasp.

There are many challenges, however, in solving task-relevant grasps. (a) Grasping success and task outcome are mutually dependent [3]. (b) Task-relevant grasping involves high-level semantic information, which cannot be easily modeled or represented. (c) 6D grasp pose annotation in 3D is more complicated than 2D image alternatives. Achieving task-relevant grasps requires additional semantic priors in the label generation process than geometric grasp generation often studied in stable grasp learning [4], [5]. (d) Generalization to highly-variant, novel instances within the category requires effective learning on category-level priors. (e) In densely cluttered industrial object picking scenarios, as considered in this work, the growing number of objects, the size of the grasp solution space, as well as challenging object properties (e.g., textureless, reflective, tiny objects, etc.), introduce additional combinatorial challenges and demand increased robustness.

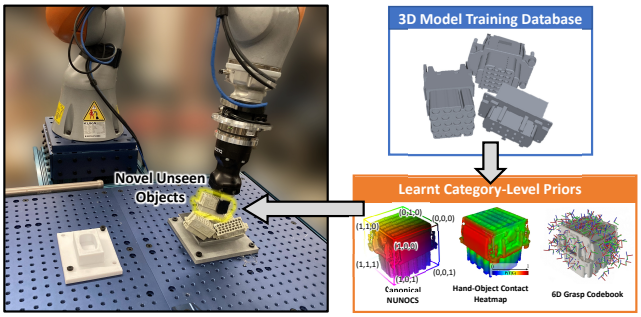


Fig. 1: Given a database of 3D models of same category, the proposed method learns: (a) an object-centric NUNOCS representation that is canonical for the object category, (b) a heatmap that indicates the task achievement success likelihood dependent on the hand-object contact region during the grasp, 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.

With recent advances in deep learning, many efforts resort to semantic part segmentation [6], [7] or keypoint detection [8] via supervised learning on manually labeled real-world data. A line of research is circumventing the complicated data collection process by training in simulation [9], [10]. It still remains an open question, however, whether category-level priors can be effectively captured through end-to-end training. Moreover, most of the existing work considers picking and manipulation of isolated objects [11]–[15]. The complexity of densely cluttered industrial object picking scenarios considered in this work, make the task-relevant grasping task even more challenging.

To tackle the above challenges, 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. This is achieved by encoding 3D properties - including shape, pose and gripper-object contact experience that is relevant to task performance - shared across diverse instances within the category. Therefore, once trained, this category-level, task-relevant grasping knowledge not only transfers across novel instances, but also effectively generalizes to real-world densely cluttered scenarios without the need for fine-tuning. In summary, the contributions of this work are the following:

- A novel framework for learning category-level, task-relevant grasping of densely cluttered industrial objects and targeted placement. To the best of the authors' knowl-

[1]Intrinsic Innovation LLC in CA, USA. {wenzhaol, sschaal}@intrinsic.ai. This research was conducted during Bowen's internship at Intrinsic.

[2]Rutgers University in NJ, USA. {bw344, kostas.bekris}@cs.rutgers.edu. Bowen Wen and Kostas Bekris were partially supported by the US NSF Grant IIS-1734492. The opinions expressed here are of the authors and do not reflect the views of the sponsor.

edge, this is the first work that effectively tackles task-relevant grasping of industrial objects in densely cluttered scenarios in a scalable manner without human annotation.
- Instead of learning sparse keypoints relevant to task-relevant manipulation, as in [8], [10], this work models dense, point-wise task relevance on 3D shapes. To do so, it leverages hand-object contact heatmaps generated in a self-supervised manner in simulation. This dense 3D representation eliminates the requirement of manually specifying keypoints.
- It introduces the "Non-Uniform Normalized Object Coordinate Space" (NUNOCS) representation for learning category-level object 6D poses and 3D scaling, which allows non-uniform scaling across three dimensions. Compared to the previously proposed Normalized Object Coordinate Space (NOCS) representation [16] developed in the computer vision community for category-level 6D pose and 1D uniform scale estimation, the proposed representation allows to establish more reliable dense correspondence and thus enables fine-grained knowledge transfer across object instances with large shape variations.
- The proposed framework is solely trained in simulation and generalizes to the real-world without any re-training, by leveraging domain randomization [17], bi-directional alignment [18], and domain-invariant, hand-object contact heatmaps modeled in a category-level canonical space. To this end, a synthetic training data generation pipeline, together with its produced novel dataset in industrial dense clutter scenarios, is presented.

II. Related Work

Stable Grasping - Stable grasping methods focus on robust grasps. The methods can be generally classified into two categories: model-based and model-free. Model-based grasping methods require object CAD models to be available beforehand for computing and storing grasp poses offline w.r.t. specific object instances. During test stage, the object's 6D pose is estimated to transform the offline trained grasp poses to the object in the scene [19]–[22]. More recently, model-free methods relax this assumption by directly operating over observations such as raw point cloud [4], [23] or images [24]–[28], or transferring the category-level offline trained grasps via Coherent Point Drift [29]. Representative works [4], [5], [30] train a grasping evaluation network to score and rank the grasp candidates sampled over the observed point cloud. For the sake of efficiency, more recent works [24], [31]–[35] develop grasp pose prediction networks, which directly output 6D grasp proposals along with their scores, given the scene observation. Differently, the proposed `CaTGrasp` aims to compute grasps that are not only stable but also task-relevant.

Task-Relevant Grasping - Task-relevant grasping requires the grasps to be compatible with downstream manipulation tasks. Prior works have developed frameworks to predict affordance segmentation [7], [36]–[40] or keypoints [41] over the observed image or point cloud. This, however, often assumes manually annotated real world data is available to perform supervised training [12], [42], [43], which is costly and time-consuming to obtain. While [6], [44] alleviates the problem via sim-to-real transfer, it still requires manual specification of semantic parts on 3D models for generating synthetic affordance labels. Instead, another line of research [9], [10], [45] proposed to learn semantic tool manipulation via self-interaction in simulation. While the above research commonly tackles the scenarios of tool manipulation or household objects, [46] shares the closest setting to ours in terms of industrial objects. In contrast to the above, our work considers more challenging densely cluttered scenarios. It also generalizes to novel unseen object instances, without requiring objects' CAD models for pose estimation or synthetic rendering during testing as in [46].

Category-Level Manipulation - In order to generalize to novel objects without CAD models, category-level manipulation is often achieved by learning correspondence shared among similar object instances, via dense pixel-wise representation [47]–[49] or semantic keypoints [8], [10]. In particular, sparse keypoint representations are often assigned priors about their semantic functionality and require human annotation. While promising results on household objects and tools have been shown [8], [10], it becomes non-trivial to manually specify semantic keypoints for many industrial objects (Fig. 4), where task-relevant grasp poses can be widely distributed. Along the line of work on manipulation with dense correspondence, [47] developed a framework for learning dense correspondence over 2D image pairs by training on real world data, which is circumvented in our case. [48], [49] extended this idea to multi-object manipulation given a goal configuration image. Instead of reasoning on 2D image pairs which is constrained to specific view points, this work proposes NUNOCS representation to establish dense correspondence in 3D space. This direct operation in 3D allows to transfer object-centric contact experience, along with a 6D grasp pose codebook. Additionally, the task-relevant grasping has not been achieved in [47]–[49].

III. Problem Statement

We assume novel unseen objects of the same type have been collected into a bin, forming a densely cluttered pile as in common industrial settings. The objective is to compute task-relevant 6D grasp poses $\xi_G \in SE(3)$ that allow a downstream constrained placement task. The grasping process is repeated for each object instance until the bin is cleared. 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 4 left). This does not include any testing instance in 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 5), including a matching receptacle and the criteria of placement success.
- A depth image I_D of the scene for grasp planning at test stage.

IV. Approach

Fig. 2 summarizes the proposed framework. Offline, given a collection of models $\mathcal{M}_C$ of the same category, synthetic

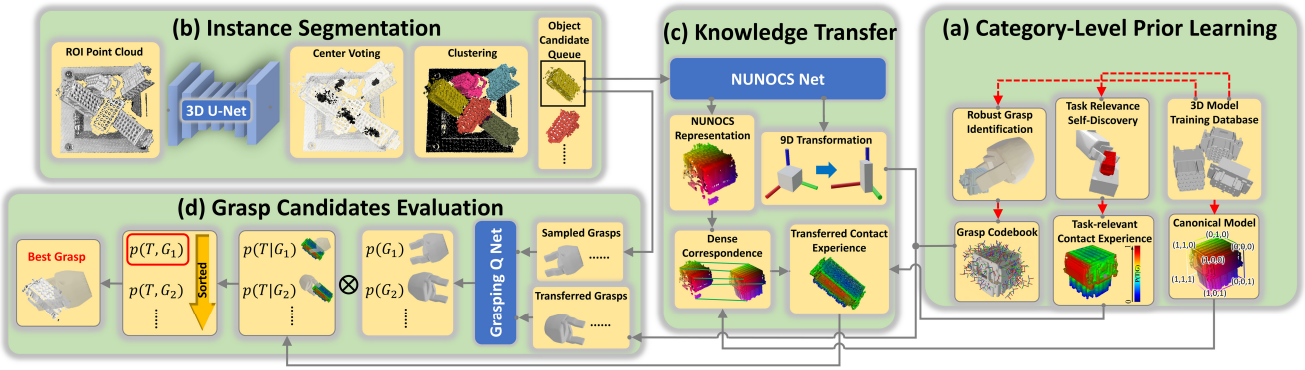


Fig. 2: Overview of the proposed framework. **Right: (a)** Given a collection of CAD models for objects of the same category, the NUNOCS representation is aggregated to generate a canonical model for the category. The CAD models are further utilized in simulation to generate synthetic point cloud data 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 [50]) 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.

data are generated in simulation (Sec. IV-E) for training the *NUNOCS Net* (Sec. IV-A), *Grasping Q Net* (Sec. IV-B) and *3D U-Net* (Sec. IV-D). Then, self-interaction in simulation provides hand-object contact experience, which is summarized in task-relevant heatmaps for grasping (Sec. IV-C). 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.

A. Category-level Canonical NUNOCS representation

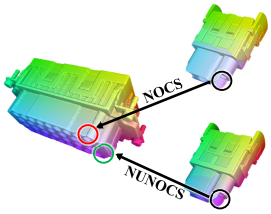
Previous work [47] learned dense correspondence between object instances using contrastive loss. It requires training on real-world data and operates over 2D images from specific viewpoints. Instead, this work establishes dense correspondence in 3D space to transfer knowledge from a trained model database $\mathcal{M}_C$ to a novel instance M_C^{test}.

Inspired by [16], this work presents the Non-Uniform Normalized Object Coordinate Space (NUNOCS) representation. Given an instance model M, all the points are normalized along each dimension, to reside within a unit cube:

$$p^d_{\mathbb{C}} = (p^d - p^d_{min})/(p^d_{max} - p^d_{min}) \in [0,1]; d \in \{x,y,z\}.$$

The transformed points exist in the canonical NUNOCS $\mathbb{C}$ (Fig. 1 bottom-right). In addition to being used for synthetic training data generation (Sec. IV-E), the models $\mathcal{M}_C$ are also used to create a category-level canonical template model, to generate a hand-object contact heatmap (Sec. IV-C) and a stable grasp codebook (Sec. IV-B). To do so, each model in $\mathcal{M}_C$ is converted to the space $\mathbb{C}$, and the canonical template model is represented by the one with the minimum sum of Chamfer distances to all other models in $\mathcal{M}_C$. The transformation from each model to this template is then utilized for aggregating the stable grasp codebook and the task-relevant hand-object contact heatmap.

For the *NUNOCS Net*, we aim to learn $\Phi : \mathcal{P}_o \rightarrow \mathcal{P}_{\mathbb{C}}$, where $\mathcal{P}_o$ and $\mathcal{P}_{\mathbb{C}}$ are the observed object cloud and the canonical space cloud, respectively. $\Phi(\cdot)$ is built with a PointNet-like architecture [51] given it is light-weight and efficient. The learning task is formulated as a classification problem by discretizing $p^d_{\mathbb{C}}$ into 100 bins. Softmax cross entropy loss is used as we found it more effective than regression by reducing the solution space [16]. Along with the predicted dense correspondence, the 9D object pose $\xi_o \in \{SE(3) \times R^3\}$ is also recovered. It is computed via *RANSAC* [52] to provide an affine transformation from the predicted canonical space cloud $\mathcal{P}_{\mathbb{C}}$ to the observed object segment cloud $\mathcal{P}_o$, while ensuring the rotation component to be orthonormal.



Compared to the original NOCS representation [16], which recovers a 7D pose of novel object instances, the proposed NUNOCS allows to scale independently in each dimension when converting to the canonical space. Therefore, more fine-grained dense correspondence across object instances can be established via measuring their similarity (L_2 distance in our case) in $\mathbb{C}$. This is especially the case for instances with dramatically different 3D scales, as shown in the wrapped figure, where colors indicate dense correspondence similarity in $\mathbb{C}$ and one example correspondence for NUNOCS and NOCS respectively. A key difference from another related work on VD-NOC [53], which directly normalizes the scanned point cloud in the camera frame, is that the proposed NUNOCS representation is object-centric and thus agnostic to specific camera parameters or viewpoints.

B. Stable Grasp Learning

During offline training, grasp poses are uniformly sampled from the point cloud of each object instance, covering the feasible grasp space around the object. For each grasp G, the grasp quality is evaluated in simulation. To compute a continuous score $s_G \in [0,1]$ as training labels, 50 neighboring grasp poses are randomly sampled in the proximity of $\xi_G \in SE(3)$ and executed to compute the empirical grasp success rate. The intuition is that grasp stability should be

continuous over its 6D neighborhood. Once the grasps are generated, they are then exploited in two ways.

First, given the relative 9D transformation from the current instance to the canonical model, the grasp poses are converted into the NUNOCS space and stored in a *stable grasp codebook* $\mathcal{G}$. During test time, given the estimated 9D object pose $\xi_o \in \{SE(3) \times R^3\}$ of the observed object's segment relative to the canonical space $\mathbb{C}$, grasp proposals can be generated by applying the same transformation to the grasps in $\mathcal{G}$. Compared with traditional online grasp sampling over the raw point cloud [4], [5], this grasp knowledge transfer is also able to generate grasps from occluded object regions. In practice, the two strategies can be combined to form a robust hybrid mode for grasp proposal generation.

Second, the generated grasps are utilized for training the *Grasping Q Net*, which is built based on PointNet [51]. Specifically, in each dense clutter generated (as in Sec. IV-E), the object segment in the 3D point cloud is transformed to the grasp's local frame given the object and grasp pose. The *Grasping Q Net* takes the point cloud as input and predicts the grasp's quality $P(G)$, which is then compared against the discretized grasp score s_G to compute softmax cross entropy loss. This one-hot score representation has been observed to be effective for training [30].

C. Affordance Self-Discovery

In contrast to prior work [7], which manually annotates parts of segments, or uses predefined sparse keypoints [8], this work discovers grasp affordance via self-interaction. In particular, the objective is to compute $P(T|G) = P(T,G)/P(G)$ automatically for all graspable regions on the object. To achieve this, a dense 3D point-wise hand-object contact heatmap is modeled. For each grasp in the codebook $G \in \mathcal{G}$ (generated as in Sec. IV-B), a grasping process is first simulated. The hand-object contact points are identified by computing their signed distance w.r.t the gripper mesh. If it's a stable grasp, i.e., the object is lifted successfully against gravity, the count $n(G)$ for all contacted points on the object are increased by 1. Otherwise, the grasp is skipped. For these stable grasps, a placement process is simulated, i.e., placing the grasped object on a receptacle, to verify the task relevance (Fig. 2 top-right). Collision is checked between the gripper and the receptacle during this process. If the gripper does not obstruct the placement and if the object can steadily rest in the receptacle, the count of joint grasp and task success $n(G,T)$ on the contact points is increased by 1. After all grasps are verified, for each point on the object point cloud, its task relevance can be computed as $P(T|G) = n(G,T)/n(G)$. Examples of self-discovered hand-object contact heatmaps are shown in Fig. 3. Interestingly, these heatmaps achieve similar performance to human annotated part-segmentation [36] but can be interpreted as a "soft" version.

Eventually, for each of the training objects within the category, the hand-object contact heatmap $P(T|G)$ is transformed to the canonical model. The task-relevant heatmaps over all training instances are aggregated and averaged to be the final canonical model's task-relevance heatmap. During testing, due to the partial view of the object's segment, the antipodal contact points p_c are identified between the gripper mesh and the transformed canonical model (Fig. 2 bottom-left). For each grasp candidate, the score $P_G(T|G) = \frac{1}{|p_c|}\sum_{p_c} P_{p_c}(T|G)$ is computed. It is then combined with the predicted $P_G(G)$ from *Grasping Q Net* (Sec. IV-B) to compute the grasp's task-relevance score: $P_G(T,G) = P_G(T|G)P_G(G)$.

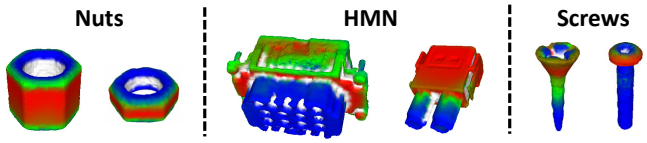


Fig. 3: Examples of task-relevant hand-object contact heatmaps $P(T|G)$. Warmer color indicates higher values of $P(T|G)$. The white areas are the small concave regions for which the rigid parallel-jaw gripper can't touch. They remain unexplored and set to the default $P(T|G) = 0.5$ though they are also unlikely to be touched during testing. The collected contact heatmap is object-centric and domain-invariant. Once identified in simulation, it is directly applied in the real world.

D. Instance Segmentation in Dense Clutter

This work employs the Sparse 3D U-Net [54], [55] due to its memory efficiency. The network takes as input the entire scene point cloud $\mathcal{P} \in R^{N\times 3}$ voxelized into sparse volumes and predicts per point offset $\mathcal{P}_{\text{offset}} \in R^{N\times 3}$ w.r.t. to predicted object centers. The training loss is designed as the L_2 loss between the predicted and the ground-truth offsets [56]. The network is trained independently, since joint end-to-end training with the following networks has been observed to cause instability during training.

During testing, the predicted offset is applied to the original points, leading the shifted point cloud to condensed point groups $\mathcal{P}+\mathcal{P}_{\text{offset}}$, as shown in Fig. 2. Next, DBSCAN [57] is employed to cluster the shifted points into instance segments. Additionally, the segmented point cloud is back-projected onto the depth image I_D to form 2D segments. This provides an approximation of the per-object visibility by counting the number of pixels in each segment. Guided by this, the remaining modules of the framework prioritize the top layer of objects given their highest visibility in the pile during grasp candidate generation.

E. Training Data Generation in Simulation

The entire framework is trained solely on synthetic data. To do so, synthetic data are generated in PyBullet simulation [58], aiming for physical plausibility [18], [59], while leveraging domain randomization [17] and bi-directional domain alignment on the depth modality [18]. At the start of each scene generation, an object instance type and its scale is randomly chosen from the associated category's 3D model database $\mathcal{M}_C$. The number of object instances in the bin, the camera pose relative to the bin, and physical parameters (such as bounciness and friction) are randomized. To generate dense clutter, object poses are randomly initialized above the bin. The simulation is executed until the in-bin objects are stabilized. The ground-truth labels for NUNOCS, grasping quality and instance segmentation are then retrieved from the simulator.

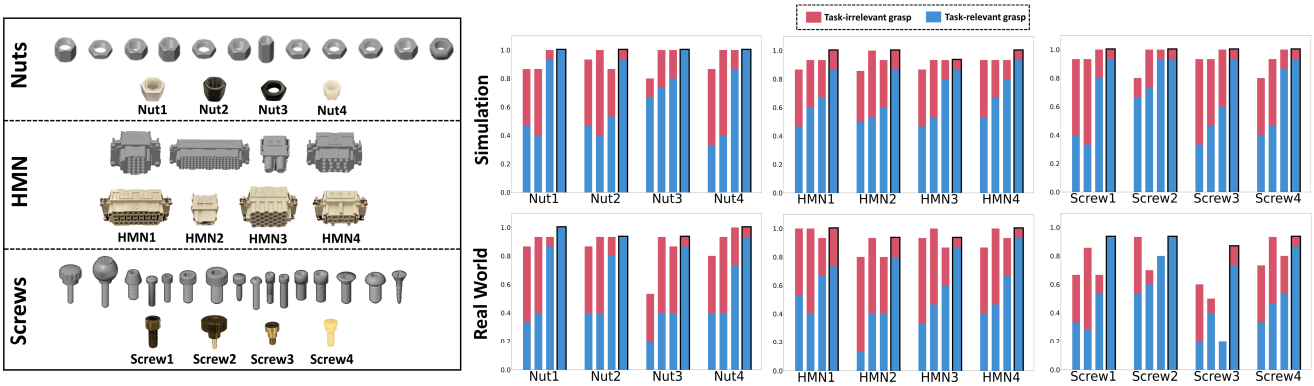


Fig. 4: **Left:** The 3 object categories: *Nuts*, *HMN* and *Screws*. For each category, the first row is a collection of 3D models used for learning in simulation. The second row is the novel unseen instances with challenging properties (tiny, texture-less, glossy, etc), used during testing both in simulation and in the real-world. **Right:** Instance-level grasp performance evaluated in simulation (top) and real-world (bottom). For each object instance, the group of 4 stacked bars from left to right are `PointNetGPD` [30], `Ours-NA`, `Ours-NOCS` and `Ours`, where the column corresponding to `Ours` is marked with a black boundary. Stable grasps include both task-irrelevant and task-relevant grasps, while the missing blanks are grasp failures.

V. Experiments

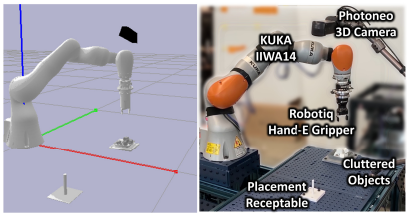
This section aims to experimentally evaluate 3 questions: i) Is the proposed dense correspondence expressive and reliable enough to represent various object instances within a category? ii) How well does the proposed model, only trained in simulation, generalize to real-world settings for task-relevant grasping? iii) Does the proposed category-level knowledge learning also benefit grasp stability? Our proposed method is compared against:

- `PointNetGPD` [30]: A state-of-the-art method on robust grasping. Its open-source code[1] is adopted. 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 [30].
- `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 [16] for solving the category-level pose, while the remainings are the same as our framework. This serves as an ablation to study the effectiveness of capturing cross-instance large variations by using the proposed NUNOCS representation.

Some other alternatives are not easy to compare against directly. For instance, `KETO` [10] and `kPAM` [8] focused on singulated household object picking from table-top, which is not easy to adapt to our setting. In addition, `kPAM` requires human annotated real-world data for training.

A. Experimental Setup

In this work, 3 different industrial object categories are included: *Nuts*, *Screws* and *HMN* series connectors. Their training and testing splits are depicted in Fig. 4(left). For each category, the first row is a collection of 3D models crawled online. This inexpensive source for 3D model training is used to learn category-level priors. The second row is 4 novel unseen object instances used for testing. Different from the training set, the testing object instances are real industrial objects purchased from retailers[2] for realistic evaluation. They are examined and ensured to be novel unseen object instances separated from the training set. The testing object instances are chosen so as to involve sufficient variance to evaluate the cross-instance generalization of the proposed method, while being graspable by the gripper in our configuration. The CAD models of the testing object instances are solely used for setting up the simulation environment.



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 (between 4 to 6) of object instances 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

[1] `https://github.com/lianghongzhuo/PointNetGPD`

[2] `https://www.digikey.com`; `https://www.mcmaster.com`

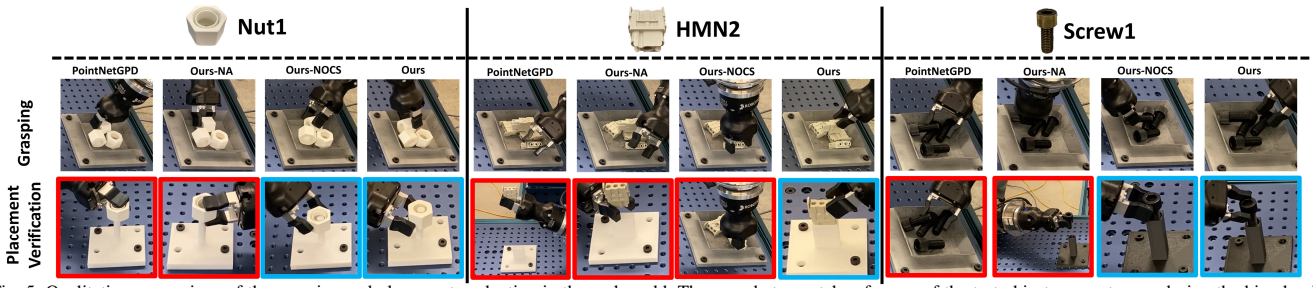


Fig. 5: Qualitative comparison of the grasping and placement evaluation in the real-world. The snapshots are taken for one of the test objects per category during the bin-clearing process, where the initial pile configuration is similar across different methods. For the placement verification images (second row), red boxes indicate either failure grasps, or stable but task-irrelevant grasps. Blue boxes indicate task-relevant grasps resulting in successful placement. Note that given the similar pile configuration, the methods do not necessarily choose the same object instance as the target due to different grasp ranking strategies. See the supplementary media for the complete video.

planning is performed based on manually annotated 6D in-hand object pose post-grasping. This effort is beyond the scope of this work.

	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%**

TABLE I: Results of **task-relevant** grasp percentage out of the total grasp attempts in simulation. For each method, approximately 600 grasps are conducted.

	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%**

TABLE II: Results of **task-relevant** grasp percentage out of the total grasp attempts in real-world. For each method, approximately 180 grasps are conducted.

B. Results and Analysis

The quantitative results in simulation and real-world are shown in Table I and II respectively. 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. Example real-world qualitative results are shown in Fig. 5.

Fig. 4 (right) decomposes the semantic grasping attempts of all the methods at the object instance level. Thanks to the task-relevant hand-object contact heatmap modeling and knowledge transfer via 3D dense correspondence, most of the grasps generated by `Ours` and `Ours-NOCS` are task-relevant semantic grasps. In comparison, a significant percentage of grasps planned by `PointNetGPD` and `Ours-NA` are not semantically meaningful, i.e., the objects, though stably grasped, cannot be directly manipulated for the downstream task without regrasping.

The fact that `Ours` reaches comparable or better performance than `Ours-NOCS` indicates that the proposed NUNOCS is a more expressive representation for 3D dense correspondence modeling and task-relevant grasp knowledge transfer. In particular, for the object "HMN2" (Fig. 4 left), the performance gap is more noticeable as its 3D scales vary significantly from the canonical model along each of the 3 dimensions. Additionally, the number of 3D training models available for the HMN category is more limited compared to the *Screws* or *Nuts* category. This requires high data-efficiency to capture the large variance. Despite these adverse factors, `Ours` is able to learn category-level task-relevant knowledge effectively, by virtue of the more representative NUNOCS space.

Although our proposed method targets at task-relevance, it also achieves a high stable grasp success rate, as shown in Fig. 4 (right). This demonstrates the efficacy of the proposed hybrid grasp proposal generation, where additional grasps transferred from the category-level grasp codebook span a more complete space around the object including occluded parts. This is also partially reflected by comparing `Ours-NA` and `PointNetGPD`, where `PointNetGPD` generates grasp candidates by solely sampling over the observation point cloud.

Comparing the overall performance across simulation and real-world experiments indicates a success rate gap of a few percent. This gap is noticeably larger for *Screws*, of which the instances are thin and tiny. Therefore, when they rest in a cluttered bin, it is challenging to find collision-free grasps and thus requires high precision grasp pose reasoning. In particular, as shown in Fig. 4, the instance "Screw3" challenges all evaluated methods in terms of grasp stability. Improving the gripper design [60] for manipulating such tiny objects is expected to further elevate the performance. In addition, during gripper closing, the physical dynamics of *Screws* is challenging to model when they roll inside the bin in simulation. With more advanced online domain adaptation techniques [61], the performance is expected to be boosted.

VI. Conclusion

This work proposed `CaTGrasp` to learn task-relevant grasping for industrial objects in simulation, avoiding the need for time-consuming real-world data collection or manual annotation. With a novel object-centric canonical representation at the category level, dense correspondence across object instances is established and used for transferring task-relevant grasp knowledge to novel object instances. Extensive experiments on task-relevant grasping of densely cluttered industrial objects are conducted in both simulation and real-world setups, demonstrating the method's effectiveness. In future work, developing a complete task-relevant framework with visual tracking [62] in the feedback loop for manipulating novel unseen objects is of interest.

References

[1] 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.

[2] 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.

[3] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: from sensory–motor coordination to imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.

[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," *RSS*, 2017.

[5] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[6] F.-J. Chu, R. Xu, and P. A. Vela, "Learning affordance segmentation for real-world robotic manipulation via synthetic images," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1140–1147, 2019.

[7] R. Monica and J. Aleotti, "Point cloud projective analysis for part-based grasp planning," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4695–4702, 2020.

[8] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "KPAM: Keypoint affordances for category-level robotic manipulation," *ISRR*, 2019.

[9] 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," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

[10] Z. Qin, K. Fang, Y. Zhu, L. Fei-Fei, and S. Savarese, "KETO: Learning keypoint representations for tool manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7278–7285.

[11] R. Wang, Y. Miao, and K. E. Bekris, "Efficient and high-quality prehensile rearrangement in cluttered and confined spaces," *Arxiv*, 2021.

[12] D. Song, C. H. Ek, K. Huebner, and D. Kragic, "Task-based robot grasp planning using probabilistic inference," *IEEE transactions on robotics*, vol. 31, no. 3, pp. 546–561, 2015.

[13] N. Mavrakis, M. Kopicki, R. Stolkin, A. Leonardis *et al.*, "Task-relevant grasp selection: A joint solution to planning grasps and manipulative motion trajectories," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 907–914.

[14] K. Gao, S. Feng, and J. Yu, "On minimizing the number of running buffers for tabletop rearrangement," *Robotics: Science and Systems XVII*, Jul 2021. [Online]. Available: http://dx.doi.org/10.15607/RSS.2021.XVII.033

[15] K. Gao, D. Lau, B. Huang, K. E. Bekris, and J. Yu, "Fast high-quality tabletop rearrangement in bounded workspace," *arXiv*, 2021.

[16] 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.

[17] 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*.

[18] B. Wen, C. Mitash, B. Ren, and K. E. Bekris, "se(3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2020. [Online]. Available: http://dx.doi.org/10.1109/IROS45743.2020.9341314

[19] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1386–1383.

[20] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *RSS*, 2018.

[21] C. Mitash, B. Wen, K. Bekris, and A. Boularias, "Scene-level pose estimation for multiple instances of densely packed objects," in *Conference on Robot Learning*. PMLR, 2020, pp. 1133–1145.

[22] B. Wen, C. Mitash, S. Soorian, A. Kimmel, A. Sintov, and K. E. Bekris, "Robust, occlusion-aware pose estimation for objects grasped by adaptive hands," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6210–6217.

[23] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, 2019.

[24] D. Park, Y. Seo, and S. Y. Chun, "Real-time, highly accurate robotic grasp detection using fully convolutional neural network with rotation ensemble module," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9397–9403.

[25] H. Cheng, D. Ho, and M. Q.-H. Meng, "High accuracy and efficiency grasp pose detection scheme with dense predictions," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3604–3610.

[26] B. Huang, S. D. Han, J. Yu, and A. Boularias, "Visual foresight trees for object retrieval from clutter with nonprehensile rearrangement," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, p. 231–238, Jan 2022. [Online]. Available: http://dx.doi.org/10.1109/lra.2021.3123373

[27] B. Huang, S. D. Han, A. Boularias, and J. Yu, "Dipn: Deep interaction prediction network with application to clutter removal," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021. [Online]. Available: http://dx.doi.org/10.1109/ICRA48506.2021.9561073

[28] B. Huang, T. Guo, A. Boularias, and J. Yu, "Self-supervised monte carlo tree search learning for object retrieval in clutter," *arXiv*, 2022.

[29] D. Rodriguez and S. Behnke, "Transferring category-based functional grasping skills by latent space non-rigid registration," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2662–2669, 2018.

[30] 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.

[31] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[32] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, "S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes," in *Conference on robot learning*. PMLR, 2020, pp. 53–65.

[33] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[34] M. Breyer, J. J. Chung, L. Ott, S. Roland, and N. Juan, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning*, 2020.

[35] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[36] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 5882–5889.

[37] R. Detry, J. Papon, and L. Matthies, "Task-oriented grasping with semantic and geometric scene understanding," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3266–3273.

[38] L. Antanas, P. Moreno, M. Neumann, R. P. de Figueiredo, K. Kersting, J. Santos-Victor, and L. De Raedt, "Semantic and geometric reasoning for robotic grasping: a probabilistic logic approach," *Autonomous Robots*, vol. 43, no. 6, pp. 1393–1418, 2019.

[39] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic, "Affordance detection for task-specific grasping using deep learning," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 91–98.

[40] P. Ardón, E. Pairet, Y. Petillot, R. P. Petrick, S. Ramamoorthy, and K. S. Lohan, "Self-assessment of grasp affordance transfer," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9385–9392.

[41] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, "An affordance keypoint detection network for robot manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2870–2877, 2021.

[42] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3352–3359, 2020.

[43] A. Allevato, M. Pryor, E. S. Short, and A. L. Thomaz, "Learning labeled robot affordance models using simulations and crowdsourcing," in *Robotics: Science and Systems (RSS)*, 2020.

[44] C. Yang, X. Lan, H. Zhang, and N. Zheng, "Task-oriented grasping in object stacking scenes with crf-based semantic model," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6427–6434.

[45] D. Turpin, L. Wang, S. Tsogkas, S. Dickinson, and A. Garg, "Gift: Generalizable interaction-aware functional tool affordances without labels," *RSS*, 2021.

[46] J. Zhao, D. Troniak, and O. Kroemer, "Towards robotic assembly by predicting robust, precise and task-oriented grasps," *CoRL*, 2020.

[47] P. R. Florence, L. Manuelli, and R. Tedrake, "Dense object nets: Learning dense visual object descriptors by and for robotic manipulation," *CoRL*, 2018.

[48] S. Yang, W. Zhang, R. Song, J. Cheng, and Y. Li, "Learning multi-object dense descriptor for autonomous goal-conditioned grasping," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4109–4116, 2021.

[49] C.-Y. Chai, K.-F. Hsu, and S.-L. Tsao, "Multi-step pick-and-place tasks using object-centric dense correspondences," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4004–4011.

[50] 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.

[51] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[52] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[53] T. Patten, K. Park, and M. Vincze, "Dgcm-net: dense geometrical correspondence matching network for incremental experience-based robotic grasping," *Frontiers in Robotics and AI*, vol. 7, 2020.

[54] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "PointGroup: Dual-set point grouping for 3d instance segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[55] B. Graham and L. van der Maaten, "Submanifold sparse convolutional networks," *arXiv preprint arXiv:1706.01307*, 2017.

[56] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Transactions on Robotics*, 2021.

[57] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[58] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2021.

[59] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *CoRL*, 2018.

[60] H. Ha, S. Agrawal, and S. Song, "Fit2Form: 3D generative model for robot gripper form design," in *Conference on Robotic Learning (CoRL)*, 2020.

[61] R. Antonova, M. Kokic, J. A. Stork, and D. Kragic, "Global search with bernoulli alternation kernel for task-oriented grasping informed by simulation," *CoRL*, 2018.

[62] B. Wen and K. Bekris, "Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models," *IROS*, 2021.