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

Bowen Wen\textsuperscript{1,2}, Wenzhao Lian\textsuperscript{1}, Kostas Bekris\textsuperscript{2} and Stefan Schaal\textsuperscript{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 \url{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 \cite{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 \cite{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 \cite{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.

With recent advances in deep learning, many efforts resort to semantic part segmentation \cite{6, 7} or keypoint detection \cite{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 \cite{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 \cite{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-

\textsuperscript{1}Intrinsic Innovation LLC in CA, USA. {wenzhao, sschaal}@intrinsic.ai. This research was conducted during Bowen’s internship at Intrinsic.

\textsuperscript{2}Rutgers University in NJ, USA. {bw3344, 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 \( M_C \) belonging to category \( C \) for training (e.g., Fig 4 left). This does not include any testing instance in the same category, i.e., \( M_C^{\text{test}} \notin M_C \).
- 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 \( M_C \) of the same category, synthetic
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 \(M_C\) to a novel instance \(M_C^{\text{test}}\).

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

The transformed points exist in the canonical NUNOCS \(C\) (Fig. 1 bottom-right). In addition to being used for synthetic training data generation (Sec. IV-E), the models \(M_C\) 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 \(M_C\) is converted to the space \(C\), and the canonical template model is represented by the one with the minimum sum of Chamfer distances to all other models in \(M_C\). 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 \(\Phi: P_o \rightarrow P_C\), where \(P_o\) and \(P_C\) 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 \(p^d\) into 100 bins. Softmax cross entropy loss is used as it is found more effective than regression by reducing the solution space [16]. Along with the predicted dense correspondence, the 9D object pose \(\xi_o \in SE(3)\) is also recovered. It is computed via RANSAC [52] to provide an affine transformation from the predicted canonical space cloud \(P_C\) to the observed object segment cloud \(P_o\), 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 (L2 distance in our case) in \(C\). 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 \(C\) and one example correspondence for NUNOCS and NOCS respectively. A key difference from another related work on VD-NOCS [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 to the NUNOCS space and stored in a stable grasp codebook \( G \). 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 \( C \), grasp proposals can be generated by applying the same transformation to the grasps in \( G \). 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 key-points \([8]\), 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 \( G \in G \) (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 \( P_G(T|G) = \frac{1}{|p_c|} \sum_{p_c} P(T|G) \) 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: \( P_G(T,G) = P_G(T|G)P_G(G) \).

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 \( P \in R^{N \times 3} \) voxelized into sparse volumes and predicts per point offset \( \mathcal{P}_{offset} \in R^{N \times 3} \) 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 \( \mathcal{P}+\mathcal{P}_{offset} \), 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 \( M_C \). 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.
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 [4] 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 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
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.
