Same Object, Different Grasps: Data and Semantic Knowledge for Task-Oriented Grasping

Adithyavairavan Murali[1], Weiyu Liu[2], Kenneth Marino[1], Sonia Chernova[2,3], and Abhinav Gupta[1,3]

[1]The Robotics Institute, Carnegie Mellon University
[2]Institute for Robotics and Intelligent Machines, Georgia Institute of Technology
[3]Facebook AI Research

Abstract: Despite the enormous progress and generalization in robotic grasping in recent years, existing methods have yet to scale and generalize task-oriented grasping to the same extent. This is largely due to the scale of the datasets both in terms of the number of objects and tasks studied. We address these concerns with the TaskGrasp dataset which is more diverse both in terms of objects and tasks, and an order of magnitude larger than previous datasets. The dataset contains $250K$ task-oriented grasps for 56 tasks and 191 objects along with their RGB-D information. We take advantage of this new breadth and diversity in the data and present the GCNGrasp framework which uses the semantic knowledge of objects and tasks encoded in a knowledge graph to generalize to new object instances, classes and even new tasks. Our framework shows a significant improvement of around 12% on held-out settings compared to baseline methods which do not use semantics. We demonstrate that our dataset and model are applicable for the real world by executing task-oriented grasps on a real robot on unknown objects.

Keywords: Robotic Grasping, Task-Oriented Grasping, Knowledge Graphs

1 Introduction

We have seen tremendous progress in the fundamental task of robotic grasping in recent years. State-of-the-art grasping algorithms have shown generalization to object instances [1, 2, 3, 4], viewpoints [5], DOF constraints [6, 7, 8], unknown environments [9] and even adversarial objects [10]. The key reason for the success of these approaches is large-scale learning. Typically data is sampled from analytical approaches in simulation [1, 7] or using a self-supervised framework [4, 5]. Despite these recent successes, there is still a significant gap between how humans grasp objects and how robots perform picking. Most techniques plan for stable grasps assuming grasping to be the end goal. However, when humans grasp an object, we do so with a particular purpose in mind and grasping is just the first step as a means to that end. For example, when humans grasp a cup, we use the handle to drink from it though several other stable grasps exist. Humans also use objects creatively, such as scooping with a bowl or hammering with a heavy mug. Different tasks may require completely different grasps for the same object. To effectively operate in human homes and complete multiple tasks, a personal robot would have to learn from humans to generalize grasping to several tasks and skills beyond a tool's prototypical use. For instance, if the robot is cooking and needs to stir a pot of pasta but doesn't have a spoon at hand, it can use an alternate tool, such as a knife. To truly get to human-level grasping, we must study not just stable grasping or grasping for an object's primary use-case but rather how to grasp depending on both the task and the object.

What are the bottlenecks in task-oriented robotic grasping? The biggest hurdle is the need for human-labeled data. Unlike self-supervised or analytical approaches for which force sensing or contact models can provide labels for stable grasps, here we need humans to identify how an object can be grasped for multiple tasks. There has been a lot of recent work in this area, including [11, 12, 13]. Brahmbhatt et al. [11] used thermal imaging in a curated setup to study human grasping contacts on 50 3D printed objects for two tasks. Fang et al. [13] proposed to jointly learn a task-oriented grasping network and manipulation policy in simulation with reinforcement learning and demonstrated the

framework on two-goal tasks with two object categories. Liu et al. [12] proposed a data-driven approach to learning the complex relationships between grasps, objects, tasks, and broadened semantic contexts. However, their approach required pixel-wise affordance segmentation [14] for a small set of known object categories, which is challenging to generalize and get supervision for. Despite this progress in learning from human grasping, there are still significant gaps, both from a data and methods perspective. On the data side, existing datasets are limited in terms of the number of object instances, but especially in the number of tasks and object classes collected. Yet, even if we scale the datasets, it is unclear if current approaches will generalize to new object categories and tasks in the real world. We tackle both problems: first, we collect a dataset that is diverse both in terms of objects and tasks and an order of magnitude larger than previous datasets. Second, we exploit the semantic knowledge of objects and tasks to present a system that can generalize to new object instances, classes, and new tasks. To the best of our knowledge, this paper is one of the first efforts in demonstrating robust generalization in task-oriented grasping, especially with semantic knowledge.

More specifically, our first key contribution of this work is the collection of a large-scale dataset which we call TaskGrasp. We increase the number of real objects from the current best of 50 in prior works [11] to 191, and collect RGB-D point cloud observations and object-centric 6-DOF grasps for the task-oriented grasping problem. We also scale the number of object classes from 40 [11] to 75 and resolve each of these to the standard WordNet ontology [15]. And perhaps most importantly, we scale the number of tasks from $2-7$ in prior works [12, 11, 13] to 56. This expanded dataset both gives a better benchmark for task-oriented grasping and allows us to study generalization by expanding the number of object categories and tasks. TaskGrasp will be publicly released upon publication.

In order to generalize to a new object or task, we need to have some prior semantics about it. For instance, if we knew that mugs and bowls were both containers, we might infer that we should apply the scoop action in a similar way. To this end, and for our second main contribution, we propose a method, called GCNGrasp, that incorporates semantic knowledge into the end-to-end learning of task-oriented grasping from object point clouds. In particular, we use a Graph Convolutional Network (GCN) [16] to reason about a knowledge graph that encodes relations between objects and tasks, and further leverage word embeddings trained on large-scale language tasks to provide additional prior information. Our GCNGrasp model shows a significant improvement of 12% and 3.5% on held-out tasks and object categories, respectively, compared to baselines which do not incorporate semantics. We also show that our method and dataset are applicable for actual robots by executing task-oriented stable grasps on a 7-DOF Sawyer Robot on unknown objects.

2 Related Work

Task-Oriented Grasping: Prior work in Task-Oriented Grasping can be grouped into analytic methods, data-driven approaches using object state information, and frameworks learning from observations. Early work in analytic grasping proposed task wrench spaces with task-oriented grasp quality metrics [17]. Data-driven approaches have been proposed to improve generalization, though a large body of work has relied on object state information. Song et al. [18] used generative Bayesian Networks to model the relations between objects, grasps and tasks; Antanas et al. [19] and Ardón et al. [20] leveraged probabilistic logic languages to reason about grasp regions affording different tasks through semantic relations. However, both methods require grounding geometric information about objects to semantic representations and can only reason about semantic knowledge alone. A related line of work has used object parts and affordance detection [14, 21, 22, 23]. Do et al. [14] leveraged the affordances of object parts to define the correspondences between affordances and grasp types (e.g., rim grasp for parts with contain or scoop affordance). Detry et al. [21] trained a separate affordance detection model using synthetic data to detect suitable grasp regions for each task. While we do not provide explicit supervision for object affordance, we demonstrate that our model achieves an implicit understanding.

More recent works have learned task-oriented grasping from just RGB-D observations of objects. Dang and Allen [24] proposed an example-based approach which learns task-oriented grasps by storing visual and tactile data of grasps. Hjelm et al. [25] proposed a discriminative model based on visual features of objects. Jang et al. [26] proposed an end-to-end learning method of grasping objects from specific categories in a bin. To accelerate learning from observations, there have been efforts in scaling datasets as discussed previously [11, 12, 13]. The computer vision community has also focused on annotating datasets for inferring human grasp pose estimation from visual data

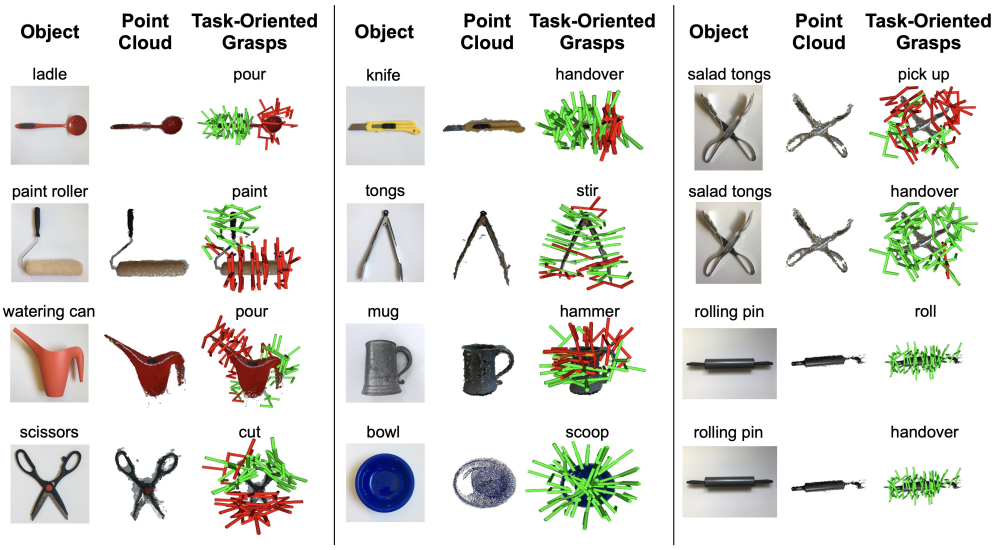

Figure 1: Example point clouds and grasps from our TaskGrasp dataset. Column 7-9 shows how grasps vary with tasks for a salad tongs (with higher diversity) and a rolling pin (with lower diversity). Green and Red means successful and incorrect task-oriented grasps respectively.

[27, 28, 29] with the aim that it could be adapted to robotic grasping with kinematic retargeting. In this work, we propose an expanded dataset in terms of the number of object categories and tasks to study generalization. We also present a unified framework that jointly learns from semantic knowledge and geometric observations.

Semantic Knowledge in Vision: The use of knowledge and knowledge graphs for visual reasoning has been well studied. Word embeddings from language has been used extensively [30]. Class hierarchies, such as WordNet [15], have often been used to aid in image recognition [31]. More generally, knowledge graphs have found extensive use in visual classification and detection [32], as well as zero-shot classification [33]. We draw on many of the ideas from these works in Computer Vision, especially those related to word embeddings and graphs, and apply them to a robotics task and to 3D point cloud data.

Semantic Knowledge in Robotics: In robotics, semantic knowledge has been used to help robots adapt to diverse and changing environments by providing abstractions that generalize across similar situations. Large-scale robotic knowledge bases, such as KnowRob [34], RoboBrain [35], and RoboCSE [36], aimed to provide robots with extensive knowledge about objects, spaces, tasks, actions, and agents. Other methods leveraged more specific knowledge in a variety of robotic tasks, such as affordance learning [37] and visual-semantic navigation [38]. Similar to Antanas et al. [19] and Ardón et al. [20], we reason about semantic knowledge for task-oriented grasping, but we leverage semantic knowledge for generalization to novel object classes and tasks.

3 Dataset

In this section we describe our dataset: TaskGrasp, specifically its properties, collection and annotation methodology. As shown in Table 1, TaskGrasp is the largest and most diverse dataset for task-oriented grasping to date with respect to number of objects, categories and tasks.

TaskGrasp contains 191 individual household and kitchen objects comprising 75 distinct object categories and varying in size, geometry, material, and visual appearance. Figure 2 shows the class of each object and its proportion in the dataset. We collect RGB-D pointclouds for each object, and automatically annotate $250K$ stable grasps. We also curate a list of 56 everyday tasks that impose different semantic constraints on grasping and annotate for each grasp whether that grasp is appropriate for each particular task.

Table 1: Comparing recent Task-Oriented Grasping Datasets

	ContactDB [11]	SG14000 [12]	TOG-Net [13]	TaskGrasp (Ours)
Semantic Knowledge	✗	✗	✗	✓
Object Categories	40	5	2	75
Objects	50	44	18K (synthetic)	191
Tasks	2	7	2	56
Grasps	3750	14K	1.5M	250K
Grasp Type	Contact Map	$SE(3)$	Planar	$SE(3)$

3.1 Data Acquisition on a Robot

After selecting our 191 objects by browsing various homegoods stores, we scan the objects to acquire their point clouds. A Realsense D415 eye-in-hand camera mounted on a LoCoBot [39] is used for 3D scanning. The object is placed on a transparent mount in front of the robot, which is commanded to different poses along the object approach direction to capture point clouds from multiple viewpoints. This setup helps to capture more of the object geometry under self-occlusion, which in turn increases the coverage of grasp samples. The multi-view observations are registered using robot kinematics and further refined with the iterative closest point algorithm. After table plane segmentation, 600 object-centric stable grasps are then sampled [40] from the object point cloud. 25 grasps are selected with farthest point sampling (to maximize grasp coverage) for annotation. These grasps are chosen as a representative, albeit limited, grasp set for the object to trade off between dataset size and budget.

3.2 Data Annotation by Crowdsourcing

We use Amazon Mechanical Turk (AMT) to crowdsource labels for the 250K stable grasps. Instead of exhaustively labelling each task-object combination ($\sim 10K$) , we reduce the annotation cost with a two-stage procedure. We use the insight that the pre-condition for a task-oriented grasp is that the object has to be capable of the task in the first place. First, we gather labels for whether a task is suitable for each object. Second, for this filtered subset of task-object combinations, we collect labels for the 25 task-oriented grasps per object. To ensure annotation quality, we assign each labeling task to three annotators and use gold standard questions (questions that we know the answers to) to filter annotators with low accuracy. For both stages, we take a majority vote between the annotators. We measure agreement with Randolph's free-marginal multirater kappa [41]. Kappa values for the two stages are 0.65 and 0.62 respectively (0.0 meaning agreement equal to chance, and 1.0 indicating perfect agreement above chance), which suggests good agreement between annotators.

3.3 Analysis

In Figure 1 we show prototypical examples from TaskGrasp. We provide additional examples in the supplementary materials.

Diversity of Grasps: As a result of the large number of objects and tasks, TaskGrasp contains a wide variety of task-oriented grasps. On average, each object is suitable for 7 tasks. As shown in Figure 1, these tasks involve both prototypical (a ladle for pouring) and creative use of objects (tongs for stirring), imposing drastically different semantic constraints on grasping. These examples also demonstrate the complex geometries presented in real world objects, which pose another challenge for generalization.

We also quantitatively measure grasp diversity by analyzing the effect of tasks on grasps. Since different tasks provide different labels for the same set of stable grasps on each object, we compute Randolph's kappa [41] on these labels as a measure of agreement between tasks, i.e., how likely grasps for one task (e.g., stir) agree with grasps for another task (e.g., cut). Ranging from 0.19 to 0.93, kappa values of the objects suggest that the effect of tasks vary greatly for different objects. Column 7-9 in Figure 1 show how grasps vary with tasks for a salad tongs with a kappa value of 0.38 and a rolling pin with kappa value of 0.97. In TaskGrasp, 25% of the objects have kappa values lower than 0.5 and these objects require significantly different grasps for different tasks.

Semantic Knowledge of Objects and Tasks: We also provide semantic knowledge about objects and tasks in the dataset. Objects are manually mapped to WordNet synsets [15] which represent a semantic hierarchy, as shown in Figure 2. Each of the 75 leaf synsets in the hierarchy represents a distinct object class and is linked to 2.5 objects on average. Building on the hypernym paths from

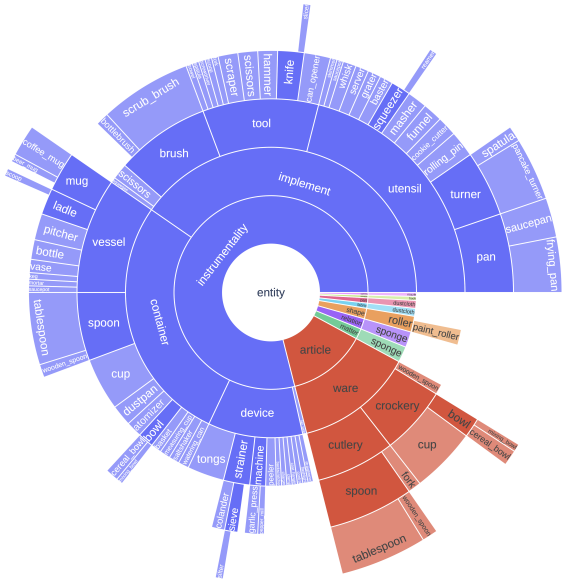

Figure 2: Semantic hierarchy of objects. Each level of the hierarchy is represented by one ring with the innermost circle as the root of the hierarchy. The angle of each segment is proportional to the number of objects.

WordNet, the semantic hierarchy includes a rich set of object concepts interlinked by "Is-A" relations. This provides useful semantic knowledge for task-oriented grasping as objects in the same subtree of the hierarchy often share similar functionalities or geometric properties. For example, mug, ladle, and bottle are in the vessel subtree and can all be used to hold liquid. In addition, we connect a task to an object class through "Used-For" relations if any object in the class is considered suitable for the task from the first stage of our crowdsourcing. We provide a thorough breakdown of object counts, class hierarchies and used-for relations in the supplementary materials.

4 Task-Oriented Grasping with Semantic Knowledge

We consider the problem of generating grasps for task-oriented grasping given the object point cloud and task constraints. Specifically, we want to estimate the grasp distribution $P(G^*|X, \mathcal{T})$, where X is the point cloud input, $\mathcal{T}$ are the constraints imposed by goal tasks, and G^* is the space of successful grasps. Following convention in related work [8, 7], we represent grasps $g \in G^*$ as the grasp pose $(R, T) \in SE(3)$ of a parallel-jaw gripper with its fingers open which when closing will lead to a stable grasp. We further factorize the estimation of $P(G^*|X, \mathcal{T})$ into 1) task-agnostic grasp sampling $P(G^*|X)$ and 2) task-oriented grasp evaluation $P(S|X, \mathcal{T}, g)$. The primary benefit of this factorization is that it allows us to leverage prior work in stable grasp generation.

In this section, we describe our method (GCNGrasp) for Task-Oriented grasping. Our method is composed of: (1) a Shape Encoder built on a PointNet++ architecture [42] to encode the object point cloud, (2) a Graph Convolutional Network [16] which takes the encoded shape as input as well as a knowledge graph $\mathcal{G}$ encoding the semantic relationships between object categories, tasks and hierarchies and (3) a Grasp Evaluator which outputs the final grasp prediction. See Figure 3.

Grasp and Object Shape Encoder: Our input observations are object point clouds and we want to reason about $SE(3)$ grasps. Qi et al. [42] proposed the PointNet++ architecture to efficiently represent 3D data which we use to learn a representation for the object point cloud and 6-DOF grasp poses. The grasp g is defined in the object frame and six control points are selected on the gripper to form a gripper point cloud X_g. Similar to Mousavian et al. [7], X_g is concatenated with the object point cloud X with an extra latent indicator vector to represent whether a point is part of the gripper or the object. The PointNet layer reasons about the relative spatial information between the grasp and the object and outputs a geometric embedding vector.

Graph Convolutional Network: We use the standard Graph Convolutional Network (GCN) model from Kipf and Welling [16], which is a neural network structured on the shape of the input graph. By

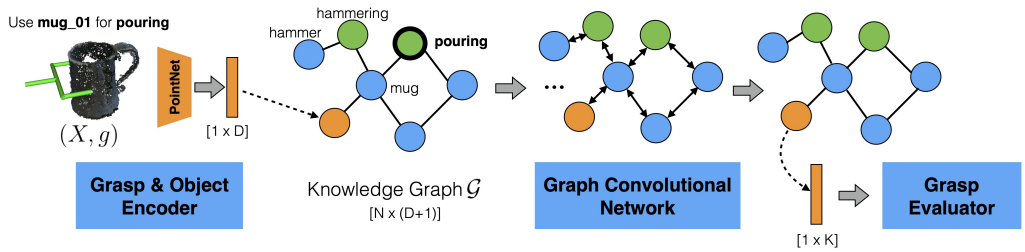

Figure 3: Overview of our Task-Oriented grasping framework using semantic knowledge graphs.

structuring a neural network to pass information between adjacent nodes, we use the input graph to correctly reason about the relationship between the object classes and the target task.

The first input of a GCN is the graph itself $\mathcal{G} = (V, E)$. In our application, we use a knowledge graph constructed from two sources: the task-object class relationships in our dataset and the object hierarchy from WordNet [15]. The graph is represented as a binary adjacency matrix A, which we normalize to obtain $\hat{A}$ following [16]. The next input to each node of the GCN is a $D-$dimensional embedding vector. The target tasks are specified using an extra indicator latent variable that is concatenated with this embedding to get the vector of size $D + 1$. The embedding vectors are stacked across nodes to get the input matrix $\mathcal{X} \in \mathcal{R}^{|V| \times (D+1)}$. We initialize the matrix with the word embeddings corresponding to each concept in the knowledge graph (e.g. "mug"). We use ConceptNet numberbatch [43] for the word embeddings. The grasp and shape encoder nodes are added online to the existing knowledge graph $\mathcal{G}$ by connecting edges to the corresponding object class nodes.

The output of the GCN are K-dimensional embeddings for each node $\mathcal{Z} \in \mathcal{R}^{|V| \times K}$. The node embeddings are propagated to their neighbours using message passing in each convolutional layer:

$$H^{(l+1)} = \sigma(\hat{A} H^{(l)} W^{(l)}) \tag{1}$$

where σ is the ReLU activation function, $H^{(0)} = \mathcal{X}$ and $H^{(L)} = \mathcal{Z}$ where L is the number of layers.

Grasp Evaluator: After the GCN, we are left with a node-level embedding $\mathcal{Z}$. We use the embedding corresponding to the grasp node z_g to train the final grasp evaluator $P(S|z_g)$, where S is the grasp score. This module has three fully connected layers with K units and a final sigmoid layer. The entire model, including the shape encoder, GCN and grasp evaluator, is optimized with ADAM using a binary cross entropy loss.

Implementation Details: The point clouds were downsampled to 4096 points during training. They were also mean centered and unit-scaled. The PointNet module consists of three set abstraction layers and the number of points sampled are 512, 128 and all points. The set abstraction layers are followed by three fully connected layers with sizes [1024, 512, D]. Each set abstraction layer has three fully connected layers to learn features. The point clouds were perturbed with random rotations, jitter and dropout for data augmentation and to build robustness when testing on novel objects in unknown poses. We choose D=300 and K=128, and L=6 as the parameters for our GCN network.

5 Experimental Evaluation

5.1 Zero-Shot Generalization

A central goal of both our dataset and our method is to show that we can learn task-oriented grasping models which generalize to novel objects, classes and tasks. In an ideal robotics system, we should be able to correctly grasp a novel object from a novel object class, or even grasp for a novel task. To test this, we measure our system and baselines in three different held-out test settings: held-out object instances, held-out object categories, and held-out tasks.

These held-out settings are of increasing difficulty in terms of zero-shot generalization. For each setting, we perform k-fold cross validation (k=4), such that each category (a task, object class, or object instance, based on the setting) will be held out exactly once. In each fold, grasps from 25% of the categories will be used for testing while remaining grasps will be used for training and validation.

Table 2: Results on TaskGrasp

(a) Object Instance Generalization

Model	Test Performance (mAP)		
	Instances	Classes	Tasks
Random	59.75	60.28	54.76
SGN [12]	78.51	75.08	68.8
SGN + word embedding	79.74	77.91	**74.36**
GCNGrasp (ours)	**80.25**	**77.94**	73.71

(b) Object Class Generalization

Model	Test Performance (mAP)		
	Instances	Classes	Tasks
Random	59.32	58.73	52.27
SGN [12]	74.2	72.95	62.55
SGN + word embedding	77.21	75.51	**63.73**
GCNGrasp (ours)	**78.81**	**76.57**	57.36

(c) Task Generalization

Model	Test Performance (mAP)		
	Instances	Classes	Tasks
Random	59.06	58.24	52.37
SGN [12]	75.17	71.59	63.35
SGN + word embedding	78.06	74.49	70.55
GCNGrasp (ours)	**81.5**	**79.56**	**76.01**

Table 3: Ablation on Semantic Knowledge

Model	Graph		Held-out Setting		
	Nodes	Edges	Task	Class	Instance
GCN + tasks + WordNet	345	989	76.01	**76.57**	80.25
GCN + tasks	131	693	**77.54**	75.86	**81.46**
GCN + WordNet	155	106	71.77	70	78.66

In all experiments, we only evaluate tasks that are valid for a given input object class. This makes sense from an evaluation perspective as it separates the problem of predicting applicable tasks for objects from task-driven grasping. It also makes the comparison to methods using object-task information fair since the models do not have to decide whether the object-task pair is valid.

Evaluation Metrics: Since k-fold cross validation in any held-out setting will evaluate all grasps in the dataset, we can compute Average Precision (AP) scores for any category, i.e., any object instance, object class, or task. We then compute an mAP averaged over object instances, mAP averaged over object classes, and mAP over tasks. We show all three metrics for each of our three settings in Tables 2a,2b,2c, but emphasize the mAP metric that corresponds to what category is being held out.

Baselines: We compare our approach to the following models: (1) Random, which represents grasping strategies that focus on grasp stability and ignore task constraints. Results are averaged over five random seeds. (2) Semantic Grasp Network (SGN), which learns to reason about context of grasps (e.g., constraints imposed by objects and tasks) from data. This model is adapted from [12], with the difference that the input to the model is replaced with geometric embedding from our shape encoder and word embeddings of the task and the object class. Note that embeddings of tasks and object classes are both learned from training data. (3) SGN + *word embedding*, which uses ConceptNet [43] numberbatch as pretrained word embeddings for object classes and tasks.

5.2 Analysis

First, to get context for our results in Table 2, we see that random grasp prediction achieves approximately 50-60% accuracy, establishing a floor for the other methods. Because the number of positive and negative grasps in the dataset is about even, random guessing is able to achieve a seemingly high mAP. In a dataset with more negatives we would expect this number to be much lower.

Our method outperforms baselines in all three settings. This confirms that our method can effectively leverage the knowledge graph to generalize to novel object instances, object classes, and tasks. SGN + *word embedding* also outperforms SGN, suggesting that implicit distributional knowledge provides a prior that is useful for generalization. Despite the benefit of distributional knowledge, it still only represents semantic similarities between concepts. In contrast, the knowledge graph directly stores relations between the relevant objects and tasks, and exploiting this additional structured knowledge allows our model to achieve better zero-shot generalization than SGN + *word embedding*.

When comparing our method with SGN and SGN + *word embedding*, we observe increasingly larger margins in performance from the held-out instance to the held-out class setting. As objects from different classes have more variance in terms of geometric and visual features than objects from the same class, semantic knowledge becomes more important in unifying these objects. The difference in performance between our method and these two baselines on the held-out task setting reached 12.6%

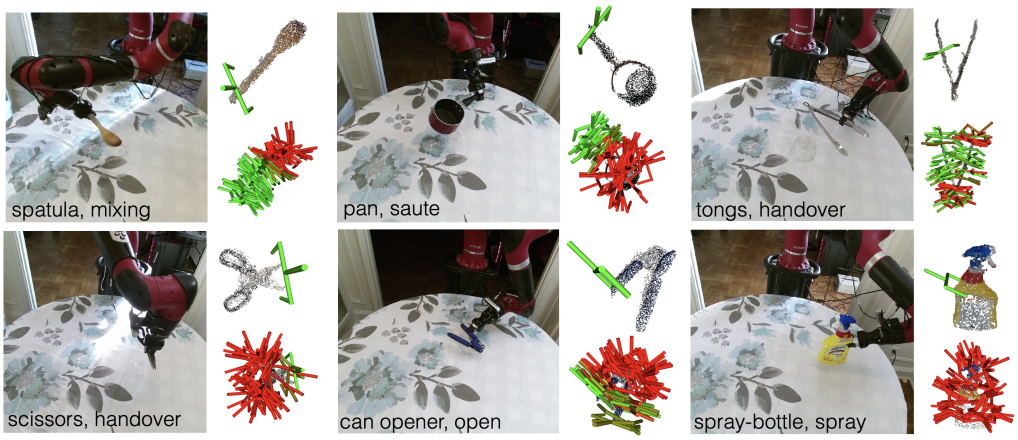

Figure 4: Robot executions of example task-oriented grasps on unknown objects. For each execution, the top 3D visualization shows the grasp that was executed (which had the best evaluator score) and the bottom shows all the stable grasp candidates colored by their scores (green is higher).

and 5.46% respectively, affirming that semantic knowledge is especially crucial for generalizing disparate constraints from different tasks.

Ablations on Knowledge Graph: We investigated how performance is affected by changing the knowledge graph used in our model. Specifically, we compared the default knowledge graph with a knowledge graph containing only the semantic hierarchy of objects and a knowledge graph containing only the relations between object classes and tasks. The results from the three held-out settings are summarized in Table 3 (we only show the mAP metrics corresponding to the held-out category). From these results, we observe that edges between object classes and tasks were the most important knowledge for generalizing to novel tasks and instances, though every task we tested was valid for the target object class. This suggests that knowledge about which objects could generally be used for which tasks provide important information for discovering similarities between tasks. In the held-out object class setting, additional knowledge from the object hierarchy helped generalize to novel object classes by associating known classes and novel classes through the WordNet hierarchy.

5.3 Real Robot Evaluation

We run experiments to show that our approach and dataset transfer to a real robot. We test our approach on novel objects not from the dataset and in unknown poses. We place each object (without clutter) on a table in front of the robot. After table plane segmentation to obtain the object point cloud, 600 stable grasps are sampled and 50 candidates are selected using farthest point sampling for evaluation. We evaluate the grasps on our best performing GCNGrasp model from the held-out task ablations (Table 2). Our hardware setup comprises of a 7-DOF Sawyer Robot with a 2-fingered Robotiq gripper and a Intel Realsense D415 RGB-D camera mounted on the gripper wrist. Inference for the 50 grasps takes around $3s$ on a desktop with an NVIDIA GTX 1080 Ti GPU and the grasp with the best score is executed on the robot. Fig 4 shows the executed task-oriented grasps on unknown objects. Even though our dataset objects were collected only in one canonical pose, our approach is able to generalize to new grasps and in unknown poses due to data augmentation during training. Based on the grasp evaluator scores from Fig 4, our model is also able to interpolate between modes in the continuous $SE(3)$ space to reason about task-oriented grasping. One failure mode of our work is that it does not generalize to categories (like the spray bottle in Fig 4 in the bottom right) with limited training data. A future work is to balance the dataset in terms of object categories.

6 Conclusion

We present the TaskGrasp dataset to study generalization in Task-Oriented grasping. The dataset is diverse and an order of magnitude larger than previous datasets. We also present a framework for jointly learning from geometric observations and semantic knowledge to generalize to new object instances, classes and even new tasks. Future work could explore recent techniques in automatic knowledge graph generation [44] for grasping tasks. While we collected real point cloud data of

objects, we could convert the point clouds to meshes or acquire shape models from large online repositories to use in physics simulators. This could expand the scope of the dataset for sim2real transfer and to even learn task policies in simulation conditioned on the task-oriented grasps like in prior work [13].

References

[1] 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.

[2] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *Conference on Robot Learning*, 2018.

[3] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[4] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[5] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *International Symposium on Experimental Robotics (ISER)*, 2016.

[6] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox. 6-dof grasping for target-driven object manipulation in clutter. *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[7] A. Mousavian, C. Eppner, and D. Fox. 6-DOF GraspNet: Variational grasp generation for object manipulation. *International Conference on Computer Vision*, 2019.

[8] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13-14):1455–1473, 2017.

[9] A. Gupta, A. Murali, D. Gandhi, and L. Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. *Neural Information Processing Systems (NeurIPS)*, 2018.

[10] D. Wang, D. Tseng, P. Li, Y. Jiang, M. Guo, M. Danielczuk, J. Mahler, J. Ichnowski, and K. Goldberg. Adversarial grasp objects. In *Conference on Automation Science and Engineering*. IEEE, 2019.

[11] S. Brahmbhatt, C. Ham, C. Kemp, and J. Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] W. Liu, A. Daruna, and S. Chernova. Cage: Context-aware grasping engine. *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[13] 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. *Robotics Science and Systems*, 2018.

[14] T.-T. Do, A. Nguyen, and I. Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 ICRA*, pages 1–5. IEEE, 2018.

[15] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[16] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*, 2017.

[17] C. Borst, M. Fischer, and G. Hirzinger. Grasp planning: How to choose a suitable task wrench space. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 1, pages 319–325. IEEE, 2004.

[18] D. Song, K. Huebner, V. Kyrki, and D. Kragic. Learning task constraints for robot grasping using graphical models. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1579–1585. IEEE, 2010.

[19] 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*, pages 1–26, 2018.

[20] P. Ardón, È. Pairet, R. P. Petrick, S. Ramamoorthy, and K. S. Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019.

[21] 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)*, pages 3266–3273. IEEE, 2017.

[22] S. R. Lakani, A. J. Rodríguez-Sánchez, and J. Piater. Exercising affordances of objects: A part-based approach. *IEEE Robotics and Automation Letters*, 3(4):3465–3472, 2018.

[23] 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)*, pages 91–98. IEEE, 2017.

[24] H. Dang and P. K. Allen. Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1311–1317. IEEE, 2012.

[25] M. Hjelm, C. H. Ek, R. Detry, and D. Kragic. Learning human priors for task-constrained grasping. In *International Conference on Computer Vision Systems*, pages 207–217. Springer, 2015.

[26] E. Jang, S. Vijayanarasimhan, P. Pastor, J. Ibarz, and S. Levine. End-to-end learning of semantic grasping. *Proceedings of Machine Learning Research*, 78:119–132, 2017.

[27] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] D.-A. Huang, M. Ma, W.-C. Ma, and K. Kitani. How do we use our hands? discovering a diverse set of common grasps. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[29] M. Kokic, D. Kragic, and J. Bohg. Learning task-oriented grasping from human activity datasets. In *IEEE Robotics and Automation Letters*. IEEE, 2020.

[30] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[31] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014.

[32] K. Marino, R. Salakhutdinov, and A. Gupta. The more you know: Using knowledge graphs for image classification. In *CVPR*, 2017.

[33] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018.

[34] M. Tenorth and M. Beetz. Representations for robot knowledge in the knowrob framework. *Artificial Intelligence*, 247:151–169, 2017.

[35] A. Saxena, A. Jain, O. Sener, A. Jami, D. K. Misra, and H. S. Koppula. Robobrain: Large-scale knowledge engine for robots. *arXiv preprint arXiv:1412.0691*, 2014.

[36] A. Daruna, W. Liu, Z. Kira, and S. Chetnova. Robocse: Robot common sense embedding. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9777–9783. IEEE, 2019.

[37] B. Moldovan, P. Moreno, M. Van Otterlo, J. Santos-Victor, and L. De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. In *2012 ICRA*, pages 4373–4378. IEEE, 2012.

[38] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*, 2018.

[39] A. Murali, T. Chen, K. V. Alwala, D. Gandhi, L. Pinto, S. Gupta, and A. Gupta. Pyrobot: An open-source robotics framework for research and benchmarking. 2019. URL https://arxiv.org/abs/1906.08236.

[40] A. ten Pas and R. Platt. Using geometry to detect grasp poses in 3d point clouds. In *Robotics Research*, pages 307–324. Springer, 2018.

[41] J. J. Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.

[42] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[43] R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. pages 4444–4451, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

[44] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi. Comet: Commonsense transformers for automatic knowledge graph construction. 2019. URL https://arxiv.org/abs/1906.05317.