I. INTRODUCTION

Grasping is a fundamental robotic task, but is challenging in practice due to imperfections in perception and control. Most commonly, grasp planning involves generating gripper pose configurations (3D position and orientation) that maximize a grasp quality metric on a target object in order to find a stable grasp. There are several factors that affect grasp stability, including object geometry, material, gripper contacts, surface friction, mass distribution, amongst several others [1], [2]. Most traditional approaches to grasping assume a separate perception system that can perfectly [1], or with some uncertainty [3], infer object information such as pose and shape. This is followed by physics-based grasp analysis [1], [4] or nearest-neighbour lookup on a database of pre-computed grasps [5]. These methods are slow [6], prone to perception error and do not generalize to novel objects.

Grasp synthesis is much harder in clutter, such as the example in Fig 1. The target object has to be grasped without any unwanted collisions with surrounding objects or the environment. In a real world application, a personal robot might be commanded to grasp a specific beverage from a narrow kitchen cabinet packed with other items. Grasps sampled agnostic of the clutter could end up in collision with the environment. Even if the gripper pre-shape is not in collision, it may be challenging to plan a collision-free and kinematically feasible path for the manipulator to achieve the gripper configuration. One would have to generate a diverse set of grasps since not all the grasps will be kinematically feasible to execute in the environment. Most model-based approaches in the grasping and task and motion planning literature assume perfect object knowledge or use an occupancy-grid representation for collision checking, which may not be reliable or practical in real-world settings [1], [7]–[9].

A large part of the difficulty lies in perception. In clutter, large and important parts of object geometry are occluded by other objects. Traditional shape matching techniques will find it extremely challenging to operate in such conditions, even when object geometry is known. In addition, getting quality 3D information is challenging and previous methods resort to using high quality depth sensors [10] or using multiple views [11], which would require observation-gathering exploratory movements impossible in confined spaces. This limits the deployment of these systems outside of controlled environments.

Recent works have explored data-driven methods for grasping unknown objects [10]–[16]. However, they mainly focus on the limited setting of planar-grasping and bin-picking. Some recent methods tackle the more difficult problem of generating grasps in $SE(3)$ from 2D (image) [17], 2.5D (depth, multi-view) [11], [18], [19] and 3D (point cloud) [20]–[22] data. These works primarily consider the problem from an object-centric perspective or in bin-picking settings. We consider the problem of 6-DOF grasp generation in structured clutter using a learning-based approach. Our method uses instance segmentation and point cloud observation from just a single view. We follow a cascaded approach to grasp generation in clutter, first reasoning about grasps at an object level and then checking the cluttered environment for collisions. We use a learned collision checker, which evaluates grasps for collisions from just raw partial point cloud observations and works under varying degrees of occlusion. Specifically, we present the following contributions:

- A learning-based approach for 6-DOF grasp synthesis for novel objects in structured clutter, which uses a learned collision checker conditioned on the gripper information
and on the raw point cloud of the scene.

- Show that our approach, trained only with synthetic data, achieves a grasp accuracy of 80.3% with 23 real-world test objects in clutter. It also outperforms a clutter-agnostic baseline approach of 6-DOF GraspNet [20] with state-of-the-art instance segmentation [23] by 17.6%.
- Demonstrate an application of our approach in moving blocking objects away out of the way to grasp a target object that is initially occluded and impossible to grasp.

II. RELATED WORK

Grasping is a widely studied field in robotics ([11], [2], [12]). In the following we will focus our comparison on existing approaches that are data-driven and the aspects in which they differ from the proposed method.

Grasping in clutter vs. isolated objects: Among learning-based methods for grasping a significant amount focuses on dealing with isolated objects on planar surfaces ([18], [20], [24]–[26]). Our approach specifically tackles the problem of grasping objects from a cluster of multiple objects. This problem is significantly harder since the accessible workspace around the target object is severely limited, occlusions are more likely to hamper perception and predicting outcomes might be more difficult due to contact interactions between objects. Although multiple learning-based approaches for dealing with grasping in clutter exist ([11], [13], [14], [27]) we will show in the following that they differ from our approach in multiple aspects.

Bin-picking vs. structured clutter: Most learning-based grasping approaches for clutter deal with rather small and light objects that are spread in a bin ([13], [14], [27], [28]). In contrast our approach focuses on structured clutter. We define structured clutter as packed configurations of mostly larger, heavier objects. Examples include kitchen cupboards or supermarket shelves. Compared to the bin-picking setup successful grasps are more sparsely distributed in structured clutter scenarios. Collisions and unintended contact is often more catastrophic since objects have fewer stable equilibria when they are not located on a pile. Since avoiding collision becomes more important, structured clutter is more prominent in evaluations of model-based task-and-motion-planning. Our approach explicitly predicts grasp configurations that are in collision and can do so despite occlusions.

Planar vs. spatial grasping in clutter: Many learning-based grasp algorithms for clutter are limited to planar grasps, representing them as oriented rectangles or pixels in the image ([13], [14], [29], [30]). As a result, grasps lack diversity and picking up an object might be impossible given additional task or arm constraints. This limitation is less problematic in bin-picking scenarios where objects are small and light. In structured clutter, spatial grasping is unavoidable, otherwise pre-grasp manipulations are needed [31]. Those learning-based approaches that plan full grasp poses are either based on hand-crafted features ([32]–[34]) or have non-learned components [11]. Our approach uses a learned grasp sampler that predicts the full 6D grasp pose and accounts for unseen parts due to occlusions.

Model-based vs. model-free: A lot of planning approaches exist that tackle scenarios of grasping in structured clutter ([7], [31], [35]–[37]). These approaches rely on full observability and prior object knowledge. In contrast, our method does not require any object models and poses; grasps are planned based on raw depth images. In that regard, it is similar to other data-driven methods for clutter ([11], [13], [14], [29], [30]) but differs from techniques that use hand-engineered features ([32]–[34], [38]).

Target-agnostic vs. target-driven: Few grasping approaches for clutter focus on grasping specific objects ([39], [40]). Our method is target-driven as it uses instance segmentation [23] to match grasps with objects.

III. 6-DOF GRASP SYNTHESIS FOR OBJECTS IN CLUTTER

We consider the problem of generating 6-DOF grasps for unknown objects in clutter. The input to our approach is the depth image of the scene and a binary mask indicating the target object. We define a grasp as the pose \( \in SE(3) \) of an opened parallel-yaw gripper that results in a robust grasp when closing its fingers. In particular, we aim to estimate the posterior grasp distribution \( P(G^*|X) \), where \( X \) is the point cloud observation and \( G^* \) successful grasps of the object.

The distribution of successful grasps is complex, multimodal and discontinuous. The number of modes for a new object is not known a-priori and is determined by the geometry, size, and physics of the object. Additionally, small perturbations of a robust grasp could lead to failure due to collision or slippage from poor contact. Finally, cluttered scenes limit the robot workspace significantly. Although a part of an object might be visible it could be impossible to grasp if the gripper itself is a large object (such as the Franka Panda robot we use in our experiments) that leads to collisions with surrounding objects.

A. Overview of Approach

Our cascaded grasp synthesis approach factors the estimation of \( P(G^*|X) \) by separately learning the grasp distribution for a single, isolated object \( P(G^*|X) \) and a discriminative model \( P(C|X,g) \) which we call CollisionNet that captures collisions \( C \) between gripper at pose \( g \) and clutter observed as \( X \). The advantage of this factorization is twofold. First, it allows us to build upon prior work [20] which successfully infers 6-DOF grasp poses for single, unknown objects. Second, by explicitly disentangling the reasons for grasp success, i.e., the geometry of the target object and a collision-free/reachable gripper pose, we can reason beyond simple pick operations. As shown in a qualitative experiment in Sec. IV-C we can use our approach to infer which object to remove from a scene to maximize grasp success of the target object.

Fig. 2 shows an overview of our approach. During inference, a target object can be selected based on a state-of-the-art segmentation algorithm [23]. Given this selection we infer possible successful grasps for the object ignoring clutter, and combine it with the collision results provided by CollisionNet.
In the following two sections, we will present both of these models. Note that our particular design decisions are based on comparisons with alternative formulations. In Sec. IV-A we will show how our approach outperforms variants that do not distinguish between grasp failures due to collisions and target geometry, or use non-learned components.

B. 6-DOF Grasp Synthesis for Isolated Objects

We first want to learn a generative model for the grasps given the point cloud observation of the cluttered scene. Though this generative model is learned from a reference set of positive grasps, it is not completely perfect due to several reasons. As a result, we follow the approach presented in [20] to have a second module to evaluate and further improve these generated grasps. Conditioned on the point cloud and grasp, the evaluator predicts a quality score for the grasp. This information could also be used to incrementally refine the grasp. We also explore the importance of object instance information in all stages of the 6-DOF grasping pipeline, from grasp generation to evaluation.

Variational Grasp Sampling: The grasp sampler is a conditional Variational Autoencoder [41] and is a deterministic function that predicts the grasp given a point cloud $X$ and a latent variable $z$. $P(z) = \mathcal{N}(0, I)$ is a known probability density function of the latent space. The likelihood of the grasps can be written as such:

$$P(G|X) = \int P(G|X, z)P(z)dz \quad (1)$$

Optimizing Eqn 1 is intractable as we need to integrate over all the values of the latent space [41]. To make things tractable, an encoder $Q(z|X, g)$ is used to map each pair of point cloud $X$ and grasp $g$ to the latent space $z$ while the decoder reconstructs the grasp given the sampled $z$. The encoder and decoder are jointly trained to minimize the reconstruction loss $\mathcal{L}(\hat{g}, g)$ between the ground truth grasps $g \in G^+$ and predicted grasps $\hat{g}$, with the KL-divergence penalty between the distribution $Q$ and the normal distribution $\mathcal{N}(0, I)$:

$$\mathcal{L}_{VAE} = \sum_{z \sim Q, g \sim G^+} \mathcal{L}(\hat{g}, g) - \alpha D_{KL}[Q(z|X, g), \mathcal{N}(0, I)] \quad (2)$$

Note that the input to the VAE is the point cloud of the target object segmented from the scene with instance mask.

To combine the orientation and translation loss, we define the reconstruction loss as $\mathcal{L}(\hat{g}, g) = \frac{1}{n} \sum ||\mathcal{T}(g; p) - \mathcal{T}(\hat{g}; p)||$ where $\mathcal{T}$ is the transformation of a set of predefined points $p$ on the robot gripper. During inference, the encoder $Q$ is discarded and latent values are sampled from $\mathcal{N}(0, I)$. Both the encoder and decoder are based on the PointNet++ architecture [42], where each point has a feature vector along with 3D coordinates. The features of each input point of the point cloud are concatenated to the grasp $g$ and the latent variable $z$ in the encoder and decoder respectively.

Though instance information can give a strong prior about the object, it is not perfect in practice. This is especially the case in cluttered scenarios where objects are occluded or very close to each other, resulting in noisy under and over-segmentation. When rendering the segmentation in simulation, we add random salt-and-pepper noise to the object boundaries and randomly merge partially occluded objects to neighboring ones in image space, to mimic the imperfections of instance segmentation methods on the real images.

Grasp Evaluation: Though the grasp sampler is trained with only positive grasps, it may still contain failed grasps which need to be identified and removed. We train an evaluator that predicts a grasp score $P(S|X_i, g)$, with the training data consisting of positive $G^+_S = G^+$ and negative $G^-_S = G^-$ grasps. The evaluator’s input is the point cloud observation $X_i = \mathcal{M}_i(X)$ of the target object, where $X$ is the cropped point cloud of the full scene and $\mathcal{M}_i$ is the instance mask of the object. Since the space of all possible 6-DOF grasp poses is large, it is not possible to sample all the negative grasps for training the grasp evaluator $P(S|X_i, g)$. Therefore, during training we sample from true negatives but also sample hard negative grasps by perturbing positive grasps with a small translation and orientation and choosing those that are in collision with the object or are too far from the object to grasp the object. At test time on the robot, the grasps are ranked by their evaluator scores and only those above a threshold are selected.

Grasp Refinement: A significant proportion of the grasps rejected by the evaluator are actually in close proximity to robust grasps. This insight could be exploited to perform a local search in the region of $g$ to iteratively improve the evaluator score. We concretely want to sample $\Delta g$ to increase the probability of success, i.e., $P(S|\Delta g + g, X) > P(S|g, X)$. The refinement was found using gradient descent in [20].
In practice, computing gradients is not fast. Instead, we use Metropolis-Hastings sampling where a random $\Delta g$ is sampled and with probability of $\frac{P(S|g+\Delta g,X)}{P(S|g,X)}$ grasp $g + \Delta g$ is accepted. We observe that this sampling scheme yields similar performance to the gradient-based one while it is computationally twice as fast.

C. Collision Detection for Grasps in Clutter: CollisionNet

CollisionNet predicts a clutter-centric collision score $P(C|X,g)$ given the full scene information $X$. The training data for CollisionNet is $G_+ = \{g | g \in G_{ref}, -\Psi(g,x) \}$ and $G_- = \{g | g \in G_{ref}, \Psi(g,x) \}$. The ground truth labels are generated in simulation with a collision checker $\Psi$ assuming full state information $x$. In each batch, we used balanced sampling of grasps within the subsets of the reference set $G_{ref}$, which consists of the positive and negative sets ($G^+, G^-$), hard-negatives generated by perturbing positive grasps ($G_{nn}^+$) and grasps in free space $G_{free}$. We observed that balanced sampling improved the stability of training and generalization at test time over uniform sampling from $G^+ \cup G^-$. Similar to the grasp evaluator, the scene/object point cloud $X_i$ and gripper point cloud $X_g$ are combined into a single point cloud by using an extra indicator feature that denotes whether a point belongs to the object or to the gripper. The PointNet++ [42] architecture then uses the relative information between grasp pose $g$ and object point cloud $X$ for classifying the grasps. CollisionNet is optimized using cross entropy loss.

D. Implementation Details

Training data is generated online by arranging multiple objects randomly at their stable poses. Objects are added to the scene with rejection sampling poses to ensure they are not colliding with existing clutter. In order to generate grasps for the scenes, we combine the positive and negative grasps of each object from [20]. From each scene we take multiple 3D crops centered on the object (with some noise) along with grasps that are inside the crop. The cropped point cloud of the 3D box is down-sampled to 4096 points. All the samplers and VAEs are based on PointNet++ [42] architecture and the dimension of latent space is set to 2. During inference, object instances are segmented with [23]. The VAE sampler generates the grasps given the point cloud of the target object by sampling 200 latent values. Grasps are further refined with 25 iterations of Metropolis-Hastings. The whole inference takes 2.5s on a desktop with NVIDIA Titan XP GPU.

IV. EXPERIMENTAL EVALUATION

A. Ablation analysis and Discussion

Evaluation Metrics: Following [20], we used two metrics for evaluating the generated grasps: success rate and coverage. Success rate is the percentage of grasps that succeed grasping the object without colliding and coverage is the percentage of sampled ground truth grasps that are within $2\text{cm}$ of any of the generated grasps. The ablations were done in simulation with a held-out test set of unknown objects. Synthetic point clouds are rendered from cluttered scenes generated with these test objects in randomized stable poses. Physical interactions are simulated using FleX [43]. Area under curve (AUC) of the success-coverage plot is used to compare different variation of the methods in the ablation studies. The success-coverage curves are computed by ranking the grasp scores predicted by the grasp evaluator.

Learned vs. Surface Normal Based Grasp Sampler: The first ablation study we consider is the effect of using a learned VAE to sample grasps in comparison with a geometric baseline. This baseline generates grasps by sampling random points on the object along surface normals, with random standoff, and random orientation along the surface normal. Fig. 3 shows that our learned VAE sampler yields more grasp coverage. It is worth noting that the surface-normal based sampler performed well for simpler shapes like boxes but failed to generate grasps for more complex geometry with rim, handles, etc.

CollisionNet vs Voxel-Based Collision Checking: We compared the effectiveness of CollisionNet with a voxel-based heuristic commonly used (such as in MoveIt! [44]) for obstacle avoidance in unknown 3D environments. In our case, from each object, 100 points are sampled using farthest point sampling. Each sampled point is represented by a voxel cube of size $2\text{cm}$. Collision checking is done by checking the
intersection of the gripper mesh with any of the voxels. As shown in Fig. 4, CollisionNet outperforms the voxel-based heuristic in terms of precision and coverage. Qualitatively, we observed that the voxel-based representation fails to capture collision when the gripper mesh intersects with occluded parts of objects, or if there is missing depth information (see Fig. 5). In cases where the voxel-based collision checking fails, CollisionNet has 89.7% accuracy in classifying the collisions correctly.

The voxel-based approach also has several false negatives by rejecting good grasps that are slightly penetrating voxels corresponding to points on the target object, as the voxels expand the spatial region for collision checking. Without considering the voxels on the target object for collisions, the coverage decreases marginally (blue curve in Fig. 4). The grasp success also decreases as grasps that are actually colliding with the target object are not pruned out. CollisionNet does not suffer from such biases and can reason about relative spatial information in the partial point clouds.

**Single-stage vs. Cascaded Evaluator:** Instead of a cascaded grasp generation approach, one could also use a single-stage sampler and evaluator with object instance information. Once the grasps are sampled, there is only a single evaluator that directly estimates $P(S, \neg C|X, g)$. The positive training set is $G^+_SC = \{g|g \in G^+, \neg\Psi(g, x)\}$ while the negative set is $G^-SC = \{g|g \in G^+, \Psi(g, x)\} \cup \{g|g \in G^-\}$. As a result, some positive grasps $g \in G^+$ will be in collision resulting in lower scores. An example of the input data to this baseline is shown in Fig. 7(b), where the indicator mask of the target object is passed as an additional feature to the PointNet++ architecture. We found that the cascaded model outperformed the single-stage model, as shown in Fig 6.

This improvement is due to two factors. First, the VAE is far more proficient in learning grasps from an object-centric observation than from scene-level information. Second, the cascaded architecture imposes an abstraction between having a grasp evaluator that is singly proficient in reasoning about grasp robustness and CollisionNet that is proficient in predicting collisions.

**Role of Object Instance Segmentation:** We compared our cascaded grasp sampling approach to an instance-agnostic baseline. Without instance information, the baseline is a single-stage grasp planner that uses the point cloud of the scene, since we cannot get a object-centric input. An example of the input data to this baseline is shown in Fig. 7(a). From the ablation shown in Fig. 6, we found that our cascaded grasp sampler (using instance information and CollisionNet) had a AUC of 0.22 and outperformed the object instance-agnostic baseline in terms of both success and coverage, which had a AUC of 0.02. A common failure mode of the instance-agnostic model is that the variational sampler gets confused as to which object to grasp in the scene, with the latent space being potentially mapped to grasps for multiple objects and degrading the overall grasp quality for all the objects.

**B. Real Robot Experiments**

In our robot experiments, we wanted to show that our cascaded grasp synthesis approach (1) transfers to the real world despite being trained only in simulation; (2) has competitive performance for target-driven grasping in real clutter scenes and (3) outperforms baseline methods using the clutter-agnostic 6-DOF GraspNet implementation [20] with instance segmentation and voxel-based collision checking. Our physical setup consists of a Franka Panda robot with a parallel-jaw gripper. We used a Intel RealSense D415 RGB-D camera mounted on the gripper wrist for perception. We selected a set of common household and kitchen objects with
Fig. 8: Application of our approach in retrieving a partly occluded mug (highlighted in (a)). The blocking objects are ranked (colored in (b), red being most inhibiting) and removed from the scene. The target object is finally grasped in (f).

![Image](image_url)

**TABLE I: Real Robot experiments**

| Approach                                      | Trials | Performance (%) |
|-----------------------------------------------|--------|-----------------|
| 6-DOF GraspNet [20] + Ins. Segmentation [23]  | 32/51  | 62.7            |
| Object Instance                               | 31/51  | 60.7            |
| Object Instance + CollisionNet (Ours)         | 41/51  | 80.3            |

C. Application: Removing Blocking Objects

Consider scenarios such as that shown in Fig. 8, where the target object is being blocked by other objects and none of the sampled grasps are kinematically feasible. To accomplish this task, the model needs to generate potential grasps for the target object even though the target object is not physically reachable (detected by low scores from CollisionNet). Given the potential grasps, we can identify which objects are interfering with the generated grasps for the target object. The blocking object $j$ is chosen to be the one with the largest increase in collision scores when removing the corresponding object points from the scene point cloud $i.e. \alpha_j = P(C | \hat{X}_j, g) - P(C | X, g)$. The objects are colored by this ranking metric $\alpha_j$ in Fig. 8(b), with the red object being the most blocking. The modified point cloud $\hat{X}_j$, which hallucinates the scene without object $j$, is implemented by merging the object’s instance mask with that of the table and projecting corresponding points to the table plane. Grasps are then generated for the blocking object and removed from the scene. Collision-free grasps can now be generated for the unoccluded target object for the robot to recover it.

V. CONCLUSION

We present a learning-based framework for 6-DOF grasp synthesis for novel objects in cluttered environments. This problem is especially challenging due to occlusions and collision avoidance which is critical. Our approach achieves a grasp accuracy of 80.3% in grasping novel objects in clutter on a real robotic platform despite being only trained in simulation. A key failure mode of our approach is that it only considers gripper pre-shape collisions by design and hence motion planning could still fail on generated grasps. In future work, we hope to consider trajectory generation in grasp generation and explore the use of our approach in task planning applications.

REFERENCES

[1] D. Prattichizzo and J. Trinkle, “Grasping,” Springer handbook of robotics, pp. 671–700, 2008.
A. Bicchi and V. Kumar, “Robotic grasping and contact: a review,” in IEEE International Conference on Robotics and Automation (ICRA), 2000.

J. Mahler, S. Patil, B. Kehoe, J. van den Berg, M. Ciocarlie, P. Abbeel, and K. Goldberg, “GP-GPS-OPT: Grasp planning with shape uncertainty using gaussian process implicit surfaces and sequential convex programming,” IEEE International Conference on Robotics and Automation (ICRA), 2015.

F. Pokorny and D. Kragic, “Classical grasp quality evaluation: New algorithms and theory,” in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2003.

H. D. Corey Goldfeder, Matei Ciocarlie and P. K. Allen, “The columbia grasp database,” in IEEE International Conference on Robotics and Automation (ICRA), 2009.

C. Goldfeder and P. K. Allen, “Data-driven grasping,” Autonomous Robots, vol. 31, no. 1, pp. 1–20, 2011.

N. Kitaev, I. Mordatch, S. Patil, and P. Abbeel, “Physics-based trajectory optimization for grasping in cluttered environments,” in IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015, pp. 3102–3109.

M. Dogar and S. Srinivasa, “Push-grasping with dexterous hands: Mechanics and a method,” IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2010.

D. Berenson, R. Diankov, K. Nishiwaki, S. Kagami, and J. Kuffner, “Grasp planning in complex scenes,” International Conference on Humanoid Robots, 2007.

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.

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.

J. Bohg, A. Morales, T. Asfour, and D. Kragic, “Data-driven grasp synthesisa survey,” IEEE Transactions on Robotics, 2014.

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.

I. 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.

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.

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.

A. Murali, L. Pinto, D. Gandhi, and A. Gupta, “CASSSL: Curriculum accelerated self-supervised learning,” IEEE International Conference on Robotics and Automation, 2018.

Q. Lu, K. Chenna, B. Sundaralingam, and T. Herrmans, “Planning multi-fingered grasps as probabilistic inference in a learned deep network,” arXiv preprint arXiv:1804.03289, 2018.

X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, “Learning 6-dof grasping interaction via deep 3d geometry-aware representations,” in IEEE International Conference on Robotics and Automation (ICRA), May 2018.

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

C. Choi, W. Schwarting, J. DelPreto, and D. Rus, “Learning object grasping for soft robotic hands,” in IROS Robotics and Automation Letters, IEEE, 2018, pp. 2370 – 2377.

H. Liang, X. Ma, S. Li, M. Gorner, S. Tang, B. Fang, F. Sun, and J. Zhang, “Pointnetgpd: Detecting grasp configurations from point sets,” in IEEE International Conference on Robotics and Automation, 2019.

C. Xie, Y. Xiang, A. Mousavian, and D. Fox, “The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation,” in Conference on Robot Learning (CoRL), 2019.

I. Lenz, H. Lee, and A. Saxena, “Deep learning for detecting robotic grasps,” IJRR, 2015.

D. Kappler, J. Bohg, and S. Schaal, “Leveraging big data for grasp planning,” in IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015, pp. 4304–4311.