

Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes

Martin Sundermeyer^{1,2,3}Arsalan Mousavian¹Rudolph Triebel^{2,3}Dieter Fox^{1,4}

Abstract—Grasping unseen objects in unconstrained, cluttered environments is an essential skill for autonomous robotic manipulation. Despite recent progress in full 6-DoF grasp learning, existing approaches often consist of complex sequential pipelines that possess several potential failure points and run-times unsuitable for closed-loop grasping. Therefore, we propose an end-to-end network that efficiently generates a distribution of 6-DoF parallel-jaw grasps directly from a depth recording of a scene. Our novel grasp representation treats 3D points of the recorded point cloud as potential grasp contacts. By rooting the full 6-DoF grasp pose and width in the observed point cloud, we can reduce the dimensionality of our grasp representation to 4-DoF which greatly facilitates the learning process. Our class-agnostic approach is trained on 17 million simulated grasps and generalizes well to real world sensor data. In a robotic grasping study of unseen objects in structured clutter we achieve over 90% success rate, cutting the failure rate in half compared to a recent state-of-the-art method. Video of the real world experiments and code are available at https://research.nvidia.com/publication/2021-03_Contact-GraspNet%3A--Efficient.

I. INTRODUCTION

The ability to grasp objects is one of the fundamental capabilities required in most robot manipulation tasks. Grasping involves reasoning about the 3D geometry and physics properties of the object such as mass and friction, and also reasoning about complex contact physics. It is studied in two main directions: Model-based grasping where the 3D model or category of the object is known and model-free grasping where there is no prior knowledge about the object. Model-based grasping circumvents reasoning about the physics of contact and grasp generation by pre-defining a set of grasps in the object frame and transform those grasps according to the 6-DoF object pose [1, 2, 3, 4] or detected keypoints of the objects [5, 6]. The downside of model-based approaches is that they only work on a limited subset of known objects or categories, and any errors in detecting 6-DoF object pose or object keypoints degrade the grasping performance.

Model-free approaches do not make any strong assumptions about the category or shape of the object, and they learn a shared representation for all object shapes and sizes. However, having one shared representation for all objects in addition to the large $SE(3)$ space for the grasp poses makes the learning problem quite challenging. As a result, a large body of work in data-driven grasping constraints the

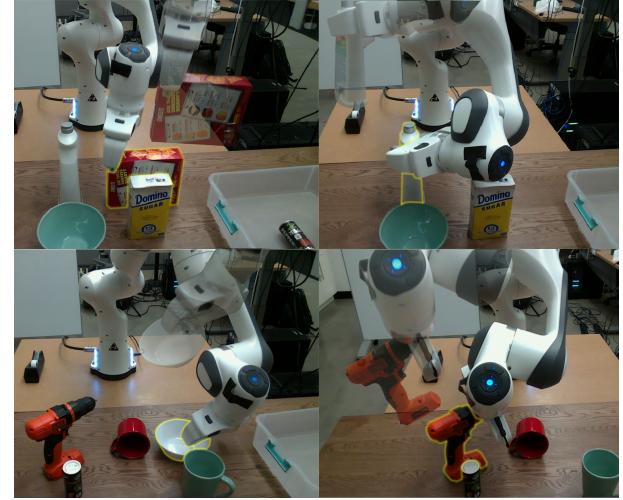


Fig. 1. Contact-GraspNet efficiently predicts diverse and stable grasps in cluttered scenes while avoiding collisions.

space of possible grasps to planar grasping, where grasps are represented by oriented rectangles around each pixel that define the grasp frame [7, 8, 9]. Such a representation needs the camera to view the scene perpendicularly and thus limits 3D reasoning and applications significantly. A large number of possible grasps and the full kinematic capabilities of the robot are also neglected. To address the limitations of planar grasping, there has been a recent interest in tackling the problem of 6-DoF grasping of unknown objects [10, 11, 12, 13, 14]. In this paper, we tackle 6-DoF grasping of unknown objects in cluttered space from a partial point cloud observation of the scene.

Grasping objects from cluttered scenes with structure introduces extra challenges. The target objects must be grasped successfully, while at the same time any collision with other objects must be avoided to prevent damages or transformations into other undesired states. This is particularly important in home robotics and healthcare applications. Additionally, it is crucial to generate a diverse set of grasps for the object due to robot kinematic constraints. Depending on the relative pose between the object and the robot, a different subset of grasps is kinematically feasible.

Our method is closely related to the work of Murali *et al.* [12], where the goal is to generate collision-free diverse grasps for a designated target object from a partial point cloud of the scene, and the objects are segmented using a pre-trained unknown object instance segmentation model [15, 16]. Murali *et al.* [12] use a multi-stage process that

*This work is done while the first author was an intern at NVIDIA.
¹NVIDIA (amousavian,dieterf)@nvidia.com, ²German Aerospace Center (DLR) <first>.<last>@dlr.de, ³Technical University of Munich (TUM), ⁴University of Washington

synthesizes grasps for the target objects from the segmented object point cloud with no context around it, and then filters out the colliding grasps using another learned model. This leads to three issues: 1) Sensitivity to instance segmentation errors. 2) Grasps are generated just from the target object point cloud and do not leverage geometric cues in the scene such as table points and surrounding object points. 3) Grasps are predicted in the large, unconstrained 6-DoF pose space. To address these issues, our method instead directly processes a full scene point cloud or a local region around a target object. Therefore, the quality of our generated grasps is not depending on an accurate mask and collisions can be directly taken into account during generation. Instance segmentation can then subsequently be used to filter grasps belonging to a target object. Thus, our main contributions are the following:

- A new end-to-end method for 6-DoF grasping of unknown objects in cluttered real world scenes where we achieve 90% grasp success rate. This is 10% higher than [12] in equal settings.
- A new grasp pose representation that projects 6-DoF grasps to their contact points in an observed point cloud. Our representation has only 4-DoF which facilitates the learning problem significantly.
- Comprehensive ablation studies in a physics simulator to evaluate the effects of different loss functions and training data.

II. RELATED WORK

As a fundamental problem in robotics, grasping has been studied for decades [17, 18, 19, 20]. We review related literature in the context of data-driven methods.

End-to-end policy learning: One line of work for grasping and manipulation of objects employs an end-to-end policy that learns to generate actions from raw input pixel values [21, 22]. This results in a monolithic model that concurrently reasons about perception, planning, grasping, and controlling the robot. A large group of these works learn from interactions of the robot with the environment through reinforcement learning. These approaches have mostly shown promise in bin picking, in (quasi-) planar grasping and in small, insensible workspaces that do not require complicated motion planning in the robot configuration space. Few works [23] have demonstrated iterative 6-DoF grasping approaches with a monolithic policy by combining imitation learning and reinforcement learning. A common drawback of these methods is the limited generalization to novel environments, because the perception and control are learned indirectly at the same time. In addition, these methods are not easily steerable towards grasping a specific object as the reward function encourages grasping any object. In contrast, our method learns to generate diverse 6-DoF grasps on novel objects and scenes for specifiable target objects while just using simulated training data. Additionally, it can be integrated with other perception and motion planning algorithms.

3D reconstruction: A complete 3D reconstruction enables traditional grasp planning. However, learned single-view re-

constructions are often ambiguous, coarse and require class-conditioning [24, 25, 26]. Multiple views for 3D scanning are beneficial [27] but not always obtainable, take additional time and typically assume a static scene. In our approach a full explicit 3D reconstruction is not required.

Discriminative methods: Discriminative methods for grasping train a classifier that evaluates the quality of existing grasps [28, 29, 7]. They use different sampling strategies to generate potential candidates. For planar grasping, cross entropy is widely used since it can converge to the final grasp location by iteratively evaluating the quality of grasps in different locations [7]. However, the cross-entropy method does not work well in the higher dimensional 6-DoF grasp space. To overcome the sampling complexity issue, grasp locations are often sampled using geometric heuristics [10, 29].

Generative methods: Learning-based generative grasp methods aim to overcome the limitations of geometric heuristics and generate meaningful 6-DoF grasps often from experience in a physics simulator [11, 12]. The main challenge is the large, multi-modal search space of 6-DoF grasps. Instead of sampling some potential candidates using heuristics and ranking them, these models directly predict a per-point graspability score and approach direction in $SO(3)$ space [30, 14, 31]. One problem with predicting approach directions is that they cannot easily capture high curvature areas such as mug rims or handles and also can not represent grasps encompassing hollow structures. Furthermore, successful approach directions are quite ambiguous to learn as multiple ones are possible for a single contact. Instead we propose to predict 6-DoF grasps densely projected to their much less ambiguous contact points. While grasps without full surface contact are plausible, e.g. through the handle of a mug, the knowledge about the object state and therefore the ability to steadily place the object again is lost. Therefore, in this work we are aiming to generate stable grasps for unknown objects with full surface contact. Our novel loss formulation further improves convergence by accounting for the discontinuities, imbalance and multi modality of the grasp distribution. Unlike other methods [31], our proposed method is independent of category labels and has no assumption of grasps being always perpendicular to a surface. Instead we learn a grasp semantic purely from a wide variety of grasp annotated training shapes [32].

III. METHOD

We consider the problem of generating 6-DoF grasps from any viewpoint on structured clutter consisting of unknown objects. Our approach takes in a raw depth image, optionally with object masks, and generates 6-DoF grasp proposals together with corresponding grasp widths. Our goal is to predict grasps that are robust, diverse and non-colliding from an only partially observable scene.

From a learning perspective, generating the distribution of successful 6-DoF grasps is quite challenging, because the distribution is multi-modal, discontinuous, imbalanced and ambiguous due to (self-) occlusions. Furthermore, direct regression in high dimensional output spaces like $SE(3)$ has

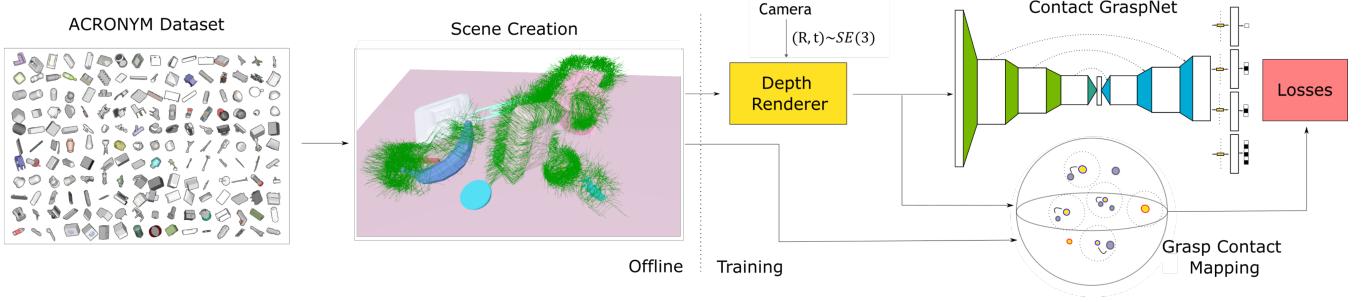


Fig. 2. Training Data Pipeline. We place object meshes with dense grasp annotations from the ACRONYM dataset [32] at random stable poses in scenes. Grasp poses that produce gripper model collisions are removed. Resulting grasps are mapped to their contacts on the mesh surface. During training, we sample virtual cameras to render point clouds from the scenes. We consider recorded points (yellow) as positive contacts if there exists a mesh contact (blue) in a 5mm radius and associate the grasp transformation belonging to the closest mesh contact to them. These per-point annotations are used to supervise the Contact Grasp Network.

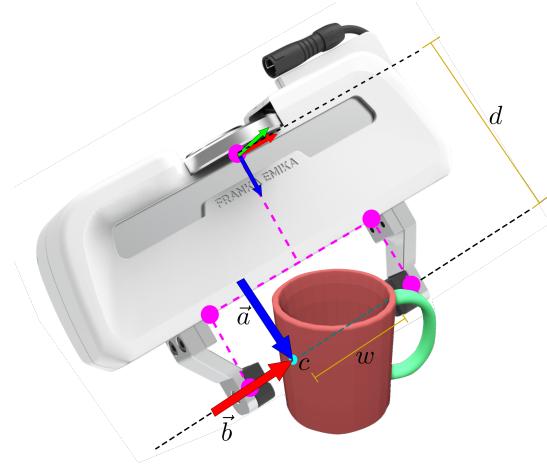


Fig. 3. Our grasp representation: c depicts an observed contact point. \mathbf{a} and \mathbf{b} constitute the 3-DoF rotation, w is the predicted grasp width, d the distance from baseline to base frame. In pink we show the five gripper points \mathbf{v} that we used in the l_{add-s} loss.

been shown to be difficult in grasping [11] and also in related fields such as object pose estimation [33].

A. Grasp Representation

For these reasons, finding an efficient grasp representation is crucial to solve this task using learning-based methods. This representation should generalize well to unseen objects and handle the high-dimensional output space well.

Contact Grasp Representation: We observe that for most predictable two-finger grasps at least one of the two contacts is visible prior to grasping. In contrast, grasps without any visible contact are often ambiguous or do not preserve the initial object pose after grasping. Therefore, we map a distribution of successful 6-DoF ground truth grasps $g \in G$ to their corresponding contact points $c \in \mathbb{R}^3$. Since visible contact points are bound to lie on surfaces that we can observe with a depth sensor, we can represent their 3D location by nearby points in a recorded point cloud.

Given that we can predict whether observed points are suitable grasp contacts, we can thus reduce the 6-DoF grasp

learning problem to estimating the 3-DoF grasp rotation $R_g \in \mathbb{R}^{3 \times 3}$ and grasp width $w \in \mathbb{R}$ of a parallel-yaw gripper.

Starting from a contact point $\mathbf{c} \in \mathbb{R}^3$, where the gripper baseline intersects the mesh, we depict a 6-DoF grasp pose $g \in G$ defined by $(R_g, t_g) \in SE(3)$ and grasp width $w \in \mathbb{R}$ as

$$\mathbf{t}_g = \mathbf{c} + \frac{w}{2} \mathbf{b} + da \quad (1)$$

$$R_g = \begin{bmatrix} | & | & | \\ \mathbf{b} & \mathbf{a} \times \mathbf{b} & \mathbf{a} \\ | & | & | \end{bmatrix}, \quad (2)$$

where $\mathbf{a} \in \mathbb{R}^3, \|\mathbf{a}\| = 1$ is the approach vector, $\mathbf{b} \in \mathbb{R}^3, \|\mathbf{b}\| = 1$ is the grasp baseline vector, and $d \in \mathbb{R}$ is the constant distance from the gripper baseline to the gripper base. Our grasp representation is depicted in Figure 3.

The reduced dimensionality greatly facilitates the learning process compared to methods that estimate grasp poses in unconstrained $SE(3)$ space. It also increases the pose accuracy of predicted grasps as they are bound to the geometry of the observed scene. In contrast to axis-angle representations, our rotation representation has neither ambiguities nor discontinuities. Moreover, at test time we can sample grasp proposals by sampling contact points that cover the whole observable surface of the scene/object and thus represent the modes of the 6-DoF grasp distribution well. While a 3D view on the scene is preferable, even a frontal view on a box produces reasonable grasps due to the radial mapping.

Point Set Networks such as PointNet++ [34] effectively process point clouds and hierarchically aggregate points and their feature representations in local 3D neighborhoods. Their predictions can be directly associated to 3D points in the input point cloud and our proposed grasp representation exploits this ability.

B. Data Generation

To learn the full distribution of stable 6-DoF grasps, diverse and dense grasp pose annotations are required. We used the ACRONYM dataset [32], which consists of 8872 meshes from the Shapenet dataset [35] and 17.7 million

simulated grasps under varying friction. An overview of our offline and online training data generation is given in Fig. 2.

During training we render a scene point cloud $\mathcal{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\} \subset \mathbb{R}^3$ and assign a point-wise grasp success

$$\forall i = 1, \dots, n \quad s_i = \begin{cases} 1 & \min_j \|\mathbf{p}_i - \mathbf{c}_j\|_2 < r \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where $\mathbf{c}_j \in \mathcal{P}$ are the mesh contact points of non-colliding ground truth grasps $g_j \in G$ in camera coordinates and $r \in \mathbb{R}$ is their maximum propagation radius. Thus, \mathcal{P} can be split into points $\mathcal{P}^- := \{\mathbf{p}_i | s_i = 0\}$, where no feasible grasp contact is found within a radius of $r = 5\text{mm}$, and $\mathcal{P}^+ := \{\mathbf{p}_i | s_i = 1\}$, containing points suitable for a contact. To the latter ones $\mathbf{p}_i^+ \in \mathcal{P}^+$ we assign the closest grasp as

$$\begin{bmatrix} w_{g,i} \\ R_{g,i} \\ \mathbf{t}_{g,i} \end{bmatrix} = \begin{bmatrix} w_{g,j} \\ R_{g,j} \\ \mathbf{p}_i^+ + \frac{w_j}{2}\mathbf{b}_j + d\mathbf{a}_j \end{bmatrix} \quad (4)$$

with

$$j = \arg \min_k \|\mathbf{p}_i^+ - \mathbf{c}_k\|_2 \quad (5)$$

Given sufficient coverage we can thereby project the ground truth distribution of 6-DoF grasps densely on the recorded point cloud.

C. Network

We employ the set abstraction and feature propagation layers proposed in PointNet++ [34] to build an asymmetric U-shaped network. The network takes $n=20000$ random points $p \in \mathbb{R}^{20000 \times 3}$ as input and predicts grasps for only $m=2048$ farthest points of the input to make sure the inference fits in GPU memory and predicted grasps have good coverage over the scene. The network has four heads with two 1D-Conv layers each and per-point outputs $s \in \mathbb{R}$, $\mathbf{z}_1 \in \mathbb{R}^3$, $\mathbf{z}_2 \in \mathbb{R}^3$, $\mathbf{o} \in \mathbb{R}^{10}$, from which we form our grasp representation. The predicted grasp width $\hat{\mathbf{w}}_i \in [0, w_{max}]$ is split into 10 equidistant grasp width bins $\hat{\mathbf{o}} \in \mathbb{R}^{10}$ to counteract data imbalance. Then, $\hat{\mathbf{w}}_i$ is represented by the center value of the bin(s) with the highest confidence. The approach direction $\mathbf{a} \in \mathbb{R}^3$ and the baseline direction $\mathbf{b} \in \mathbb{R}^3$ are orthonormal by definition. We inject this property into training by coupling the predictions $\hat{\mathbf{a}}, \hat{\mathbf{b}}$ through an in-network Gram Schmidt orthonormalization

$$\hat{\mathbf{b}} = \frac{\mathbf{z}_1}{\|\mathbf{z}_1\|} \quad \hat{\mathbf{a}} = \frac{\mathbf{z}_2 - \langle \hat{\mathbf{b}}, \mathbf{z}_2 \rangle \hat{\mathbf{b}}}{\|\mathbf{z}_2\|} \quad (6)$$

Thus, we perform a projection and only predict $\hat{\mathbf{a}}$ as the component that is orthonormal to $\hat{\mathbf{b}}$. The orthonormalization further reduces the dimensionality of our predicted grasp representation and facilitates the regression of 3D rotations [36].

D. Target Losses

The contact grasp success predictions $\hat{s} \in \mathbb{R}$ are evaluated at all output points $\mathbf{p}_i \in \mathbb{R}^3 : \forall i \in [0, m]$ using binary cross entropy. We only backpropagate the top-k point predictions with the largest errors $l_{bce,k}$, with $k=512$, to counteract data imbalance. The other predictions concerning the geometry of grasps are only evaluated at positive contact points \mathbf{p}_i^+ . Instead of supervising all network heads in isolation, we propose to combine the predictions to the 6-DoF grasp pose $\hat{g} \in G$ given in Eq. (1) and (2) already during training. We define five 3D points $\mathbf{v} \in \mathbb{R}^{5 \times 3}$ representing the 6-DoF gripper pose, as shown in Fig. 3, and transform these using all ground truth and predicted grasp poses defined in Eq. (4)

$$\mathbf{v}_i^{gt} = \mathbf{v} R_{g,i}^T + \mathbf{t}_{g,i} \quad \mathbf{v}_i^{pred} = \mathbf{v} \hat{R}_{g,i}^T + \hat{\mathbf{t}}_{g,i} \quad (7)$$

We formulate the 6-DoF grasp loss l_{add-s} as a weighted minimum average distance between gripper points \mathbf{v}^{gt} and \mathbf{v}^{pred} where we take the symmetry of the gripper into account.

$$l_{add-s} = \frac{1}{n^+} \sum_i^{n^+} \hat{s}_i \min_u \|\mathbf{v}_i^{pred} - \mathbf{v}_u^{gt}\|_2, \quad (8)$$

where n^+ is the size of \mathcal{P}^+ . We weight each distance to the closest ground truth grasp points with the predicted contact success confidence \hat{s}_i .

Our proposed loss formulation has several advantages: (1) We can learn the different modes of the ground truth grasp distribution, e.g. different predicted grasp approach directions $\hat{\mathbf{a}}$ can produce a small error. (2) The point-wise weighting with \hat{s}_i couples the contact point classification with the grasp pose predictions. Contact confidence can only increase if the network predicts a 6-DoF grasp pose close to a ground truth pose. (3) Wrongly predicted grasps in regions far away from any ground truth grasp, e.g. at artificial edges from occlusions, produce a high loss and are thus avoided.

On the grasp width bin predictions, we optimize a weighted, multi-label binary cross entropy loss l_{width} . Since small grasp widths are highly over-represented, we weight the bin losses anti-proportional to bin size. Our total loss is $l = \alpha l_{bce,k} + \beta l_{add-s} + \gamma l_{width}$ with $\alpha = 1, \beta = 10, \gamma = 1$.

E. Implementation Details

We use the Adam optimizer with an initial learning rate of 0.001 and a step-wise decay to 0.0001. Our set abstraction layers have 3 parallel branches with query ball radii [0.02, 0.04, 0.08], [0.04, 0.08, 0.16] and [0.08, 0.16, 0.32]. For inference the point cloud is centered at its mean in camera coordinates. For training we generate 10000 table top scenes by placing 8-12 grasp annotated ShapeNet models [32] at random stable poses. We use rejection sampling to avoid collisions. We train with a batch size of 3 for 144.000 iterations which takes ~ 40 hours on a single Nvidia V100 GPU. Convergence is significantly faster than on previous methods [12, 14, 11] which take up to one week on a single GPU for training. This also reflects the effectiveness of our proposed grasp representation.

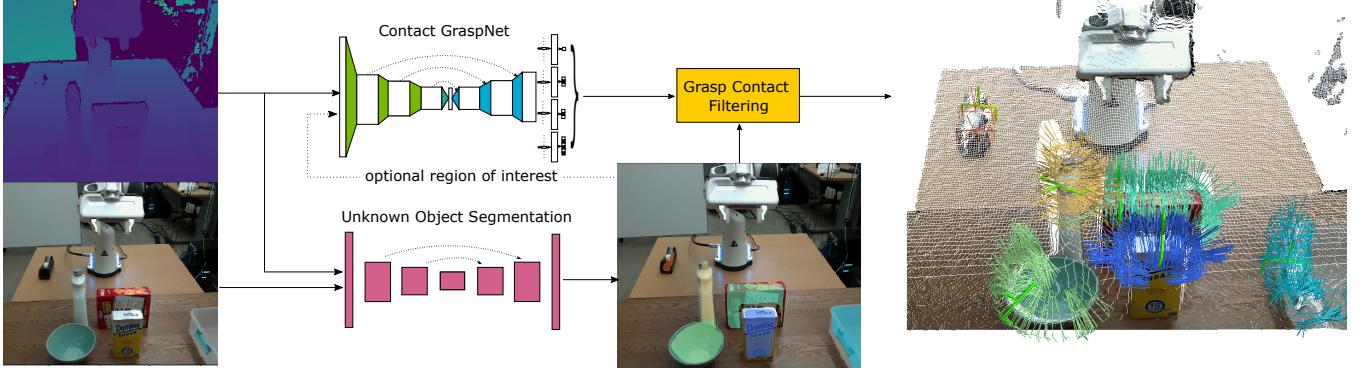


Fig. 4. Full Inference Pipeline: We segment unknown objects from an RGB-D image using [15]. Our Contact-GraspNet processes the full scene point cloud or a local region of interest around a target object. Predicted 6-DoF grasps are then associated to object segments by filtering their contact points. On the right we show the predicted 6-DoF grasp distribution and, in bold, the most confident grasp per segment.

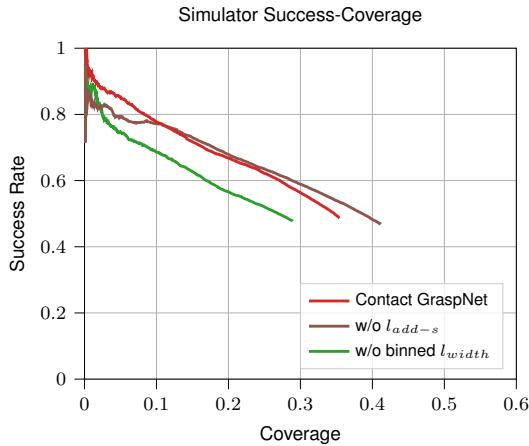


Fig. 5. Loss Ablations: Without weighted binning in the grasp width loss l_{width} both, success rate and coverage decrease. The l_{add-s} loss leads to increased success rates at high confidence contacts ($\text{Coverage} \in [0, 0.1]$) and to slightly decreased success rate in the low-confidence regime. This confidence calibration is important, since it determines which grasp is eventually executed.

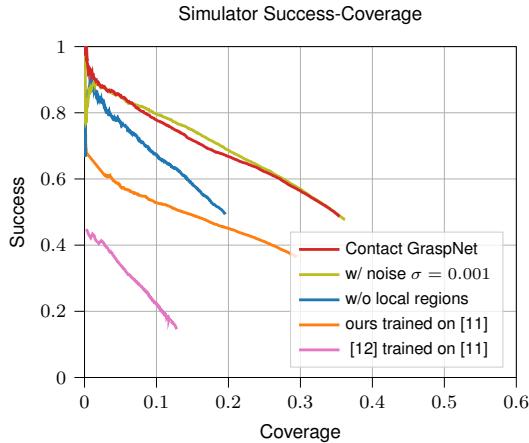


Fig. 6. Data Ablations: Training with Gaussian noise has similar performance in simulation but helps generalization to noisy sensor data. Predicting grasps directly on full scenes without extracting local regions yields a similar average success rate, but significantly lowers grasp coverage. Training on the small grasp dataset from [11] with 5 categories is not sufficient to generalize to arbitrary objects and shows the importance of ACRONYM [32].

IV. EXPERIMENTAL EVALUATION

We evaluate our method in a grasping study with a Franka robot where we pick unknown objects in cluttered scenes. We also compare different variations of our method and of our data by executing a large number of predicted grasps in the FleX physics simulator [37].

A. Inference

Our inference pipeline is shown and described in Fig. 4. The Contact-GraspNet can also be applied to raw depth images by itself, but most robotic tasks require some kind of instance detection/segmentation to specify a target.

Local regions of interest can be optionally extracted around the 3D centroid of point cloud segments in order to maximize the number of potential contact points. In our experiments, we extract cubes with an edge length set to twice the largest spanning dimension, but at least 0.3m and at most 0.6m.

Run time: The Contact-GraspNet has a run time of 0.28s for a full scene or $\sim 0.19s$ for a local region around a target object. Compared to other 6-DoF grasp generation methods this is quite fast and enables applications requiring reactive closed loop grasping.

Grasp Selection: At test time we select grasps by setting a contact confidence threshold of 0.23 and then use farthest point sampling on the (filtered) contact points to ensure broad grasp coverage. If the number of predicted grasps for an object is too low, we reduce the confidence threshold to 0.19. In the end we execute the most confident grasp that is kinematically reachable and where the robot does not collide with the scene [38].

B. Evaluation Metrics

In our robotic experiments we report the number of successful grasps and the number of trials. The latter is often disregarded when picking small objects from a bin. However, grasping in only one or two trials is crucial in cluttered scenes (e.g. in households) with large, densely packed objects where collisions should be avoided and stable grasp opportunities can vanish after objects tip over. We

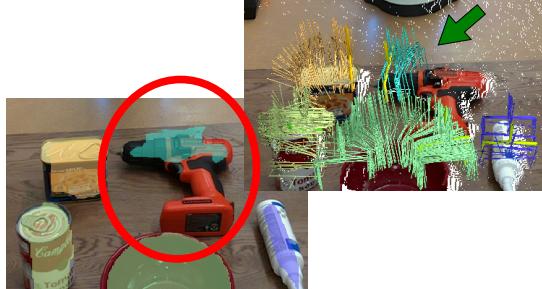


Fig. 7. One advantage of our method is that it does not rely on an accurate segmentation of unknown objects. Here, successful grasp contacts are still found on the driller despite severe under-segmentation.

limit ourselves to a maximum of two grasp trials per object without rearrangements and report the success rate after a single trial as well.

Our simulator experiments allow us to also evaluate the diversity of grasps and ablate variations of our method. Here, we evaluate the success rate and coverage of the generated grasps following [11]. A grasp is considered successful if (1) the open gripper does not collide with the object/scene and (2) the object is still in the gripper after grasping and a shaking motion. This is a conservative measure, as most real world grasps can slightly collide and do not undergo a shaking motion. Coverage is the percentage of ground truth grasps (including occluded ones) whose base coordinates are within 2cm of any of the generated grasps.

C. Real robot grasp experiments

Setup: Our physical setup consists of a 7-DoF Franka Panda robot with a parallel-jaw gripper. We closely replicate the 9 cluttered scenes defined in [12] with a total of 51 unseen objects. The task is to pick the objects from the cluttered scene and place them into a bin. We manually select target objects and grasp them in the same random order as in [12]. In our experiments, we use the Intel Realsense L515 LiDAR camera mounted on a tripod for both RGB and depth data. Robot motions are generated using [38].

Results: Table I shows our grasp evaluation results on the robot. We observe a significantly higher grasp success rate of our method compared to [11] and [12] which themselves outperform other learning-based methods and analytic/heuristic baselines. Furthermore, our method strongly improves the grasp success at first trial and thereby reduces the number of re-grasps. We also addressed the shortcomings of cropping objects from the point cloud using potentially imprecise segmentation masks. Fig. 7 shows an imprecise segmentation example where cropping would be catastrophic but where our grasp filtering method can still extract successful grasps.

D. Ablations

Optimization Targets: In Fig. 5 we first investigate the effect of our loss targets. The weighted loss on the grasp width bins l_{width} is crucial to deal with the imbalanced widths in our grasp dataset. Without weighting the bins, the predictions mostly collapse into narrow grasp widths.

TABLE I

CLUTTERED SCENE GRASPING: WE ACHIEVE A CLEAR IMPROVEMENT OVER RECENT STATE-OF-THE-ART GRASPING PIPELINES

	Success	First attempt	#Attempts
6-DOF GraspNet [11]	62.7	-	-
[11] +CollisionNet [12]	80.39	68.63	67
Contact-GraspNet	90.20	84.31	59

Weighting also performs better than oversampling in our experiments. The average distance loss l_{add-s} improves the success rate of high confidence contacts which is important because most grasps that we execute lie in the first decimal of coverage. The connection of contact confidence with the grasp pose results in an overall improved calibration.

Data: In Fig. 6 we examine the effects of different training and test data. Zooming into local regions allows the network to concentrate potential contact points on the object and thus increases coverage. We also show the importance of a large and diverse grasp dataset like ACRONYM [32]. Training on a small grasp datasets with 110 objects from 5 categories [11] is not sufficient for out-of-category generalization irrespective of the method.

Failure Cases: We observe some failure cases for thick objects that only allow grasps almost at maximum grasp width. Here, grasp predictions are less confident presumably because of the discontinuous decision boundary. Injecting noise during training reduces this effect. Finally, small objects sometimes have contact points with low confidence possibly because of their small impact on the total loss.

V. CONCLUSIONS

We considered the fundamental problem of grasping unknown objects in structured clutter with a parallel jaw gripper. We proposed an efficient, accurate and simplifying 6-DoF grasp generation method called Contact-GraspNet. By transforming the hardly tractable 6-DoF grasp estimation problem into a grasp contact point classification and a grasp rotation estimate, we greatly limit the predicted pose space and facilitate the learning process. Through tailored optimization targets that take into account the multi-modality, imbalance and sparsity of the 6-DoF grasp distribution, our network learns to generate diverse grasps covering the whole graspable surface in a recorded scene. Gripper collisions are effectively avoided by considering them during training and by predicting grasps directly in scenes. Our approach can incorporate segmentation predictions as well but is not dependent on accurate masks itself. It is also complementary to grasp ranking methods that use gripper and/or robot models as input. Grasping successfully with a single attempt is crucial in sensible environments. Our method showed strong advances in that regard and is a step towards reaching the required grasp reliability.

REFERENCES

- [1] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, “Poserbpf: A rao-blackwellized particle filter for 6d object pose tracking,” in *Robotics: Science and Systems (RSS)*, 2019.
- [2] M. Sundermeyer, Z.-C. Marton, M. Durner, and R. Triebel, “Augmented autoencoders: Implicit 3d orientation learning for 6d object detection,” *International Journal of Computer Vision*, vol. 128, no. 3, pp. 714–729, 2020.
- [3] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, “Self-supervised 6d object pose estimation for robot manipulation,” in *International Conference on Robotics and Automation (ICRA)*, 2020.
- [4] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, “Densefusion: 6d object pose estimation by iterative dense fusion,” 2019.
- [5] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, “kpam: Keypoint affordances for category-level robotic manipulation.” *International Symposium on Robotics Research (ISRR)*, 2019.
- [6] K. Fang, Y. Zhu, A. Garg, A. Kuryenkov, V. Mehta, L. Fei-Fei, and S. Savarese, “Learning task-oriented grasping for tool manipulation from simulated self-supervision,” *Robotics: Science and Systems (RSS)*, 2018.
- [7] 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.
- [8] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, “Learning ambidextrous robot grasping policies,” *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.
- [9] I. Lenz, H. Lee, and A. Saxena, “Deep learning for detecting robotic grasps,” *IJRR*, 2015.
- [10] 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.
- [11] A. Mousavian, C. Eppner, and D. Fox, “6-dof grapsnet: Variational grasp generation for object manipulation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2901–2910.
- [12] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, “6-dof grasping for target-driven object manipulation in clutter,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.
- [13] S. Song, A. Zeng, J. Lee, and T. Funkhouser, “Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations,” *Robotics and Automation Letters*, 2020.
- [14] H.-S. Fang, C. Wang, M. Gou, and C. Lu, “Grasnet-1billion: A large-scale benchmark for general object grasping,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11444–11453.
- [15] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, “Learning rgb-d feature embeddings for unseen object instance segmentation,” *Conference on Robotic Learning(CoRL)*, 2020.
- [16] 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.
- [17] D. Prattichizzo and J. J. Trinkle, *Grasping*, 01 2008, pp. 671–700.
- [18] A. Bicchi and V. Kumar, “Robotic grasping and contact: a review,” in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, vol. 1, 2000, pp. 348–353 vol.1.
- [19] K. Hang, J. A. Stork, and D. Kragic, “Hierarchical fingertip space for multi-fingered precision grasping,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 1641–1648.
- [20] G. Du, K. Wang, S. Lian, and K. Zhao, “Vision-based robotic grasp detection from object localization, object pose estimation to grasp estimation: A review,” *arXiv preprint arXiv:1905.06658*, 2019.
- [21] 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.
- [22] 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.
- [23] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, “Learning synergies between pushing and grasping with self-supervised deep reinforcement learning,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4238–4245.
- [24] X. Yan, M. Khansari, J. Hsu, Y. Gong, Y. Bai, S. Pirk, and H. Lee, “Data-efficient learning for sim-to-real robotic grasping using deep point cloud prediction networks,” 2019.
- [25] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, “Learning 6-dof grasping interaction via deep geometry-aware 3d representations,” 2018.
- [26] W. Agnew, C. Xie, A. Walsman, O. Murad, C. Wang, P. Domingos, and S. Srinivas, “Amodal 3d reconstruction for robotic manipulation via stability and connectivity,” 2020.
- [27] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, “Volumetric grasping network: Real-time 6 dof grasp detection in clutter,” 2021.
- [28] D. Fischinger, A. Weiss, and M. Vincze, “Learning grasps with topographic features,” *The International Journal of Robotics Research*, vol. 34, no. 9, pp. 1167–1194, 2015. [Online]. Available: <https://doi.org/10.1177/0278364915577105>
- [29] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, “PointNetGPD: Detecting grasp configurations from point sets,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.
- [30] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, “S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes,” in *Conference on robot learning*. PMLR, 2020, pp. 53–65.
- [31] P. Ni, W. Zhang, X. Zhu, and Q. Cao, “Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds,” *arXiv preprint arXiv:2003.09644*, 2020.
- [32] C. Eppner, A. Mousavian, and F. Dieter, “Acronym: A large-scale grasp dataset based on simulation,” *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.
- [33] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, “Implicit 3D orientation learning for 6D object detection from RGB images,” in *ECCV*, 2018.
- [34] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Neural Information Processing Systems (NeurIPS)*, 2017.
- [35] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., “Shapenet: An information-rich 3d model repository,” *arXiv preprint arXiv:1512.03012*, 2015.
- [36] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, “On the continuity of rotation representations in neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.
- [37] M. Macklin, M. Müller, N. Chentanez, and T.-Y. Kim, “Unified particle physics for real-time applications,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–12, 2014.
- [38] M. Danielczuk, A. Mousavian, C. Eppner, and D. Fox, “Object rearrangement using learned implicit collision functions,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.