Learning Object Grasping for Soft Robot Hands

Changhyun Choi, Wilko Schwarting, Joseph DelPreto, and Daniela Rus

***Abstract*—We present a three-dimensional deep convolutional neural network (3D CNN) approach for grasping unknown objects with soft hands. Soft hands are compliant and capable of handling uncertainty in sensing and actuation, but come at the cost of unpredictable deformation of the soft fingers. Traditional model-driven grasping approaches, which assume known models for objects, robot hands, and stable grasps with expected contacts, are inapplicable to such soft hands, since predicting contact points between objects and soft hands is not straightforward. Our solution adopts a deep CNN approach to find good caging grasps for previously unseen objects by learning effective features and a classifier from point cloud data. Unlike recent CNN models applied to robotic grasping which have been trained on 2D or 2.5D images and limited to a fixed top grasping direction, we exploit the power of a 3D CNN model to estimate suitable grasp poses from multiple grasping directions (top and side directions) and wrist orientations, which has great potential for geometry-related robotic tasks. Our soft hands guided by the 3D CNN algorithm show 87% successful grasping on previously unseen objects. A set of comparative evaluations shows the robustness of our approach with respect to noise and occlusions.**

***Index Terms*—Perception for grasping and manipulation, deep learning in robotics and automation.**

I. Introduction

IN ROBOTIC manipulation, robust object grasping is an important prerequisite for advanced autonomous manipulation tasks. While object grasping with robotic manipulators has been actively studied for decades [1], reliable grasping of previously unseen objects is still a challenging problem. The main challenges are the uncertainties in *perception* and *action*. Earlier work has leveraged prior knowledge of object shape, manipulator, stable grasps, etc. [2]. These model-driven approaches, however, are problematic when prior knowledge is partial or not available. Recent work has focused more on learning from data with hope of generalizing to novel situations by learning a mapping function from raw sensory data to a grasp representation [3]. However, these learned grasp representations are rather limited as they often use a 2D grasp location and 1D wrist orientation with a fixed grasping direction, which does not generalize to 6-DoF grasp reasoning and thus does not utilize the full workspace of the robot arm for grasp planning. Robot hands with hard fingers require careful positioning to achieve closure grasps, and the placement of the fingers is usually sensitive to uncertainties. To overcome this limitation, soft robot hands have been actively studied and fabricated using soft materials [4]. The main advantages of soft hands include compliance with external perturbation and tolerance of uncertainties in actuation and perception [5], which enable soft hands to be more suitable for manipulating unknown objects. Moreover, manufacturing of soft hands is faster and less expensive than their hard counterparts [4].

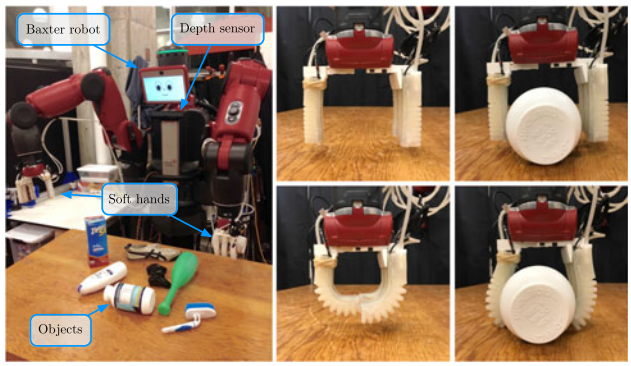

Fig. 1. Baxter with soft hands. Our Baxter robot has two soft hands on its end effectors. A depth sensor is affixed to the upper torso of the robot, and point clouds from the sensor are used to predict suitable grasps for soft hands so as to successfully grasp the objects on the table. The right four figures show our four-finger soft hand in action. Each finger is controlled by an external pneumatic actuator and the Baxter's original parallel gripper actuator is further controlled to maximize the acquisition region.

In this letter, we design a soft robotic manipulation system which is capable of grasping previously unseen objects. Fig. 1 shows our Baxter robot setup with two soft hands mounted on its end effectors. A depth sensor, which is affixed to the upper torso of the robot, obtains partial point clouds of the objects on the table. Given the input clouds, a 3D CNN model predicts the likelihood of success of a set of suitable grasps. Together with the grasp poses, the compliance and adaptability of the soft hands yield successful grasping of novel objects. The work described in this letter uses neither proprioceptive sensors nor 3D object models; it learns appropriate grasps from partial point cloud data and generalizes well to new objects the robot has never seen before. The main contributions of this letter are as follows:

Manuscript received September 9, 2017; accepted February 5, 2018. Date of publication February 28, 2018; date of current version March 28, 2018. This letter was recommended for publication by Associate Editor F. L. Hammond III and Editor H. Ding upon evaluation of the reviewers' comments. This work was supported in part by the Boeing Company, in part by the National Science Foundation IIS 1226883, and in part by the National Science Foundation Graduate Research Fellowship 1122374. *(Corresponding Author: Changhyun Choi.)*

The authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: cchoi@csail.mit.edu; wilkos@csail.mit.edu; delpreto@csail.mit.edu; rus@csail.mit.edu).

This letter has supplemental downloadable multimedia material available at http://ieeexplore.ieee.org, provided by the authors. The Supplementary Materials contain a video briefly explaining the main idea of the work and showing the soft robot hand grasping previously unknown objects. This material is 24.2 MB in size.

Digital Object Identifier 10.1109/LRA.2018.2810544

- *3D CNN-based grasp prediction:* Our approach exploits the power of a 3D CNN model to predict a set of suitable grasp poses from a partial 3D point cloud of an object. While many learning-based approaches have focused on predicting wrist orientations with a fixed top grasping direction, our approach predicts both grasping directions and wrist orientations which determine a set of suitable grasp poses.
- *Vision-based soft hands:* Unlike most soft hands demonstrating object grasping with human operation or known object pose, we propose an end-to-end system that combines vision and soft actuation. In particular, we combine soft hands with the 3D CNN grasping prediction to reliably grasp previously unseen objects.

The rationale behind our approach is that the 3D CNN grasp prediction and soft hands complement each other. A set of discretized grasp poses from learning-based methods requires adaptable grasping as there is always a discrepancy between the predicted grasp pose and real object pose. At the same time, soft hands necessitate good grasping guidance in spite of their compliance and adaptability. To the best of our knowledge, this is the first work to apply a 3D CNN-based grasp prediction to soft hands. While the work in [6] presented a learning from demonstration approach for soft hands, the work employed a marker system to obtain object trajectories, and thus no learning occurs on the visual perception side. In addition, their object grasping capability has hinged upon a set of human demonstrations for a known object. Our work is different in that it learns a suitable grasp policy from the partial point cloud of a previously unknown object.

The paper is organized as follows. Section II reviews prior work in robotic object grasping. After our problem is formalized in Section III, the details of our approach are described in Section IV. Section V presents experimental results on grasp pose prediction and object grasping with our soft manipulator.

II. Related Work

Robotic grasping and grasp synthesis have been actively studied in robotics literature [1]. While there are many ways to categorize this literature [1], [7], the research approaches of object grasping can be roughly divided into *model-driven* and *data-driven* approaches.

A. Model-driven Approaches

Classical object grasping approaches rely on prior knowledge. Such knowledge includes known stable grasp and contact information, 3D models of objects and manipulators, and their physical properties such as weights, centers of mass, friction coefficients, etc. [2]. The goal of these approaches is to find a set of stable force closures to grasp the known objects [8], [9]. Since the models of objects are given, the grasping approaches are based on object recognition and pose estimation with known object grasps [10], [11] or grasp candidates sampled from the known model or simpler geometric primitives [12], [13]. As reviewing extensive model-driven approaches is beyond the scope of this letter, we refer the reader to comprehensive literature surveys [1], [7].

B. Data-driven Approaches

While model-driven approaches assume rich prior knowledge, data-driven approaches gain knowledge of objects and grasping from data. The key idea is to map directly from visual sensor data to grasp representations. The most popular representation is the grasping rectangle describing suitable grasping in the image plane [3], [14], which is of lower dimension than the traditional grasping parameters such as the grasping point, the approach vector, the wrist orientation, and the initial finger configuration [1]. While a simple logistic regression was proven to be an effective learning algorithm for object grasping [14], more recently deep neural network models outperformed the previous method [3]. In particular, convolutional neural networks (CNNs) have been successful in generic object recognition [15] due to their end-to-end feature learning in a hierarchical structure. Following the success, the CNN models have recently been applied to robotic graping [16]–[21]. Common approaches employ CNN models to classify feasibility of a set of grasping hypotheses. Training data has been generated via crowdsourcing [16], physics simulation with 3D model databases [16]–[18], [21], or trial-and-error [19], [22]. Since our approach uses soft hands which are hard to model in physics simulation and crowdsourcing, we adopt the trial-and-error scheme to collect the training data with the ground truth grasp labels annotated manually.

While the most common modality for CNN models is monocular images, visual perception for object grasping can potentially benefit from depth data. The main advantages of using depth data include 1) invariance to photometric variations such as color or texture, and 2) exploiting geometric information closely related to object grasping. Although some prior works employed depth as a sensory modality, their usages were restricted to object proposal [19] or 2.5D depth information [3], [21], [23] without exploiting the full 3D shape information. Robust 3D reasoning is important for the object grasping problem, as the problem is closely related to geometric characteristics and constraints of objects and their surrounding environments. The work in [23] employed a CNN model that learns a grasp quality from a depth image. While their system uses point clouds as a visual input, the CNN model treats it as 2.5D image not 3D, and thus the object grasp pose is limited to top grasping. Recently, full 3D CNN models have been studied and show state-of-the-art performance for shape-based object recognition tasks [24], [25]. These 3D CNNs are relatively new models and have great potential for geometry-related robotic perception.

C. Soft Hands

Robust object manipulation in unstructured environments is a challenging problem due to the uncertainty associated with complex and unpredictable environments. Conventional robot hands, requiring multiple articulated fingers and sensors, are expensive to manufacture and control, and are often fragile in these unstructured environments. More adaptive and compliant robot hands were explored that use underactuation [26].

Recently, new types of robot hands have been designed and fabricated using soft materials [4], [27]. The main advantage of soft hands is compliance, which is well suited for manipulation tasks handling delicate, irregularly shaped, or unknown objects. In addition, soft hands are more tolerant of uncertainties in perception and actuation [5].

III. Problem Formulation

The grasping problem we solve in this letter can be formalized as follows:

Definition 1: Given a point cloud $\mathcal{P} \subset \mathbb{R}^3$, the goal is to find an appropriate grasp pose $\mathbf{X}_g \in SE(3)$ for a previously unseen object $\mathbf{o} \in \mathcal{O}$ that is placed with arbitrary pose in $\mathcal{P}$ within the field of view of the robot.

The grasp pose $\mathbf{X}_g$ is in the Special Euclidean group $SE(3)$, which represents the 3D rigid body transformation, and is defined with respect to the robot coordinate frame. The point cloud $\mathcal{P}$ is obtained via a depth sensor affixed to a robot with a known extrinsic parameter $\mathbf{X}_s^r$, by which the cloud $\mathcal{P}$ is transformed from the sensor coordinate frame to the robot coordinate frame. In $\mathcal{P}$, if multiple objects $\mathcal{O}$ exist, a grasping pose $\mathbf{X}_g$ should be estimated for each object $\mathbf{o} \in \mathcal{O}$. An important assumption is that

Assumption 1: There is no prior knowledge of the objects $\mathcal{O}$ (e.g., no shape model, weight distribution, center of mass, friction coefficients, stable grasp configurations, etc.).

We wish to learn to predict $\mathbf{X}_g$ directly from data $\mathcal{P}$. As there are infinite poses for a given object $\mathbf{o}$, we constrain $\mathbf{X}_g$ so that our 3D CNN learns effectively as follows:

Constraint 1: The grasping pose $\mathbf{X}_g \in SE(3)$ is constrained such that the grasping direction $\hat{\delta}$ is one of $N_\delta = 6$ directions (top, left, left-front, front, right-front, and right), and the wrist orientation $\hat{\omega}$ is one of $N_\omega = 4$ orientations ($0^\circ, 45^\circ, 90^\circ$, and 135°).

Fig. 2 depicts six grasping directions and four wrist orientations. The rationale behind this grasping direction and wrist orientation is in [28] which has shown that grasps orthogonal to objects' principal axes tend to be more stable than randomly sampled grasps. Discretization of grasping orientation is common in CNN-based grasping approaches since CNNs perform better on classification rather than regression problems [19], [21]. Although we chose six grasping directions and four wrist orientations in this work, this framework can be easily adapted to a different number of grasping directions and wrist orientations depending on task requirements. In general, the more outputs the CNN has, the more training data is required. The training data amount will increase linearly with respect to the number of outputs, and hence the computation cost will increase linearly as well. However, it was recently shown that most of the energy, about 80% of the entire computational effort, is consumed by the convolution layers [29]. If we add additional approaching directions to the output layers, it will change only the last fully connected layer. Therefore, the computation cost will be increased at worst sub-linear.

Algorithm 1: 3D CNN Object Grasping.

Data: point cloud $\mathcal{P}$, 3D CNN model $\mathcal{N}$
Result: the set of grasp poses $\mathcal{X} \subset SE(3)$

$\mathcal{S} \leftarrow$ `PlanarSegmentation(`$\mathcal{P}$`)`
for $\mathbf{s} \in \mathcal{S}$ **do**
  $\mathcal{G} \leftarrow$ `Voxelization(`$\mathbf{s}$`)`
  $\mathbf{p}(\delta) \leftarrow \mathcal{N}$`.FeedForward(`$\mathcal{G}$`)`
  $\hat{\delta} \leftarrow \arg\max_{\delta \in \mathbb{N}^+} \mathbf{p}(\delta)$
  $\mathcal{G}' \leftarrow$ `VoxelTransformation(`$\mathcal{G}$, $\hat{\delta}$`)`
  $\mathbf{p}(\omega) \leftarrow \mathcal{N}$`.FeedForward(`$\mathcal{G}'$`)`
  $\hat{\omega} \leftarrow \arg\max_{\omega \in \mathbb{N}^+} \mathbf{p}(\omega)$
  $\mathbf{t} \leftarrow$ `VoxelCoordinates(`$\mathcal{G}$, $\hat{\delta}$`)`
  $\mathbf{X}_g \leftarrow \begin{pmatrix} \mathrm{Rot}(\hat{\delta}, \hat{\omega}) & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \in SE(3)$
  $\mathcal{X} \leftarrow \mathcal{X} \cup \{\mathbf{X}_g\}$

IV. Proposed Approach

A. System Overview

Our grasping system is composed of one Baxter robot and two soft hands attached to its end effectors as shown in Fig. 1. A depth sensor is affixed to the upper body of the robot looking down on the table. The flow of our grasping system is described in Fig. 3. When there are objects on the table, our system obtains a point cloud $\mathcal{P}$ and finds a set of segmented object point clouds $\mathcal{S}$ by removing the planar background in $\mathcal{P}$[1]. Each segmented point cloud $\mathbf{s} \in \mathcal{S}$ is then voxelized to a 3D voxel grid $\mathcal{G} \in \mathbb{Z}^{N_g \times N_g \times N_g}$ where each voxel in the grid is either -1 (not occupied) or 1 (occupied) and N_g is the edge length of the cubic voxel grid. During the voxelization, the point cloud is aligned to the lower center of $\mathcal{G}$. Given $\mathcal{G}$, our 3D CNN model determines the most likely grasping direction $\hat{\delta}$ and wrist orientation $\hat{\omega}$, and the chosen grasp is then executed with our soft hand robot manipulator. Algorithm 1 explains the details of the grasping prediction procedure. The algorithm takes the point cloud $\mathcal{P}$ and the trained 3D CNN model $\mathcal{N}$ as inputs and returns the set of grasp poses $\mathcal{X} \subset SE(3)$ for each segmented object cloud $\mathbf{s} \in \mathcal{S}$, i.e., $|\mathcal{X}| = |\mathcal{S}|$. The direction $\hat{\delta}$ and orientation $\hat{\omega}$ determine the most likely rotation of the grasp pose $\mathbf{X}_g$, while the translation $\mathbf{t}$ of $\mathbf{X}_g$ (i.e., wrist location) is estimated via the voxel coordinates contacting with $\hat{\delta}$. One may sample the voxel along the principal axis of the voxel grid, but picking the center voxel has proven to be effective in our system.

B. 3D Convolutional Neural Network

To determine appropriate grasping directions and wrist orientations given an input point cloud, we train a 3D convolutional neural network (CNN). Inspired by [25], our 3D CNN model is composed of convolution, pooling, and dense layers. The architecture of our model is shown in Fig. 4. The input layer is a $32 \times 32 \times 32$ 3D voxel grid $\mathcal{G}$ which is voxelized from the raw 3D point cloud. There are two convolution layers where the first and second layers have 32 filters

[1]While the tabletop manipulation assumption for object segmentation is considered in this work, our pipeline can easily accommodate advanced segmentation approaches, such as [30], [31], in order to relax the tabletop assumption.

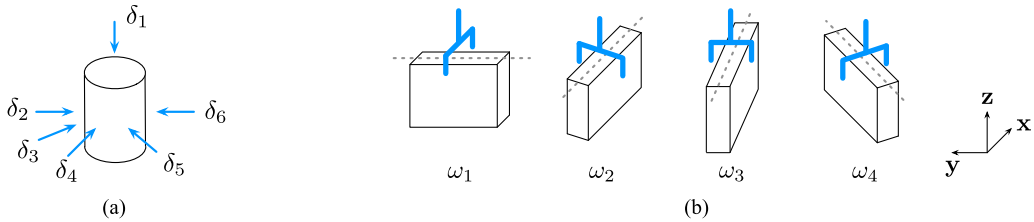

Fig. 2. Grasping directions and wrist orientations. Given an object, our approach discretizes grasping directions to six directions and wrist orientations to four orientations. The total number of grasp orientations is thus $6 \times 4 = 24$. The grasping directions include both top grasp δ_1 as well as side grasps $\delta_2, \ldots, \delta_6$. Each wrist orientation corresponds to the principal axis (dotted gray line) of the box-shaped object. The discretization step is $45°$. Although these grasping directions and wrist orientations are quite coarse, our soft hands are compliant enough to adapt to the discrepancy in object orientation. (a) Grasping directions (b) Wrist orientations

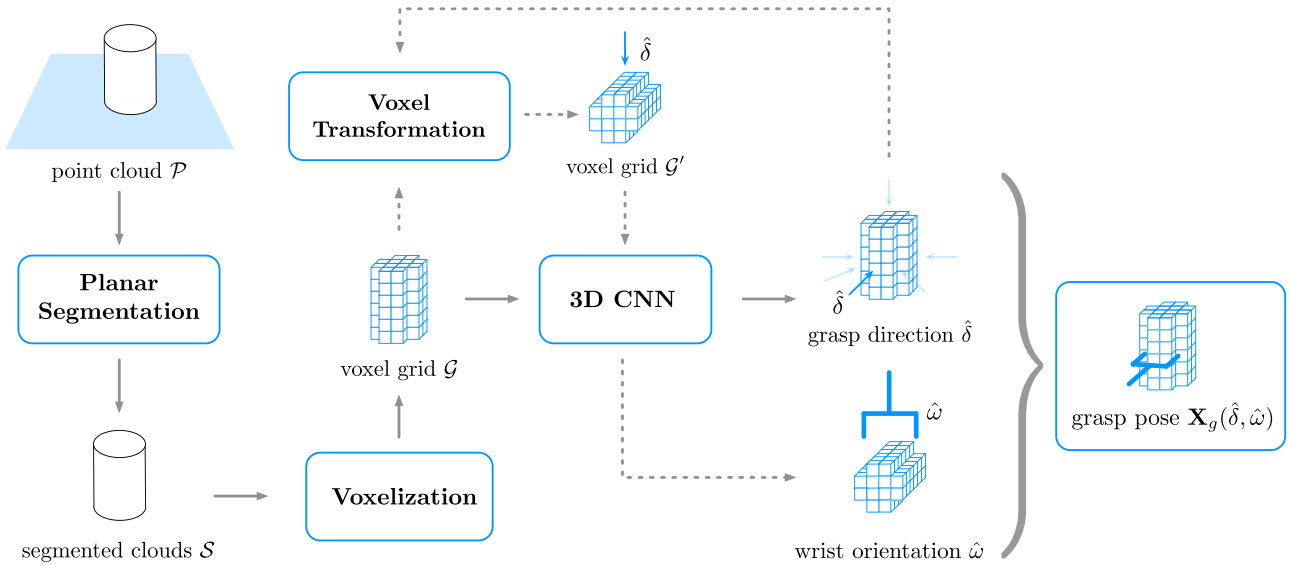

Fig. 3. Grasping Pipeline. The pipeline starts from the raw point cloud $\mathcal{P}$ and segments object clouds $\mathcal{S}$ from $\mathcal{P}$. Each segment cloud is voxelized to generate a voxel grid $\mathcal{G}$. Our approach is two-fold. First, it predicts the most likely grasping direction $\hat{\delta}$ from $\mathcal{G}$ (solid arrows). Second, given $\hat{\delta}$, the voxel grid is transformed so that the chosen direction $\hat{\delta}$ is from the top of the transformed voxel grid $\mathcal{G}'$. The 3D CNN then estimates the most likely wrist orientation $\hat{\omega}$ (dotted arrows). Finally, the chosen grasping direction $\hat{\delta}$ and wrist orientation $\hat{\omega}$ determine the rotation part of the grasp pose, and the translation part of the pose is determined by the contacting voxel with $\hat{\delta}$.

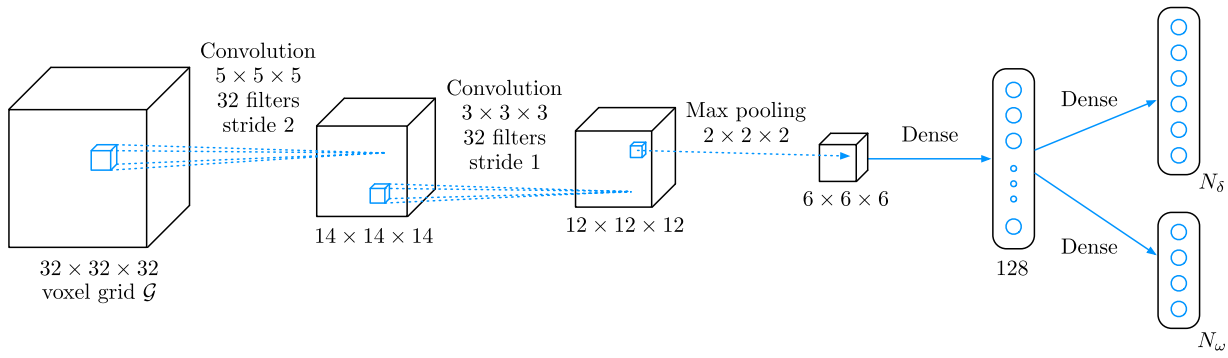

Fig. 4. 3D CNN Architecture. Our network is composed of two convolution layers, one max pooling layer, and two dense (fully connected) layers. The input layer is the voxel grid $\mathcal{G}$ of $32 \times 32 \times 32$ size, in which each voxel grid has either -1 (unoccupied) or 1 (occupied). The output layer returns the probabilities of N_δ and N_ω classes. In our problem, $N_\delta = 6$ directions and $N_\omega = 4$ orientations are considered. Details of convolution and max pooling layers are described in each layer.

of $5 \times 5 \times 5$ and $3 \times 3 \times 3$ size, respectively. After the convolution layers, the data is fed into the max pooling layer of $2 \times 2 \times 2$ followed by two dense layers, 128 and $N_\delta + N_\omega$ each. Unlike the model in [25], the output layer of our model is activated via the sigmoid function instead of the softmax function because the output should be $N_\delta + N_\omega$ independent probabilities, not the probability distribution over N_δ grasping directions and N_ω wrist orientations (i.e., $0 \leq p(\delta_i), p(\omega_j) \leq 1$ where $i = 1, 2, \ldots, N_\delta, j = 1, 2, \ldots, N_\omega$, not $\sum_{i=1}^{N_\delta} p(\delta_i) = 1$ and $\sum_{j=1}^{N_\omega} p(\omega_j) = 1$). Hence, our loss function is defined by

the binary cross-entropy instead of the categorical cross-entropy. We use the binary voxel grid in the input layer, which has only a binary state (occupied or unoccupied) in each voxel, as [25] reported that the performance difference between binary, hit, and density grid for object recognition tasks is negligible. Although we designed this model for grasping with soft hands, we believe that this approach is general. So it should be applicable to arbitrary robot end effectors, hard and soft hands and parallel jaw grippers, by adjusting N_δ and N_ω depending on the compliance of the end effector.

C. Soft Hand Control

Once the grasp pose is determined by the 3D CNN, the robot arm with our soft hand approaches a target object and grasps it. Our soft hand in action is shown in Fig. 1, and its detailed design and fabrication methods are explained in [27]. Each hand has four soft fingers controlled by a set of four external pneumatic actuators. Two parts connect these fingers to the wrist of the Baxter robot arm; one part connects three fingers in a row to the wrist, and the other part connects a single finger opposing them. The wrist has a linear sliding actuator which controls the distance between the two parts. In total, there are five control inputs for one soft hand. While it is possible to learn to control these 5D control parameters, it requires substantial grasp data for training. Instead, we employ a fixed two-state control policy which has open and close states[2]. This simpler control leverages a main benefit of soft hands, namely their innate compliance.

D. Training With Grasping Data

To prepare the object grasping dataset, we chose 10 training objects as shown in Fig. 5a. Each data entry has a voxel grid converted from a partial point cloud, and its associated ground truth label of grasping success obtained by executing with our soft hands.[3] We have collected 719 labeled data entries for the 10 training objects. We further augment the training dataset by transforming the voxel grids, such as by mirroring and translating, and the total number of data entries is 21,570.[4] The ground truth labels were also adjusted when the voxel grids were mirrored. The network was trained with the `adadelta` optimizer.

V. Experiments

To evaluate the effectiveness of our approach in object grasping tasks, we run a set of comparative evaluations. Five approaches, including our approach, are compared. 1) **3DCNN**, our approach described in this letter, predicts the probability of grasping directions and wrist orientations via our 3D CNN.

Fig. 5. Training and testing objects. We use the 10 training objects to train our CNN and evaluate its performance on the 10 test objects. Note the shape differences between the training and test objects. (a) Training objects (b) Test objects

2) **RAND** randomly chooses one of the grasping directions and wrist orientations instead of estimating them from the CNN. It serves as a baseline showing the effectiveness of the soft hands without the use of visual perception. 3) **SVM** is an approach using Support Vector Machine (SVM). A voxel grid is flattened and used as a ($32 \times 32 \times 32 = 32,768$ dimensional) feature vector. As there are multiple feasible grasping directions and wrist orientations for a given voxel grid, a set of multiple binary SVM classifiers was separately trained on each grasping direction and wrist orientation. In this experiment, $N_\delta + N_\omega = 10$ SVM classifiers were trained on the training dataset. 4) **PCA** is an approach using Principal Component Analysis (PCA). Given a voxel grid, this approach estimates the first principal component (PC) which is often aligned with the principal axis of the voxel grid. If the PC is upright (i.e., the orientation of PC is more than 45° from the ground surface), its grasping direction is one of the side grasps ($\delta_2, \delta_3, \ldots, \delta_6$). Otherwise, its grasping direction is the top grasp δ_1. In the former case, the voxel grid is transformed as shown in Fig. 3, and the wrist orientation is determined by the $\mathbf{x}, \mathbf{y}$ values of the PC, which are corresponding to the wrist orientation as shown in Fig. 2b. This is an example of human engineered approaches. Since the PCA is a purely geometric approach, it does not require a training phase and predicts only the best grasping direction and wrist orientation. 5) **FCN**, an approach based on a fully connected network (FCN), has two hidden layers followed by two dropout layers for regularization. This is a baseline highlighting the performance difference between convolutional networks and multi-layer neural networks.

A. Grasp Pose Prediction

In this experiment, we evaluate the accuracy of grasp pose prediction in a test dataset. The test dataset is composed of 638

[2]In the open state, there is no air pressure applied to the soft fingers and the distance between one and three fingers is the longest. In the close state, maximum air pressure is applied to the fingers and the distance between fingers is the shortest.

[3]Each ground truth label has two binary vectors $\mathbf{l}_\delta \in \mathbb{R}_+^{N_\delta}, \mathbf{l}_\omega \in \mathbb{R}_+^{N_\omega}$ where $\mathbf{l}_\delta(i) = 1, \mathbf{l}_\omega(j) = 1$ if δ_i, ω_j are successful grasps. Otherwise, $\mathbf{l}_\delta(i) = 0, \mathbf{l}_\omega(j) = 0$.

[4]For each voxel grid, we flipped left and right followed by 14 translations (4 translations in $\mathbf{x}$ and $\mathbf{y}$ $(-2, -1, 1, 2)$ and 6 translations in $\mathbf{z}$ $(1, 2, \ldots, 6)$). In total, $719 \times 2 \times (1 + 14) = 21,570$ training voxel grids were obtained.

TABLE I
GRASPING PREDICTION ACCURACY ON THE TEST DATASET

	RAND	**PCA**	**SVM**	**FCN**	**3DCNN**
Grasping Direction (%)	26.67	96.55	93.09	100.00	97.60
Wrist Orientation (%)	31.67	97.49	80.84	72.73	99.77

voxel grids from the 10 test objects and their corresponding grasping direction and wrist orientation labels. For each approach, if the chosen grasping direction and wrist orientation belong to the labels, they are regarded as an accurate prediction, otherwise inaccurate. As the **RAND** and **PCA** approaches only return one hypothesis for a given input, for fair comparison the best hypothesis is chosen for the other approaches and compared with the labels. The grasping accuracy on the test dataset is reported in Table I. Among the five approaches, the **RAND** approach reports the worst performance in terms of accuracy. The random choice of grasping direction is slightly better than 25% since the expected chance is $\frac{1}{N_\delta}$ where $N_\delta = 6$ and some objects allow multiple grasping directions due to their symmetry. The **PCA** works reasonably well, but we noticed that it returns a wrong prediction when the partial voxel grid does not give a clue to its complete shape. The **SVM** is worse than the **PCA**, in particular in its wrist orientation. The **FCN** shows perfect prediction in grasping direction, but it turns out that the **FCN** was overfitted to top grasping direction δ_1. Since all examples in both the training and test datasets allow δ_1 direction, it always predicts the top direction as the highest probability rather than considering the side grasps. Moreover, its wrist orientation prediction is the second worst among the five approaches. By comparing the **FCN** and **3DCNN**, our 3D CNN is much more capable of predicting the correct wrist orientation, although both approaches use deep neural networks. We ascribe this to its 3D structure reasoning. Whereas the **FCN** simply treats the voxel grids as real valued features, the **3DCNN** examines the geometric structures of the voxel grids with learned 3D voxel filters. This difference leads to the significant distinction in the prediction accuracy. The **PCA** is the second best approach among the five approaches, but we will see in the next section how it degrades with noise and occlusions.

B. Robustness to Noise and Occlusions

Since voxel grids are obtained via segmentation, it is common to have unexpected noise, wrong segments, or occlusions. In this section, we compare the robustness of the five approaches with respect to noise and occlusions. To this end, we add artificial noise to voxel grids or randomly remove some voxel planes to simulate occlusions. The prediction accuracies of the five approaches with noise and occlusions are reported in Fig. 6. For statistically meaningful results, we ran 30 trials for each noise level and calculated the mean and standard deviation of the prediction outcome.

For the grasping direction, both the **SVM** and **PCA** approaches are increasingly inaccurate as the number of noise voxels and the number of occluded voxel planes increase. In particular, the **PCA** is substantially affected by the noise voxels. Since the **PCA** mainly relies on the principal axis of objects to reason about the grasping direction, a few number of noise voxels are critical to the approach. The **SVM** is relatively less sensitive than the **PCA**, but the accuracy of the **SVM** monotonously decreases as noise and occlusions increase. The **FCN** is not affected by noise or occlusions because the grasping direction of the **FCN** is overfitted to the top grasp, while the **3DCNN** approach is slightly disrupted by occlusions.

For the wrist orientation, the **3DCNN** approach clearly outperforms all other approaches. The **FCN** is not encouraging for predicting wrist orientation, and it consistently performs worse than the **SVM** baseline. The **PCA** is seriously affected by noise, and when occlusions are severe its prediction is even worse than guessing randomly, **RAND**. From this evaluation, we notice that the **3DCNN** approach is more robust than the other approaches. We attribute the superior performance of the **3DCNN** to the hierarchical structure of CNN wherein the convolution with the learned filters effectively suppress noise voxels and the amalgamation of multi-layer responses enables our approach to predict robustly even with serious occlusions.

C. Object Grasping With a Soft Robot Hand

In this experiment, we run an object grasping experiment in which the goal is to pick up a given object on the table. We placed each object on the table with a random location and pose. The robot system and its experiment setting is shown in Fig. 1. If the system can grasp and lift the test object for more than 3 seconds, it is regarded as *success*. If the system cannot grasp the object or the object slips from the hand within 3 seconds, it is counted as *failure*. While in the previous experiments each approach chose one best grasp pose, in this experiment each approach examines a set of grasp poses whose probability is over a certain threshold value $\tau_p = 0.5$. The set of grasp poses is sorted in decreasing order of their probability. The system tries the best grasp pose first to check for feasibility. If the best grasp pose is not feasible due to the kinematic constraints of the robot, it tries the next best pose until it is able to find a valid trajectory plan. We have performed 10 trials for each object with varying locations and orientations. Since there are 10 test objects and 5 prediction approaches, the total number of grasping trials for this experiment was $10 \times 10 \times 5 = 500$.

Fig. 7 presents the grasping success rate of the five approaches on the test objects, and the average grasping accuracy of the approaches are reported in the rightmost bars. Depending on the type and shape of the objects, the grasping rate of these approaches varies. However, the **3DCNN** approach clearly outperforms the other approaches in terms of success rate. The **3DCNN** approach achieves 87% successful grasping for the previously unseen objects, while the **RAND** approach shows about 15% chance of successful grasping. The performance of the **PCA** and **SVM** are similar at about 62%, while the **FCN** shows slightly inferior performance. Unlike the previous experiments, this experiment considers further challenges, such as kinematic constraints and workspace limitations of robot arms and the feasibility of trajectory plans. As the **RAND** and **PCA** return only

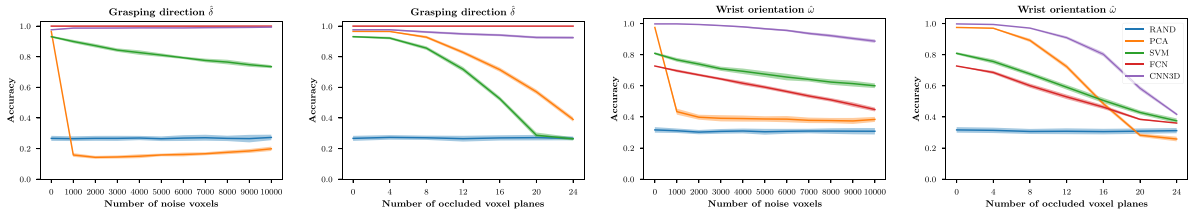

Fig. 6. Prediction accuracy with respect to noise and occlusions. The prediction accuracies of grasping direction $\hat{\delta}$ and wrist orientation $\hat{\omega}$ for five approaches are reported with different degrees of noise and occlusions. We added artificial random noise voxels to the set of test voxel grids. We randomly removed some consecutive voxel planes to mimic occlusions. The solid lines represent means and the shaded areas depict standard deviation.

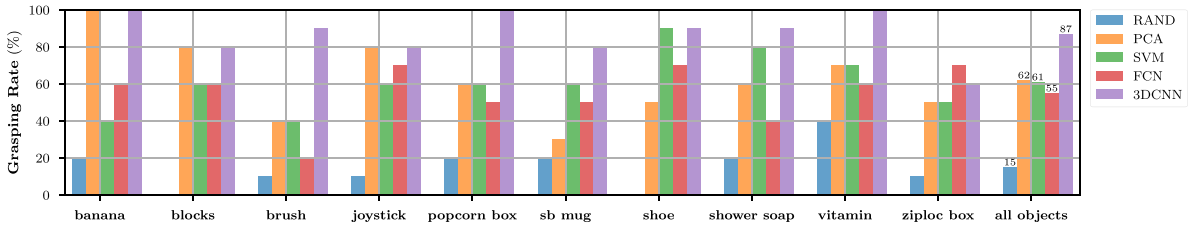

Fig. 7. Grasping success rates using a real robot. The successful grasping rates of five approaches on the 10 test objects. The plot clearly shows the effectiveness of our approach, **3DCNN**, which achieves 87% of successful grasping for previously unseen objects.

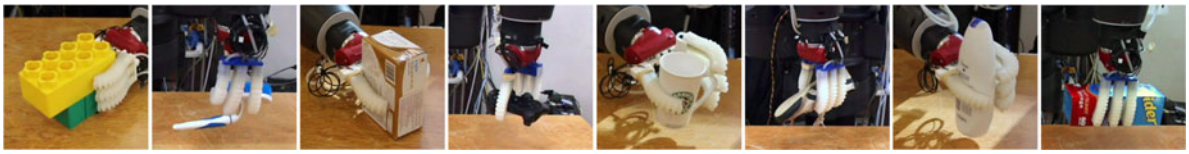

Fig. 8. Successful grasps of the **3DCNN** approach. Our approach reliably grasps various objects even if these objects are unknown to the robot. The 3D CNN model generalizes for these previously unseen objects and enables our soft hands to approach them using correct directions and wrist orientations for grasping.

Fig. 9. Unsuccessful grasps of the **RAND** approach. Though our soft hands are compliant and adaptable to a certain degree of discrepancy, random trials often result in poor grasping as shown here. It clearly shows that an appropriate grasp pose is crucial even for these compliant soft hands.

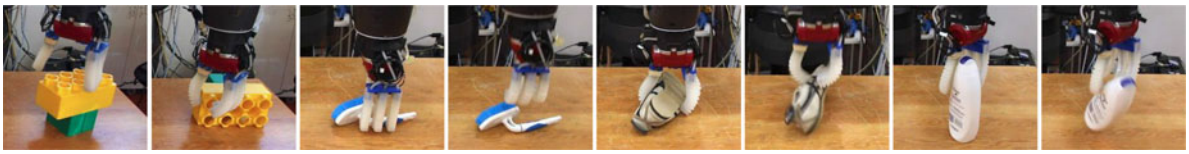

Fig. 10. Unsuccessful grasps of the **3DCNN** approach. For each object, two images depict the pre-grasping and post-grasping situations. Although grasping directions and wrist orientations are right, final grasps are unsuccessful due to either the offsets in the gripper locations or challenging object poses.

one grasp pose, they are rather handicapped if their chosen grasp poses are infeasible due to these constraints. The **FCN** approach is also limited not only by its inaccurate wrist orientation prediction but also by its overfitted grasping direction, and hence it is easily affected by these constraints. This emphasizes the importance of model capability which is able to generalize to multiple grasping directions, and it turns out that the **3DCNN** approach is more capable of grasping under these constraints.

Some successful grasps of our approach are shown in Fig. 8. Our 3D CNN learns partial-view invariance from the training data and generalizes to new objects. Moreover, we can see the synergy effect between our CNN grasping prediction and soft hands. Thanks to the compliance of the soft hands, the acquisition region of the soft hands for successful grasping is large. This empowers the 3D CNN model to focus on learning the coarsely sampled grasping directions and wrist orientations, not worrying about other grasping parameters such as detailed shapes of objects, minor offsets in hand pose, or more complex hand control. Although the grasping directions and wrist orientations are coarsely discretized, our flexible soft hands can grasp objects with a discrepancy in orientation. We also notice the importance of guided grasping direction and wrist orientation information for the soft hands. Fig. 9 shows some unsuccessful grasps of the **RAND** approach. Even though our soft hands are flexible and compliant, a good enough grasp pose is an important prerequisite for successful grasping. By

comparing the accuracies of the **3DCNN** and **RAND** approaches in Fig. 7, we notice that the 3D CNN improves the grasping performance of the soft hands by 72 percentage points. Some failure cases of our approach are presented in Fig. 10. For each object, two images depict the pre-grasping and post-grasping situations to show how these grasps failed. Note that grasping directions are correct, but the final grasps are unsuccessful due to either the offsets in the gripper locations or challenging object poses.

VI. Conclusion

A deep learning powered grasping approach was presented. A 3D CNN model was trained with the dataset obtained by executing grasping with soft hands. Our soft hands with the 3D CNN model achieved 87% successful grasping on unknown objects, which outperforms the other compared approaches including one other deep neural network baseline. We noticed the synergy between our CNN grasping algorithm and soft hands. Our compliant soft hands were able to perform reliable grasping with the grasp poses determined by our CNN model, and the grasp prediction by the CNN significantly increased the success rate by about 72 percentage points compared to the approach without the 3D CNN-based grasp prediction.

References

[1] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis–a survey," *IEEE Trans. Robot.*, vol. 30, no. 2, pp. 289–309, Apr. 2014.

[2] A. T. Miller and P. K. Allen, "Graspit! A versatile simulator for robotic grasping," *IEEE Robot. Autom. Mag.*, vol. 11, no. 4, pp. 110–122, Dec. 2004.

[3] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, no. 4–5, pp. 705–724, Apr. 2015. [Online]. Available: http://ijr.sagepub.com/content/34/4-5/705

[4] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *Int. J. Robot. Res.*, vol. 35, no. 1–3, pp. 161–185, Jan. 2016. [Online]. Available: http://ijr.sagepub.com/content/35/1-3/161

[5] C. Choi, J. DelPreto, and D. Rus, "Using vision for pre- and post-grasping object localization for soft hands," in *Proc. Int. Symp. Exp. Robot.*, Tokyo, Japan, 2016, pp. 601–612.

[6] A. Gupta, C. Eppner, S. Levine, and P. Abbeel, "Learning dexterous manipulation for a soft robotic hand from human demonstrations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 3786–3793. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7759557/

[7] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3d object grasp synthesis algorithms," *Robot. Auton. Syst.*, vol. 60, no. 3, pp. 326–336, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889011001485

[8] C. Goldfeder, M. Ciocarlie, J. Peretzman, H. Dang, and P. K. Allen, "Data-driven grasping with partial sensor data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2009, pp. 1278–1283. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5354078

[9] J. Weisz and P. K. Allen, "Pose error robust grasping from contact wrench space metrics," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 557–562. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6224697

[10] A. Collet, M. Martinez, and S. S. Srinivasa, "The MOPED framework: Object recognition and pose estimation for manipulation," *Int. J. Robot. Res.*, vol. 30, no. 10, pp. 1284–1306, 2011.

[11] C. Choi and H. I. Christensen, "RGB-D object pose estimation in unstructured environments," *Robot. Auton. Syst.*, vol. 75, pp. 595–613, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889015002158

[12] K. Huebner and D. Kragic, "Selection of robot pre-grasps using box-based shape approximation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2008, pp. 1765–1770. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4650722

[13] M. Przybylski, T. Asfour, and R. Dillmann, "Unions of balls for shape approximation in robot grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 1592–1599. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5653520

[14] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008. [Online]. Available: http://ijr.sagepub.com/content/27/2/157.short

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-w

[16] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 4304–4311. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7139793

[17] J. Varley, J. Weisz, J. Weiss, and P. Allen, "Generating multi-fingered robotic grasps via deep learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 4415–4420.

[18] J. Mahler *et al.*, "Dex-net 1.0: A cloud-based network of 3D objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2016, pp. 1957–1964. [Online]. Available: http://goldberg.berkeley.edu/pubs/icra16-submitted-Dex-Net.pdf

[19] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50 k tries and 700 robot hours," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 3406–3413. [Online]. Available: http://arxiv.org/abs/1509.06825

[20] M. Gualtieri, A. t. Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 598–605.

[21] E. Johns, S. Leutenegger, and A. J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 4461–4468.

[22] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," in *Proc. Int. Symp. Exp. Robot.*, 2016. [Online]. Available: http://arxiv.org/abs/1603.02199

[23] J. Mahler *et al.*, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. Robot.: Sci. Syst.*, 2017.

[24] Z. Wu *et al.*, "3d ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, June 2015, pp. 1912–1920.

[25] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 922–928. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7353481

[26] A. M. Dollar and R. D. Howe, "The highly adaptive SDM hand: Design and performance evaluation," *Int. J. Robot. Res.*, vol. 29, no. 5, pp. 585–597, 2010. [Online]. Available: http://ijr.sagepub.com/content/29/5/585.short

[27] B. S. Homberg, R. K. Katzschmann, M. R. Dogar, and D. Rus, "Haptic identification of objects using a modular soft robotic gripper," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2015, pp. 1698–1705.

[28] R. Balasubramanian, L. Xu, P. D. Brook, J. R. Smith, and Y. Matsuoka, "Human-guided grasp measures improve grasp robustness on physical robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 2294–2301.

[29] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[30] S. Christoph Stein, M. Schoeler, J. Papon, and F. Worgotter, "Object partitioning using local convexity," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 304–311. [Online]. Available: http://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Stein_Object_Partitioning_using_2014_CVPR_paper.html

[31] A. Ecins, C. Fermller, and Y. Aloimonos, "Cluttered scene segmentation using the symmetry constraint," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 2271–2278.