

Pixels to Plans: Learning Non-Prehensile Manipulation by Imitating a Planner

Tarik Tosun*, Eric Mitchell*, Ben Eisner, Jinwook Huh, Bhoram Lee, Daewon Lee, Volkan Isler, H. Sebastian Seung, and Daniel Lee

Abstract—We present a novel method enabling robots to quickly learn to manipulate objects by leveraging a motion planner to generate “expert” training trajectories from a small amount of human-labeled data. In contrast to the traditional sense-plan-act cycle, we propose a deep learning architecture and training regimen called PtPNet that can estimate effective end-effector trajectories for manipulation directly from a single RGB-D image of an object. Additionally, we present a data collection and augmentation pipeline that enables the automatic generation of large numbers (millions) of training image and trajectory examples with almost no human labeling effort.

We demonstrate our approach in a non-prehensile tool-based manipulation task, specifically picking up shoes with a hook. In hardware experiments, PtPNet generates motion plans (open-loop trajectories) that reliably (89% success over 189 trials) pick up four very different shoes from a range of positions and orientations, and reliably picks up a shoe it has never seen before. Compared with a traditional sense-plan-act paradigm, our system has the advantages of operating on sparse information (single RGB-D frame), producing high-quality trajectories much faster than the expert planner (300ms versus several seconds), and generalizing effectively to previously unseen shoes.

I. INTRODUCTION

Despite many advances in robotic manipulation in industrial settings, manipulating general objects in unstructured environments remains challenging. The traditional approach for manipulation relies on the *sense-plan-act* paradigm which decouples these three components [1]. A common example comprises of a camera module that captures camera input and processes it generate an intermediate geometric representation of the object to be manipulated, a trajectory planner which generates a path based on this representation, and a path-following controller that executes the path.

Decoupling these components allows for independent progress in complementary areas. For example, recent advances in object detection and segmentation in images can be directly incorporated in the sensing module. Similarly, state-of-the-art planners can be used for generating trajectories in high-dimensional configuration spaces. However, hand-designing the interface between these components can introduce brittleness at the system level. For example, even though the planner can effectively generate a trajectory when given a complete three-dimensional model and the pose of the object to be manipulated, it might be too difficult for the sensing system to generate a precise 3D model and pose for a given input image.

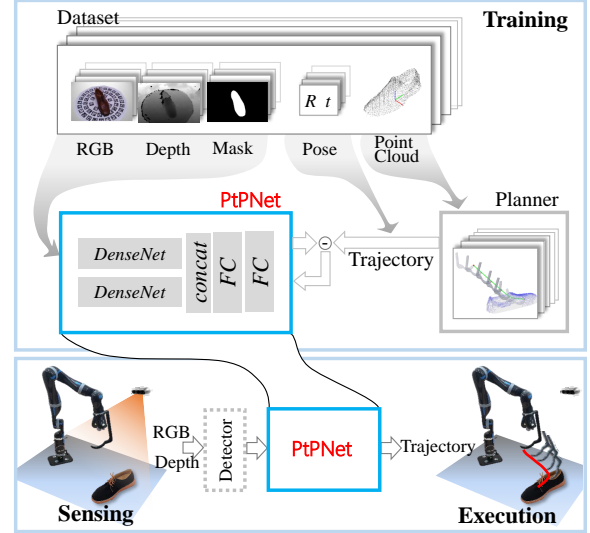


Fig. 1. Conceptual System Overview. During training, we leverage a high-quality point cloud model of a shoe and an RRT motion planner to generate ground truth trajectories for each training example used to train PtPNet. At inference time, we rely on only a single RGB-D camera sample and an attention mechanism/object detector to directly output an end-effector trajectory in the camera frame.

Recent “pixels-to-actions” methods have shown promise in addressing these challenges posed by rigid interfaces. Rather than requiring an explicit intermediate representation, “pixels-to-actions” methods estimate effective actions (in the form of joint angles or torques) directly from raw sensor data, without any explicit intermediate state [2], [3]. However, current pixels-to-actions techniques often suffer from both high sample complexity and brittleness in the presence of deviations from the learning environment, which are particularly significant in robotic applications. So far, these methods have mostly been used in very task-specific environments because learning the dynamics of the task at the controller level and discovering appropriate actions requires numerous training examples. The direct coupling of sensor input to controller actions may be too restrictive and leads to bad generalization performance.

In this paper, we present a new approach in which we train a deep neural network called PtPNet to generate a motion plan (represented as a sequence of trajectory waypoints) from a single RGB-D image. We demonstrate this approach in the context of a tool-based manipulation task, specifically picking up shoes with a hook. Manipulation with a pas-

*These authors contributed equally to this work
All authors are with the Samsung AI Center NY, 123 West 18th Street, New York, New York 10011

sive tool presents a challenging motion planning problem, because it requires moving the tool through a potentially complex sequence of positions and orientations with respect to the object being manipulated, as opposed to selecting a single grasp pose for a gripper. The range of shapes of the shoes in our training and test sets require a variety of qualitatively different hooking trajectories to manipulate them all effectively.

Provided with only partial information about the pose and geometry of a shoe (in the form of a single RGB-D image), PtPNet is trained to closely replicate example trajectories generated by an “expert” motion planner that has access to detailed information about the pose and geometry of the shoe. Core to our training paradigm is a dataset of 3D-scanned shoes that registers many individual RGB-D views to a single dense point cloud for each shoe, a trajectory generation framework that employs a motion planner to generate example trajectories from shoe point clouds, and a robust data augmentation procedure that automatically generates millions of data samples in the form of input image/ground-truth trajectory pairs over the course of training.

PtPNet’s training corpus is based on a relatively small number of 7335 images of 34 shoes, generated by an automated 3d capture system. Using the augmentation of procedure described in Section IV-D we can generate thousands of new images (and matching trajectories) for each original image and thereby effectively increase the training size to millions of images over the course of training. In hardware experiments, we demonstrate that the network successfully generates open-loop trajectories that reliably (89% success over 189 trials) pick up four very different shoes from a range of positions and orientations within the camera view, and generalizes to reliably pick up a shoe it has never seen before. Our results demonstrate that PtPNet has learned to infer from a single RGB-D image *what kind* of shoe it is seeing, *where* the shoe is with respect to the camera, and ultimately *how to move the hook* to capture the shoe. Compared with a traditional sense-plan-act paradigm, it has the advantages of operating on **sparse information** in the partially-observed setting (single RGB-D frame rather than a complete 3D model), producing high-quality trajectories much **faster** than the “expert” planner (300ms versus several seconds), and effective **generalization** to shoes it has never seen before (for which dense 3D information is not available).

Compared to many ‘pixels-to-actions’ paradigms, our method achieves robust manipulation without the need for any training on a real robot and with only very few human-labeled annotations. Our proposed system can also be implemented (as we do) in a way that generalizes across different robotic and camera hardware and conditions, making it desirable for use as a general-purpose manipulation learning method.

II. RELATED WORK

Several approaches have been proposed for applying deep learning to the problem of robotic manipulation. Levine et al. [2] trained a convolutional net to map raw images and

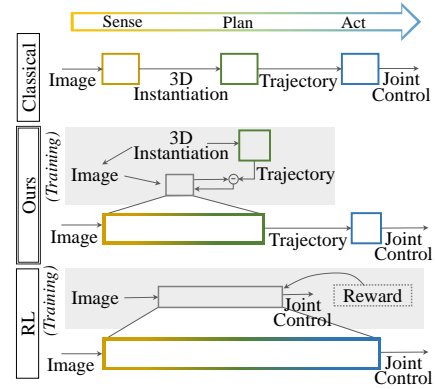


Fig. 2. Conceptual comparison of our method (middle) with the classical sense-plan-act paradigm (top) and end-to-end reinforcement learning methods (bottom). Our method mitigates the need for a structured representation of the scene at inference time (as in the classical paradigm), but produces a generic plan that any controller can follow instead of directly outputting actions for a specific controller.

joint angles to joint torques. The net was trained through policy search guided by a linear-Gaussian controller. The work demonstrated that a pixels-to-torques mapping could be learned, but was limited to manipulation tasks that could be solved by a linear-Gaussian controller.

In later work, Levine et al. [3] trained a convolutional net to predict grasp success given a raw image and a one-step motion plan. At run time, the motion plan at each time step was chosen to maximize the net’s prediction of grasp success. This approach was formalized as a kind of Q-learning [4], and required a large number of manipulation trials to learn an accurate predictor of grasp success. The trials were generated by building an “arm farm” with many robotic arms learning to manipulate in parallel.

Zhang et al. [5] trained a convolutional net to generate a motion plan from raw images and end-effector pose. In this deep imitation learning approach, the target values for the motion plan were provided by virtual reality teleoperation of the manipulator by a human teacher. A potential disadvantage is that collection of large amounts of training data may require significant labor by human teachers.

Here we propose an alternative deep imitation learning approach in which the targets for supervised training are provided by a trajectory planner rather than a human teacher. A similar approach has been applied to autonomous driving [6], but learning to imitate a planner is novel in the domain of manipulation as far as we know. Contemporary planning algorithms such as RRT [7] are quite powerful; however, these algorithms may be slow. As our empirical results will show, training a convolutional net to imitate the planner yields performance that is fast, accurate, and generalizes to objects that were not seen during training. Furthermore the net operates with sparser information (single view RGB-D) than the planner (full 3D collision geometry).

The prior works mentioned above were used to train closed-loop systems that use sensory feedback while generating motions. In the present work, we train an open-loop

system to generate a motion plan using sensory information from a single image. This was done for simplicity; extensions to closed-loop manipulation control will be sketched in the Discussion.

III. PROBLEM STATEMENT AND OVERVIEW

A. Problem Statement

We consider a setup which consists of a manipulator (whose kinematics are known) and an external fixed RGB-D camera. Throughout the paper, \mathcal{R} denotes the base frame of the manipulator, \mathcal{E} denotes the frame of the robot end-effector (which in this case is a hook tool), \mathcal{C} is the camera frame and \mathcal{S} is the shoe frame. The pose of the camera with respect to the robot ${}^{\mathcal{R}}T_{\mathcal{C}}$ is assumed to be known. Given a single RGB-D image of the shoe from the camera, the objective is for the robot to pick up the shoe using the hook. Picking is considered successful if at the end of the robot's motion, the shoe has been lifted completely off the table and hangs stably on the hook, and the shoe and hook are not damaged during the motion.

B. Method Overview

Our method trains a neural network that accepts as input a single RGB-D image of a shoe and outputs a *camera-frame end-effector trajectory* ${}^{\mathcal{C}}W = \{w_1, w_2 \dots w_N\}$ consisting of N waypoints, where each waypoint $w_k = {}^{\mathcal{C}}T_{\mathcal{E}_k}$ defines a pose of the end-effector with respect to the camera. Since the camera-to-robot calibration ${}^{\mathcal{R}}T_{\mathcal{C}}$ is assumed to be known, this camera-frame trajectory can be transformed into the robot frame for execution by the robot: ${}^{\mathcal{R}}W = {}^{\mathcal{R}}T_{\mathcal{C}} {}^{\mathcal{C}}W$.

Our network is trained via imitation learning to closely replicate example trajectories generated by an “expert” motion planner that has access to detailed information about the pose and geometry of the shoe. Figure 1 shows an overview of our training framework. Our dataset consists of dense 3D point cloud models, RGB-D images, masks, and poses of real shoes, generated using a data capture system consisting of a turntable and three Intel Realsense RGB-D cameras (Fig. 4). For each shoe, an “expert” end-effector trajectory ${}^{\mathcal{C}}W^*$ is generated by an RRT motion planner that has access to a 3D model of the end-effector tool, a dense point cloud of the shoe, a desired goal position for the hook within the shoe, and the pose of the shoe in the camera frame.

The images from the capture system coupled with these trajectories provide a core set of training data. The network is provided with an RGB-D image of the shoe as input, and produces a camera-frame end-effector trajectory ${}^{\mathcal{C}}W$ as output, which is then compared with the expert trajectory ${}^{\mathcal{C}}W^*$, and the network is iteratively trained to minimize a loss function which measures the similarity of the two trajectories.

In Section IV, we describe our neural network architecture, example trajectory generation process, and data collection and augmentation procedures in detail. In Section V, we present hardware experiments that benchmark our system in a shoe-picking task, characterize its ability to generalize across shoe poses, and test its ability manipulate shoes that

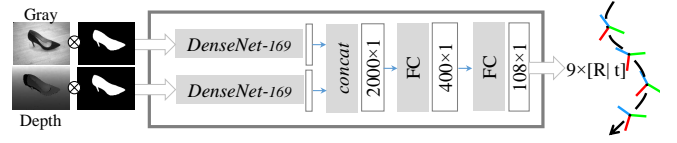


Fig. 3. PtPNet Architecture. From an input consisting of an (I^g, I^d) pair, corresponding to the grayscale and depth images from an RGB-D camera, we compute ${}^{\mathcal{C}}W_{\mathcal{E}} = \{\hat{w}_1, \hat{w}_2, \dots, \hat{w}_N\}$. I^g and I^d are fed into two separate, identical DenseNet-169 instantiations, after which they are combined into several fully-connected layers and then output as a trajectory.

were not in the training set. In Section VI, we discuss lessons learned, and conclude.

IV. METHODS

A. Neural Network Architecture and Training

1) *Architecture*: The fundamental building block of our neural network architecture is the successful DenseNet paradigm introduced in [8], specifically the variant DenseNet-169¹ (a particular instantiation of the DenseNet concept) proposed in [8] and implemented in the deep learning framework PyTorch [9]. DenseNet-169 contains 168 convolutional layers and a single fully connected layer following the learned features. The general DenseNet architecture was chosen as the backbone of our network because it has been shown to be a powerful, parameter-efficient architecture that learns quickly, yielding state-of-the-art results on a variety of computer vision tasks including classification, segmentation, and real-time object detection on mobile devices [10] [11] [12].

The network architecture used in this paper, which we call PtPNet, uses two separate DenseNet-169 modules to process the grayscale and depth measurements from the camera in separate streams, which are ultimately joined by a sequence of two fully connected layers at the end of the network. Weights are *not* shared across these streams as in some “Siamese” architectures [13]. A diagram of this architecture is given in Figure 3.

The input to the network is a pair of aligned grayscale and depth images (I^g, I^d) . The network assumes that both images have been foreground-masked, that is, non-shoe pixels and depth values have been set to zero. Several possible techniques for acquiring such masks exist, including simply estimating a plane and filtering depth points or even training a separate convolutional neural network for object segmentation. We implemented and tested both strategies for this work, and both were effective. In our experiments, we performed shoe segmentation with a convolutional neural network based on UNet [14] that was trained on a small subset (roughly 15%) of our trajectory dataset; we chose to use a multi-scale convolutional network over other segmentation techniques (e.g. plane subtraction from depth image) because its output was substantially more robust to noise

¹Source code of the DenseNet-169 building block is located here: <https://pytorch.org/docs/1.0.0/modules/torchvision/models/densenet.html#densenet169>

from the RGB-D camera, and is robust to non-shoe objects in the frame.

In the forward pass, each image is first processed separately by one of two DenseNet-169 networks to extract 1000 features each, for a total of 2000 features. Two more hidden layers, coupled with the ReLU activation function [15], compute ‘mixed’ features, and a final output layer directly regresses the trajectory plan estimate ${}^C\hat{W} = \{\hat{w}_i : i \in 1, \dots, n\}$ as a single vector $\hat{t} \in \mathbb{R}^{(12N)}$ for N trajectory waypoints. In this paper, we use $N = 9$, but our framework is general to other trajectory lengths. This output vector is interpreted as a sequence of sub-vectors $\hat{t}_i \in \mathbb{R}^{12}$, each corresponding to a homogeneous transform $\hat{w}_i = {}^C T_{\mathcal{E}}$ from the camera coordinate frame to the desired coordinate frame of the end effector of the robot at waypoint i . For each waypoint vector t_i , the first 3 values (t_i^1, t_i^2, t_i^3) represent the desired (x, y, z) position of the robot’s end effector in the camera’s coordinate frame. The last 9 values represent a serialized 3D rotation matrix giving the relative orientation of the robot’s end effector relative to the camera for that waypoint. Learning end-effector trajectories in the camera frame decouples the learned solution to the task from the position of the robot; as long as the camera-to-robot calibration ${}^R T_C$ is known, we can use the trained network with arbitrary robot position and camera positions with **no** retraining. Provided they have similar dexterous workspaces and identical end-effectors, the method should also generalize without retraining to an entirely different robots (there is no notion of a robot during training, only an end-effector trajectory).

2) *Training*: During training, the network is presented with input/output pairs $((I^g, I^d), {}^C W^*)$, where I^g and I^d are grayscale and depth images from the same scene, respectively, and ${}^C W^*$ is the corresponding ground-truth trajectory generated by the RRT motion planner. Ground-truth trajectories are serialized as a single vector, as described in Section IV-A.1.

For each input (I^g, I^d) seen during training, the network makes a prediction ${}^C\hat{W} = \{\hat{w}_i : i \in 1, \dots, N\}$ and a loss is computed from the ground truth trajectory ${}^C W$. The loss for the trajectory estimate ${}^C\hat{W}$ is the weighted sum of the individual trajectory waypoint losses:

$$L_{traj}(\hat{w}, w^*) = \sum_{i=1}^n \alpha_i \ell(\hat{w}_i, w_i^*) \quad (1)$$

This trajectory waypoint loss, given its corresponding ground truth trajectory waypoint w_i^* , is computed by first decomposing each waypoint into its representative translation and rotation sub-components. For any 12 dimensional waypoint vector w , we define the functions $\text{Trans}(w) : \mathbb{R}^{12} \rightarrow \mathbb{R}^3$ and $\text{Rot}(w) : \mathbb{R}^{12} \rightarrow \mathbb{R}^{3 \times 3}$, which extract the position and rotation matrix of a given waypoint relative to the camera coordinate frame, respectively. Using this notation, the trajectory waypoint loss is given as:

$$\ell(\hat{w}_i, w_i^*) = \lambda \ell_T(\hat{w}_i, w_i^*) + \gamma \ell_R(\hat{w}_i, w_i^*) \quad (2)$$

where $\ell_T(\hat{w}_i, w_i^*)$ is the squared Euclidean distance loss:

$$\ell_T(\hat{w}_i, w_i^*) = \|\text{Trans}(\hat{w}_i) - \text{Trans}(w_i^*)\|^2 \quad (3)$$

and $\ell_R(\hat{w}_i, w_i^*)$ is the squared deviation of the product of the predicted rotation matrix and the transpose of the ground truth rotation matrix:

$$\ell_R(\hat{w}_i, w_i^*) = \|\text{Rot}(\hat{w}_i)\text{Rot}(w_i^*)^T - I\|^2 \quad (4)$$

In this work, we use $\lambda = \gamma = 1$ and $\alpha_i = 1$ for all i ; that is, we weight the rotation matrix deviation loss and waypoint coordinate loss equally and weight all individual waypoint losses equally within a trajectory. We train with the Adam optimizer [16] with learning rate 1e-4, batch size 64, and weight decay coefficient 1e-4 for 1000 epochs on 4 NVIDIA 1080 Ti GPUs.

B. Dataset

In order to train P_tPNet, we have compiled a dataset comprised of point cloud models, RGB-D images, masks, and poses of real shoes. The capture setup includes three low-cost RGB-D cameras (Intel RealSense D435²), a controllable turn-table, and an AprilTag [17] pattern board (Fig.4(a)).

The cameras capture images of the exact same scene from different camera poses relative to the target object when the turn-table stops at a certain interval while rotating. After the images are collected, the camera pose of each frame is computed using AprilTags. The depth point cloud of the target is obtained by removing all other points except the target area above the pattern board and by filtering using a voxel-based simplification method. The point cloud can be re-projected onto each RGB image to generate the mask and its pose can be computed with respect to the corresponding camera frame (Fig.4(b)). In this way, each model includes 216 views (5 deg interval \times 3 cameras) recorded in RGB, IR, and depth, as well as the shoe pose for each view. At the moment, the dataset include 45 shoe instance models, 34 of which were used for training P_tPNet.

Note that we do not impose a clean surface structure to build a mesh model which is difficult to obtain dynamically. This approach enables gathering images of real objects with masks and poses without the need for 3D surface models. The capture and labeling process is automated without any manual annotation.

C. Example Trajectories

1) *Trajectory Representation*: We define a *shoe-frame trajectory* ${}^S W$ as a set of N waypoints of the end effector with respect to the shoe frame. Let a waypoint w_k be $[x_k, y_k, z_k, \phi_k, \theta_k, \psi_k]$ in $SE(3)$ space and $w_k \in {}^S W$ for $k \in 1, \dots, N$, where $[x_k, y_k, z_k]$ is a position, and $[\phi_k, \theta_k, \psi_k]$ are Euler angles of the hook. For trajectory generation, we use the Euler angle representation of rotations because it is stable and relatively computationally efficient. For training of P_tPNet, we represent $SO(3)$ as a rotation matrix \mathbf{R}_k since Euler angle representation has an ambiguity due to multiple

²<https://realsense.intel.com/depth-camera/>

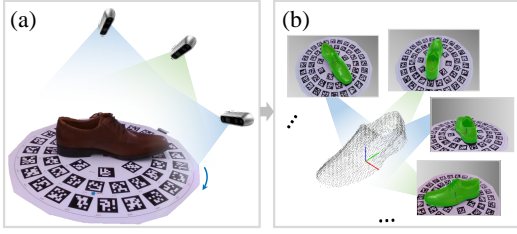


Fig. 4. Training dataset capture procedure. (a) shows the physical collection configuration, with a single shoe mounted on a turntable marked with AprilTags, and three cameras placed at different poses around the turntable. The turntable is rotated a full 360° , with RGB-D images captured at 5° intervals. (b) shows the computational procedure used to extract a complete point cloud of the shoe by synthesizing multiple RGB-D views into a single model.

parameter values for the same rotation representation [18]. In addition, we fix the number of waypoints N as 9 in this paper. The complete trajectory is thus represented by a sequence of (9) 12-dimensional keypoints, or a single 108-dimensional vector.

In order to train PtPNet, we need ground truth end-effector trajectories *in the camera frame* for each grayscale/depth image pair in our dataset. However, because we collect the shoe to camera transform ${}^C T_S$ for each sample of our dataset, we only need to generate a single trajectory for each *shoe* ($N=34$) rather than for each *shoe image* ($N=7335$). Using a motion planner, we generate a **single** example trajectory ${}^S W$ for each shoe in the shoe’s own coordinate frame. For each image pair and corresponding shoe-to-camera transform $((I^g, I^d), {}^C T_S)$ in our dataset, we can then generate an appropriate camera-frame trajectory ${}^C W$ by simply transforming ${}^S W$ into the camera frame: ${}^C W = {}^C T_S {}^S W$. Thus, our dataset is generated from exactly 34 human-labeled annotations.

2) *RRT Trajectory Generation*: To achieve robust shoe hooking across a wide class of shoes (that vary significantly in color, texture, and shape), a robot must necessarily utilize distinct manipulation strategies for sufficiently distinct shoes. For example, the actions required to pick up a high heel are fundamentally different from those required to pick up a sneaker, as the shape and mass distributions are quite different. Fig. 5 demonstrates the distinct trajectories required to successfully hook four different shoes. In this paper, we apply a sampling based planner to generate appropriate hooking trajectories for each shoe based on its point cloud. We define a goal pose inside a shoe manually for the sampling based planning, and we an RRT (Rapidly exploring Random Trees) motion planner to generates a trajectory from a fixed initial pose near the shoe to the specified target goal pose without colliding with the shoe point cloud. We employ the uniform sampling method of Euler angles and distance metric in [18] to effectively sample $SE(3)$. In addition, we use a bidirectional approach and 10% goal biased sampling to improve performance.

Collision checks are computed by approximating the geometry of the hook end-effector as a point cloud, and

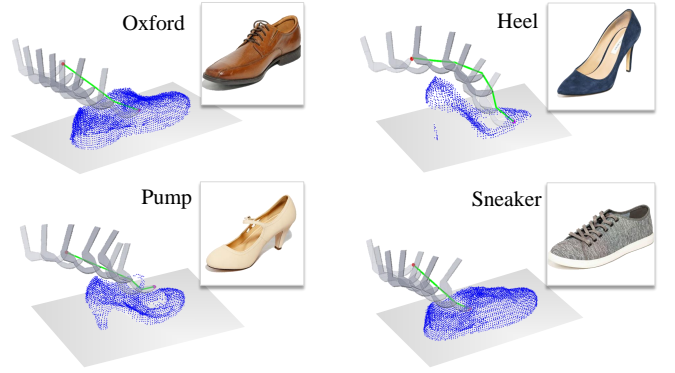


Fig. 5. RRT planning from 3D point clouds. In this figure, we demonstrate qualitatively different hooking trajectories output by the RRT algorithm when run on shoes with different morphologies. Notice how the trajectories approach at different angles depending on what occlusions the shoes themselves introduce.

comparing the minimum distance between the end-effector point cloud and shoe point cloud. Once a path to the goal point is found, one additional waypoint is added directly above the goal point to lift the shoe. This waypoint is $20cm$ higher in the z direction from the goal pose.

Sampling-based planners sometimes generate jerky and unnatural paths, so after planning the RRT we apply a path smoothing heuristic. Possible path smoothing techniques include the shortcut or spline algorithms [19], [20]; we chose to apply the shortcut smoothing method due to its empirically determined effectiveness on our task and its low computational complexity. Functionally, this algorithm repeats 100 iterations in which it randomly chooses two configurations and linearly interpolates between the two if no collisions are detected. After smoothing the trajectory, we choose 9 evenly-spaced waypoints from the trajectory to represent the ground truth trajectory for training.

D. Data Augmentation

When trained solely on raw images obtained by the acquisition system and their corresponding correct trajectories, PtPNet learns a trajectory-generating function that fails to produce successful trajectories when the test case deviates from the training conditions, e.g., when the shoe is not carefully centered in the image or the camera is closer or farther from the shoe than in training by more than a few centimeters. This is to be expected, since the raw dataset contains only images and point clouds of shoes in relatively homogeneous configurations, and thus the only trajectories the net is exposed to are trajectories that hook shoes at a specific distance from the camera and at the center of the image. To mitigate these geometric test-time limitations, we used the detailed 3-dimensional data collected for each shoe to generate new examples during training that span the entire field of view of the camera at a wide range of depths.

First, we define a sample tuple as $(I^g, I^d, {}^C W, {}^C T_S)$, corresponding to the grayscale image, depth image, ground truth trajectory of the tool in the camera frame, and pose

of the shoe in the camera frame, \mathcal{C} , respectively. Applying augmentation to a sample requires applying a single augmentation transformation ${}^{\mathcal{C}'}T_{\mathcal{C}}$ to each attribute of the sample, which is interpreted as re-capturing the sample from a new camera pose \mathcal{C}' . We achieve this by simulating a random rotation about the x and y axes of the camera frame, forming the rotation matrix $\mathbf{R}_{\theta\phi} = \mathbf{R}_y(\phi)\mathbf{R}_x(\theta)$, as well as a random displacement in the direction of the camera frame z axis, Δz . We bound these random parameters such that the all shoes remain within the camera’s field of view. The intuition behind this camera rotation and z -displacement is that it will *shift* and *scale* the image of the shoe without deforming its appearance.

We form the augmentation transformation ${}^{\mathcal{C}'}T_{\mathcal{C}}$ as:

$${}^{\mathcal{C}'}T_{\mathcal{C}} = \begin{bmatrix} \mathbf{R}_{\theta\phi} & \mathbf{t} \\ 0 & 1 \end{bmatrix}, \quad \mathbf{t} = [0 \quad 0 \quad \Delta z]^T$$

Augmenting I^d , ${}^{\mathcal{C}}T_{\mathcal{E}}$, and ${}^{\mathcal{C}}T_{\mathcal{S}}$ thus involves applying a single homogeneous coordinate frame transformation (augmenting the depth image requires projecting the points with the inverse camera projection matrix, K^{-1} , applying the ${}^{\mathcal{C}'}T_{\mathcal{C}}$, and then re-projecting with K).

We approximate the application of this transformation to the image I^g by approximating the scaling effect of the depth shift first and then applying the rotation as a shift in the image plane. To account for the depth offset Δz , we scale the image size by a factor of $\frac{z_0}{z_0 - \Delta z}$, where z_0 is the original z position of the shoe frame in the camera frame ${}^{\mathcal{C}}T_{\mathcal{S}}$. We then crop or pad the image symmetrically so that it is the same size as the original. To implement the shifting augmentation, we shift the image by $i = \frac{\theta}{fov_v}h$ and $j = \frac{\phi}{fov_u}w$ pixels, where h and w are the height and width of I^g in pixels, and fov_u and fov_v are the field of view of the camera in radians in the u and v directions.

When trained on a dataset augmented in this fashion, trajectory prediction becomes significantly more robust, and hook-success increases dramatically across camera poses and shoe poses. To make the case for augmentation more concrete, we report that the network trained without augmentation reliably fails when the shoe position deviates from the center of the image by more than roughly $.1w$ pixels (where w is the width of the image in pixels) or the shoe is closer to or farther away from the camera than the narrow range of z distances present in the training set (roughly 55-65 cm). As evidenced in the experiments in the following section, PtPNet trained with augmentation performs similarly no matter where the shoe is in the image (as long as it is completely visible) and in a much wider band of z values (roughly 40-100cm in our experiments).

V. EXPERIMENTS

We perform four experiments in order to motivate our method and characterize its performance on the shoe hooking task. In Experiment 1, we compare the generalization ability of RRT-generated trajectories to that of the learned network when the shoe is varied, demonstrating that as with shoes themselves, there is no *one size fits all* shoe-hooking

	Oxford	Heel	Pump	Sneaker
Oxford Trajectory	100%	0%	0%	10%
Heel Trajectory	0%	100%	80%	0%
Pump Trajectory	0%	0%	100%	0%
Sneaker Trajectory	80%	0%	0%	100%
PtPNet Trajectory	90%	90%	90%	80%

TABLE I. Experiment 1: Comparison of Training Trajectories. In this table, we show how effective each predefined trajectory and PtPNet are at hooking each type of shoe. In general, the static trajectories only work for the shoe they were generated from, while PtPNet demonstrates good performance on all shoes. that was designed for a given shoe is effective at picking up that shoe, and ineffective at picking up other shoes.

trajectory. In Experiment 2, we characterize on the general level of performance of our learned system when both shoe and shoe pose are varied. In Section V-C, we evaluate the consistency of performance of PtPNet as shoes are moved to different positions in the camera frame, characterizing the effectiveness of our camera-view augmentation strategy in generalizing network performance to shoes not in the center of the frame. Finally, in Experiment 4, we examine the network’s ability to generalize to shoes that are not sitting upright on the table by introducing an artificial roll angle to the shoe’s pose.

The experimental setup (Figure 7) consists of an RGB-D camera, a table, and a robot equipped with a hook tool. All experiments measure shoe picking performance; a shoe picking attempt is considered successful if the shoe is lifted completely off the table and remains hanging on the hook at the end of the robot motion.

A. Experiment 1: Comparison of Training Trajectories

This experiment tests whether qualitatively different hooking motions are actually necessary to hook different shoes. For the set of four shoes shown in Figure 5 and their four corresponding hooking trajectories generated by the RRT, we test whether each trajectory is able to pick up each shoe. In each test run, one of the four shoes is placed in a known, fixed location on the table, and one of the four RRT trajectories is executed by the robot at that location to attempt to pick up the shoe. Each shoe and trajectory combination was run 10 times. For comparison, PtPNet was also run 10 times per shoe with the shoe in the center of the camera frame.

Table I shows the experimental results. As expected, each RRT trajectory succeeds 100% of the time for the shoe for which it was designed, and is generally unsuccessful at picking other dissimilar shoes, indicating that different trajectories are indeed needed to pick up different shoes. In contrast, PtPNet succeeds about 90% of the time on all shoes.

B. Experiment 2: Shoe Hooking Task Performance

This experiment tests the overall performance of PtPNet at the shoe picking task. Once again, each of the four shoes shown in Figure 5 is tested. The Heel, Pump,

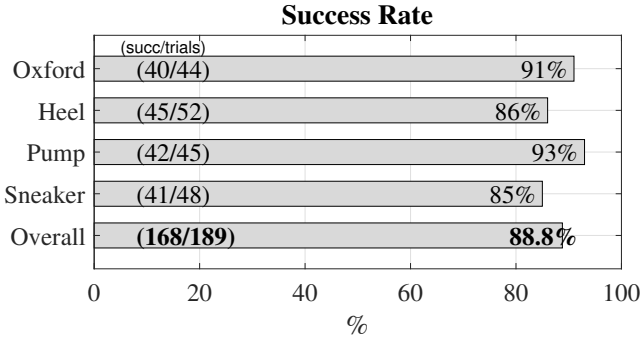


Fig. 6. Experiment 2: Overall success rate of PtPNet at the shoe picking task. The position and orientation of each shoe is varied across the table between trials. The Oxford shoe was held out of the training set (PtPNet has never seen it before). PtPNet successfully picks up all four of these diverse shoes at least 85% of the time.

and Sneaker were all included in the data used to train PtPNet, while the Oxford shoe was withheld as a test set shoe (PtPNet has never seen this shoe or its corresponding RRT-generated trajectory before).

In each trial, the test shoe is placed in a random orientation at one of the five positions on the table shown in Figure 7. A single RGB-D image of the table is captured, segmented, and provided to the network as input. The network outputs a hooking trajectory (sequence of waypoints) in the camera frame cW_E . This trajectory is then transformed to the robot frame via \mathcal{R}_{T_C} , and executed by the robot to attempt to pick up the shoe. Each shoe is tested in each position about ten times.

Figure 6 shows the results for each shoe averaged over all trials. The network consistently hooks all shoes at least 85% of the time, and successfully generalizes to the previously-unseen Oxford shoe, picking it up 91% of the time.

C. Experiment 3: Generalization Across Pose

To test how well the network generalizes to shoes in different positions and orientations within the field of view of the camera, we compare its performance against what we refer to as the ‘pose estimation system’, which explicitly estimates shoe pose and then attempts to pick up the shoe using a human-selected appropriate hooking motion pre-generated by our RRT planner.

To generate picking trajectories, our RRT planner requires a full 3D point cloud of a shoe, as well as a human-specified goal pose for the hook within the shoe. Consequently, the input to the network (a single-view RGB-D image of a shoe) is not an adequate input to run the RRT planner on live data. Instead, the pose estimation-based system attempts to pick up shoes by estimating the pose of the shoe on the table and then running the (human-selected) pre-computed RRT trajectory for that shoe, taking advantage of the fact that the RRT trajectories are defined with respect to the shoe body frame. The position of the shoe is estimated by computing the centroid of the (segmented) shoe point cloud in the table frame. The orientation is estimated by computing the major axis of the shoe point cloud via PCA.

Position	1	2	3	4	5	Overall
Network	100%	83%	100%	83%	100%	93.2%
Pose Estimation	83%	100%	100%	50%	67%	80%

TABLE II. Experiment 3: Generalization Across Poses. We compare the success rate of two algorithms (PtPNet and a pose estimation-based method) at picking the Sneaker across 5 different starting positions. PtPNet outperforms the pose-estimation method, even though the pose-estimation method automatically uses the correct (human-selected) RRT-generated trajectory to attempt to pick the sneaker.

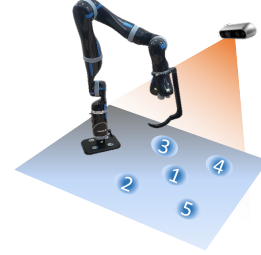


Fig. 7. Experimental Setup: Five pre-defined locations on the table. During testing, we placed shoes with a random z-axis rotation at one of these five positions, to maintain consistency across testing runs.

The experimental procedure is the same as the shoe-picking task, except we select a single shoe (the Sneaker), testing only the ability of the pose estimation and learned systems to pick it up when it was placed in each of 5 poses.

Table II compares the performance of the pose estimation system to the network. We see that PtPNet slightly outperforms the pose estimation system in most positions, indicating that the PtPNet’s estimation of shoe pose is more accurate than that of the pose estimation system.

D. Experiment 4: Generalization to Untrained Shoe Orientations

PtPNet is trained only with images of shoes sitting upright on a flat surface; any variations in shoe roll present in the dataset occur because of the variation in the angle of inclination of the three cameras used. However, our framework is not inherently limited to picking up shoes in these conformations. In this experiment, we test the ability of the PtPNet to generalize to shoe orientations that are not well represented in the training data. Specifically, we characterize the sensitivity of performance at the shoe picking task (Sec. V-B) when the shoe is resting unevenly on a small block that introduces a rotation about the shoe’s heel-toe axis, testing roll angles of 0, 17, and 32 degrees. Table III shows the results of this experiment. The network successfully picks up shoes with approximately 17 degrees of roll, but fails once the roll angle approaches 30 degrees.

VI. DISCUSSION

The results of our experiments suggest that imitation learning using a kinematic motion planner as a supervision signal is a robust, data-efficient method for single-view estimation of end-effector trajectories for the examined manipulation

Rotation (Degrees)	0°	17°	32°
Success Fraction	41 / 48	28 / 36	2 / 35
Percentage	85%	78%	6%

TABLE III. Experiment 4: Generalization to Untrained Shoe Orientations. The success rate of PtPNet when asked to lift shoes that have been rotated about their principal axis (and thus are no longer perpendicular to the plane). The larger the roll angle, the worse the performance.

task, using only 34 human-generated annotations. Further, we posit that the method should generalize to other manipulation tasks in which collision avoidance is important but only partial state observations are available. Our method could also be applied for closed-loop control where the network is applied to subsequent image inputs during the motion execution and used to modulate the initial planned trajectory.

However, the learned system has several clearly defined, repeatable failure modes. For example, we observed that system performance was sensitive to the quality of the segmentation: poor shoe segmentation results often lead the system to failure. There are two primary avenues to addressing this issue, including a) increasing the robustness of PtPNet to bad segmentations and/or b) eliminate the need for masking. Significant progress can be made on both fronts through data augmentation, assuming that the ground truth masks for the dataset image pairs (I^g, I^d) exist. However, segmentations serve as an effective attention mechanism, allowing the network to estimate trajectories for scenes containing multiple objects of interest by selectively masking them. Further, as demonstrated in Experiment 4, the network has only a limited ability to generalize to unseen shoe poses. This is largely due to the range of camera positions used during data collection and could be mitigated by acquiring more complete coverage of the hemisphere of possible camera positions during data collection. In general, when the system fails, it fails because the end-effector collides with the outside of the shoe (a near miss).

VII. CONCLUSION

We presented a self-supervised system which can generate a shoe hooking trajectory directly from a single RGB-D image. Our system is trained using 3D models which are used for generating training instances composed of input images and corresponding hooking trajectories. The hooking trajectories in turn are generated using an RRT planner which takes the 3D model along with goal points as input. The only manual labelling required for our method is the annotation of each of the 34 shoe data bundles (all images and pointclouds from all angles) with a goal point for the calculated RRT trajectories. We also presented a novel augmentation method which was used to generate millions of images over the course of 1000 training epochs from the 7335 training tuples of these 34 data bundles. Hardware experiments demonstrate that the network can successfully hook different types of shoes across a wide range of poses.

Our results suggest several possible paths for future work. A more comprehensive test containing broader manipulation tasks and classes of objects would help better demonstrate the

generality of our method. In addition, more experimentation is needed to determine an effective trade-off between neural network architecture size and computational demands; for true real-time trajectory estimation, a more compact architecture backbone is needed.

REFERENCES

- [1] S. Chitta, E. G. Jones, M. Ciocarlie, and K. Hsiao, "Perception, planning, and execution for mobile manipulation in unstructured environments," *IEEE Robotics and Automation Magazine, Special Issue on Mobile Manipulation*, vol. 19, no. 2, pp. 58–71, 2012.
- [2] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *CoRR*, vol. abs/1504.00702, 2015.
- [3] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *CoRR*, vol. abs/1603.02199, 2016.
- [4] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.
- [5] T. Zhang, Z. McCarthy, O. Jowl, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 ICRA*. IEEE, 2018.
- [6] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. Theodorou, and B. Boots, "Agile autonomous driving using end-to-end deep imitation learning," *RSS*, 2018.
- [7] S. M. LaValle and J. J. Kuffner, "Randomized kinodynamic planning," *The International Journal of Robotics Research*, vol. 20, no. 5, pp. 378–400, 2001.
- [8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [9] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
- [10] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense convolutional networks for efficient prediction," *CoRR*, vol. abs/1703.09844, 2017.
- [11] S. Jégou, M. Drozdal, D. Vázquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," *CoRR*, vol. abs/1611.09326, 2016.
- [12] R. J. Wang, X. Li, S. Ao, and C. X. Ling, "Pele: A real-time object detection system on mobile devices," *CoRR*, vol. abs/1804.06882, 2018.
- [13] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *International Conference on Machine Learning*, 2015.
- [14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.
- [15] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814.
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
- [17] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *IEEE ICRA*, 2011.
- [18] J. J. Kuffner, "Effective sampling and distance metrics for 3d rigid body path planning," in *ICRA*. Citeseer, 2004, pp. 3993–3998.
- [19] K. Hauser and V. Ng-Thow-Hing, "Fast smoothing of manipulator trajectories using optimal bounded-acceleration shortcuts," in *2010 IEEE ICRA*. IEEE, 2010.
- [20] R. Geraerts and M. H. Overmars, "Creating high-quality paths for motion planning," *The International Journal of Robotics Research*, vol. 26, no. 8, pp. 845–863, 2007.

