

SAPIEN: A SimulAted Part-based Interactive ENvironment

Fanbo Xiang¹ Yuzhe Qin¹ Kaichun Mo² Yikuan Xia¹ Hao Zhu¹
Fangchen Liu¹ Minghua Liu¹ Hanxiao Jiang³ Yifu Yuan⁵ He Wang² Li Yi⁴
Angel X. Chang³ Leonidas Guibas² Hao Su¹

¹UC San Diego ²Stanford University ³Simon Fraser University ⁴Google Research ⁵UC Los Angeles

<https://sapien.ucsd.edu>

Abstract

Building home assistant robots has long been a pursuit for vision and robotics researchers. To achieve this task, a simulated environment with physically realistic simulation, sufficient articulated objects, and transferability to the real robot is indispensable. Existing environments achieve these requirements for robotics simulation with different levels of simplification and focus. We take one step further in constructing an environment that supports household tasks for training robot learning algorithm. Our work, SAPIEN, is a realistic and physics-rich simulated environment that hosts a large-scale set for articulated objects. Our SAPIEN enables various robotic vision and interaction tasks that require detailed part-level understanding. We evaluate state-of-the-art vision algorithms for part detection and motion attribute recognition as well as demonstrate robotic interaction tasks using heuristic approaches and reinforcement learning algorithms. We hope that our SAPIEN can open a lot of research directions yet to be explored, including learning cognition through interaction, part motion discovery, and construction of robotics-ready simulated game environment.

1. Introduction

To achieve human-level perception and interaction with the 3D world, home-assistant robots must have the capability to use perception to interact with 3D objects [11, 59]. For a robot to help put away groceries, it must be able to open the refrigerator by locating the door handle, pulling the door and fetching the target objects.

One direct way to address the problem is to train robots by interacting with the real environment [29, 4, 26]. However, training robots in the real world could be very time consuming, costly, unstable, and potentially unsafe. Moreover, a slight perturbation in hardware or environment setup can result in different outcomes in the real world, thus inhibiting reproducible research. Researchers, therefore,

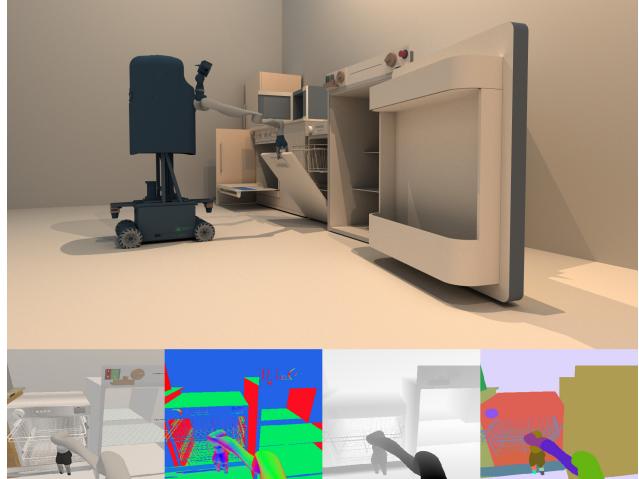


Figure 1: **Robot-object Interaction in SAPIEN.** We show the ray-traced scene (top) and robot camera views (bottom): RGB image, surface normals, depth and semantic segmentation of motion parts, while a robot is learning to operate a dishwasher.

have long been pursuing simulated environments for tasks such as navigation [41, 56, 42, 1, 3, 54, 14, 56] and control [24, 40, 48, 10].

Constructing simulated environments for robot learning with transferability to the real world is a non-trivial task. It faces challenges from four major aspects: 1) The environment needs to reproduce the real-world physics to some level. As it is still infeasible to simulate real-world physics exactly, any physical simulator needs to decide the level-of-details and accuracy it operates on. Some approximate physics by simulating rigid bodies and joints [35, 48, 10]; some handle soft deformable objects [48, 10]; and others simulate fluid [48, 43]. 2) The environment should incorporate the simulation of real robots, being able to reproduce the behaviors of real robotics manipulators, sensors and controllers [34]. Only this can enable seamless transfer to the real-world after training. 3) The environment needs

Environment	Level	Physics	Rendering	Tasks	Interface
Habitat [42]*	Scene	Static+	Real Photo	Navigation, Vision	Python, C++
AI2-THOR [25]*	Scene-Object	Dynamic	Unity	Navigation+, Vision	Python, Unity
OpenAI Gym MuJoCo [2]	Scene-Object	Dynamic	OpenGL(fixed)	Learning, Robotics	Python
RLBench[22]	Scene-Object	Dynamic	V-REP, PyRep	Learning, Vision, Robotics	Python, V-REP
SAPIEN	Scene-Object-Part	Dynamic	Customizable	Learning, Vision, Robotics	Python, C++

Table 1: **Comparison to other Simulation Environments.** Habitat [42] is a representative for navigation environments, which include Gibson [56, 55], Minos [41]; they primarily use static physics but are starting to add interactions very recently. AI2-THOR [25] is a representative for game-like interactive environments; these environments usually support navigation with limited object interactions. OpenAI Gym [2] and RLBench [22] provide interactive environments, but the use of commercial software limits their customizability.

to produce physically accurate renderings to mitigate the visual domain gap. 4) Most importantly, the environment requires sufficient content, scenes and objects for the robot to interact with, since data diversity is always critical for training and evaluating learning-based algorithms. The content also determines how much we shall address challenges in the previous tasks: data with soft objects such as cloth requires deformable body simulation; translucent objects require special rendering techniques, and specific robot requires a specific interface.

Existing environments achieve these requirements for robotics simulation with different levels of simplification and focus. For example, OpenAI Gym [2] provides an interactive and easy-to-use interface; Gibson [56] and AI Habitat [42] use photorealistic rendering for semantic navigation tasks. A more detailed discussion of popular environment features can be found in Sec 2. These environments can support the benchmarking and training of down-stream tasks such as navigation, low-level control, and grasping. However, from the perspective of tasks, there still lacks environments that target at object manipulation of daily objects, a basic skill of household robots. In a household environment, a great portion of daily objects are articulated and require manipulation: bottles with caps, ovens with doors, electronics with switches and buttons. Notably, RLBench [22] (unpublished) provides well-defined robotics tasks and realistic controller interface with detailed manipulation demonstration, but it lacks diversity in its simulated scenarios.

We take one step further in constructing an environment that supports the manipulation of diverse articulated objects. Our system, SAPIEN, is a realistic and physics-rich simulated environment that hosts a large set for articulated objects. At the core of SAPIEN are three main components: 1) SAPIEN Engine, an interaction-rich and physics-realistic simulation environment integrating PhysX physical engine and ROS control interface; this engine supports accurate simulation of rigid body and joint constraints for simulation of articulated objects. 2) SAPIEN Asset, including PartNet-

Mobility dataset, which contains 14K movable parts over 2,346 3D articulated models from 46 common indoor object categories, richly annotated with kinematic part motions and dynamic interactive attributes; 3) SAPIEN Renderer, with both fast-frame-rate OpenGL rasterizer and more photorealistic ray-tracing options. We demonstrate that our SAPIEN enables a large variety of robotic perception and interaction tasks by benchmarking state-of-the-art vision algorithms for part detection and motion attribute recognition. We also show a variety of robotic interaction tasks that SAPIEN supports by demonstrating heuristic approaches and reinforcement learning algorithms.

2. Related Work

Simulation Environments. In recent years, there has been a proliferation of indoor simulation environments primarily designed for navigation, visual recognition and reasoning [41, 54, 56, 42, 1]. Static environments, based on synthetic scenes [54] or real-world RGB-D scans [1] and reconstructions [41, 56, 42], are able to provide images that closely resembles reality, minimizing the domain gap in the visual aspect. However, they usually offer very limited or no object interactions, failing to capture the dynamic and interactive nature of the real world.

In order to allow for more interactive features to the environment, researchers leverage partial functionalities of game engines or physics engine to provide photorealistic rendering together with interactions [52, 33, 25, 37, 3, 57, 13]. When agents interact with objects in these environments, it is via high level state changes triggered by explicit commands (*e.g.* “open refrigerator”), or proximity (*e.g.* refrigerator door opens when the robot or robot arm is next to the trigger region). In addition, the underlying physics is often over-simplified such as direct exertion of force and torque. While they enable research on high-level object interactions, they cannot close the gap between high-level instructions and the low-level dynamics for not including accurate simulation of articulated robots and objects by design. This limits the use of such simulators for learning

Dataset	#Categories	#Models	#Motion Parts
Shape2Motion[51]	45	2,440	6,762
RPM-Net[58]	43	949	1,420
Hu <i>et al.</i> [18]	-	368	368
RBO*[30]	14	14	21
Ours	46	2,346	14,068

Table 2: **Comparison of Articulated Part Datasets.**
*RBO is collected in real-world with long video sequences.

of detailed low-level robot-object interactions.

Finally, there are environments that integrate full-featured physics engines. These environments are favored in continuous control and reinforcement learning tasks. OpenAI Gym [2], RLLAB [12], DeepMind Control Suite [47] and DoorGym [49] integrate MuJoCo physical engine to provide RL environments. Arena [45], a platform that supports multi-agent environments, is built on top of Unity [23]. PyBullet [10], a real-time physics engine with Python interface, powers a series of projects focusing on robotics tasks [60, 27]. Gazebo [24], a high-level visualization and modeling package, is widely used in robotics community [31, 20]. Recently, RL Bench [22], a benchmark and physical environment for robot learning, uses V-REP [40] as the backend to provide diverse tasks for robot manipulation. Our environment, SAPIEN engine, is directly based on the open-source Nvidia PhysX API [35], which has comparable performance and interface with PyBullet, avoiding the unnecessary complication introduced by game engine infrastructures, or any barriers from commercial software such as MuJoCo and V-REP. Table 1 provides a brief summary of several representative environments.

One bottleneck of these robotic simulators is their limited rendering capability, which causes a gap between simulation and the real world. Another constraint of many of these environments, including RL Bench [22] and DoorGym [49], is that they are very task-centric, designed to work for only a few predefined tasks. Our SAPIEN simulator, equipped with 2,346 3D interactive models from 46 object categories and flexible rendering pipelines, provides robot agents a virtual environment for learning a large set of complex, diverse and customizable robotic interaction tasks.

Simulation Content. Navigation environments typically use datasets providing real-world RGB-D scans [56, 6, 46], and/or high-quality synthetic scenes [44]. Simulation environments that leverage game engines [52, 33, 3, 13, 25] come with manually designed or procedurally generated game scenes. For environments with detailed physics and reinforcement learning support [2, 47, 12], they usually support very few scenarios with simple objects and robot

agents. Notably, RL Bench [22] provides a relatively large robot learning dataset with varied tasks. To address the lack-of-content problem, our work provides a large-scale simulation-ready dataset, PartNet-Mobility dataset, that is constructed from 3D model datasets including PartNet [32] and ShapeNet [7].

There are also shape part datasets with part articulation annotations. Table 2 summarizes recent part mobility datasets. The RBO dataset [30] is a collection of 358 RGB-D video sequences of humans manipulating 14 objects which are reconstructed as articulated 3D meshes. The meshes have detailed part motion parameters and have realistic textures. Other datasets annotate 3D synthetic CAD models with articulation information. Hu *et al.* [18] introduced a dataset of 368 mobility part articulations with diverse types. RPM-Net [58] provides another dataset with 969 objects and 1,420 mobility units. Shape2Motion [51] provides a dataset of 2,440 objects and 6,762 movable parts for mobility analysis, but it does not provide RGB textures and motion limits that hinders physical simulation. Compared to these datasets, our dataset contains comparable number of objects (2,346), but with much more movable part annotations (14,068). Besides, our models have textures and motion range limits, which are crucial for the dataset to be simulatable.

3. SAPIEN Simulation Environment

SAPIEN aims to integrate state-of-the-art physical simulators, modern graphics rendering engines, and user-friendly robotic interfaces into a unified framework (Figure 2), to support a diverse set of robotic perception and interaction tasks. We develop the environment with C++ for efficiency and provide Python wrapper API for ease-of-use at the user end. Below we detailedly introduce the three main components: SAPIEN engine, SAPIEN asset and SAPIEN renderer.

3.1. SAPIEN Engine

We use the open-source Nvidia PhysX physical engine to provide detailed robot-object interaction simulation. The system provides Robot Operating System (ROS) supports that are easy-to-use for end-stream robotic research. We provide both synchronous and asynchronous modes of simulation to support reinforcement learning training and robotics tasks.

Physical Simulation. We choose PhysX 4.1 [35] to provide rigid body kinematics and dynamics simulation, since it is open-source, simplistic, and provides functionalities designed for robotics. To simulate articulated bodies, we provide 3 different body-joint systems: kinematic joint system, dynamic joint system, and PhysX articulation. The kinematic joint system provides kinematic objects with

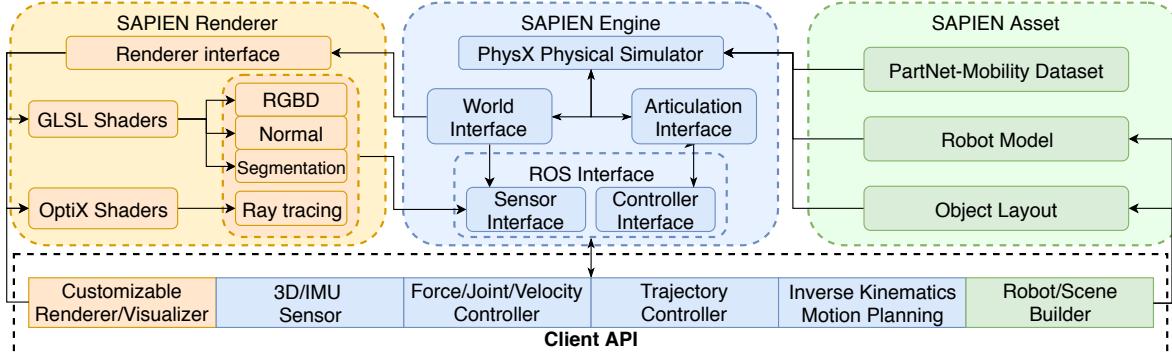


Figure 2: **SAPIEN Simulator Overview.** The left box shows SAPIEN Renderer, which takes custom shaders and scene information to produce images such as RGB-D and segmentation. The middle box shows SAPIEN Engine, which integrates PhysX simulator and ROS control interface that enables various robot actions and utilities. The right box shows SAPIEN Asset, which contains the large-scale PartNet-Mobility dataset that provides simulatable models with part-level mobility.

parent-child relations, suitable for simulating very heavy objects that are not affected by small forces. Dynamic joint systems use PhysX joints to drive rigid bodies towards constraints, suitable for simulating complicated objects that do not require accurate control. PhysX articulation is a system specifically designed for robot simulation. It natively supports accurate force control, P-D control and inverse dynamics with the cost of relatively low speed.

ROS Interfaces. Robot Operating System (ROS) [39] is a generic and widely-used framework for building robot applications. Our ROS interface bridges the gap between ROS and physical simulator, as well as provides a set of high-level APIs for interacting with robots in the physics world. It supports three levels of abstractions: direct force control, ROS controllers and motion planning interface.

In the lowest level control, forces and torques are directly applied on joints, similar to OpenAI Gym [2]. This control method is simple and intuitive, but rather difficult to transfer to real environments, since real-world dynamics are quite different from the simulated ones, and the continuous nature present in real-robots are fundamentally different from the discretized approximation in simulations. For high-level control, we provide joint space and Cartesian coordinate space control APIs. We build various controllers (Figure 2) based upon [8] and implement standard interface. A typical use case is to move the robot arm to a desired 6-DoF pose with specific path constraints. Thus, at the highest level, we provide motion planning support based on the popular MoveIt framework [9], which can generate motion plans that effectively move the robot around without collision.

Synchronous and Asynchronous Modes. Our SAPIEN Engine (see Figure 2 middle) can support both synchronous and asynchronous simulation modes. In synchronous mode, the simulation step is controlled by the client, which is common in training reinforcement learning algorithm [2]. For

example, the agent receives observations from simulated environments and uses a customized policy model, often a neural network, to generate the corresponding action. Then the simulation runs forward for a step. In this synchronous mode, the simulation and client algorithms are integrated together.

However, for real-world robotics, the simulation and client response need to be asynchronous [24] and separated. The simulation should run independently, like the real world, while the client uses the same API as a real robot to interact with the simulation backend. To build such a framework, we create multiple sensors and controllers following the ROS API. After simulation starts, the client receives information from sensors and uses the controller interface (see Figure 2) to command robots via TCP/IP communication. The timestamp is synchronized from simulation to the client side, acting as a proxy for the real-world clock time. Under the framework, the simulated robots can use the same code as their real counterparts because most real robot controllers and sensors have exactly the same interface as our simulator API. This provides one important advantage: it enables robot researchers to migrate between simulated robots and real robots without any extra setup.

3.2. SAPIEN Asset

SAPIEN Asset is our simulation content, shown in the right box in Figure 2. It contains the large-scale ready-to-simulate PartNet-Mobility dataset, the simulated robot models and scene layouts.

PartNet-Mobility Dataset. We propose a large-scale 3D interactive model dataset that contains over 14K articulated parts over **2,346** object models from **46** common indoor object categories. All models are collected from 3D Warehouse^{*} and organized as in ShapeNet [7] and

^{*}<https://3dwarehouse.sketchup.com/>

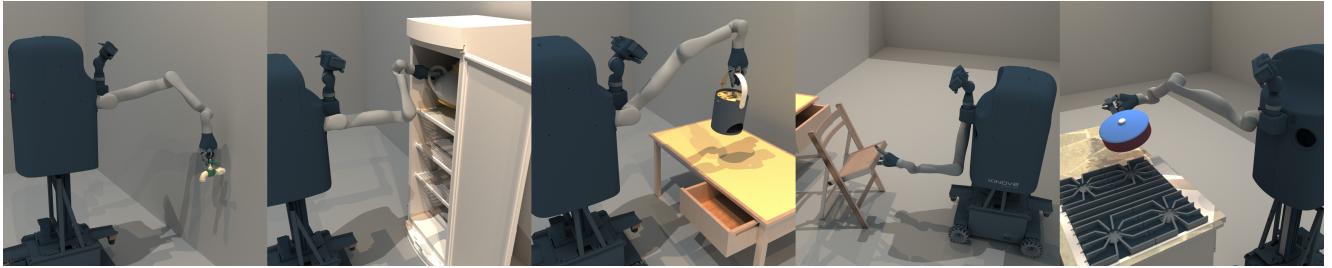


Figure 3: **SAPIEN Enables Many Robotic Interaction Tasks.** From left to right, we show five examples: faucet manipulation, object fetching, object lifting, chair folding, and object placing.

	All	Bottle	Box	Bucket	Cabinet	Camera	Cart	Chair	Clock	Coffee	DishWsh.	Dispenser	Door	Eyegls	Fan	Faucet
#M	2,346	57	28	36	345	37	61	80	31	55	48	57	36	65	81	84
#P	14,068	114	94	74	1,174	341	232	1,235	106	374	112	162	103	195	172	228
	Chair	Fridge	Globe	Kettle	Keybrd	Knife	Lamp	Laptop	Lighter	MicWav	Monitor	Mouse	Oven	Pen	Phone	Pliers
#M	26	44	61	29	37	44	45	56	28	16	37	14	30	48	17	25
#P	58	118	130	66	3,593	149	165	112	86	85	93	61	214	97	271	59
	Pot	Printer	Remote	Safe	Scissors	Stapler	Stcase	Switch	Table	Toaster	Toilet	TrashCan	USB	Washer	Window	
#M	25	29	49	30	47	23	24	70	101	25	69	70	51	17	58	
#P	53	376	1,490	202	94	69	101	195	420	116	229	208	103	144	195	

Table 3: **Statistics of PartNet-Mobility Dataset.** #M and #P shows the number of models and movable parts respectively.

PartNet [32]. We annotate 3 types of motions: hinge, slider, and screw, where hinge indicates rotation around an axis (*e.g.* doors); slider indicates translation along an axis (*e.g.* drawers), and screw indicates a combined hinge and slider (*e.g.* bottle caps, swivel chairs). For hinge and slider joints, we annotate the motion limit (*i.e.* angles, lengths). For screw, we annotate the motion limits and whether the 2 degrees of freedom are coupled. Each joint has a parent and a child, and the collection of connected bodies and joints is called an articulation. We require the joints of an articulation to follow a tree structure with a single root, since most physical simulator handles tree-structured joint system well. Next, for each movable part, we assign a category-specific semantic label. Table 3 summarizes the dataset statistics. Please see the supplementary for more details about the data annotation pipeline.

SAPIEN Asset Loader Unified Robot Description Format (URDF) is a common format for representing a physical model. For each object in the SAPIEN Asset, including PartNet-Mobility models and robot models, we provide an associated URDF file, which can be loaded in simulation. For accurate simulation of contact, we decompose meshes into convex parts [28, 19]. We randomize or manually set the physical properties, *e.g.* friction, damping, density, to appropriate ranges. For robot models, we also provide C++/Python APIs to create a robot piece by piece to avoid complications introduced by URDF.

3.3. SAPIEN Renderer

SAPIEN Renderer, shown in the left box of Figure 2, renders simulated scenes with OpenGL 4.5 and GLSL

shaders, which are exposed to the client application for maximal customizability. By default, the rendering module uses a deferred lighting pipeline to provide RGB, albedo, normal, depth, and segmentation from camera space, where lighting is computed with OrenNayar diffuse model [53] and GGX specular model [50]. Our customizable rendering interface can suit special rendering needs, and even allow completely different rendering pipelines. We demonstrate this by replacing the fast OpenGL framework with our ray tracer coded with Nvidia OptiX [36] to produce physically accurate images at the cost of rendering time (see Figure 1).

3.4. Profiling Analysis

Our SAPIEN engine can run at about 5000Hz on the manipulation task we will describe in Sec. 4.2 and can render at about 700Hz with OpenGL mode. Tests were performed on a laptop with Ubuntu 18.04, on 2.2 GHz Intel i7-8750 CPU and an Nvidia GeForce RTX 2070 GPU.

4. Tasks and Benchmarks

We demonstrate the versatile abilities of our simulator by demonstrating robotic perception and interaction tasks.

4.1. Robotic Perception

SAPIEN simulator, equipped with the PartNet-Mobility dataset, provides a platform for several robotic perception tasks. In this paper, we study the tasks of movable part detection and part motion estimation, which are two important vision tasks supporting downstream robotic interaction.

		Cabinet			Table						Faucet	Fan	All			
Algorithm	Inputs	rot. door	body	drawer	trans. door	drawer	body	wheel	door	caster	switch	base	spout	rotor	frame	mAP
Mask-RCNN [16]	2D (RGB)	62.0	94.2	66.4	27.7	54.3	88.0	3.4	6.3	0.0	52.5	47.9	99.7	54.4	67.5	53.0
	2D (RGB-D)	61.7	93.0	63.0	26.3	58.6	89.9	1.4	13.2	0.0	52.1	55.8	98.9	39.4	67.4	52.8
PartNet InsSeg [32]	PC (XYZ)	20.6	65.9	35.1	9.8	15.7	71.3	1.7	1.0	0.0	34.4	55.9	64.2	50.9	74.8	36.1
	PC (XYZRGB)	17.4	64.3	23.6	5.0	16.4	81.8	1.3	2.0	1.0	29.9	64.1	78.0	42.0	63.5	37.1

Table 4: **Movable Part Detection Results.** (AP% with IoU threshold 0.5) 2D and PC denote 2D images and point clouds as different input modalities for the two algorithms. We show the detailed results for four objects categories and summarize the mAP over all categories. See supplementary for the full table.

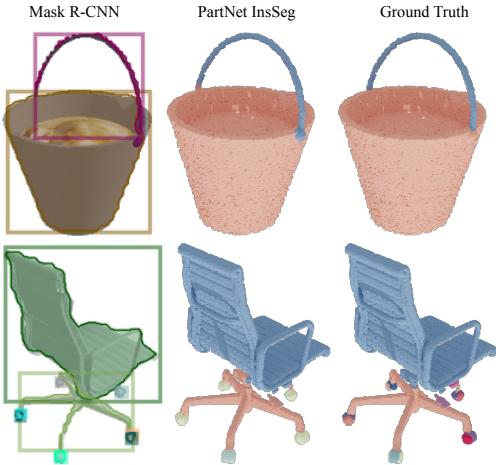


Figure 4: **Movable Part Detection Results.** The left column shows the results of Mask R-CNN [16], where each bounding box indicates a detected movable part. The middle and the right columns show the results of PartNet-InsSeg [32] and the ground truth point clouds respectively, where different parts are in different color.

Movable Part Detection Before interacting with objects by parts, robotic agents need to first detect the parts of interest. Therefore, we define the task of movable part detection as follows. Given a single 2D image snapshot or 3D RGB-D partial scan of an object as input, an algorithm should produce several disjoint part masks associated with their semantic labels, each of which corresponds to an individual movable part of the object.

Leveraging the rich assets from the PartNet-Mobility dataset and the SAPIEN rendering pipeline, we evaluate two state-of-the-art perception algorithms for object or part detection in literature. Mask R-CNN [16] takes a 2D image as input and uses a region proposal network to detect a set of 2D part mask candidates. PartNet-InsSeg [32] is a 3D shape part instance segmentation approach that uses PointNet++ [38] to extract geometric features and proposes panoptic segmentation over shape point clouds.

We render each object in the PartNet-Mobility dataset

into RGB and RGB-D images from 20 randomly sampled views, with resolution 512×512 . The camera positions are randomly sampled over the upper hemisphere to ensure space coverage. Simple ambient and directional lighting without shadows are provided for RGB rendering. With known camera intrinsics, we lift the 2.5D RGB-D images into 3D partial scans for PartNet-InsSeg experiments. We use all 2,346 objects over 46 categories from the PartNet-Mobility dataset for this task. We use 75% of data (1,772 shapes) for training and 25% (574 shapes) for testing. For quantitative evaluation, we report per-part-category Average Precision (AP) scores as commonly used for object detection tasks and average across all part categories to compute mAP for each algorithm.

Table 4 shows the quantitative results of Mask R-CNN on RGB and RGB-D settings and PartNet-InsSeg on the XYZ (depth-only) and XYZRGB (RGB-D images) settings. We observe that both methods perform poorly on detecting small parts (*e.g.*, table wheel and table caster), and the phenomenon is less severe for object categories that have relatively balanced sizes (*e.g.*, fan and faucet). Small movable parts (*e.g.*, button, switch, and handle) often play critical roles in robot-object interaction, and will demand more well-designed algorithms in the future. Figure 4 visualizes the Mask-RCNN and PartNet-InsSeg part detection results on two example RGB-D partial scans.

Motion Attributes Estimation Estimating motion attributes for articulated parts gives strong priors for robots before interacting with objects. In this section, we perform the motion attributes estimation task that jointly predicts the motion type, motion axis, and part state for articulated parts.

We consider two types of rigid part motions: 3D rotation and translation. Some parts, such as bottle cap, may have both rotation and translation motions. For translation motions, we use a 3-dim vector to represent the direction. For rotation motions, we parameterize the outputs as two 3-dim vectors to specify rotation axis direction and a pivot point on the axis. We define relative positions of the articulated part with respect to its semantic rest positions as part states. For example, the rest position for drawers

Setting	Algorithm	H acc.	S acc.	H_o err (m)	H_a err ($^\circ$)	S_a err. ($^\circ$)	door err. ($^\circ$)	drawer err. (m)
RGB-D	ResNet50	95.5%	95.5%	0.168	18.9	6.35	14.4	0.0645
RGB-pc	PointNet++	95.4%	95.5%	0.195	18.5	7.75	20.8	0.0918

Table 5: **Motion recognition results.** H acc. and S acc. denotes classification accuracy for hinge and slider respectively. H_o err. denotes average distance from predicted hinge origin to ground truth axis. H_a/S_a denotes average hinge/slider angle difference from predicted axis to ground truth. door err. is average angle difference from predicted door pose to ground truth. drawer err. is average length difference from predicted drawer pose to ground truth.

and doors is when they are closed. However, defining part rest states has intrinsic ambiguities. For example, round knobs with rotation symmetry do not present a detectable rest position. Thus, we use a subset of 640 models over 10 categories, which consists of 779 doors and 529 drawers for this task, following the same train and test splits used in the previous section.

We evaluate two baseline algorithms, ResNet-50 [17] and PointNet++ [38], that deals with the input RGB-D partial scans using either 2D or 3D formats. For ResNet-50, we input RGB-D images augmented with target part mask (5-channel in total). For PointNet++, we substitute the 5-channel image with its camera-space RGB point cloud. We train both networks to output a 14-dim motion vector $m = (T_r, T_t, p_1^r, p_2^r, p_3^r, d_1^r, d_2^r, d_3^r, d_1^t, d_2^t, d_3^t, x_{\text{door}}, x_{\text{drawer}})$, where T_r and T_t respectively output the probability of this joint being rotational and translational, (p_1^r, p_2^r, p_3^r) and (d_1^r, d_2^r, d_3^r) indicate pivot point and rotation axis for hinge joints, (d_1^t, d_2^t, d_3^t) represents the direction of a proposed slider axis, and finally, x_{door} and x_{drawer} regress the part poses for doors and drawers respectively. The part pose is a number normalized within $[0, 1]$ indicating the current joint position. See supplementary for more details about network architectures, loss designs, and training protocols.

We summarize the experimental results in Table 5. The classification of different motion types achieves quite high accuracy, and the axis prediction for sliders (translational joints) achieves lower error than for hinges (rotational joints). In our experiments, ResNet50 achieves better performance than PointNet++. This could be explained by the much higher number of network parameters in ResNet. However, intuition suggests that such 3D information should be more easily predicted directly on 3D data. Future research should focus more on how to improve 3D axis prediction with 3D grounding.

4.2. Robotic Interaction

With the large-scale PartNet-Mobility dataset, SAPIEN also supports various robotic interaction tasks, including solving low-level control tasks, such as button pushing, handle grasping, and drawer pulling, and planning tasks that require long-horizon logical planning and low-level controls, *e.g.*, removing the mug from a microwave oven and then putting it on a table. Having both diverse object



Figure 5: **Robotic Interaction tasks.** We study two robotic interaction tasks: door-opening and drawer-pulling.

categories and rich intra-class instance variations allows us to perform such tasks on multiple object instances at category levels. Figure 3 shows a rich variety of robotic interaction tasks that SAPIEN enables.

In SAPIEN, we enable two modes for robotic interaction tasks: 1) using perception ground-truth (*e.g.*, part mask, part motion information, and 3D locations) to accomplish the task. In this way, we factor out the perception module and allow algorithms to focus on robotic control and interaction tasks; 2) using the raw image/point-cloud as inputs, the method needs to develop its own perception, planning and control modules, which is our end-goal for the home-assistant robots to achieve. Also, this mode enables end-to-end learning for perception and interactions (*e.g.*, learning perception with a specific interaction target).

Door-opening and Drawer-pulling. We perform two manipulation tasks: door-opening and drawer-pulling, as shown in Figure 5. We use a flying gripper (Kinova Gripper 3 [5]) that can move freely in the workspace. All dynamics properties except gravity, (*e.g.*, contact, friction, and damping) are simulated in our environment. We perform our drawer-pulling tasks on 108 cabinet instances and door-opening tasks on 77 cabinet instances.

In our tasks, if the gripper can move a given joint (*e.g.*, slider joint of the drawer, hinge joint of the door) through 90% of its motion range, then it will be regarded as a success. If the agent cannot move the joint to the given threshold or move in the opposite direction, then it fails. The input of the agent consists of point clouds, normal maps

and segmentation masks captured by three fixed cameras mounted on the left, right and front of the arena respectively. The agent can also access all information about its self (*e.g.*, 6 DoF pose).

Heuristic Based Manipulation. To demonstrate our simulator in manipulation tasks, we first use manually designed heuristic pipelines to solve the tasks. For drawer-pulling, we use point cloud with ground-truth segmentation to detect a valid grasp pose for drawer handle. Then we use velocity controller to pull it to the joint limit. Using ground-truth visual information, we can achieve a 95.3% success rate. As for the door-opening task, we first open the door with a small angle using a similar approach (grasp a handle at first). Then we use Position Based Visual Servoing (PBVS) [21] to track and clamp the edge of the door. Finally, the door is opened by rotating the edge. This method (PBVS) achieves an 81.8% success rate for door opening. A more detailed illustration of this heuristic-based pipeline can be found in our supplementary video.

Learning Based Manipulation. We also demonstrate the above two tasks using reinforcement learning. We test the generalizability of the RL agent by training on limited objects and testing on unseen objects with different size, density, and motion properties. We adopt Soft Actor-Critic(SAC) [15], which is one of the SOTA reinforcement learning algorithms, trained on 2, 4, 8, 16 doors or drawers, and test on the rest unseen models.

We provide three different state representations: 1) raw state of the whole scene (**raw-exp**), consisting of current positions and velocities of all the parts; 2) mobility-based representation (**mobility-exp**), with 6D pose of motion axis and average normal, and current joint angles and velocities of the target part; 3) visual inputs (**visual-exp**), where we set a front-view camera capturing RGB-D images for the object every time step, augmented with segmentation mask for the target part.

We use the same flying gripper and initialize it on the handle. The grasp pose is generated by the heuristic method as described in the above section. During training, agents receive positive rewards when the target part approaches the joint limit with the opening door/drawer, while obtaining negative rewards when the gripper falls off the handle. We interact with multiple objects simultaneously during training, and use a shared replay buffer to collect experiences to train SAC. After 1M interaction steps, we evaluate the performance on the unseen objects, each for 20 episodes.

For doors, the evaluation metric is the average achieved degree. For drawers, we report the success rate of opening 80% of joint limits. Table 6 shows our experimental results. For door-opening, the RL agent tends to overfit the training objects, as when the number of training objects grows, the performance drops. However, training on more

Tasks		Door (Final Angle Degree)				Drawer (Success Rate)			
		2	4	8	16	2	4	8	16
raw-exp	train	85.4	70.5	50.5	38.4	0.84	0.82	0.77	0.75
	test	14.7	18.7	21.2	27.3	0.61	0.63	0.66	0.66
mobility-exp	train	88.7	78.6	59.2	41.1	0.83	0.81	0.79	0.78
	test	22.9	27.3	27.5	32.8	0.65	0.65	0.69	0.68
visual-exp	train	90.2	65.2	56.7	32.1	0.80	0.72	0.69	0.63
	test	21.7	24.5	28.1	29.6	0.59	0.60	0.61	0.60

Table 6: SAC results on door and drawer opening.

scenarios will improve the generalization capability with increased test performance. For drawer-pulling, although the performance follows the same pattern as the door, it is relatively stable across the number of training objects. This is because drawers are relatively easier to pull out, as the movement for the gripper almost follows the same pattern every time step.

Among all the representations, **mobility-exp** gives the best performance. For doors, **visual-exp** representation also performs close to **mobility-exp**; however for drawers, **raw-exp** is better than **visual-exp**. This is because the camera is fixed during the interaction. For drawer-opening, the visual features remain almost the same every time step from the front view, so it provides little information about state changing. These observations lead us to some interesting future work. First, we need proper vision methods to encode the geometric information of the scene, which may change during interaction procedures. Second, although these tasks are not hard for heuristic algorithms, RL-based approaches fail to perform well on all the objects. Future works may study how to enhance the transferability and efficiency of RL on the tasks.

5. Conclusion

We present SAPIEN, a simulation environment for robotic vision and interaction tasks, which provides detailed part-level physical simulation, hierarchical robotics controllers and versatile rendering options. We demonstrate that our SAPIEN enables a large variety of robotic perception and interaction tasks.

Acknowledgements

This research was supported by NSF grant IIS-1764078, NSF grant IIS-1763268, a Vannevar Bush Faculty Fellowship, the Canada CIFAR AI Chair program, gifts from Qualcomm, Adobe, and Kuaishou Technology, and grants from the Samsung GRO program and the SAIL Toyota Research Center.

References

- [1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2
- [2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016. 2, 3, 4
- [3] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, and Aaron Courville. HoME: A household multimodal environment. *arXiv preprint arXiv:1711.11017*, 2017. 1, 2, 3
- [4] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36(3):261–268, 2017. 1
- [5] Alexandre Campeau-Lecours, Hugo Lamontagne, Simon Latour, Philippe Fauteux, Véronique Maheu, François Boucher, Charles Deguire, and Louis-Joseph Caron L’Ecuyer. Kinova modular robot arms for service robotics applications. In *Rapid Automation: Concepts, Methodologies, Tools, and Applications*, pages 693–719. IGI Global, 2019. 7
- [6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 3
- [7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3, 4
- [8] Sachin Chitta, Eitan Marder-Eppstein, Wim Meeussen, Vijay Pradeep, Adolfo Rodríguez Tsouroukdissian, Jonathan Bohren, David Coleman, Bence Magyar, Gennaro Raiola, Matthias Lüdtke, et al. ros_control: A generic and simple control framework for ros. 2017. 4
- [9] Sachin Chitta, Ioan Sucan, and Steve Cousins. Moveit![ros topics]. *IEEE Robotics & Automation Magazine*, 19(1):18–19, 2012. 4
- [10] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. *GitHub repository*, 2016. 1, 3
- [11] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063, 2018. 1
- [12] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016. 3
- [13] Xiaofeng Gao, Ran Gong, Tianmin Shu, Xu Xie, Shu Wang, and Song-Chun Zhu. VRKitchen: an interactive 3D virtual environment for task-oriented learning. *arXiv*, abs/1903.05757, 2019. 2, 3
- [14] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2017. 1
- [15] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018. 8
- [16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7
- [18] Ruizhen Hu, Wenchao Li, Oliver Van Kaick, Ariel Shamir, Hao Zhang, and Hui Huang. Learning to predict part mobility from a single static snapshot. *ACM Transactions on Graphics (TOG)*, 36(6):227, 2017. 3
- [19] Jingwei Huang, Hao Su, and Leonidas Guibas. Robust watertight manifold surface generation method for shapenet models. *arXiv preprint arXiv:1802.01698*, 2018. 5
- [20] Louis Hugues and Nicolas Bredeche. Simbad: an autonomous robot simulation package for education and research. In *International Conference on Simulation of Adaptive Behavior*, pages 831–842. Springer, 2006. 3
- [21] Seth Hutchinson, Gregory D Hager, and Peter I Corke. A tutorial on visual servo control. *IEEE transactions on robotics and automation*, 12(5):651–670, 1996. 8
- [22] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *arXiv preprint arXiv:1909.12271*, 2019. 2, 3
- [23] Arthur Juliani, Vincent-Pierre Berges, Esh Vckay, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*, 2018. 3
- [24] Nathan Koenig and Andrew Howard. Design and use paradigms for Gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2149–2154. IEEE, 2004. 1, 3, 4
- [25] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *arXiv:1712.05474*, 2017. 2, 3
- [26] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for

- robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018. 1
- [27] Michael Lutter, Christian Ritter, and Jan Peters. Deep lagrangian networks: Using physics as model prior for deep learning. *arXiv preprint arXiv:1907.04490*, 2019. 3
- [28] Khaled Mamou, E Lengyel, and Ed AK Peters. Volumetric hierarchical approximate convex decomposition. *Game Engine Gems 3*, pages 141–158, 2016. 5
- [29] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. ROBOTURK: A crowdsourcing platform for robotic skill learning through imitation. *arXiv preprint arXiv:1811.02790*, 2018. 1
- [30] Roberto Martín-Martín, Clemens Eppner, and Oliver Brock. The RBO dataset of articulated objects and interactions. *The International Journal of Robotics Research*, 38(9):1013–1019, 2019. 3
- [31] Johannes Meyer, Alexander Sendobry, Stefan Kohlbrecher, Uwe Klingauf, and Oskar Von Stryk. Comprehensive simulation of quadrotor uavs using ROS and Gazebo. In *International conference on simulation, modeling, and programming for autonomous robots*, pages 400–411. Springer, 2012. 3
- [32] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 4, 6
- [33] Matthias Müller, Vincent Casser, Jean Lahoud, Neil Smith, and Bernard Ghanem. Sim4cv: A photo-realistic simulator for computer vision applications. *International Journal of Computer Vision*, 126(9):902–919, 2018. 2, 3
- [34] Adithyavairavan Murali, Tao Chen, Kalyan Vasudev Alwala, Dhiraj Gandhi, Lerrel Pinto, Saurabh Gupta, and Abhinav Gupta. Pyrobot: An open-source robotics framework for research and benchmarking. *arXiv preprint arXiv:1906.08236*, 2019. 1
- [35] Nvidia. PhysX physics engine. <https://www.geforce.com/hardware/technology/physx>. 1, 3
- [36] Steven G Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, et al. Optix: a general purpose ray tracing engine. In *Acm transactions on graphics (tog)*, volume 29, page 66. ACM, 2010. 5
- [37] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2
- [38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017. 6, 7
- [39] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009. 4
- [40] Eric Rohmer, Surya PN Singh, and Marc Freese. V-REP: A versatile and scalable robot simulation framework. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1321–1326. IEEE, 2013. 1, 3
- [41] Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multi-modal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017. 1, 2
- [42] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2
- [43] Connor Schenck and Dieter Fox. Spnets: Differentiable fluid dynamics for deep neural networks. *arXiv preprint arXiv:1806.06094*, 2018. 1
- [44] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3
- [45] Yuhang Song, Jianyi Wang, Thomas Lukasiewicz, Zhenghua Xu, Mai Xu, Zihan Ding, and Lianlong Wu. Arena: A general evaluation platform and building toolkit for multi-agent intelligence. *arXiv preprint arXiv:1905.08085*, 2019. 3
- [46] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3
- [47] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. Technical report, DeepMind, Jan. 2018. 3
- [48] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 1
- [49] Yusuke Urakami, Alec Hodgkinson, Casey Carlin, Randall Leu, Luca Rigazio, and Pieter Abbeel. DoorGym: A scalable door opening environment and baseline agent. *arXiv preprint arXiv:1908.01887*, 2019. 3
- [50] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction

- through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206. Eurographics Association, 2007. 5
- [51] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinpeng Zhao, and Kai Xu. Shape2Motion: Joint analysis of motion parts and attributes from 3D shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019. 3
- [52] Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, Yizhou Wang, Alan Yuille, Weichao Qiu, Fangwei Zhong. UnrealCV: Virtual worlds for computer vision. *ACM Multimedia Open Source Software Competition*, 2017. 2, 3
- [53] Lawrence B Wolff, Shree K Nayar, and Michael Oren. Improved diffuse reflection models for computer vision. *International Journal of Computer Vision*, 30(1):55–71, 1998. 5
- [54] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3D environment. *arXiv preprint arXiv:1801.02209*, 2018. 1, 2
- [55] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive Gibson: A benchmark for interactive navigation in cluttered environments. *arXiv preprint arXiv:1910.14442*, 2019. 2
- [56] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 1, 2, 3
- [57] Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. CHALET: Cornell house agent learning environment. *arXiv:1801.07357*, 2018. 2
- [58] Zihao Yan, Ruizhen Hu, Xingguang Yan, Luanmin Chen, Oliver van Kaick, Hao Zhang, and Hui Huang. RPM-Net: recurrent prediction of motion and parts from point cloud. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)*, 2019. 3
- [59] Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2040–2050, 2019. 1
- [60] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. TossingBot: Learning to throw arbitrary objects with residual physics. *arXiv preprint arXiv:1903.11239*, 2019. 3

SAPIEN: a SimulAteD Part-based Interactive ENvironment

Supplementary Material

Fanbo Xiang¹ Yuzhe Qin¹ Kaichun Mo² Yikuan Xia¹ Hao Zhu¹
Fangchen Liu¹ Minghua Liu¹ Hanxiao Jiang³ Yifu Yuan⁵ He Wang² Li Yi⁴
Angel X. Chang³ Leonidas Guibas² Hao Su¹

¹UC San Diego ²Stanford University ³Simon Fraser University ⁴Google Research ⁵UC Los Angeles

<https://sapien.ucsd.edu>



Figure 1: Diverse manipulation tasks supported by SAPIEN

Table of Contents

- **Appendix A** Details on PartNet-Mobility Annotation System.
- **Appendix B** Experiment details on movable part segmentation and motion recognition tasks.
- **Appendix C** Terminologies

Appendix A: Annotation System

We developed a web interface (Figure 2) for mobility annotation. This tool is a question answering (QA) system, which proposes questions based on current stage of annotation. It exploits the hierarchical structures of PartNet to propose objects without relative mobility, and generates new questions based on past annotations. Using this tool, annotators will not miss any movable parts if they answer every question correctly, and they will not face any redundant questions by design. The output mobility annotations are guaranteed to satisfy tree properties suitable for simulation.

The annotation procedure has the following steps:

- We start with a PartNet semantic tree, and traverse the tree nodes. Annotators are prompted with questions asking if current subtree has relative motion. If it does not, all parts in this tree will be fixed together; otherwise, the same question is asked again on the child nodes of this subtree.
- When the PartNet semantic tree traversal is finished, annotators are asked to choose parts that are fixed together.
- Next, annotators are asked to choose parts that are connected with a hinge (rotational) joint. They will then choose parent-child relation, and annotate axis position/motion limit with our 3D annotation tool.
- Next, annotators are asked to choose parts that are connected with a slider (translational) joint. They will similarly choose motion parameters and decide if this axis also bears rotation (screw joint).
- Finally, annotators will annotate each separate object in the scene as “fixed base”, “free”, or “outlier”.

The procedure is summarized in the following pseudo-code block.

Appendix B: Movable Part Segmentation and Motion Recognition

Movable Part Segmentation: complete results

Table 1 shows the movable part segmentation results for all categories in PartNet-Mobility dataset.

Annotating PartNet-Mobility dataset

```

1: Propose fixed parts based on PartNet tree
2: for There are parts can be fixed together do
3:   Select a group of relatively fixed parts
4: end for
5: for Rotation relationship exists do
6:   Select parent and child
7:   Pick rotation axis
8:   Input motion range
9: end for
10: for Translation relationship exists do
11:   Select parent and child
12:   Pick translation axis
13:   Input motion range / whether it can also rotate
14: end for
15: Choose whether root nodes are fixed/free

```

Motion Recognition: experiment details

For this task, we normalize the $[0, 2\pi]$ hinge joint range to $[0, 1]$. For sliders, we normalize by the maximum motion range over the dataset to make the motion range prediction within $[0, 1]$.

Algorithm. The baseline algorithm we use is a a ResNet[1] classification and Regression network. The input is the ground truth RGB-D image and the segmentation mask for the target movable part. The output has 7 terms:
 $\hat{T}_r \in \{0, 1\}$, whether this part has a rotational joint.
 $\hat{T}_t \in \{0, 1\}$, whether this part has a translational joint.
 $\hat{\mathbf{p}}^r \in \mathbb{R}^3$, pivot of a predicted rotational axis.
 $\hat{\mathbf{d}}^r \in [-1, 1]^3$, direction of a predicted rotational axis.
 $\hat{\mathbf{d}}^t \in [-1, 1]^3$, direction of a predicted translational axis.
 $\hat{x}_{\text{door}} \in [0, 1]$, predicted joint position for a door.
 $\hat{x}_{\text{drawer}} \in [0, 1]$, predicted joint position for a drawer.

In the following, letters without hat indicates their corresponding ground-truth labels.

In our experiment, we modify the input layer of a ResNet50 network to accept 5 channels, and output layer to output 13 numbers. In addition, we apply tanh activation to produce $\hat{\mathbf{d}}^r, \hat{\mathbf{d}}^t$, and sigmoid activation to produce $\hat{x}_{\text{door}}, \hat{x}_{\text{drawer}}$. The loss has 7 terms:

Axis alignment loss, measured by cosine distance:

$$L_{dr} = \sum_{T_r=1} 1 - \left| \frac{\mathbf{d}^r \cdot \hat{\mathbf{d}}^r}{\|\mathbf{d}^r\| \|\hat{\mathbf{d}}^r\|} \right| \quad L_{dt} = \sum_{T_t=1} 1 - \left| \frac{\mathbf{d}^t \cdot \hat{\mathbf{d}}^t}{\|\mathbf{d}^t\| \|\hat{\mathbf{d}}^t\|} \right|$$

Pivot loss, measured by the distance from predicted pivot to ground truth joint axis:

$$L_p = \sum_{T_r=1} \|\hat{\mathbf{p}}^r - \mathbf{p}^r - ((\hat{\mathbf{p}}^r - \mathbf{p}^r) \cdot \mathbf{d}^r) \mathbf{d}^r\|_2^2$$

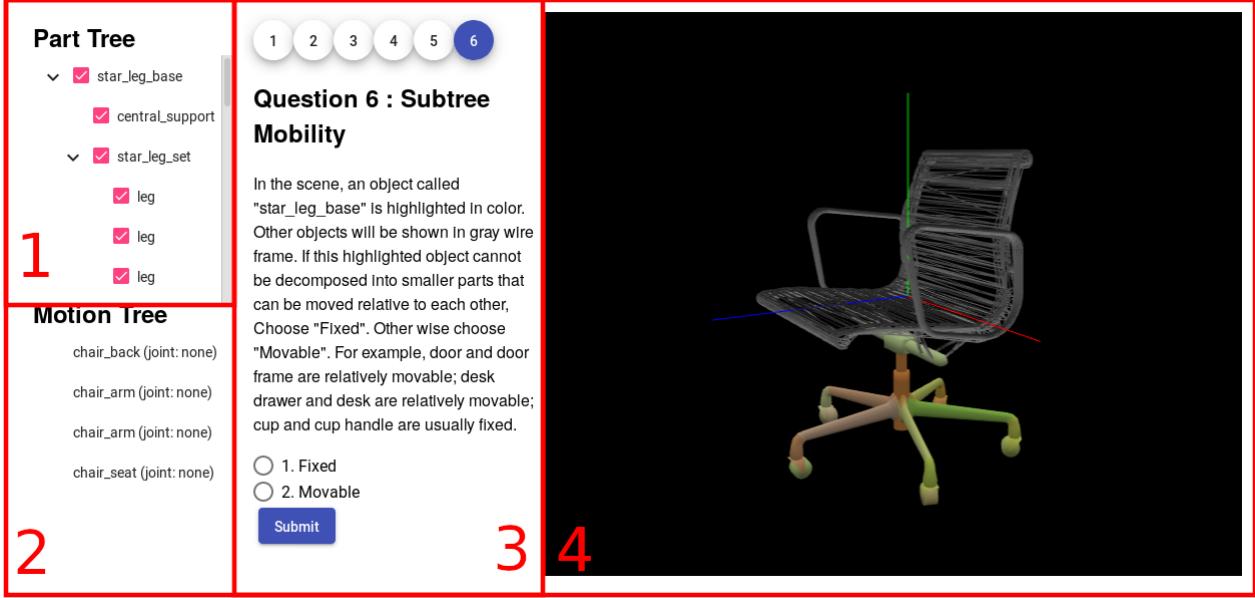


Figure 2: **Annotation interface.** 1) Part Tree: PartNet semantic tree that proposes fixed parts. 2) Motion tree: annotated movable parts. 3) Question: auto-generated exhaustive questions. 4) Visualization for current question and for motion axis annotation.

Joint type prediction loss:

$$L_{Tr} = - \sum T_r \log \hat{T}_r + (1 - T_r) \log(1 - \hat{T}_r)$$

$$L_{Tt} = - \sum T_t \log \hat{T}_t + (1 - T_t) \log(1 - \hat{T}_t)$$

Joint position loss, L_2 loss between predicted position and ground truth position.

$$L_{door} = \sum_{\text{valid hinge}} (x_{door} - \hat{x}_{door})^2$$

$$L_{drawer} = \sum_{\text{valid slider}} (x_{drawer} - \hat{x}_{drawer})^2$$

The final loss is a summation of all the losses above:

$$L = L_{dr} + L_{dt} + L_p + L_{Tr} + L_{Tt} + L_{door} + L_{drawer}$$

This objective is optimized on mini-batches using proper masking based on H and S values.

We repeat this experiment with PointNet++[4] operating on 3D RGB-point cloud produced by the same images. For each image, we sample 10,000 points from the partial point cloud (create random copies if the total number of points is less than 10,000). Figure 3 shows the network structure for the motion recognition tasks.

Appendix C: Terminology

SAPIEN Engine

- **Articulation:** An articulation is composed of a set of links connected together with translational or rotational joints [3]. The most common articulation is a robot.
- **Kinematic/Dynamic joint system:** Both joint systems are an assembly of rigid bodies connected by pairwise constraints. Kinematic system does not respond to external forces while dynamic objects do.
- **Force/Joint/Velocity Controller:** Controller which can control the force/position/velocity of one or multiple joints at once. Like real robot, controller may fail depending on whether the target is reachable.
- **Inertial Measurement Unit(IMU):** A sensor which can measure the orientation, acceleration and angular velocity of the mounted link.
- **Trajectory Controller:** A controller which receives trajectory command and execute to move through the trajectory points. Note that trajectory consists of a sequence of position, velocity and acceleration, while path is simply a set of points without a schedule for reaching each point [2].
- **End-effector:** End-effector is a manipulator that performs the task required of the robot, The most common

		Bottle			Box		Bucket			Cabinet				Camera			Cart		Chair					
Algorithm	Setting	tr. lid	body	rot. lid	rot. lid	body	handle	body	door	body	door	drawer	lens	button	body	knob	wheel	body	wheel	seat	leg			
Mask-RCNN	RGB	0.0%	57.4%	69.3%	49.3%	65.7%	2.7%	91.7%	62.0%	94.2%	27.7%	66.4%	26.7%	20.9%	79.0%	4.8%	54.6%	95.6%	25.1%	97.0%	88.3%			
	RGB-D	13.9%	68.3%	67.8%	51.5%	66.5%	1.6%	100.0%	61.7%	93.0%	26.3%	63.0%	26.4%	17.0%	92.6%	8.1%	55.3%	93.9%	23.1%	99.0%	85.2%			
PartNet	XYZ	24.5%	47.7%	53.5%	27.6%	46.2%	63.4%	99.7%	20.6%	65.9%	9.8%	35.1%	17.0%	0.0%	51.4%	0.0%	6.2%	71.7%	1.2%	93.0%	86.4%			
InsSeg	XYZRGB	5.9%	41.3%	54.8%	24.2%	36.8%	60.7%	98.9%	17.4%	64.3%	5.0%	23.6%	10.5%	0.0%	46.1%	1.0%	9.4%	77.3%	1.9%	95.7%	89.2%			
		Chair			Clock		CoffeeMachine						Dishwasher	Dispenser	Display			Door						
Algorithm	Setting	knob	caster	lever	hand	body	button	lid	body	lever	knob	container	rot.	Door	body	lid	body	rot.	screen	base	button	frame rot.	door	
Mask-RCNN	RGB	0.0%	2.5%	20.0%	11.4%	61.4%	14.7%	73.4%	65.7%	0.0%	43.0%	100.0%	70.4%	90.0%	74.9%	90.1%	74.4%	34.7%	0.0%	39.4%	40.7%			
	RGB-D	0.0%	3.6%	13.4%	12.5%	68.3%	10.4%	61.4%	67.4%	1.0%	35.6%	98.0%	66.8%	87.8%	73.2%	88.1%	71.3%	33.4%	0.0%	35.7%	54.6%			
PartNet	XYZ	0.0%	1.0%	0.0%	0.0%	77.0%	0.0%	43.6%	62.4%	0.0%	0.0%	94.0%	50.5%	67.0%	49.1%	57.6%	66.1%	37.1%	0.0%	49.2%	35.3%			
InsSeg	XYZRGB	0.0%	1.0%	0.0%	0.0%	79.4%	0.0%	81.2%	45.8%	0.0%	0.0%	85.1%	58.2%	73.3%	27.4%	39.5%	58.2%	39.1%	0.0%	34.6%	24.6%			
		Eyeglasses		Fan		Faucet		FoldingChair		Globe		Kettle		Keyboard		KitchenPot		Knife						
Algorithm	Setting	leg	body	rotor	frame	switch	base	spout	seat	leg	sphere	frame	lid	body	base	key	lid	body	blade	body	blade			
Mask-RCNN	RGB	51.2%	85.2%	54.4%	67.5%	52.5%	47.9%	99.7%	90.6%	46.1%	98.0%	71.1%	75.2%	99.4%	15.0%	17.5%	99.0%	94.5%	11.7%	88.5%	33.4%			
	RGB-D	49.2%	84.9%	39.4%	67.4%	52.1%	55.8%	98.9%	93.8%	47.2%	96.0%	69.6%	94.1%	100.0%	8.8%	5.1%	100.0%	95.0%	10.0%	77.8%	34.5%			
PartNet	XYZ	62.1%	93.8%	50.9%	74.8%	34.4%	55.9%	64.2%	91.2%	79.4%	83.0%	77.6%	71.1%	74.1%	6.8%	1.0%	94.6%	94.4%	3.1%	80.1%	9.4%			
InsSeg	XYZRGB	80.6%	92.4%	42.0%	63.5%	29.9%	64.1%	78.0%	86.3%	75.6%	79.0%	82.0%	87.2%	90.7%	4.0%	1.0%	93.5%	95.0%	5.0%	82.7%	10.1%			
		Lamp			Laptop		Lighter			Microwave				Mouse		Oven								
Algorithm	Setting	base	rot. bar	head	base	screen	wheel	button	body	rot. lid	door	body	button	button	wheel	body	door	knob	body	tr. tray	button			
Mask-RCNN	RGB	54.6%	14.6%	64.5%	51.9%	93.1%	35.0%	80.8%	96.8%	97.0%	53.8%	94.0%	0.0%	0.0%	46.5%	98.0%	54.0%	49.9%	86.8%	1.0%	0.0%			
	RGB-D	48.8%	10.8%	69.5%	47.2%	92.8%	57.2%	94.1%	89.2%	92.1%	49.5%	97.1%	0.0%	1.0%	45.3%	95.2%	53.4%	42.3%	93.3%	1.0%	0.0%			
PartNet	XYZ	51.8%	8.8%	38.5%	93.0%	97.7%	1.0%	0.0%	77.4%	80.9%	25.9%	45.8%	0.0%	1.0%	0.0%	76.0%	23.1%	0.0%	36.6%	1.0%	0.0%			
InsSeg	XYZRGB	50.6%	9.3%	39.7%	89.8%	96.1%	9.5%	61.4%	82.5%	84.9%	24.3%	48.7%	0.0%	1.0%	61.1%	1.0%	26.9%	0.0%	49.1%	0.0%	0.0%			
		Pen			Phone		Pliers		Printer		Refrigerator		Remote		Safe		Scissors		Stapler					
Algorithm	Setting	cap	body	button	button	base	base	leg	button	body	body	door	button	base	knob	button	body	door	leg	body	lid	base		
Mask-RCNN	RGB	94.1%	91.0%	52.8%	18.4%	51.4%	79.9%	2.8%	87.1%	83.0%	60.7%	35.6%	75.2%	34.1%	0.0%	88.5%	68.5%	34.2%	32.1%	60.2%	84.6%			
	RGB-D	94.1%	96.2%	57.6%	12.8%	50.2%	78.7%	1.5%	72.3%	81.2%	55.0%	25.6%	78.2%	24.5%	0.0%	92.1%	74.6%	57.4%	33.6%	75.0%	90.5%			
PartNet	XYZ	67.9%	98.0%	53.0%	1.0%	38.0%	37.9%	0.0%	34.8%	30.0%	16.2%	1.0%	63.2%	0.0%	0.0%	40.5%	30.5%	20.6%	31.7%	49.2%	76.7%			
InsSeg	XYZRGB	15.0%	96.2%	25.4%	0.0%	27.0%	46.0%	0.0%	48.5%	40.2%	27.7%	1.0%	75.9%	0.0%	0.0%	60.8%	42.3%	36.4%	28.5%	83.3%	89.5%			
		Suitcase			Switch		Table			Toaster				Toilet										
Algorithm	Setting	rot. handle	body	tr. handle	wheel	caster	frame	lever	button	slider	drawer	body	wheel	door	casters	knob	slider	body	button	lever	lid			
Mask-RCNN	RGB	25.5%	81.7%	74.3%	6.2%	0.0%	85.9%	24.3%	73.6%	60.8%	54.3%	88.0%	3.4%	6.3%	0.0%	40.1%	39.0%	90.1%	5.9%	51.6%	98.3%			
	RGB-D	36.4%	97.3%	70.0%	18.1%	0.0%	74.0%	26.0%	65.8%	22.8%	58.6%	89.9%	1.4%	13.2%	0.0%	40.6%	33.0%	94.1%	4.0%	36.4%	98.0%			
PartNet	XYZ	3.7%	53.7%	63.6%	1.4%	0.0%	52.3%	2.3%	4.9%	1.0%	15.7%	71.3%	1.7%	1.0%	0.0%	0.0%	9.9%	79.3%	0.0%	0.0%	69.3%			
InsSeg	XYZRGB	4.3%	53.2%	64.5%	2.0%	0.0%	21.8%	75.4%	73.5%	3.9%	5.0%	23.9%	42.9%	12.2%	14.5%	0.0%	0.0%	22.9%	24.0%	85.2%	37.1%			
		Toilet			TrashCan		USB			WashingMachine				Window										
Algorithm	Setting	body	lid	seat	button	pad	lid	body	door	wheel	rotation	body	lid	door	knob	button	body	window	frame	mAP				
Mask-RCNN	RGB	95.3%	64.3%	61.1%	8.9%	43.4%	68.1%	85.6%	35.9%	73.7%	59.8%	65.7%	71.3%	52.0%	6.8%	4.4%	53.5%	55.9%	12.2%		53.0%			
	RGB-D	91.8%	64.4%	62.5%	3.0%	37.1%	69.7%	84.9%	29.7%	69.3%	74.4%	62.8%	68.6%	41.4%	4.0%	0.0%	73.3%	48.7%	13.4%		52.8%			
PartNet	XYZ	83.2%	17.6%	1.4%	0.0%	12.7%	67.1%	57.7%	12.1%	5.5%	30.0%	27.1%	22.2%	22.4%	0.0%	0.0%	30.5%	22.6%	83.5%		36.1%			
InsSeg	XYZRGB	86.5%	25.1%	5.2%	0.0%	21.8%	75.4%	73.5%	3.9%	5.0%	23.9%	42.9%	12.2%	14.5%	0.0%	0.0%	22.9%	24.0%	85.2%	37.1%				

Table 1: Movable part segmentation results for all categories

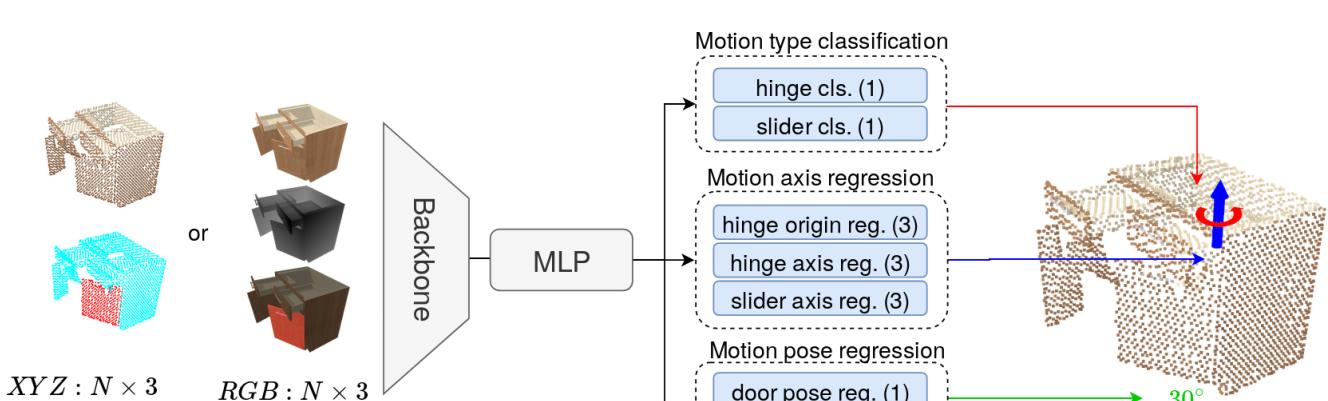


Figure 3: **Vision Tasks.** Show 2 vision task definitions: inputs + outputs.

end-effector is gripper.

- **Inverse Kinematics:** Determine the joint position corresponding to a given end-effector position and orientation [5].
- **Inverse Dynamics:** Determining the joint torques which are needed to generate a given motion. Usually, the input of inverse dynamics is the output of inverse kinematics or motion planning.

SAPIEN Renderer

- **GLSL** is OpenGL's shading language which describes how the GPU draws visuals.
- **Rasterization** is the process of converting shapes to pixels. It is the pipeline used by most real-time graphics applications.
- **Ray tracing** is a rendering technique by simulating light-rays, reflections, refractions, etc. It can achieve physically accurate images at the cost of rendering time. OptiX is Nvidia's GPU based ray-tracing framework.

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2
- [2] Seth Hutchinson, Gregory D Hager, and Peter I Corke. A tutorial on visual servo control. *IEEE transactions on robotics and automation*, 12(5):651–670, 1996. 3
- [3] Nvidia. PhysX physics engine. <https://www.geforce.com/hardware/technology/physx>. 3
- [4] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017. 3
- [5] Bruno Siciliano, Lorenzo Sciavicco, Luigi Villani, and Giuseppe Oriolo. *Robotics: modelling, planning and control*. Springer Science & Business Media, 2010. 5