

# Close the Optical Sensing Domain Gap by Physics-Grounded Active Stereo Sensor Simulation

Xiaoshuai Zhang<sup>ID</sup>, Rui Chen<sup>ID</sup>, Member, IEEE, Ang Li<sup>ID</sup>, Fanbo Xiang, Yuzhe Qin<sup>ID</sup>, Student Member, IEEE, Jiayuan Gu, Zhan Ling, Minghua Liu, Peiyu Zeng<sup>ID</sup>, Songfang Han<sup>ID</sup>, Zhiao Huang, Tongzhou Mu<sup>ID</sup>, Jing Xu<sup>ID</sup>, Member, IEEE, and Hao Su<sup>ID</sup>, Member, IEEE

**Abstract**—In this article, we focus on the simulation of active stereovision depth sensors, which are popular in both academic and industry communities. Inspired by the underlying mechanism of the sensors, we designed a fully physics-grounded simulation pipeline that includes material acquisition, ray-tracing-based infrared (IR) image rendering, IR noise simulation, and depth estimation. The pipeline is able to generate depth maps with material-dependent error patterns similar to a real depth sensor in real time. We conduct real experiments to show that perception algorithms and reinforcement learning policies trained in our simulation platform could transfer well to the real-world test cases without any fine-tuning. Furthermore, due to the high degree of realism of this simulation, our depth sensor simulator can be used as a convenient testbed to evaluate the algorithm performance in the real world, which will largely reduce the human effort in developing robotic algorithms. The entire pipeline has been integrated into the SAPIEN simulator and is open-sourced to promote the research of vision and robotics communities.

**Index Terms**—Active stereovision, depth sensor, robot simulation, sensor simulation, sim-to-real.

## I. INTRODUCTION

RECENTLY, researchers in the robotics community are becoming increasingly interested in training robots in simulation environments, due to the convenience of conducting large-scale experiments (e.g., grasp label generation [1], [2], [3], [4], [5], [6] and object manipulation by deep reinforcement learning [7], [8], [9]).

In this work, we aim to achieve low optical sensing domain gap (not physical domain gap) of robot simulators with

Manuscript received 16 November 2022; accepted 4 January 2023. Date of publication 27 January 2023; date of current version 7 June 2023. This paper was recommended for publication by Associate Editor J. Kober and Editor F. Chaumette upon evaluation of the reviewers' comments. (*Xiaoshuai Zhang and Rui Chen contributed equally to this work.*) (*Corresponding authors:* Jing Xu; Hao Su.)

Xiaoshuai Zhang, Ang Li, Fanbo Xiang, Yuzhe Qin, Jiayuan Gu, Zhan Ling, Minghua Liu, Songfang Han, Zhiao Huang, Tongzhou Mu, and Hao Su are with the Department of Computer Science and Engineering, University of California, San Diego, CA 92122 USA (e-mail: xiz040@eng.ucsd.edu; a3li@ucsd.edu; fxiang@eng.ucsd.edu; y1qin@eng.ucsd.edu; jigu@eng.ucsd.edu; z6ling@ucsd.edu; minghua@ucsd.edu; s5han@eng.ucsd.edu; z2huang@eng.ucsd.edu; tongzhou.mu@gmail.com; haosu@eng.ucsd.edu).

Rui Chen, Peiyu Zeng, and Jing Xu are with the Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China (e-mail: callmeray@163.com; zengyuyi0307@gmail.com; jingxu@tsinghua.edu.cn).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TRO.2023.3235591>.

Digital Object Identifier 10.1109/TRO.2023.3235591

a light-weight pipeline. We drive the exploration by important applications: 1) Given computer-aided design (CAD) models of objects, the robot simulator generates a large amount of simulated data, on which learning-based perception models (e.g., object detection, pose estimation) and policy learning algorithms (e.g., object grasping) are trained and can work in the real world without any fine-tuning; 2) Before carrying out real-world experiments, researchers can evaluate perception or policy learning algorithms in simulation, which requires that the performance in the real world and the simulator should be consistent. We choose to narrow the sim-to-real optical sensing domain gap by studying the simulation of depth sensors, more precisely, active stereovision depth sensors. This choice distinguishes us from most recent efforts except for few ones. We list five reasons to simulate active stereovision depth sensors instead of RGB sensors (cameras) or other depth sensors. The first two reasons justify the choice of depth sensor over RGB camera, and the last three justify the choice of active stereovision sensor over other depth sensors.

- 1) Depth map is more suitable for robotic tasks. For many robotic tasks that require interaction, 3-D geometry information of the physical world is of utmost importance [3], [10], [11], [12]. However, this information cannot be retrieved easily from RGB cameras.
- 2) Depth maps do not involve color information, which is extremely challenging to align between the simulator and the real world. A few recent papers [13], [14], [15] for depth map simulation are also motivated by this observation.
- 3) Active stereovision depth sensor has wide adoption. Compared with other kinds of depth sensors, active stereovision depth sensors have relatively high accuracy, high spatial resolution, low cost, and robustness to lighting conditions [16], and therefore, are widely used in academic and industry scenarios [17].
- 4) Active stereovision depth sensor simulation requires the estimation of fewer parameters. Active stereovision depth sensors usually work at a specific wavelength, mostly infrared (IR). In an indoor environment, the energy within the narrow IR band is quite limited, and IR cameras will filter out most of the light in other spectra. Therefore, when modeling the environment, passive environment light is negligible, allowing us to ignore the complex environment illumination, which is the main challenge if we otherwise simulate passive stereovision sensors or RGB sensors.

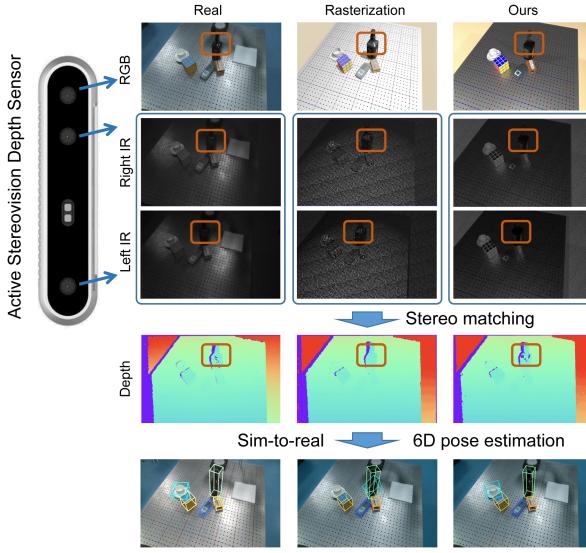


Fig. 1. Depth measurement of real active stereovision depth sensors is greatly affected by the object materials. The projected pattern is not well reflected on the translucent or transparent surface, resulting in incomplete measurement. However, conventional rasterization rendering cannot simulate this phenomenon. In this article, we aim to close the optical sensing domain gap by building an active stereovision depth sensor simulator in a physics-grounded fashion, so that it can simulate the material-dependent depth errors more accurately and improve the sim-to-real performance of 6-D pose estimation algorithms. (Zoom in to better see the IR patterns and pose estimation results).

- 5) Real-time realistic ray tracing techniques have matured just recently. Active stereovision sensors use IR lights to probe the environment and form stereo image pairs. State-of-the-art ray tracing pipelines allow us to simulate the transport of IR lights with high fidelity in real time, which is only available recently with hardware acceleration and learning-based denoising.

However, depth sensing by active stereovision faces unique challenges relevant to the materials of objects, which needs special attention when building a simulation pipeline. As shown in Fig. 1, the measurement results on objects with optically challenging materials (e.g., transparent) are inferior, due to the complex lighting effects of the projected pattern on these surfaces. To generate simulation data with a relatively low domain gap, *the simulation process must also create material-dependent error patterns similar to real sensors*.

We explore how to build a light-weight yet effective pipeline to model environments and conduct active stereovision depth sensor simulation that is able to simulate visual observations with a low domain gap. As shown in Fig. 2, the pipeline includes the acquisition of object material parameters, the simulation of the IR pattern projector, the transport of IR lights by ray tracing, the simulation of sensor noise, and lastly, stereo matching, all grounded on the physics of the process. We highlight two innovations in the pipeline. The first is how we acquire object material parameters. We propose a multispectral loss function to acquire object material parameters, which include visual appearance loss and neural network based perceptual loss to help eliminate the mismatch in brightness and color in both visible spectrum and IR spectrum. Our second highlighted innovation is

in adding textured light support. Existing ray tracing rendering systems usually do not support the rendering of textured lights. We include textured light support to simulate the IR pattern projector. In order to achieve real-time simulation of the depth sensor, we further integrate learning-based denoisers into the renderer such that we can generate high-fidelity IR images with hundreds of frame per second (FPS) using a small number of samples per pixel (SPP). Another time-consuming part is stereo matching. We build a GPU-accelerated stereo matching module consists of common algorithms in real-world depth sensors. The accelerated module can achieve 200+ FPS under common settings while usual CPU implementations can hardly achieve 1 FPS.

In order to validate the effectiveness of our depth simulator for various down-stream applications in vision and robotics communities, we conducted experiments including generating training data for object detection, 6-D pose estimation and robot grasping, and estimating the ranking of pose estimation algorithms. The proposed depth simulator is able to generate large-scale training data for different perception and policy learning algorithms. It enables the trained algorithms to transfer to the real world without any fine-tuning, even for optically challenging objects whose depth measurements are noisy and incomplete. It can also decrease the cost of comparing different algorithms' performances.

In summary, the contributions of this work are as follows.

- 1) We present a physics-grounded active stereovision depth sensor simulator that has been validated effective for various applications by real-world experiments. The simulator has been integrated with the SAPIEN<sup>1</sup> robot physics simulation platform [18] and open-sourced for community usage.
- 2) We build a dataset with precisely aligned simulated data and real data, which can be used to evaluate the transferability of object detection and 6-D pose estimation algorithms.
- 3) We propose a material acquisition method based on the multispectral matching loss and the pixel-wise alignment between the simulation and the real data.

The rest of this article is organized as follows. In Section II, we discuss related works. Section III introduces the mechanism of active stereovision depth sensor and the concept of physically based rendering (PBR). Section IV gives an overview of our depth sensor simulation method. Section V describes the material acquisition method in detail. Section VI presents the active stereovision depth sensor simulation pipeline and key technical contributions. Section VII describes the experimental setup in simulation and the real world. Section VIII shows the quantitative and qualitative results of various algorithms on domain transferability for various applications. Section IX validates our design choices through ablation studies. Finally, Section X concludes this article.

<sup>1</sup>SAPIEN simulator: <https://github.com/haosulab/SAPIEN>

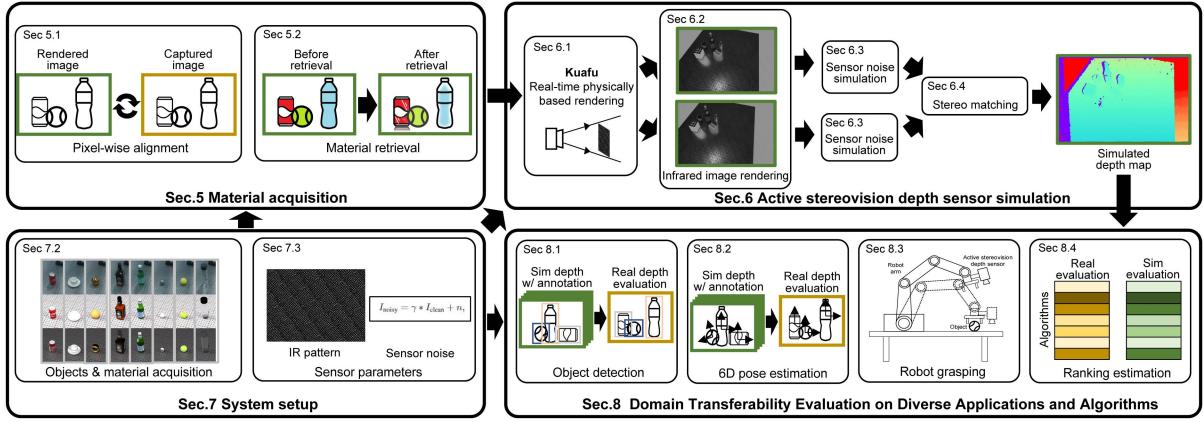


Fig. 2. Paper roadmap. In this article, we study how to build an active stereovision depth sensor simulator with a low optical sensing domain gap. In Section V, we describe the material acquisition method, which includes pixel-wise alignment between the simulation and the real world and PBR material retrieval based on a novel multispectral loss. In Section VI, we establish an active stereovision depth sensor simulator by following the same mechanism of real sensors, which includes IR pattern projection, transport of IR lights by ray tracing, IR camera sensor noise, and depth generation by stereo matching. Our new customizable renderer Kuafu is covered in that section. In Section VII, we describe the real-world and simulated system setup and the preparation for experiments. In Section VIII, we conduct experiments of four applications of our depth sensor simulator.

## II. RELATED WORK

### A. Sim-to-Real

In computer vision and robotics, using synthetic data for training and transferring model to the real world has been a common approach. A first idea is to use rendered CAD models to train algorithms for solving vision tasks, such as view point estimation [19] and object detection [20]. Sim-to-real approaches can be divided into two main categories: domain randomization and domain adaptation. Domain randomization applies random augmentation on training data to enforce the learning-based methods to extract task-relevant features and be more robust to data corruption [21], [22], [23]. However, the distribution of data augmentation has significant influence on the model's transferability [24], and it is difficult to make the augmentation distribution cover that of real data while not being too wide for learning methods. Domain adaptation aims to align data between simulation with real domains through manually-designed rules or learned mapping [25], [26], [27]. Compared to domain adaptation for RGB images, domain adaptation for depth images is far less studied [28], [29]. The goal of our work is similar to domain adaptation, which is to close the sim-real domain gap by making the simulated data more realistic. However, we choose to achieve this goal through a physics-grounded simulation method, rather than learning-based domain adaptation.

### B. Depth Sensor Simulation

For rasterization and ray-tracing renderers, a *clean* depth map is generated as a by-product. The simplest way to simulate the depth sensor is to just use this depth. But it cannot reflect the error characteristics of sensor depth in the real world. Thus, using the clean depth in the training data may harm the sim-to-real performance for learning-based methods.

Keller and Kolb [30] applied a physical model that simulates the time-of-flight (ToF) phase image generation for depth map synthesis. Meister et al. [31] proposed to simulate the multipath

interference of ToF sensors via ray tracing. These two methods are designed for ToF sensors and cannot be used for active stereovision sensors. Landau et al. [32] simulated the Kinect depth sensor by perturbing the simulated IR image with empirical noise models and performing stereo matching, and the authors validated the effectiveness on flat surfaces and edges. However, the geometry and material of the objects are not taken into account for simulation. Planche et al. [33] simulated the depth image from CAD models and validated that their method could be used for learning-based pose estimation and classification on real chairs. However, the depth error caused by specularity and transparency cannot be accurately simulated by rasterization, and they did not perform thorough studies on the most advanced learning algorithms.

Bousmalis et al. [25] employed generative adversarial networks (GAN) to generate simulated data and decoupled the task-specific content and domain appearance to improve the generalizability of GAN. Sweeney et al. [34] learned to predict no-depth-return pixels for depth simulation in a supervised fashion; however, their method to generate the required real training data is time-consuming and prone to error. Gu et al. [35] used hole prediction network and color-guided GAN to degrade the synthetic depth. Shen et al. [36] proposed to preserve the geometric information in the generated depth by enforcing the differential features to be invariant through the generation process. One common drawback of these statistical learning depth sensor simulation methods is that the physical mechanisms of light transport, such as specular reflection and refraction, are not taken into account, so they are not able to generate material-dependent depth errors. Pashevich et al. [37] proposed to search for the optimal sequence of augmentation policies (e.g., white noise, salt noise, etc.) through Monte Carlo tree search (MCTS). However, the complex material-dependent depth errors cannot be imitated by combining simple augmentation policies.

Most recently, Planche and Singh [38] built an end-to-end differentiable depth simulation pipeline based on differentiable ray-tracing. They optimize the rendering and stereo matching

parameters by using the loss between the simulated depth and the real depth. However, their simulation speed is slow due to current differentiable ray-tracers. Moreover, the differentiable rendering technique they used [39] does not support rendering of transparent materials, which limits the depth simulation fidelity on these materials. Also, in order to make the pipeline differentiable, the authors build a differentiable stereo matching module, but this pipeline greatly compromises customization capability. For example, it cannot incorporate the directional cost calculation procedure, which is widely used in commercial active stereovision depth sensors and can significantly improve the matching quality.

### C. Depth Completion

Besides depth sensor simulation, depth completion is another technique to improve detecting, grasping, and manipulating optically challenging objects. It aims to improve the quality of captured incomplete and noisy depth maps [40]. ClearGrasp [41] used the RGB image to predict surface normals, occlusion boundaries, and object masks before using global optimization to reconstruct the depth map. However, global optimization is impractical for real-time robot applications due to its high time cost. Zhu et al. [42] proposed to use a local implicit neural representation based on ray-voxel pairs to complete the depth and then iteratively refined the depth estimation. To accelerate the computation, they developed a customized CUDA kernel, which requires advanced GPU programming skills. Xu et al. [43] proposed to combine depth completion with object point cloud completion to better utilize the geometry priors and achieved improved performance. But it requires object masks for point cloud completion, which may be unavailable in certain scenarios. TransCG [44] constructed a large-scale real-world dataset for transparent object depth completion and trained a 2-D convolutional neural network (CNN)-based depth completion network on the dataset. However, constructing the real-world dataset with ground-truth depth and expanding it to include novel objects required a considerable amount of time.

Although depth completion methods are able to complete the depth map, they do not take into account the noise pattern in measurable areas and may introduce additional artifacts, limiting their ability to narrow the domain gap. Moreover, depth completion increases the computation cost when used for downstream tasks. In contrast, depth sensor simulation is able to improve the network's performance on real corrupted depth maps without introducing any additional burden at inference time.

### D. Material Acquisition

The goal of material acquisition is to estimate material properties from images with known geometry, and is a long-standing problem in the field of computer graphics. To adequately sample the surface function, classical methods typically require complex controlled camera and lighting setups, e.g., light stage [45] or structured lights [46], [47]. A few recent methods utilize neural network to extract bidirectional reflectance distribution function (BRDF) maps from captured images, and have enabled material acquisition from more casual setups, such as flash

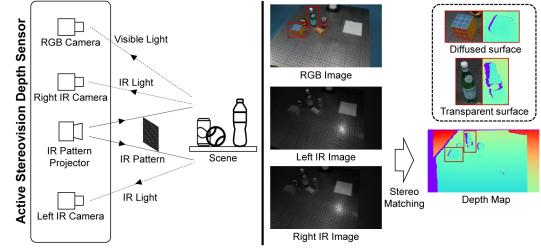


Fig. 3. Mechanism of active stereovision depth sensors.

photographs [48], [49], [50]. More recently, Zhang et al. [51] proposed employing a neural radiance field [52] based pipeline to jointly recover a fully factorized 3-D model from images captured under a single illumination. These methods represent materials in the form of spatially varying BRDF, which typically includes diffuse albedo, normal, roughness, and metallic maps. The majority of these methods do not consider transparent materials. Consequently, they cannot be applied to many common real objects such as bottles and glasses. Our setting differs from these previous works in that we assume a known CAD model with a diffuse base color, which is very common in the field of robotics. With this information, we can assume a PBR material for each object part and search for material parameters efficiently.

### E. Robot Simulation

Thanks to the recent progress of physics simulation [53], [54], [55], it has drawn increasing interest to build full-physics robot simulation environments [9], [56], [57], [58], [59]. Compared to robot simulation with abstract action [60], [61], [62], full-physics robot simulation supports low-level policy learning that could be transferred to the real world. Therefore, we integrate our depth simulator with state-of-the-art robot simulator SAPIEN [18], targeting at vision-based robot policy learning. To the best of authors knowledge, our work is the first physics-grounded depth simulator integrated into a robot simulation environment.

## III. PRELIMINARIES

### A. Active Stereovision Depth Sensor

Existing commercial optical depth sensors can be broadly categorized into the following four groups: passive stereovision, active stereovision, structured light, and ToF. In this section, we will introduce the mechanism of active stereovision depth sensors. We refer to [17] for a detailed introduction of the principles of other depth sensing techniques.

As shown in Fig. 3, an active stereovision depth sensor consists of an IR pattern projector, two IR cameras, and an RGB camera. All transformations between the three cameras are known. The projector casts an IR pattern (typically random dots) onto the scene. The two IR cameras capture the scene with the reflected pattern in the IR spectra. The depth map is computed by stereo matching on the two captured IR images. One common postprocess is to warp the depth map to the RGB image from the RGB camera to generate aligned RGBD

data. Compared with passive stereovision, active stereovision can measure texture-less areas. Compared with structured light which uses coded pattern(s), active stereovision is unaffected when multiple devices measure the same area simultaneously. Compared with ToF, active stereovision has higher spatial resolution and is not affected by multipath interference.

While an active stereovision depth sensor can achieve the satisfactory performance on objects with diffused surfaces where the projected pattern is uniformly reflected, the results on objects with challenging materials (e.g., glossy or transparent) are inferior due to the complex lighting effects of the projected pattern on these surfaces. To generate simulation data with a relatively low domain gap, the simulation process must also generate the same material-dependent measurement error patterns as real sensors.

### B. Physically Based Rendering

PBR seeks to render images in a way that models the transport of light in the real world. In nature, a light source emits light rays that travel to surfaces interrupting their progress. Depending on the surface material, the rays could get absorbed, reflected, refracted, or any combination of these. A PBR pipeline seeks to model both the light transport process and the surface material in a physically correct way.

*Light transport process:* Many rendering techniques have been proposed to model the light transport process, such as ray casting, recursive path tracing, and photon mapping. These methods are covered by an umbrella term *ray tracing*. In general, a ray tracing algorithm calculates the color for each pixel by tracing a path from the camera and accumulates the material-dependent weight along the path to light sources. Compared to the traditional scan-line-based renderer (rasterizer), ray-tracing-based renderers (ray-tracers) are typically slower but can produce much more photo-realistic images with complex indirect lighting effects. They have shown promising sim-to-real transferability on vision tasks, e.g., object recognition [63], [64]. Recent years have seen great progress on accelerating ray tracing from both academia and industry. The adoption of deep-learning-based denoising [65], [66] and super-resolution [67] has made real-time ray tracing possible. In this work, we implement a new ray-tracing-based renderer that integrates many of these cutting-edge acceleration approaches and customizable features to support our depth sensor simulation pipeline.

*Surface material modeling:* Another crucial factor in producing photo-realistic images, as aforementioned, is the physically correct modeling of surface materials and light transportation. In the real world, the bidirectional scattering distribution function (BSDF) describes how light scatters from a surface. It is defined as the ratio of scattered radiance in a direction  $\mathbf{o}$  (outcoming light direction) caused per unit irradiance incident from direction  $\mathbf{i}$  (incoming light direction). We denote it as  $f(\mathbf{i}, \mathbf{o}, \mathbf{n})$  where  $\mathbf{n}$  denotes the local surface normal. The BSDF is often called BRDF or BTDF when restricted to only reflection or transmission. The BSDF of a real-world surface is very complex. However, researchers have proposed many parameterized models to approximate the function. Among all these efforts, the Disney PBR material model [68] (referred to as PBR material)

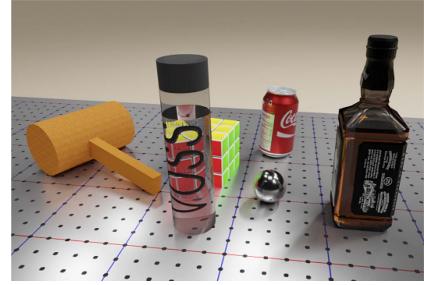


Fig. 4. Ray-tracing rendering result of a table-top scene. Our renderer features broad support for PBR materials and can produce realistic indirect lighting effects (e.g., soft shadow, reflection, and refraction).

is most widely recognized. Based on a large-scale investigation of various real materials, the model parameterizes the BSDF into a set of parameters including base color, metallic, specular, roughness, anisotropic, sheen, clearcoat, transmission, emission, etc. This model is widely adopted by industry-level softwares, such as Blender, Renderman, and Unreal Engine. A detailed demonstration of the effect of these parameters can be seen in Blender Principled BSDF Introduction.<sup>2</sup> In our rendering solution, we have adopted a version of the PBR material model that includes base color, metallic, specular, roughness, index of refraction (IOR), transmission, and emission as its parameters. These parameters suffice to model most common materials in the real world (see Fig. 4, for examples). Many recent works have employed a PBR material [60], [69], [70]. But they all neglect some of the parameters (metallic, transmission, or emission). Combined with their rasterization-based renderers, this could lead to unrealistic results for some common household objects, such as lamps, mirrors, and glasses under complex lighting conditions.

## IV. METHOD OVERVIEW

In this section, we give a more detailed description of our objective and our method to achieve it.

The objective of our work is to close the optical sensing domain gap between the depth sensor simulator and the real depth sensor. The depth sensor simulator is defined as

$$I_{\text{depth}} = D_{\psi}(\mathcal{S}_{\text{objects}}, \mathcal{S}_{\text{lighting}}) \quad (1)$$

where  $I_{\text{depth}}$  is the output depth map,  $\psi$  is the depth sensor simulator parameters,  $\mathcal{S}_{\text{objects}}$ ,  $\mathcal{S}_{\text{lighting}}$  are the object status and the lighting status in the robot simulation environment.  $\mathcal{S}_{\text{objects}}$  includes the geometries, PBR materials (base color, metallic, specular, roughness, transmission, IOR, emission), and poses of objects.  $\mathcal{S}_{\text{lighting}}$  includes the positions and intensities of light sources. In this article, we assume the geometries and base colors of objects are available from textured CAD models.<sup>3</sup> The positions of light sources are approximated in the real world. This is a common and useful setup in robotic research. In many

<sup>2</sup>[Online]. Available: [https://docs.blender.org/manual/en/latest/render/shader\\_nodes/shader/principled.html](https://docs.blender.org/manual/en/latest/render/shader_nodes/shader/principled.html)

<sup>3</sup>For example, if we use a multiview stereo 3-D scanner to reconstruct the surface of 3-D objects, we will acquire such textured CAD models.

scenarios, such as logistics, a collection of CAD models can be build beforehand for all the objects of interest.

The optical sensing domain gap can be evaluated as follows: a group of perception models are evaluated on one simulation dataset and one real dataset, which are accurately aligned. The optical sensing domain gap is considered smaller if the evaluation results on the simulation dataset are more correlated with those on the real dataset. Moreover, the effectiveness of simulation as the training domain is reflected by the algorithm performance in the real world.

In order to close the optical sensing domain gap, we will 1) acquire the unknown PBR materials of objects, including metallic, specular, roughness, and transmission, along with the intensities of light sources; and 2) construct  $D_\psi()$  by simulating an active stereovision depth sensor in a physics-grounded fashion.

## V. MATERIAL ACQUISITION

Synthesizing data from textured CAD models is a common practice in previous robotic works [3], [71], [72]. Almost all of these works assume a simple material model (e.g., Lambertian) and do not set the material parameters carefully. This is generally acceptable when using a rasterization-based renderer, which inherently cannot generate indirect lighting effects such as specular reflection, refraction, and shadows, without additional tricks. Because our goal is to count in all real-world factors including these lighting effects, a ray-tracing renderer along with a good PBR material model are needed.

In order to acquire the object material parameters, we first develop a labor-saving and markerless pipeline to achieve pixel-wise alignment between the simulation and real scenes. Furthermore, with the scene setting aligned, we can achieve material acquisition by optimizing PBR material parameters for each object part.

### A. Pixel-Wise Alignment Between Simulation and Real World

To retrieve the materials of objects, it is necessary to accurately align the scene configuration, including camera parameters, object geometries, and poses, between the simulation and the real world so that the rendered images are pixel-wise aligned with the real-world captures. Previous works achieved this goal by estimating the object poses in the initial camera frame, using markers (e.g., quick response (QR) code cardboards) to track the pose of camera frames, and computing the object pose for each frame [73], [74], but it is undesirable to leave markers in captured images. We eliminate the need for markers by first generating the rendered images and then aligning the real scene layout with the simulation instead of aligning the simulation with the real world.

Fig. 5 shows the setup for material acquisition. The active stereovision depth sensor is fixed at the robot end effector. Objects are placed on the table, and the depth sensor captures them from different angles. To align the simulation environment with the real world, we need to compute the relative transformations for depth sensors, the robot, and the table in the following steps:

- 1) the camera intrinsic matrices of the depth sensor are obtained from the firmware;

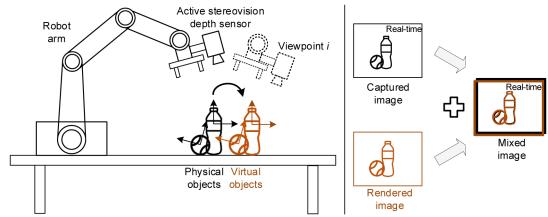


Fig. 5. Demonstration of setup for material acquisition, and pixel-wise alignment between the simulation and the real world.

- 2) the relative transformations between cameras and the table of all the viewpoints are computed using OpenCV's `solvePnP` [75], [76];
- 3) the relative transformation between the depth sensor and the robot end effector is computed using hand-eye calibration from OpenCV [77].

Once all the relative transformations are obtained, we use the aligned poses of the depth sensors, the robot, and the table to construct the simulation environment and maintain them across all scenes.

After building the simulation environment, we proceed to create aligned object layouts between the simulation and the real world. We first use the physics simulation to generate a physically plausible scene layout of objects in simulation following the pipeline in [3]. Second, we render images of the scene. With well-aligned transformations between the simulation and the real world, the pose of the table is precisely aligned in rendered images and captured images. As shown in the right part of Fig. 5, we generate the mixed image in real time by overlaying the rendered image on the captured image as the feedback, and adjust the object poses according to this real-time feedback without the aid of markers. Because the layout is physically plausible, we can accurately align the object poses in the real world with those in the simulation. We measure the displacement accuracy using a motion capture system, and the result demonstrates that we can achieve an error of less than 2 mm, 1.5°, which is sufficient for the evaluation of pose estimation.

### B. Material Acquisition

With the scene layout already aligned between the simulation and the real world, we can now retrieve the material parameters of objects in the scene. Note that we need approximated positions of main light sources in this step ( $\sim 10$  cm accuracy in our experiments), and the light intensities will be optimized automatically along with the material parameters. We first preprocess the real captured images using automatic exposure and white-balancing to align the image brightness and color, and then we search through light intensities and several important PBR material parameters, including roughness, metallic, specular, and transmission, for each object part to find the best match.

1) *Multispectral Matching Loss*: For an aligned pair of multispectral images ( $I_{\text{RGB}}^{\text{sim}}, I_{\text{IR}}^{\text{sim}}$ ) and ( $I_{\text{RGB}}^{\text{real}}, I_{\text{IR}}^{\text{real}}$ ), the matching loss  $L$  is defined as

$$L = L_{\text{RGB}} + \lambda L_{\text{IR}} \quad (2)$$

$$L_{\text{RGB}} = L_2(I_{\text{RGB}}^{\text{sim}}, I_{\text{RGB}}^{\text{real}}) + L_{\text{percept}}(I_{\text{RGB}}^{\text{sim}}, I_{\text{RGB}}^{\text{real}}) \quad (3)$$

$$L_{\text{IR}} = L_2(I_{\text{IR}}^{\text{sim}}, I_{\text{IR}}^{\text{real}}) + L_{\text{percept}}(I_{\text{IR}}^{\text{sim}}, I_{\text{IR}}^{\text{real}}) \quad (4)$$

where  $L_{\text{percept}}$  is the perceptual loss used to match the overall appearance of the objects. The perceptual loss is defined as the  $L_2$  difference between AlexNet features of the input rendered 2-D images [78]. We empirically find that by adding the perceptual loss, we can achieve better material acquisition results with no need for strict color, exposure, and lighting condition alignment, while plain  $L_2$  could give suboptimal results due to bias on brightness and color.

2) *Adaptive Grid Search*: Thanks to the real-time performance of our ray-tracing pipeline, we can perform a grid search on the PBR material parameters for each object part to find the optimal match. The parameter set  $\mathcal{P}$  we search for includes roughness, metallic, specular, and transmission. We first perform a coarse-level grid search with 10 grid samples for each parameter. After obtaining the best set  $\mathcal{P}^{\text{coarse}}$ , we perform another fine-level grid search with 10 grid samples in the neighborhood of  $\mathcal{P}^{\text{coarse}}$  to acquire the final material parameter approximation  $\mathcal{P}^{\text{fine}}$ . These identified material parameters are then applied to 3-D models and used to render images.

## VI. ACTIVE STEREOVISION DEPTH SENSOR SIMULATION

In this section, we introduce individual components of our end-to-end solution for active stereovision depth sensor simulation.

This proposed pipeline follows the same underlying mechanism of real active stereovision depth sensors. Specifically, we simulate the full processing flow from IR pattern projection and rendering to IR noise simulation and depth reconstruction. We build our pipeline based on SAPIEN [18] as it already has state-of-the-art physical simulation and user-friendly robotic interfaces integrated so that we can focus on the sensor simulation.

### A. Real-Time PBR

Originally, SAPIEN features a real-time rasterization renderer, which meets the demands of many simple robotic and vision tasks that do not require a high degree of realism. However, complex optical effects such as indirect lights, soft shadows, cannot be easily rendered by such a rasterization-based pipeline, leading to unrealistic results.

Building upon the SAPIEN platform, we add a new ray-tracing renderer named *Kuafu*<sup>4</sup> in this article. Kuafu is implemented with the Vulkan ray-tracing API, which provides the real-time performance as well as cross-vendor and cross-platform compatibility. Kuafu renderer can generate physically correct realistic rendering results (see Fig. 4) and can render in real time for common robotic scenes under reasonable settings. In addition, Kuafu supports many advanced render functionalities, including geometry instancing, environment map, and customizable render target. To support the realistic sensor simulation pipeline, we highlight the following features.

<sup>4</sup>Kuafu Real-Time Ray Tracing Renderer: <https://github.com/jetd1/kuafu>

1) *PBR Material Support*: Kuafu adopts a subset of the widely-used Disney PBR material model [68]. For the shader implementation, we follow these existing frameworks to use the Lambertian model or (optional) the Oren-Nayar model [79] for diffused component modeling and GGX microfacet model [80] for specular component modeling. Our material model includes base color, metallic, specular, roughness, IOR, transmission, and emission. All parameters can be either from a single scalar or from an assigned image texture, giving it great flexibility to model all common object materials. It is worth noting that we feature full support for transmissive and emissive materials, which is uncommon in previous simulation platforms, however supporting these materials is crucial for accurate sensor simulation, as will be shown in later sections.

The final BSDF model we use is a mix between the diffuse term and the specular term

$$f(\mathbf{i}, \mathbf{o}, \mathbf{m}) = w_d f_d(\mathbf{i}, \mathbf{o}, \mathbf{m}) + (1 - w_d) f_s(\mathbf{i}, \mathbf{o}, \mathbf{m}) \quad (5)$$

$$w_d = (1 - \mu) * (1 - \alpha) \quad (6)$$

where  $\mu$  is the material metallic,  $\alpha$  is the object transmission,  $f_d$  is the diffuse term (Lambertian or Oren-Nayar), and  $f_s$  is the specular term (GGX microfacet model).  $\mathbf{i}$ ,  $\mathbf{o}$ ,  $\mathbf{m}$  denote unit incoming light direction, outgoing light direction, and local surface normal, respectively.

2) *Textured Light Support*: Kuafu supports a wide variety of light types, including directional, point, spot, and area lights. Most lights in Kuafu feature texture map support. This is implemented in shaders by first determining which pixel of the light texture the ray passes through when tracing shadow rays, and then attenuating the light intensity according to the light texture. This feature further improves flexibility and extends to a broad range of usages. For instance, combined with the variable texture support, one can use lights with programmed textures to simulate the shape from the shading pipeline [81] by changing the light pattern across the frames. Also, the pattern projector in active stereovision sensors can be simulated by a textured spot light in Kuafu, and thus, this functionality is crucial for our proposed simulation pipeline.

3) *Denoising Support*: Images from a vanilla ray tracing pipeline typically contain significant noise due to the underlying Monte Carlo sampling process. The noise is especially severe when the scene is underilluminated or rendered with a small number of SPP, both of which can be encountered when rendering IR images. To reduce these undesired noises while using smaller SPPs, we integrate two deep-learning-based denoisers into the renderer—NVIDIA OptiX Denoiser [82] and Intel Open Image Denoise. By explicitly sharing GPU memory between Kuafu and the OptiX CUDA buffer, we eliminate user-space memory copying and thus achieve high-quality denoising with a minimal performance impact. As showed by Chaitanya et al. [66], deep-learning-based denoisers can achieve visual appearance very similar to PBR. Therefore, we feel it reasonable that our work can still be summarized as physically grounded.

TABLE I  
RENDERING FRAME RATE (FPS) WITH DIFFERENT NUMBERS OF SPP

Sample per pixel	w/o Denoiser	w/ Denoiser
2	352.80	140.63
8	146.51	100.47
32	57.36	43.70
128	27.93	24.79

The experiments are conducted on a single RTX 4090 graphics card. The rendering resolution is  $960 \times 540$ .

### B. IR Image Rendering

With the support of textured spot light, it is straightforward to render IR images in Kuafu. In our implementation, when rendering IR images, we assume that all existing lights in the environment will also cast rays in the IR spectrum, with reduced intensity  $a * l$ , where  $l$  is the light intensity in the visible spectrum and  $a$  is an attenuating factor set to 0.05 in all our experiments. We also assume a weak ambient light value simulating radiance from the environment (e.g., sunlight). We render images with this special light setting, and then take the R channel of the denoised images as the final IR rendering results. The rendering speed of an IR image is shown in Table I.

There are indeed commercial physically based solutions like Blender and 3dsMax, which feature a similar full ray-tracing pipeline. However, these solutions are often heavy and could bring in a large amount of run-time overhead. By implementing a similar full-featured pipeline ourselves, we can:

- 1) achieve performance gains by optimizing out the unnecessary steps;
- 2) implement certain techniques for the better efficiency and performance (e.g., sharing GPU memory between Kuafu and the stereo matcher, and customize on-device postprocessing);
- 3) build a unified pipeline into robotics simulators.

Users can use the Kuafu renderer and depth simulator out-of-the-box in our SAPIEN robotics simulator with only a few lines of code.

### C. Sensor Noise Simulation

After the IR projection, rendering, and denoising steps, we can acquire clean IR images with complex indirect lighting effects. However, real-world cameras and projectors are imperfect. Due to hardware or manufacturing limitations, real-world cameras, and projectors will introduce different kinds of noise into IR images.

We generally follow the model used in [32] to model the IR sensing noise. After obtaining the rendered IR image  $I_{\text{clean}}$  (in light intensity), we corrupt the image with random noise to simulate the noise encountered in real optical systems. Our noise model contains two parts: a multiplicative term  $\gamma$  modeling the laser speckle and an additive term  $n$  modeling camera thermal noise

$$I_{\text{noisy}} = \gamma * I_{\text{clean}} + n \quad (7)$$

where  $*$  and  $+$  denote element-wise operations. According to [83], the total power received by a pixel can be modeled as

the sum of exponentially distributed power (Rayleigh voltage) random variables. As a result, we can fit a gamma distribution to sample this noise

$$f_\gamma(x; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \quad (8)$$

where  $k$  and  $\theta$  are two noise parameters. The camera thermal noise can be modeled by a Gaussian distribution  $\mathcal{N}(\mu, \sigma^2)$  [84].

For a specific sensor that we choose to simulate, we estimate the parameters of its noise model by capturing multiple sets of IR images, where each set contains frames of the same static scene. In this way, we can calculate the noise-free approximation of each scene by averaging the frames, and then estimate the noise model parameters with a maximum a posteriori probability estimate.

### D. Depth Generation by Stereo Matching

After noise simulation, a series of computationally expensive stereo matching steps need to be performed to generate the final depth map. Usual CPU implementations have a computation time that is far from real time. To accelerate our pipeline, we build a high-speed GPU-based depth computing module named *SimSense*.<sup>5</sup> *SimSense* is an encapsulated module that integrates all necessary algorithms to compute depth from a pair of stereo images. It is highly configurable and has been fully accelerated with CUDA.

With a pair of stereo images as input, *SimSense* first performs a stereo rectification to project the images onto a common image plane. It then performs a center-symmetric census transform (CSCT) [85] on the input as suggested in the RealSense overview [86]. After CSCT, a four-path semiglobal matching (SGM) [87] is performed to search for the best disparity candidates, with hamming distance as the cost function. A uniqueness test is done to filter out disparities that are not better than the second best match by a threshold. Sub-pixel-level disparity is generated by performing quadratic curve fitting. Several common disparity postprocessing steps including left-right consistency check and median filtering can be done to refine the disparity map. The refined disparity map is then converted into a depth map, with an optional depth registration that aligns the depth map to the RGB camera frame.

We follow the parallel scheme proposed by Hernandez-Juarez et al. [88] to accelerate CSCT and SGM. CSCT and matching cost computation of SGM make use of shared memory to optimize for data reuse. Cost aggregation of SGM is further accelerated with warp-based optimization. Rectification, uniqueness test, subpixel interpolation, left-right consistency check, median filter, and depth registration are integrated within the parallel scheme with little overhead. Besides the original SGM, *SimSense* also supports semiglobal block matching (SGBM). While SGM computes the matching cost of two pixels solely by their hamming distance, SGBM computes the cost by the hamming distance between the local regions of the two pixels. Such implementation is common in library like OpenCV [75]. Table II shows the computation time performance of *SimSense*

<sup>5</sup>SimSense depth computing module: <https://github.com/angli66/simsense>

TABLE II  
DEPTH COMPUTING FRAME RATE (FPS) UNDER WITH DIFFERENT MAX DISPARITY VALUES

Max disparity	SGM (Ours)	SGBM (Ours)	SGBM (OpenCV)
64	414.51	342.24	0.73
96	326.79	268.98	0.71
128	281.26	232.49	0.68
256	147.13	128.56	0.60

Our SimSense algorithms run on a single RTX 4090 graphics card. The OpenCV implementation is run on a machine with i5-10400 CPU @ 2.90 GHz, and 16 G memory. The input image resolution is  $960 \times 540$  and the generated depth map resolution is  $1920 \times 1080$ .

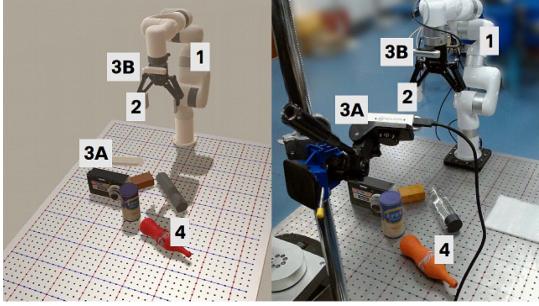


Fig. 6. Experimental setup: 1) 7-DoF robot arm, 2) parallel gripper, (3A) Intel RealSense D415 mounted on the table (base RealSense), (3B) Intel RealSense D415 mounted on the robot end effector (hand RealSense), 4) Target Objects.

with a comparison to the CPU implementation of SGBM with OpenCV.

Key parameters are designed to be adjustable throughout the pipeline. Thanks to the configurability of the module, the entire system can easily adapt to new sensors.

## VII. SYSTEM SETUP

### A. Experimental Setup

Fig. 6 shows the experimental setup in simulation and the real world. It consists of a 7 DoF robot arm (UFACTORY xArm 7), a parallel gripper (Robotiq 2F-140), 2 IR active stereovision depth sensors (Intel RealSense D415), and objects. One of the depth sensors is fixed on the table (referred to as *base RealSense*) and the other is mounted to the robot end effector (referred to as *hand RealSense*). We further build the simulation environment in SAPIEN [18] and align the real world and simulation as described in Section V-A. Fig. 7 shows the alignment result.

### B. Objects and Material Acquisition

As shown in Fig. 8, there are two kinds of objects used in our experiments: 9 3D-printed objects and 8 real-life objects. 3D-printed objects are printed using color plaster powder from CAD models, therefore, the geometry and texture consistency between physical objects and CAD models is guaranteed and researchers can print these objects by themselves. Note that “texture” here only refers to surface reflectance, also known as albedo or diffuse color of an object. Because the materials of 3D-printed objects are all Lambertian diffused, we further choose 8 real-life objects of different optical materials (specular,

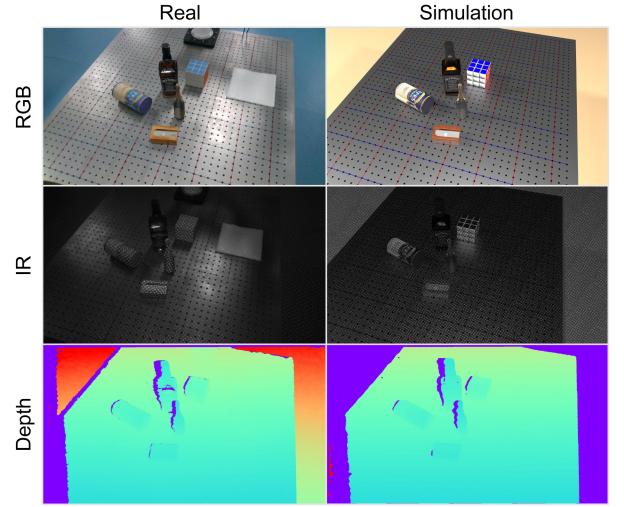


Fig. 7. Pixel-wise alignment between simulation and the real world.

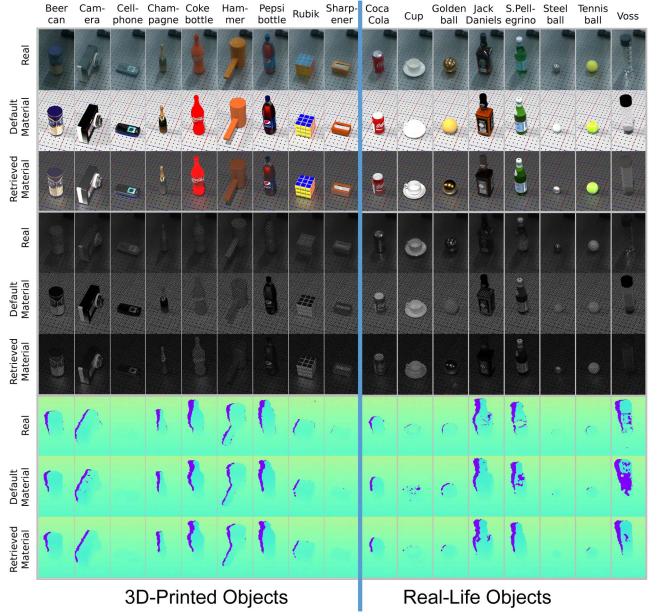


Fig. 8. Comparison of captured and rendered RGB images, IR images, and depth maps of 3D-printed objects and real-life objects.

translucent, transparent), which are challenging for active stereovision depth sensors. We approximate the material parameters of all the objects by conducting material acquisition as described in Section V. For each object, it takes less than 1 s to render the synthetic images for object pose alignment. And it takes about 2 min to control the robot arm to move the camera around the object and capture images in the real world. Finally, it takes less than 1 h to search for the best material parameters for an object.

### C. Sensor Parameters

For our Intel RealSense D415, to estimate the parameters of our noise model as described in Section VI-C, we capture 33 sets of IR images, where each set contains 100 frames of a static scene. The estimated parameters are  $\mu = -0.231$ ,  $\sigma =$

$0.83$ ,  $k = 3.98$ ,  $\theta = 0.254$ . To retrieve the IR projector pattern of D415, we capture a perpendicular wall multiple times to filter out noise and then use simple thresholding to extract the pattern.

### VIII. DOMAIN TRANSFERABILITY EVALUATION ON DIVERSE APPLICATIONS AND ALGORITHMS

We define four applications of the proposed depth sensor simulator to evaluate the domain transferability.

- 1) *Generate training data for object detection* Object detection is one of the most widely deployed perception algorithms. In Section VIII-A, we generate a training dataset using the proposed depth sensor simulator. On the simulated training dataset, learning-based object detection algorithms are trained before being evaluated on the real-world dataset.
- 2) *Generate training data for 6-D pose estimation* Pose estimation is critical for downstream robotic applications (pick-and-place, manipulation). Compared with object detection, it is more difficult to manually label accurate annotations of 6-D poses. In Section VIII-B, we choose three learning-based 6-D pose estimation algorithms and run them through the same pipeline as object detection.
- 3) *Generate training data for robot grasping* In contrast to object detection and 6-D pose estimation, where annotations can be generated simply from the state of the static virtual scene, the annotations of robot grasping require physical interaction with the environment. To do this, we integrate the proposed depth simulator with the full-physics robot simulation environment SAPIEN [18]. In Section VIII-C, we train a continuous robot control policy in the simulation environment, and directly test the policy in the real world.
- 4) *Estimate the ranking of pose estimation algorithms* in real world In addition to generating training data, the proposed depth sensor simulator can also be used to rank algorithms. It facilitates researchers to compare different algorithms in simulation without conducting experiments in the real world.

In Section VIII-D, we generate a simulated dataset using our proposed depth sensor simulator, and align it with a real-world dataset. Different pose estimation approaches are assessed on the datasets simulated by our simulator and other approaches. We compare the rankings of algorithms on simulated datasets with rankings on the real-world dataset.

#### A. Generate Training Data for Object Detection

1) *Training Dataset:* For both learning-based object detection and pose estimation, we generate a large-scale synthetic training dataset. As shown in Figs. 6 and 7, the scene in the dataset is constructed by randomly choosing 5 objects and placing them on the table, and depth maps are generated by simulating the depth sensor from different viewpoints. In order to avoid generating unreasonable object layouts, we use physics simulation to generate physically plausible object layouts following [3]. We use 41 000 images for training. We use SPP=128 for ray-tracing rendering with learning-based denoising and

TABLE III  
COMPARISON OF MAP@0.5 OF OBJECT DETECTION ON REAL DATA

Data	Depth	RGB	RGBD
Clean depth	0.730	-	-
Rasterization	0.942	0.839	0.875
<b>Ours</b>	<b>0.977</b>	<b>0.941</b>	<b>0.943</b>

The bold entities show that our method outperforms the other two methods.

sensor noise simulation. During training, we use online data augmentation, including color jittering and color drop for RGB images, Gaussian noise, and randomly setting depth to zero for depth maps.

2) *Real-World Test Dataset:* In order to collect a real-world test dataset with the ground-truth object detection and pose estimation labels, we first generate scene layouts in the simulation and build the real-world scenes using the sim-real pixel-wise alignment as described in Section V-A, so that we can acquire precise object poses without the aid of markers. We collect 504 real images of 24 different scenes using the two depth sensors as the real-world test dataset.

3) *Input Modality: Depth Versus RGB, RGBD:* Here, we evaluate the impact of input modality on object detection. We choose YOLOv3 [89] as the object detection algorithm. We use three types of input: depth, RGB, and RGBD. We use the pretrained weights on COCO [90], and the weights of the first layer are not loaded for depth and RGBD input.

4) *Experimental Results:* We use mAP@0.5 (mean of average precision with IoU threshold=0.5) as the evaluation metric. Table III shows the comparison result of object detection on real-world images. “Clean depth” represents using the depth buffer from the renderer. “Rasterization” represents simulating the process of active stereovision depth sensor by using rasterization rendering. The performance improvement on RGB images can be attributed to the improved fidelity of ray-tracing rendering, which is consistent with existing works [63], [64]. Comparing the three input modalities, we can find that depth maps have the best transferability, as they are primarily affected by geometry, whereas RGB images are also affected by surface texture, material, and environmental illumination. It can be shown that, when using the depth map as input, the detection network trained on the simulated data generated by our method has better performance than rasterization and clean depth, demonstrating that our method provides for development of better object detection algorithms.

#### B. Generate Training Data for 6-D Pose Estimation

*Datasets:* We use the same training and test datasets as we use for object detection (see Section VIII-A).

1) *Pose Estimation Algorithms:* We choose three learning-based 6-D pose estimation algorithms.

1) PVN3D [91]. PVN3D predicts the keypoint offsets and semantic segmentation for each point in the point cloud, and uses a deep Hough voting network to predict 6-D object poses.

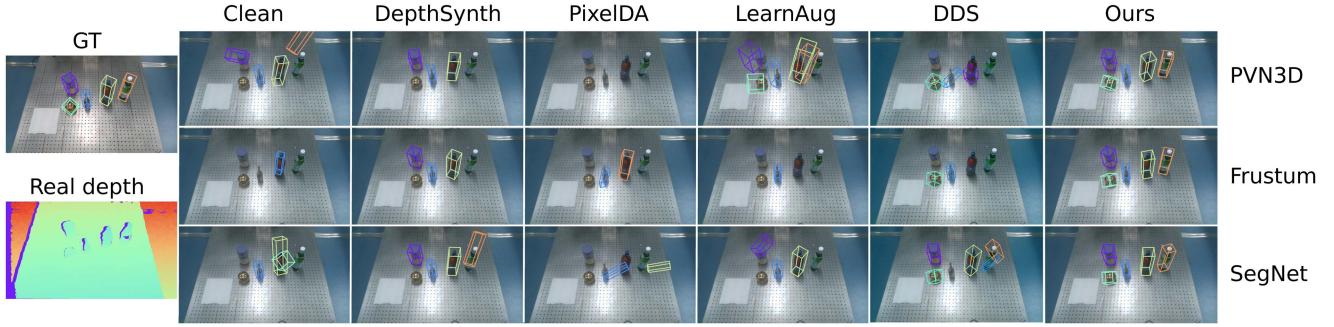


Fig. 9. Qualitative comparison of 6-D object pose estimation algorithms on real depth images. The scene is challenging for pose estimation as the depth measurement of real objects (Golden ball, S.Pellegrino) is noisy and incomplete. All the three pose estimation algorithms are able to infer accurate poses while trained solely on the simulated data generated by our method. Note that we use depth maps for the pose estimation and RGB images are only used for better visualization.

- 2) **Frustum PointNets [92].** Frustum PointNets first predict 2-D bounding boxes of objects in the depth map and further estimate the 6-D object poses using points within the 3-D frustums. In our experiments, we use YoloV3 [89] for 2-D detection and adopt the Hough voting strategy from PVN3D to improve the performance.
- 3) **SegICP [93].** SegICP first predicts per-pixel semantic segmentation in the depth map, unprojects the pixels of one object into 3-D point cloud, computes the object pose by registering the point cloud with CAD models of different initial poses and selects the result with minimal error as the prediction.

Note that we only use depth information for all the three algorithms, and no RGB information is used.

**Baseline methods:** We compare our proposed depth sensor simulation method with one simple baseline (referred to as *Clean*) that simply uses the depth buffer from the renderer and four state-of-the-art sim-to-real methods: *DepthSynth* [33], *PixelDA* [25], learning to augment (referred to as *LearnAug*) [37], and differentiable depth sensor simulation (referred to as *DDS*) [38]. We chose these four works for benchmarking because each represents a distinct type of depth sensor simulation method. *DepthSynth* simulates the process of an active stereovision depth sensor by rendering binocular images using rasterization rendering and generating the simulated depth by using stereo matching. It represents methods simulating the depth sensor's mechanism. *PixelDA* is an unsupervised GAN-based domain adaptation method that decouples the task-specific content and domain appearance, and it represents learning-based domain adaptation methods. *LearnAug* is a domain randomization method that searches for the optimal sequence of data augmentation policies through MCTS to maximize the trained detection network's performance on a real dataset with ground-truth object detection annotations. It represents domain randomization methods. *DDS* builds an end-to-end differentiable stereo matching pipeline and optimizes the rendering parameters by using the loss between the real captured depth maps and the generated depth maps. It represents differentiable-rendering-based optimization methods.

**Experimental results:** We evaluate the performance of pose estimation algorithms by using the percentage of predictions

whose rotation and translation errors are smaller than certain thresholds. In this article, we choose two thresholds:  $(10^\circ, 10 \text{ mm})$ ,  $(20^\circ, 20 \text{ mm})$ . For symmetrical objects, the rotation error is computed as the smallest error of all possible rotations. In order to analyze the influence of object materials, we further compute the metrics of real objects and 3D-printed objects separately.

Fig. 9 shows the qualitative results. As can be seen in the real depth image, the 3D-printed objects can be properly measured by the real active stereovision depth sensor, whereas the measurement of real objects is noisy and incomplete, which makes it more challenging to simulate. All three pose estimation algorithms can make accurate predictions when trained solely on the simulated data generated by our depth simulator, which validates the effectiveness of our method for generating training data for 6-D pose estimation.

Table IV shows the comparison result of different pose estimation algorithms on real-world depth images. In general, our method can significantly improve the performance of all three algorithms, with the improvement on real objects being greater than that on 3D-printed objects. Because PVN3D uses PointNet++ to extract the geometry information for pose estimation, it is more sensitive to geometry distortion. It has unsatisfactory performances when trained on PixelDA and LearnAug, demonstrating that these two methods introduce additional unrealistic geometric distortions, which is also claimed in Shen et al. [36]. Compared with 3D-printed objects, the performance of *DepthSynth* on real objects is decreased. It is because 3D-printed objects are Lambertian diffused on which light is uniformly reflected and can be well rendered by rasterization rendering technique, but the light transformation process on real objects is complicated and cannot be correctly rendered by rasterization. The probable reason that *DDS*'s performance is suboptimal is that it uses a differentiable stereo matching module whose capability is limited compared with SGBM. Moreover, the differentiable rendering used in *DDS* does not support the rendering of transparent materials.

**Comparison with depth completion methods:** We compare our proposed depth simulation method with two learning-based depth completion methods: ClearGrasp [41] and TransCG [44].

TABLE IV  
COMPARISON OF 6D OBJECT POSE ESTIMATION ON REAL DEPTH IMAGES

Pose algo.	Training data	Overall(%)		Real(%)		Printed (%)	
		10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm
PVN3D	Clean	11.54	16.47	9.59	16.62	12.82	16.03
	DepthSynth	54.06	68.93	30.49	51.52	74.09	84.07
	PixelDA	0.00	0.00	0.00	0.00	0.00	0.00
	LearnAug	0.32	0.48	0.35	0.71	0.28	0.28
	DDS	23.57	39.21	23.11	42.42	23.96	36.48
	<b>Ours</b>	<u>63.37</u>	<b>83.45</b>	<u>42.60</u>	<b>74.81</b>	<b>80.95</b>	<b>90.77</b>
Frustum	Clean	2.98	4.93	1.55	2.15	4.35	7.53
	DepthSynth	40.93	68.26	34.33	62.22	46.62	74.32
	PixelDA	15.51	26.81	12.42	27.39	18.57	27.75
	LearnAug	18.89	30.23	11.27	18.78	26.10	41.33
	DDS	20.63	35.99	21.21	39.65	20.15	32.89
	<b>Ours</b>	<u>43.25</u>	<u>75.08</u>	<u>38.53</u>	<u>71.08</u>	<u>47.25</u>	<u>78.46</u>
SegICP	Clean	5.69	7.20	2.48	3.84	8.16	9.79
	DepthSynth	55.65	68.26	39.12	60.43	69.61	75.62
	PixelDA	15.04	20.01	5.04	10.21	23.37	28.21
	LearnAug	37.07	46.50	19.37	33.17	51.70	58.16
	DDS	30.71	42.30	22.16	39.13	37.95	44.98
	<b>Ours</b>	<b>65.52</b>	<u>79.64</u>	<b>50.56</b>	<u>74.72</u>	<u>78.17</u>	<u>83.81</u>

The best result of all three pose estimation algorithms is marked as bold, and the best result of each algorithm is marked as underlined.

TABLE V  
COMPARISON OF DEPTH COMPLETION AND DEPTH SENSOR SIMULATION

Pose algo.	Training data	Test data	Overall(%)	
			10°, 10mm	20°, 20mm
PVN3D	Clean	Real	13.73	19.56
	Clean	ClearGrasp	1.67	6.43
	Clean	TransCG	0.00	0.04
	<b>Ours</b>	Real	<b>63.37</b>	<b>83.45</b>
Frustum	Clean	Real	3.85	5.12
	Clean	ClearGrasp	0.16	0.52
	Clean	TransCG	0.00	0.12
	<b>Ours</b>	Real	<b>43.25</b>	<b>75.08</b>
SegICP	Clean	Real	5.83	7.46
	Clean	ClearGrasp	1.39	3.29
	Clean	TransCG	0.48	0.71
	<b>Ours</b>	Real	<b>65.52</b>	<b>79.64</b>

The bold entities show that our method outperforms the other three methods for all the three pose estimation algorithms.

For ClearGrasp, we train the surface normal estimation module on the synthetic RGB images rendered by the ray-tracing renderer, and complete the real depth map using ground-truth boundaries and object masks. For TransCG, we use the pre-trained weight<sup>6</sup> released by the authors to complete the depth map. For these two methods, we train the pose estimation algorithms on the synthetic *clean* depth and evaluate them on the depth completion outputs. Table V shows the comparison results. For all three algorithms, the pose estimation accuracies on completed depth maps are worse than those on the original captured depth maps. This is likely due to the limited generalizability of learning-based depth completion methods to unseen objects and scenes, which may introduce more artifacts to depth maps and reduce pose estimation algorithms' transferability.

<sup>6</sup>[Online]. Available: <https://github.com/galaxies99/transcg>

### C. Generate Training Data for Robot Grasping

1) *Experimental Setting*: In this section, we choose robot grasping as the task to validate the effectiveness of our method for continuous policy learning, as it is one of the most fundamental robotic tasks. The experimental setting is similar to [13] and [11]. To evaluate the sim-to-real performance of different objects, we only focus on the grasping of a single object. The input is the depth map from the wrist-mounted depth sensor (3B in Fig. 6). The action of each step is  $a = [\Delta x_g, \Delta y_g, \Delta z_g, \Delta \theta_g] \in \mathbb{R}^4$ , where  $(x_g, y_g, z_g)$  is the gripper position,  $\theta_g$  is the rotation of the gripper around the  $z$ -axis. After  $N_g$  steps, the gripper goes down a predefined distance, closes, and lifts. The grasp is considered as success if the object is lifted stably.

We follow the pipeline of [9] to train the grasping policy: we first train an RL policy using privileged states as input, and train a depth-map-based policy by using behavior cloning (BC). We directly deploy the trained depth-map-based policy on the real robot without any finetuning.

*Experimental Details*: We choose 6 optically challenging real objects and 2 3D-printed objects for the real grasping experiment. The object is randomly placed on the table with the range of  $x$  0.15 m,  $y$  0.30 m,  $\theta$  0.6 rad.  $N_g = 10$ . For the RL part, we choose SAC [94] as the RL algorithm. We use Adam as the optimizer with a learning rate of  $1 \times 10^{-3}$ , buffer size  $10^6$ , discount coefficient 0.99, and target smoothing coefficient 0.01. The agent is trained for 540 k steps and evaluated for every 8 k steps. The network weights of the best performance are chosen for the following BC part.

For the BC part, the resolution of the depth map is  $352 \times 224$ . The number of demonstrations is  $6 \times 10^5$ . Data augmentation, including Gaussian noise and randomly setting depth to zero, is used for both our method and baseline methods. We use Adam as the optimizer with a learning rate of  $1 \times 10^{-4}$ , batch size 256, and 100 epochs. The model of the best performance in

TABLE VI  
EXPERIMENTAL RESULTS OF GRASPING IN REAL WORLD FOR CONTINUOUS POLICY LEARNING ON DIFFERENT SIMULATED DEPTHS

Object type	Object name	Simulated depth		
		Clean	DepthSynth	Ours
Real	Coca Cola	7/10	9/10	<b>10/10</b>
	Jack Daniels	4/10	7/10	<b>8/10</b>
	Golden ball	3/10	6/10	<b>10/10</b>
	S.Pellegrino	3/10	3/10	<b>7/10</b>
	Tennis ball	<b>10/10</b>	7/10	<b>10/10</b>
	Voss	8/10	8/10	<b>10/10</b>
Printed	Beer Can	5/10	8/10	<b>10/10</b>
	Sharpener	9/10	7/10	<b>10/10</b>
Overall		49/80	55/80	<b>77/80</b>

The bold entities show that our method outperforms the other two methods for all the objects except Tennis ball.



Fig. 10. 10 unseen optically challenging objects for evaluating the generalizability of the grasping policy.

simulation is used for grasping in the real world. We use SPP=8 with learning-based denoiser for a shorter rendering time cost.

*Experimental results:* For the real grasping experiment, each object is randomly placed on the Table X times within the same range as simulation training. Table VI shows the comparison results of grasping in the real world for continuous policy learning on different simulated depths. The policy trained on our simulated depth achieves an overall success rate of 96.25%, which validates the effectiveness of our method for continuous policy learning algorithm training. Note that the success rate of grasping different objects is dependent not only on the sim-to-real transferability of the policy but also on the target object's shape and friction coefficient. For instance, Golden ball cannot be grasped if there is a little offset between the gripper and the ball, whereas Voss, and Sharpener may be grasped with a larger offset. We further evaluate the trained grasping policy on 10 unseen optically challenging objects as shown in Fig. 10. It achieved a success rate of 100%, which validates the generalizability of the trained policy.

#### D. Estimate the Ranking of Algorithms in Real World

1) *Experimental Setting:* In this section, we validate the effectiveness of our method as algorithm benchmarking for 6-D object pose estimation algorithms. We choose the three pose estimation algorithms: PVN3D, Frustum, and SegICP. For each algorithm, we train it on three simulated depth image datasets generated by different methods (Clean, DepthSynth,

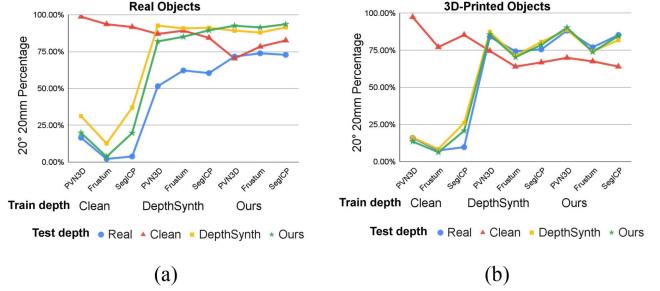


Fig. 11. Metric curve of 9 pose estimation models on (a) real objects and (b) 3D-printed objects.

ours) and make  $3 \times 3 = 9$  pose estimation models with different performances for algorithm benchmarking. We evaluate the performances of the 9 models on both the real test data and the simulated test data generated using baseline methods and our method, respectively. Because the real depth images and the simulated depth images are precisely aligned as described in Section V-A, the evaluation results on the real-world test dataset and the simulated test dataset can be compared directly.

*Experimental results:* An ideal simulated test dataset should deliver evaluation results of different models that are consistent with real-world evaluation results. Therefore, we compute the correlation coefficients of the evaluation results between the real-world test dataset and the simulated test dataset. Table VII shows the experimental results. Our method can deliver more consistent evaluation results than baseline methods on both metrics. In addition, the difference between our method and DepthSynth is greater on real objects than that on 3D-printed objects, which demonstrates that our method can simulate the depth error pattern on optically challenging objects more accurately than DepthSynth.

Fig. 11 shows the performance of the 9 models on real objects and 3D-printed objects. The evaluation results of our method are aligned with the results on real-world depth maps for both kinds of objects. For real objects, the evaluation results of models trained on DepthSynth have higher scores when tested on DepthSynth, which is inconsistent with the results on real-world depth maps. Fig. 12 shows the depth maps of different simulation methods and the real sensor and the 6-D object pose estimation results of PVN3D trained on different simulated data and tested on different simulated data along with real data. Our method can simulate the holes and noise on the specular and transparent objects as the depth map captured by the real depth sensor. For the models trained on different simulated data, the evaluation results on our simulated depth are aligned with the ones on real-world depth maps. For the model trained on DepthSynth, it can estimate the object poses accurately from DepthSynth simulated depth, but fail on both our simulated depth and the real-world depth.

#### IX. ABLATION STUDY

In this section, we conduct experiments to validate our design choices and evaluate the influence of key hyperparameters. We choose the application of generating training data for 6-D pose estimation for the ablation study. According to the order of our

TABLE VII

CORRELATION COEFFICIENTS OF 6-D OBJECT POSE ESTIMATION RESULTS BETWEEN DIFFERENT TEST DEPTH IMAGES AND REAL TEST DEPTH IMAGES

Test depth	Overall		Real		Printed	
	10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm
Clean	-0.562	-0.867	-0.689	-0.791	-0.383	-0.790
DepthSynth	0.973	0.977	0.925	0.947	0.988	0.987
Ours	<b>0.982</b>	<b>0.991</b>	<b>0.959</b>	<b>0.986</b>	<b>0.993</b>	<b>0.992</b>

The bold entities show that our method outperforms the other two methods.

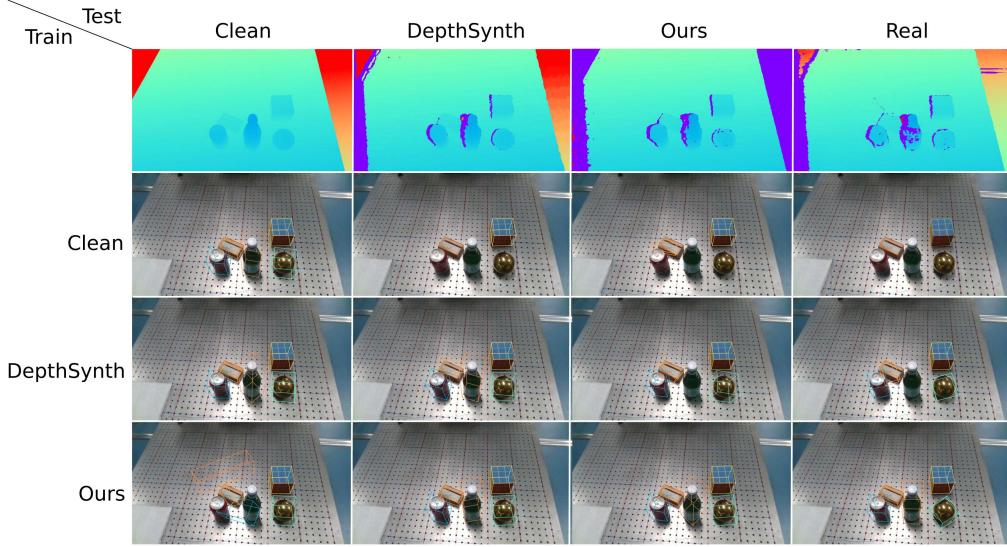


Fig. 12. Qualitative comparison of depth images of different simulation methods and real sensor and the pose estimation results of PVN3D trained and tested on different depth data. Our method can simulate the depth noise pattern on optically challenging objects (specular, transparent). The evaluation results on simulated test data generated by our method are consistent with the real evaluation results for all the three models trained with different simulated data.

TABLE VIII  
ABLATION STUDY ON MATERIAL ACQUISITION FOR GENERATING TRAINING DATA FOR 6-D POSE ESTIMATION

Pose algo.	Material parameters	Overall(%)		Real(%)		Printed (%)	
		10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm
PVN3D	Default	60.87	79.44	37.23	66.58	80.88	90.33
	Random1	62.70	80.28	42.16	70.13	80.07	88.86
	Random2	62.66	80.75	40.00	67.10	81.83	<b>92.31</b>
	Random3	62.10	79.88	40.09	67.45	80.73	90.40
	Ours w/o $L_{percept}$	<b>64.09</b>	83.02	<u>43.12</u>	72.73	<b>81.83</b>	91.72
Frustum	Ours	63.29	<b>83.25</b>	<u>42.60</u>	<b>74.63</b>	80.81	90.55
	Default	41.67	73.53	34.63	66.75	<u>47.62</u>	<u>79.27</u>
	Random1	<u>43.57</u>	<b>74.48</b>	39.05	70.82	47.40	<u>77.58</u>
	Random2	42.18	<u>73.89</u>	36.88	67.97	46.67	78.90
	Random3	40.99	71.83	37.40	68.31	44.03	74.80
SegICP	Ours w/o $L_{percept}$	42.62	72.66	36.80	67.53	47.55	77.00
	Ours	42.42	74.37	<u>39.39</u>	72.38	44.98	76.04
	Default	55.04	66.55	39.13	58.96	68.50	72.97
	Random1	42.18	49.64	36.45	50.22	47.03	49.16
	Random2	62.38	<u>76.55</u>	44.94	68.92	<u>77.14</u>	<u>83.00</u>
Ours	Random3	56.27	68.06	42.86	62.51	67.62	72.75
	Ours w/o $L_{percept}$	50.63	60.28	36.28	53.59	62.78	65.93
	Ours	<b>62.66</b>	76.43	<b>46.93</b>	71.17	75.97	80.88

The best result of all three pose estimation algorithms is marked as bold, and the better result of each algorithm is marked as underlined.

sensor simulation pipeline, we first validate the effectiveness of material acquisition (see Section IX-A), then the effect of rendering settings (see Section IX-B), and finally the sensor noise simulation (see Section IX-C).

#### A. Material Acquisition

Table VIII shows the comparison result of different pose estimation algorithms trained on different material parameters.

TABLE IX  
ABLATION STUDY ON RENDERING SETTINGS, INCLUDING SPP AND DENOISING

Pose algo.	SPP	Denoiser	Overall(%)		Real(%)		Printed (%)	
			10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm	10°, 10mm	20°, 20mm
PVN3D	2	✗	64.33	83.85	44.94	<u>75.58</u>	80.73	90.84
	2	✓	<u>65.00</u>	<b>85.91</b>	43.55	<b>77.58</b>	<b>83.15</b>	<b>92.97</b>
	8	✗	60.48	79.88	41.39	72.38	76.63	86.23
	8	✓	62.42	83.25	41.99	74.72	79.71	90.48
	32	✗	63.89	84.44	42.42	75.67	82.05	91.87
	32	✓	63.29	83.25	42.60	74.63	80.81	90.55
	128	✗	63.29	82.46	43.72	73.16	79.85	90.33
	128	✓	63.37	83.45	42.60	74.81	80.95	90.77
Frustum	2	✗	42.42	74.60	38.87	<u>73.77</u>	45.42	75.31
	2	✓	42.42	75.36	36.19	<u>72.03</u>	47.69	78.17
	8	✗	41.90	75.12	36.71	72.38	46.30	77.44
	8	✓	43.10	<u>75.63</u>	37.58	72.38	<u>47.77</u>	78.39
	32	✗	42.06	<u>74.76</u>	37.14	71.00	<u>46.23</u>	77.95
	32	✓	42.42	74.37	39.39	72.38	44.98	76.04
	128	✗	42.70	74.96	<u>38.35</u>	71.60	46.37	77.80
	128	✓	43.25	75.08	38.53	71.08	47.25	78.46
SegICP	2	✗	64.92	80.60	47.88	74.55	79.34	<u>85.71</u>
	2	✓	65.44	81.19	48.40	76.19	<u>79.85</u>	85.42
	8	✗	<b>66.23</b>	<u>81.51</u>	50.48	<u>76.54</u>	<u>79.56</u>	<u>85.71</u>
	8	✓	65.40	<u>80.79</u>	48.48	<u>74.98</u>	79.71	<u>85.71</u>
	32	✗	50.04	60.63	38.53	58.27	59.78	62.64
	32	✓	62.66	76.43	46.93	71.17	75.97	80.88
	128	✗	52.62	62.34	40.61	56.80	62.78	67.03
	128	✓	65.52	79.64	<b>50.56</b>	74.72	78.17	83.81

The best result of all three pose estimation algorithms is marked as bold, and the best result of each algorithm is marked as underlined.

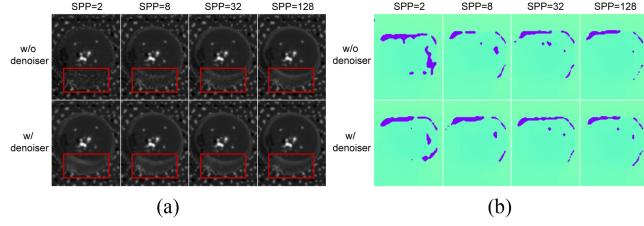


Fig. 13. (a) Rendered IR image and (b) simulated depth maps of different rendering settings.

Our method is compared to a default material parameter set and three random material parameter sets. The default material is set to roughness = 0, metallic = 0, specular = 0, transmission = 1 for Jack Daniels, S.Pellegrino, and VOSS, and transmission = 0 for all other objects. The performance on real objects are consistently improved for all the three algorithms, which demonstrates that the simulation of optically challenging objects requires more accurate material parameters. By combining plain  $L_2$  loss and the perceptual loss, we can improve the performance across the majority of metrics, because  $L_{\text{percept}}$  is more robust to exposure and lighting condition misalignment.

### B. Rendering Settings

Table IX shows the comparison result of different rendering settings: SPP and whether to use the learning-based denoiser. We set SPP to 2, 8, 32, 128. It can be seen that using the denoiser improves the performance across the majority of metrics and the impact of SPP is not remarkable for all the three algorithms. As shown in Fig. 13, when SPP is low, the IR pattern on the metallic ball is noisy and is erased by the learning-based denoiser. Since the two rendered IR images are equally affected by the denoiser

TABLE X  
ABLATION STUDY ON SENSOR NOISE SIMULATION

Pose algo.	Noise scale	Overall(%)	
		10°, 10mm	20°, 20mm
PVN3D	0.0	61.71	79.25
	0.1	62.62	81.11
	0.3	62.30	81.63
	1.0	<u>63.37</u>	<b>83.45</b>
	3.0	60.71	80.71
	10.0	46.19	61.11
Frustum	0.0	41.31	72.42
	0.1	42.66	<u>75.12</u>
	0.3	42.38	<u>73.61</u>
	1.0	<u>43.25</u>	75.08
	3.0	42.74	74.68
	10.0	36.98	60.56
SegICP	0.0	63.29	77.02
	0.1	65.12	78.97
	0.3	<b>65.56</b>	79.21
	1.0	65.35	<u>79.72</u>
	3.0	62.74	77.02
	10.0	21.59	28.81

The best result of all three pose estimation algorithms is marked as bold, and the better result of each algorithm is marked as underlined.

and SPP, the effect is decreased for stereo matching depth maps. Therefore, when focusing on the transferability of depth maps, we can use a low SPP with the denoiser to simulate the depth, whereas we better use a high SPP when aiming for high-fidelity RGB or IR images.

### C. Sensor Noise Simulation

Table X shows the comparison of different sensor noise scales. For the majority of metrics, the best performance is observed with the identified sensor scale. When using a more aggressive

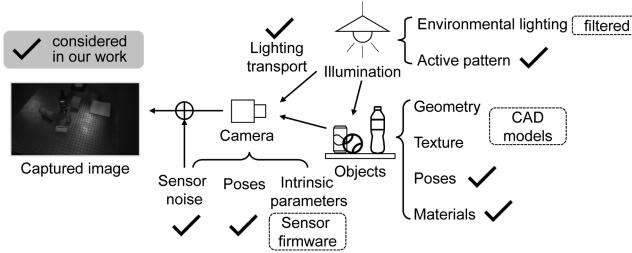


Fig. 14. Image capturing process in the real world.

noise, the trained algorithm may become more robust to noise, but performance is compromised as a result.

## X. CONCLUSION

Despite the thrive of novel algorithms developed in simulators, there has always been strong dispute on whether algorithms learned in simulators can transfer to the real world, among engineers, paper reviewers, and conference attendees. The opponents of robot simulators argue that, obtaining high-quality simulation, if possible at all, will be extremely costly in modeling the 3-D environment and synthesizing realistic data.

In this article, we have presented a physics-grounded active stereovision depth sensor simulator that can generate low optical sensing domain gap depth maps with realistic material-dependent error patterns. As validated by the experimental results, our depth sensor simulator can significantly narrow the sim-to-real optical sensing domain gap and enable the trained algorithms to transfer to the real world without any fine-tuning. We attribute this achievement to the reasons as follows.

As shown in Fig. 14, images captured by complementary metal–oxide–semiconductor (CMOS) in the real world are determined by the camera, the illumination and the objects. Our work takes into account most of these factors in a physics-grounded fashion. Most importantly, by simulating the active stereovision depth sensor instead of passive RGB cameras, we eliminate the complex influence of environmental lighting in the real world and replace it with actively projected pattern that is known and controllable. In fact, we believe that we leverage a key advantage of robot systems which deserve the exploration by many future works—compared with humans, robots can customize their sensing system and actively emit controllable signals to facilitate sensing. By taking this advantage, we can simplify the simulation process and better narrow the sim-to-real domain gap. We look forward to many future works to dig deeper in leveraging this advantage in the sim-to-real transfer field.

In the future, we would further close the sim-to-real gap of depth sensing by modeling more advanced effects, e.g., motion blur and rolling shutter effect. Also, as revealed in this work, object material capturing is a key to credible sensor simulation. Our current object material acquisition method requires the geometry to be precisely aligned between the simulation and the real world, which is more difficult and time-consuming for deformable objects. For future work, we plan to develop a novel material acquisition method that supports unpaired simulated and real images. Moreover, our current method does

not account for the fact that some material properties of real deformable objects can change due to interaction. We will also study and integrate more type of sensors in our simulation platform to facilitate more sim-to-real studies in robotics and vision community.

## REFERENCES

- [1] A. T. Miller and P. K. Allen, “Graspit! A versatile simulator for robotic grasping,” *IEEE Robot. Autom. Mag.*, vol. 11, no. 4, pp. 110–122, Dec. 2004.
- [2] C. Eppner, A. Mousavian, and D. Fox, “A billion ways to grasp—An evaluation of grasp sampling schemes on a dense, physics-based grasp data set,” in *Proc. Int. Symp. Robot. Res.*, 2019, pp. 890–905.
- [3] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, “S4G: Amodal single-view single-shot se (3) grasp detection in cluttered scenes,” in *Proc. Conf. Robot Learn.*, 2020, pp. 53–65.
- [4] H.-S. Fang, C. Wang, M. Gou, and C. Lu, “Grasnet-1billion: A large-scale benchmark for general object grasping,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11444–11453.
- [5] L. Shao et al., “Unigrasp: Learning a unified model to grasp with multifingered robotic hands,” *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2286–2293, Apr. 2020.
- [6] R. Newbury et al., “Deep learning approaches to grasp synthesis: A review,” 2022, *arXiv:2207.02556*.
- [7] K. Fang et al., “Learning task-oriented grasping for tool manipulation from simulated self-supervision,” *Int. J. Robot. Res.*, vol. 39, no. 2–3, pp. 202–216, 2020. [Online]. Available: <https://doi.org/10.1177/0278364919872545>
- [8] T. Chen, J. Xu, and P. Agrawal, “A system for general in-hand object re-orientation,” in *Proc. 5th Annu. Conf. Robot Learn.*, 2021, pp. 297–307. [Online]. Available: <https://openreview.net/forum?id=7uSBJDopT7Y>
- [9] T. Mu et al., “Maniskill: Generalizable manipulation skill benchmark with large-scal demonstrations,” in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/eda80a3d5b344bc40f3bc04f65b7a357-Abstract-round2.html>
- [10] T. Schmidt, K. Hertkorn, R. Newcombe, Z. Marton, M. Suppa, and D. Fox, “Depth-based tracking with physical constraints for robot manipulation,” in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 119–126.
- [11] D. Morrison, P. Corke, and J. Leitner, “Learning robust, real-time, reactive robotic grasping,” *Int. J. Robot. Res.*, vol. 39, no. 2/3, pp. 183–201, 2020. [Online]. Available: <https://doi.org/10.1177/0278364919859066>
- [12] Y. Cong, R. Chen, B. Ma, H. Liu, D. Hou, and C. Yang, “A comprehensive study of 3-D vision-based robot manipulation,” *IEEE Trans. Cybern.*, to be published, doi: [10.1109/TCYB.2021.3108165](https://doi.org/10.1109/TCYB.2021.3108165).
- [13] U. Viereck, A. Pas, K. Saenko, and R. Platt, “Learning a visuomotor controller for real world robotic grasping using simulated depth images,” in *Proc. Conf. Robot Learn.*, 2017, pp. 291–300.
- [14] Y. Litvak, A. Biess, and A. Bar-Hillel, “Learning pose estimation for high-precision robotic assembly using simulated depth images,” in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 3521–3527.
- [15] J. Mahler et al., “DEX-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” in *Proc. Robot.: Sci. Syst.*, 2017. [Online]. Available: <http://www.roboticsproceedings.org/rss13/index.html>
- [16] R. Chen, J. Xu, and S. Zhang, “Comparative study on 3D optical sensors for short range applications,” *Opt. Lasers Eng.*, vol. 149, 2022, Art. no. 106763. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0143816621002335>
- [17] S. Giancola, M. Valenti, and R. Sala, *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Berlin, Germany: Springer, 2018.
- [18] F. Xiang et al., “Sapien: A simulated part-based interactive environment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11097–11107.
- [19] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, “Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3 d model views,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2686–2694.
- [20] B. Sun and K. Saenko, “From virtual to reality: Fast adaptation of virtual object detectors to real domains,” in *Proc. Brit. Mach. Vis. Conf.*, 2014, vol. 1, Art. no. 3.

- [21] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 23–30.
- [22] J. Tremblay et al., "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 969–977.
- [23] O. M. Andrychowicz et al., "Learning dexterous in-hand manipulation," *Int. J. Robot. Res.*, vol. 39, no. 1, pp. 3–20, 2020.
- [24] Q. Vuong, S. Vikram, H. Su, S. Gao, and H. I. Christensen, "How to pick the domain randomization parameters for sim-to-real transfer of reinforcement learning policies?", 2019, *arXiv:1903.11774*.
- [25] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3722–3731.
- [26] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2107–2116.
- [27] K. Bousmalis et al., "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 4243–4250.
- [28] N. Patricia, F. M. Carlucci, and B. Caputo, "Deep depth domain adaptation: A case study," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 2645–2650.
- [29] A. Martínez-González, M. Villamizar, O. Canévet, and J.-M. Odobez, "Investigating depth domain adaptation for efficient human pose estimation," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 346–363.
- [30] M. Keller and A. Kolb, "Real-time simulation of time-of-flight sensors," *Simul. Model. Pract. Theory*, vol. 17, no. 5, pp. 967–978, 2009.
- [31] S. Meister, R. Nair, and D. Kondermann, "Simulation of time-of-flight sensors using global illumination," in *Proc. Vis. Model. Visual.*, 2013, pp. 33–40.
- [32] M. J. Landau, B. Y. Choo, and P. A. Beling, "Simulating kinect infrared and depth images," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3018–3031, Dec. 2016.
- [33] B. Planche et al., "Depthsynth: Real-time realistic synthetic data generation from cad models for 2.5 d recognition," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 1–10.
- [34] C. Sweeney, G. Izatt, and R. Tedrake, "A supervised approach to predicting noise in depth images," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 796–802.
- [35] X. Gu, Y. Guo, F. Deligianni, and G.-Z. Yang, "Coupled real-synthetic domain adaptation for real-world deep depth enhancement," *IEEE Trans. Image Process.*, vol. 29, pp. 6343–6356, 2020.
- [36] Y. Shen, Y. Yang, Y. Zheng, C. K. Liu, and L. J. Guibas, "DCL: Differential contrastive learning for geometry-aware depth synthesis," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4845–4852, Apr. 2022.
- [37] A. Pashevich, R. Strudel, I. Kalevatykh, I. Laptev, and C. Schmid, "Learning to augment synthetic images for sim2real policy transfer," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 2651–2657.
- [38] B. Planche and R. V. Singh, "Physics-based differentiable depth sensor simulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14387–14397.
- [39] T.-M. Li, M. Aittala, F. Durand, and J. Lehtinen, "Differentiable Monte Carlo ray tracing through edge sampling," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–11, 2018.
- [40] Z. Xie, X. Yu, X. Gao, K. Li, and S. Shen, "Recent advances in conventional and deep learning-based depth completion: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: [10.1109/TNNLS.2022.3201534](https://doi.org/10.1109/TNNLS.2022.3201534).
- [41] S. Sajjan et al., "Clear grasp: 3 d shape estimation of transparent objects for manipulation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 3634–3642.
- [42] L. Zhu et al., "RGB-D local implicit function for depth completion of transparent objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4647–4656.
- [43] H. Xu, Y. R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, "Seeing glass: Joint point-cloud and depth completion for transparent objects," in *Proc. 5th Annu. Conf. Robot Learn.*, 2021. [Online]. Available: <https://openreview.net/forum?id=tCfLLiP7vje>
- [44] H. Fang, H.-S. Fang, S. Xu, and C. Lu, "Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 7383–7390, Jul. 2022.
- [45] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *Proc. 27th Annu. Conf. Comput. Graph. Interactive Techn.*, 2000, pp. 145–156.
- [46] A. Gardner, C. Tchou, T. Hawkins, and P. Debevec, "Linear light source reflectometry," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 749–758, 2003.
- [47] A. Ghosh, T. Chen, P. Peers, C. A. Wilson, and P. Debevec, "Estimating specular roughness and anisotropy from second order spherical gradient illumination," *Comput. Graph. Forum*, vol. 28, no. 4, pp. 1161–1170, 2009.
- [48] M. Aittala, T. Aila, and J. Lehtinen, "Reflectance modeling by neural texture synthesis," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–13, 2016.
- [49] Z. Li, K. Sunkavalli, and M. Chandraker, "Materials for masses: SVBRDF acquisition with a single mobile phone image," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 72–87.
- [50] P. Henzler, V. Deschaintre, N. J. Mitra, and T. Ritschel, "Generative modelling of brdf textures from flash images," *ACM Trans. Graph.*, vol. 40, no. 6, 2021, Art. no. 284.
- [51] X. Zhang, P.P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron, "Nerfactor: Neural factorization of shape and reflectance under an unknown illumination," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–18, 2021.
- [52] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NERF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421.
- [53] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 5026–5033.
- [54] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.
- [55] T. Erez, Y. Tassa, and E. Todorov, "Simulation tools for model-based robotics: Comparison of Bullet, Havok, MuJoCo, ODE and Physx," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 4397–4404.
- [56] Y. Urakami, A. Hodgkinson, C. Carlin, R. Leu, L. Rigazio, and P. Abbeel, "Doorgym: A scalable door opening environment and baseline agent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.
- [57] S. Tunyasuvunakool et al., "dm\_control: Software and tasks for continuous control," *Softw. Impacts*, vol. 6, 2020, Art. no. 100022.
- [58] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín, "robosuite: A modular simulation framework and benchmark for robot learning," 2020, *arXiv:2009.12293*.
- [59] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "RLbench: The robot learning benchmark & learning environment," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3019–3026, Apr. 2020.
- [60] M. Deitke et al., "Robothor: An open simulation-to-real embodied AI platform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3164–3174.
- [61] M. Savva et al., "Habitat: A platform for embodied ai research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9339–9347.
- [62] F. Xia et al., "Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 713–720, Apr. 2020.
- [63] Y. Zhang et al., "Physically-based rendering for indoor scene understanding using convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5287–5295.
- [64] T. Hodaň et al., "Photorealistic image synthesis for object instance detection," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 66–70.
- [65] S. Bako et al., "Kernel-predicting convolutional networks for denoising Monte Carlo renderings," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 97.
- [66] C. R. A. Chaitanya et al., "Interactive reconstruction of Monte Carlo image sequences using a recurrent denoising autoencoder," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, 2017.
- [67] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
- [68] B. Burley and W. D. A. Studios, "Physically-based shading at disney," in *Proc. ACM SIGGRAPH*, 2012, vol. 2012, pp. 1–7.
- [69] B. Shen et al., "iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 7520–7527.
- [70] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson ENV: Real-world perception for embodied agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9068–9079.

- [71] B. Calli et al., "Yale-CMU-berkeley dataset for robotic manipulation research," *Int. J. Robot. Res.*, vol. 36, no. 3, pp. 261–268, 2017.
- [72] Z. Liu et al., "OCRTOC: A cloud-based competition and benchmark for robotic grasping and manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 1, pp. 486–493, Jan. 2022.
- [73] C. Rennie, R. Shome, K. E. Bekris, and A. F. De Souza, "A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place," *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 1179–1185, Jul. 2016.
- [74] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robot.: Sci. Syst.*, 2018. [Online]. Available: <http://www.roboticsproceedings.org/rss14/index.html>
- [75] G. Bradski, "The opencv library," *Dr. Dobb's J.: Softw. Tools Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.
- [76] G. Terzakis and M. Lourakis, "A consistently fast and globally optimal solution to the perspective-n-point problem," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 478–494.
- [77] K. Daniilidis, "Hand-eye calibration using dual quaternions," *Int. J. Robot. Res.*, vol. 18, no. 3, pp. 286–298, 1999.
- [78] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
- [79] L. B. Wolff, S. K. Nayar, and M. Oren, "Improved diffuse reflection models for computer vision," *Int. J. Comput. Vis.*, vol. 30, no. 1, pp. 55–71, 1998.
- [80] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance, "Microfacet models for refraction through rough surfaces," in *Proc. 18th Eurograph. Conf. Rendering Techn.*, 2007, vol. 2007, pp. 195–206.
- [81] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 690–706, Aug. 1999.
- [82] S. G. Parker et al., "Optix: A general purpose ray tracing engine," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 1–13, 2010.
- [83] J. W. Goodman, *Speckle Phenomena in Optics: Theory and Applications*. Greenwood Village, CO, USA: Roberts and Company Publishers, 2007.
- [84] D. V. Perekeltsa, *Johnson Noise and Shot Noise*. Cambridge, MA, USA: MIT Press, 2006.
- [85] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Eur. Conf. Comput. Vis.*, 1994, pp. 151–158.
- [86] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel(r) realsense(tm) stereoscopic depth cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1267–1276.
- [87] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.
- [88] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. Moure, and A. López, "Embedded real-time stereo estimation via semi-global matching on the GPU," *Procedia Comput. Sci.*, vol. 80, pp. 143–153, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050916306561>
- [89] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [90] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [91] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11632–11641.
- [92] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from RGB-D data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 918–927.
- [93] J. M. Wong et al., "SEGICP: Integrated deep semantic segmentation and pose estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 5784–5789.
- [94] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.



**Xiaoshuai Zhang** received the B.S. degree in computer science from Peking University, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in computer science with the University of California, San Diego, San Diego, CA, USA.

His current research interests include 3-D scene reconstruction and understanding, computer graphics and simulation.



**Rui Chen** (Member, IEEE) received the Ph.D. degree in mechatronical engineering and the B.E. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2020 and 2014, respectively.

He is currently a Research Assistant Professor with the Department of Mechanical Engineering, Tsinghua University. His research interests include three-dimensional computer vision and robot learning.



**Ang Li** is currently working toward the B.E. degree in computer science with the Department of Computer Science and Engineering, University of California, San Diego, San Diego, CA, USA.

His research interests lie in the intersection of computer vision, computer graphics, and robotics.



**Fanbo Xiang** received the B.S. degree in computer science and the B.S. degree in mathematics from the University of Illinois, Urbana-Champaign, Champaign, IL, USA, in 2018, and the M.S. degree in computer science from the University of California San Diego, San Diego, CA, USA, in 2020, where he is currently working toward the Ph.D. degree in computer science.

His research focuses on physical simulation and embodied AI platforms.



**Yuzhe Qin** (Student Member, IEEE) received the B.E. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2018. He is currently working toward the Ph.D. degree in computer science with the Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, USA.

His research interests includes robot learning for manipulation, reinforcement learning with visual input, and robot teleoperation.



**Jiayuan Gu** received the B.S. degree in computer science from Peking University, Beijing, China, in 2018. He is currently working toward the Ph.D. degree in computer science with the Department of Computer Science and Engineering, University of California, San Diego, San Diego, CA, USA.

His research interests focus on interfacing perception and policy learning algorithms for Embodied AI.



**Zhaio Huang** received the B.E. degree in computer science from Tsinghua University, Beijing, China, in 2018. He is currently fifth-year Ph.D. degree in computer science with University of California San Diego, San Diego, CA, USA, advised by Prof. Hao Su.

His research interests include soft-body manipulation with differentiable physics and reinforcement learning.



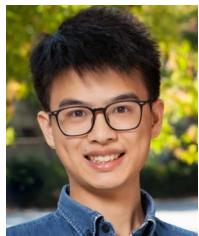
**Zhan Ling** received the B.E. degree in computer science from the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China, in 2019. He is currently working toward the Ph.D. degree in computer science with the University of California, San Diego, San Diego, CA, USA.

His research interests include deep reinforcement learning and robot manipulation.



**Tongzhou Mu** is currently working toward the Ph.D. degree in computer science with the Department of Computer Science and Engineering, University of California San Diego, San Diego, CA, USA, where he is advised by Prof. Hao Su.

His research interests include reinforcement learning/imitation learning, concept discovery and reasoning, and robotics/embodied AI.



**Minghua Liu** received the bachelor's degree in computer science from Tsinghua University, Beijing, China. He is currently working toward the Ph.D. degree in computer science with the Department of Computer Science, University of California San Diego, San Diego, CA, USA.

His research interests include 3-D computer vision and embodied AI.



**Jing Xu** (Member, IEEE) received the Ph.D. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2008.

He was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA. He is currently an Associate Professor with the Department of Mechanical Engineering, Tsinghua University, Beijing. His research interests include vision-guided manufacturing, image processing, and intelligent robotics.



**Peiyu Zeng** received the Bachelor of Engineering degree in measurement, control technology and instruments from Tsinghua University, Beijing, China, in 2021. He is currently working toward the Master of Science degree in robotics, systems and control with ETH Zurich, Zurich, Switzerland.

His current research interests include model-based planning and control.



**Hao Su** (Member, IEEE) received the B.S. degree in computer science and the Ph.D. degree in mathematics from Beihang University, Beijing, China, in 2006 and 2014, respectively and the Ph.D. degree in computer science from Stanford University, California, USA, in 2018.

He has been with the University of California San Diego, San Diego, CA, USA, as an Assistant Professor of Computer Science and Engineering since July 2017. He is affiliated with the Contextual Robotics Institute and Center for Visual Computing.

Mr. Su served on the program committee of multiple conferences and workshops on computer vision, computer graphics, and machine learning. He is the Area Chair of International Conference on Computer Vision (ICCV), 2019, Conference on Computer Vision and Pattern Recognition (CVPR), 2019, Senior Program Chair of AAAI'19, IPC of Pacific Graphics'18, Program Chair of International Conference on 3D Vision (3DV), 2017, Publication Chair of 3DV'16, and the Chair of various workshops at CVPR, European Conference on Computer Vision, and ICCV. He is also invited as keynote speakers at workshops and tutorials in Neural Information Processing Systems, 3DV, CVPR, Robotics: Science and Systems, IEEE International Conference on Robotics and Automation, S3PM, etc.



**Songfang Han** received the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2019.

She was a Postdoctoral Researcher with the University of California San Diego, San Diego, CA, USA, and is currently a Machine Learning Engineer with Snap Inc., Santa Monica, CA, USA. Her research interests include 3-D reconstruction and geometry processing.