

The RobotriX: An eXtremely Photorealistic and Very-Large-Scale Indoor Dataset of Sequences with Robot Trajectories and Interactions

Alberto Garcia-Garcia, Pablo Martinez-Gonzalez, Sergiu Oprea,
John Alejandro Castro-Vargas, Sergio Orts-Escalano, Jose Garcia-Rodriguez and Alvaro Jover-Alvarez

Abstract—Enter the RobotriX, an extremely photorealistic indoor dataset designed to enable the application of deep learning techniques to a wide variety of robotic vision problems. The RobotriX consists of hyperrealistic indoor scenes which are explored by robot agents which also interact with objects in a visually realistic manner in that simulated world. Photorealistic scenes and robots are rendered by Unreal Engine into a virtual reality headset which captures gaze so that a human operator can move the robot and use controllers for the robotic hands; scene information is dumped on a per-frame basis so that it can be reproduced offline to generate raw data and ground truth labels. By taking this approach, we were able to generate a dataset of 38 semantic classes totaling 8M stills recorded at +60 frames per second with full HD resolution. For each frame, RGB-D and 3D information is provided with full annotations in both spaces. Thanks to the high quality and quantity of both raw information and annotations, the RobotriX will serve as a new milestone for investigating 2D and 3D robotic vision tasks with large-scale data-driven techniques.

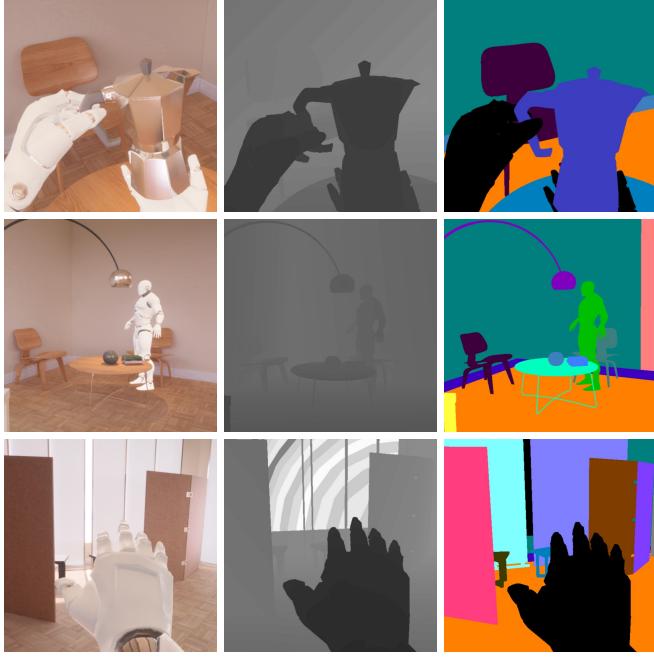


Fig. 1. The RobotriX features extremely photorealistic indoor environments in which robot movements and interactions with objects are captured from multiple points of view at high frame rates and resolutions.

Albert, Pablo, Sergiu, John, Sergio, Jose and Alvaro are with 3D Perception Lab, University of Alicante, Spain agarcia@dtic.ua.es, pmartinez@dtic.ua.es, soprea@dtic.ua.es, jacastro@dtic.ua.es, sorts@ua.es, jgarcia@dtic.ua.es, ajover@dtic.ua.es

I. INTRODUCTION

Recent years have witnessed an increasing dominance of deep learning techniques targeted at a wide range of robotic vision problems such as scene understanding, depth estimation, optical flow estimation, tracking, and visual grasping among others. Those new approaches have been slowly but surely closing the gap with traditional ones in terms of accuracy and efficiency, surpassing them in more and more cases and situations as those new techniques mature and better data is available. A key requirement for those new approaches to achieve outstanding accuracy while being able to generalize properly to unseen situations is a good dataset. Researchers need large-scale sets of images, which are representative enough for the problem at hand but at the same time include a considerable amount of variability, and complete ground truth for each one of them depending on the needs of the problem. Currently, there is a lack of such data since generating a dataset which satisfies those requirements is hard in practice due to the inherent difficulties of data collection. In this work, we focus on that challenge and aim to provide a unified benchmark for training, evaluating, and reproducing algorithms and methods on a wide spectrum of robotic vision problems.

The first matter that we must address is scale. The importance of large-scale data when working with data-hungry learning algorithms is critical in order to achieve proper results and generalization. This is due to the fact that collecting and labelling real-world data is a tedious and costly process. However, for a synthetic dataset to be useful, it must resemble reality as much as possible. In this regard, the use of photorealistic rendering solutions is a must. Other issues that we have to take into account are video-related intrinsics such as image resolution, frame rate, as well as the nature of each frame. Various works have highlighted the importance of high-resolution and high frame-rate data for various computer vision applications [1][2]. In that sense, our target is providing a high resolution dataset with fairly high frame rate (+60 FPS) and three data modalities for each frame: RGB-D, stereo, and 3D (colored point clouds) simulating the particularities of widespread sensors.

Although providing a large-scale and photorealistic dataset with decent ground truth is always highly useful for the community, there is a number of already existing works that do that but with certain shortcomings (low resolution, low frame rate, somewhat artificial scenes, or scarce modalities). Apart from iterating over those features to strengthen them,

TABLE I
SUMMARY OF FEATURES OF REALISTIC/SYNTHETIC INDOOR DATASETS AND ENVIRONMENTS.

Dataset	Scale		Photorealism		Video		Modalities			Resolution	Interaction	Trajectories	Design
	Frames	Layouts	Realism	Renderer	Seqs.	FPS	RGB	Depth	3D				
NYU-D V2 [3]	1.5K	464	Real	-	-	-	•	•	-	640 × 480	-	-	Real
Sun RGB-D [4]	10K	N/A	Real	-	-	-	•	•	-	Mix	-	-	Real
Stanford 2D-3D-S [5]	70K	270	Real	-	-	-	•	•	•	1080 × 1080	-	-	Real
Matterport 3D [6]	200K	90	Real	-	-	-	•	•	•	1280 × 1024	-	-	Real
SunCG [7]	130K	45K	••	N/A	-	-	-	•	•	640 × 480	-	-	Manual
PBR-Princeton [8]	500K	45K	••	Mitsuba	-	-	•	•	•	640 × 480	-	-	Manual
SceneNet RGB-D [9]	5M	57	•••	NVIDIA Optix	16895	1	•	•	•	320 × 240	-	-	Random
Ours	8M	16	•••••	Unreal Engine	512	+60	•	•	•	1920 × 1080	Hands	Synthetic	Manual
HoME [10]	-	45K	••	Panda3D	-	N/A	•	•	-	N/A	Physics	-	Manual
House3D [11]	-	45K	••	OpenGL	-	600	•	•	-	120 × 90	Physics	-	Manual
AI2-THOR [12]	-	120	••••	Unity	-	13	-	-	-	300 × 300	Actionable	-	Manual
MINOS (Sun CG) [13]	-	45K	••	WebGL	-	100	•	•	-	N/A	-	-	Manual
MINOS (Matterport) [13]	-	90	Real	-	-	100	•	•	-	N/A	-	-	Real

we wanted to include a set of novel ones which make our proposal truly stand out from the crowd: (1) hands, first person, and room points of view to make the dataset useful for various robotic applications, (2) realistic robot trajectories and head movements (if applicable) controlled by a human operator in Virtual Reality (VR), and (3) visually realistic interactions with objects by using VR controllers to move robotic hands with inverse kinematics. Furthermore, we release all the details and tools to allow researchers to generate custom data to fulfill their needs.

The rest of this paper is organized as follows. Firstly, Section II reviews the state of the art of each dataset-related features and ground truth information that we mentioned before. Next, Section III provides an in-depth overview of the whole dataset and its generation. At last, Section IV draws conclusions about this work and proposes some future lines of research to improve it.

II. RELATED WORKS

Synthetic image datasets have been used for a long time to benchmark vision algorithms [14]. Recently, their importance has been highlighted for training and evaluating machine learning models for robotic vision problems [10], [15], [16]. Due to the increasing demand for annotated data, fostered by the rise of deep learning, real-world datasets are having a hard time to keep up since generating ground truth for them can be tedious and error-prone. Many indoor synthetic datasets have been proposed in the literature and some of them have been successfully used to train and validate deep architectures. In certain cases, it has even been proven that artificial data can be highly beneficial and increase the accuracy of state-of-the-art models on challenging real-world benchmarks and problems [17], [18], [15]. However, synthetic datasets have their own problems and existing ones can be improved in many ways. In this section, we review the most important aspects that make an indoor dataset suitable for training deep learning architectures: scale, photorealism, video, modalities, resolution, interactions, trajectories, and its design method. In addition, we also review the ground truth provided by them to determine their quality and which problems can be addressed by using that data. At last, we put our proposal in context by briefly enumerate its contributions and how our dataset improves and extends existing works.

A. Features

The criteria that we used in this brief review of features (see Table I for a summarized view of those features for the most popular and promising indoor datasets and environments that are already public) are the following ones:

- **Scale:** Data-driven algorithms such as deep learning approaches rely on massive amount of data to achieve unprecedented accuracy levels. Furthermore, massive amounts of information are not only needed to make those systems able to learn properly but also to give them the ability to generalize their knowledge to unseen situations. We measure scale according to the number of frames and possible layouts or room configurations (note that environments do not provide frames per-se so they are potentially infinite and that quantification makes no sense).
- **Photorealism:** Recently, synthetic approaches have gained so much popularity since generating ground truth for them is an easy and automated task. Many successful stories have proven that artificial data can be used to train deep architectures for a wide variety of problems [15], [19], [20], [21], [22], [23]. Furthermore, some of them have highlighted the fact that training machine learning algorithms on virtual worlds even improves accuracy when they are applied to real-world scenarios [24] [25]. In any case, the dataset will be more useful as it is closer to reality. We quantify realism on a scale of one to five according to the combination of texturing quality, rendering photorealism, object geometry, and layout coherence.
- **Sequences:** Some problems can only be approached or at least they get much easier if video sequences are provided with a certain amount of Frames per Second (FPS) is reached. For instance, object tracking mechanisms based on recurrent networks [2] or temporally coherent segmentation models [26] benefit from higher frame rates which provide smoother changes without huge differences between frames. In this regard, we report whether the dataset provides video data or not, the number of sequences and the average framerate; for the environments the framerate indicates how many actions/renderings can be performed per frame.

TABLE II
OVERVIEW OF GROUND TRUTH INFORMATION PROVIDED BY THE REVIEWED DATASETS AND ENVIRONMENTS.

Dataset	2D BBox	2D Segm. Class	2D Segm. Inst.	3D BBox	3D Segm. Class	3D Segm. Inst.	3D Object Pose	Camera Pose	Hand Pose	Depth
NYU-D-V2 [3]		•	•					~ ¹		•
Sun RGB-D [4]		•	•							•
Stanford 2D-3D-S [5]		•	•		•	•				•
Matterport 3D [6]		•	•		•	•		•		•
Sun CG [7]					•					•
PBR-Princeton [8]		•	•							•
SceneNet RGB-D [9]		•	•				•			•
Ours	•	•	•	•	•	•	•	• (Multi)	•	•
HoMe [10]		•								•
House3D [11]		•								•
AI2-THOR [12]										•
MINOS (Sun CG) [13]	•	•	•							•
MINOS (Matterport) [13]	•	•	•							•

- **Modalities:** There is certain importance in providing as many data modalities as possible. On the one hand, some problems can only be addressed if data from a particular nature is available, e.g., depth information is needed to make a system learn to estimate depth in a supervised manner. On the other hand, even if a problem can be solved using a concrete data modality, having more sources of information available might foster the development of alternative ways of fusing that extra data to boost performance. For all the reviewed datasets, we report the kind of data modalities that they provide.
- **Resolution:** Vision systems usually benefit from larger image sizes since higher resolution means better image quality and thus more robust features to be learned. One notorious success case of high-resolution imaging in deep learning is the case of Baidu Vision's system [27] which introduced a novel multi-scale and high-resolution model to achieve the best score on the well-known ImageNet challenge [28] by the time they published the work. However, this is not always true and for some applications it is important to find balance in the tradeoff between accuracy and performance when processing large images. We indicate the image resolution for each dataset and environment.
- **Interaction:** Despite the importance of hand pose estimation and object interaction in many applications, e.g., grasping and dexterous manipulation, this aspect is often neglected. The main reason about this scarcity is the difficulty of generating annotations for the hand joints and moving objects. In some cases, interaction is reduced to simple actions with binary states. We report the kind of interaction that is performed on each dataset (physics simulations for collisions, actionable items, or full object interaction with robotic hands).
- **Trajectories:** For sequences, the way trajectories are generated plays an important role in the quality or applicability of the dataset. In order to make the dataset distribution as close as possible to the application scenario one, the camera trajectory should be similar too. Real-world datasets usually leverage handheld or head-mounted devices to generate human-like trajectories and, in the case of a robot, capture devices are usually mounted in the same place where they will work when

deployed. Synthetic datasets must devise strategies to place and orient cameras in a coherent manner. For each dataset that provides video sequences, we report the way those trajectories are generated.

- **Design:** The design of scene layouts is another factor that must be taken into account to make the dataset as similar as possible to the real-world. This means that scenes must be coherent, i.e., objects and lights must be positioned according to actual room layouts. Generating coherent rooms with plausible configurations synthetically to achieve large-scale data is a hard task and only *SceneNet*[29] and *SceneNet RGB-D*[9] approached the room design problem algorithmically; their results are plausible but oftentimes not really representative of what a real-world room would look like due to artificial object positioning and orientations. For this feature, we report whether the design is real, manual or algorithmic/synthetic.

B. Ground Truth

Although all the aforementioned features are of utmost importance for a dataset, ground truth is the cornerstone that will dictate the usefulness of the data. It determines the problems that can be solved by using the available data. Table II shows the ground truth information provided by each one of the reviewed datasets including ours, which completes and offers more annotations than the state of the art.

C. Our Proposal in Context

After analyzing the strong points and weaknesses of the most popular indoor datasets, we aimed to combine the strengths of all of them while addressing their weaknesses and introducing new features. The major contributions of our novel dataset with regard to the current state of the art are:

- **Large-scale and high level of photorealism.**
- **High frame-rate and resolution sequences.**
- **Multiple data modalities (RGB-D/3D/Stereo).**
- **Realistic robot trajectories with multiple PoVs.**
- **Robot interaction within the scene.**
- **Ground truth for many vision problems.**
- **Open-source pipeline and tools².**

¹It provides Roll, Yaw, Pitch and Tilt angle of the device from an accelerometer.

²<https://github.com/3dperceptionlab/therobotrix>

III. OVERVIEW

After reviewing the existing datasets and stating the contributions of our proposal, we will provide a detailed description of its main features and how did we achieve them: hyperphotorealism by combining a powerful rendering engine with extremely detailed scenes and realistic robot movements and trajectories. We will also provide an overview of the data collection pipeline: the online sequence recording procedure and the offline data and ground truth generation process. Furthermore, we will list the contents of the dataset and the tools that we provide as open-source software for the robotics research community.

A. Photorealistic Rendering

The rendering engine we chose to generate photorealistic RGB images is Unreal Engine 4 (UE4)³. The reasons for this choice are the following ones: (1) it is arguably one of the best game engines able to produce extremely realistic renderings, (2) beyond gaming, it has become widely adopted by VR developers and architectural visualization experts so a whole lot of tools, examples, documentation, and assets are available; (3) due to its impact, many hardware solutions offer plugins for UE4 that make them work out-of-the-box; and (4) Epic Games provides the full C++ source code and updates to it so the full suite can be easily used and modified.

Arguably, the most attractive feature of UE4 that made us take that decision is its capability to render photorealistic scenes like the one shown in Figure 2 in real time. Some UE4 features that enable this realism are: physically-based materials, pre-calculated bounce light via Lightmass, stationary lights using IES profiles, post-processing, and reflections.



Fig. 2. Snapshot of the daylight room setup for the *Realistic Rendering* released by Epic Games to show off the rendering capabilities of UE4.

It is also important to remark that we do have strict real-time constraints for rendering during sequence recording since we need to immerse a human agent in virtual reality to record the sequences. UE4 is engineered for virtual reality with a specific rendering solution for it named *Forward Renderer*. That renderer is able to generate images that meet our quality standards at 90 FPS thanks to high-quality lighting features, Multisample Anti-Aliasing (MSAA), and instanced stereo rendering.

³<http://www.unrealengine.com>

B. Scenes

To complement the rendering engine and achieve the desired level of photorealism we needed coherent indoor scenes with extreme attention to detail. UE4Arch⁴ is a company devoted to creating hyper-realistic and real-time architecture visualizations with UE4. We take advantage of various house projects and assets created by that company to populate our dataset with rich environments and layouts. Figure 3 shows a sample project from UE4Arch.



Fig. 3. Viennese Apartment archviz project snapshot by UE4Arch.

C. Robot Integration

Seamlessly integrating robots in our scenes and making them controllable in VR by a human agent to record sequences requires three issues to be solved: (1) gaze and head movement with first person Point of View (PoV), (2) inverse kinematics to be able to move them with motion controllers and reach for objects, and (3) locomotion to displace the robot within the scene.

The first issue is solved by using the Oculus Rift headset to control the robot's head movement and render its first person PoV. Inverse kinematics for the virtual robot are manually implemented with Forward And Backward Reaching Inverse Kinematics (FABRIK), a built-in inverse kinematics solver in UE4 that works on a chain of bones of arbitrary length. Locomotion is handled by thumbsticks on the Oculus Touch motion controllers. By doing this we are able to integrate any robot in FBX format such as Unreal's mannequin or the well-known Pepper by Aldebaran (see Figure 4).



Fig. 4. Pepper and Mannequin integrated with colliders and constraints.

⁴<https://ue4arch.com/>

D. Object Interaction

One key aspect of the dataset is the simulation of realistic interactions with objects. On the one hand, we need to simulate basic physics to move, push, pull, or lift objects in the scene. On the other hand, for small props we need to provide a way to grasp them if they fit in the robot's hand for a more complex interaction.

To solve this issue we leverage UE4's built-in physics engine. Complex surfaces such as the robot's hands or its body and certain objects are modeled with single convex hulls as collider primitives. Simpler geometries are just approximated with less sophisticated primitives such as spheres or boxes. Those objects which are susceptible of being manipulated are properly weighted to resemble their real-world physical behavior. Furthermore, we implemented a simple yet visually appealing grasping approach to ease object interaction and make it look as realistic as possible: firstly, each hand is animated to have a pose blending from open to closed (this blending or interpolation is controlled by the analog triggers from each Oculus Touch); second, for each object that is small enough to be grabbed, we check which hand bones collide with it; if the five fingers and the palm collide with the object, we attach that object to the hand and stop simulating its physics so that it is stable in the hand; once those collisions are no longer happening, the object is detached and its physics are enabled again. Figure 5 shows the hand colliders and examples of interaction and grasping.

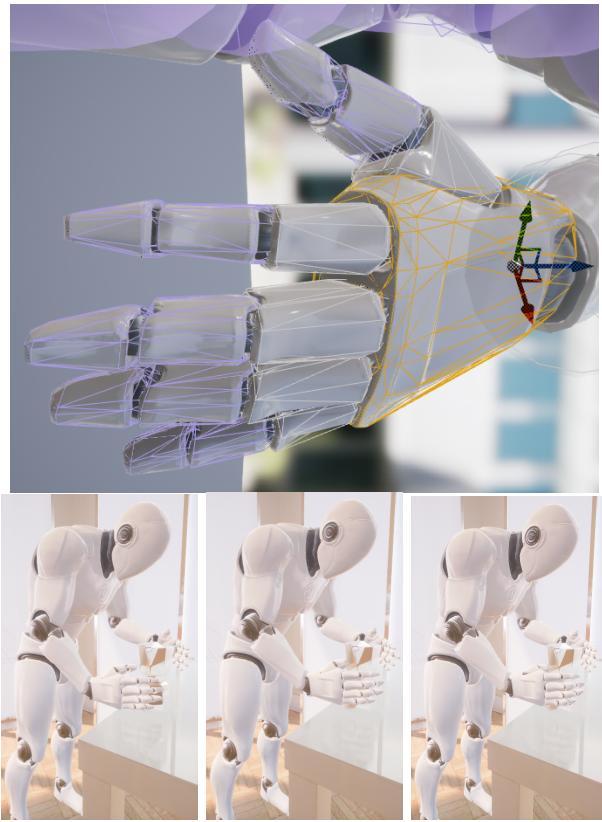


Fig. 5. Single convex hull colliders for UE4's mannequin robotic hand and a example of object interaction and grasping.

E. Sequence Recording and Data Collection

To record and generate all the data for this dataset, we made extensive use of a tool that was specifically built for this dataset: UnrealROX [30], a virtual reality environment for generating synthetic data for various robotic vision tasks. In such environment, a human operator can be embodied, in virtual reality, as a robot agent inside a scene to freely navigate and interact with objects as if it was a real-world robot. Our environment is built on top of UE4 to take advantage of its advanced VR, rendering, and physics capabilities. That system provides the following features: (1) a visually plausible grasping system for robot manipulation which is modular enough to be applied to various finger configurations, (2) routines for controlling robotic hands and bodies with commercial VR setups such as Oculus Rift and HTC Vive Pro, (3) a sequence recorder component to store all the information about the scene, robot, and cameras while the human operator is embodied as a robot, (4) a sequence playback component to reproduce the previously recorded sequence offline to generate raw data such as RGB, depth, normals, or instance segmentation images, (5) a multi-camera component to ease the camera placement process and enable the user to attach them to specific robot joints and configure their parameters (resolution, noise model, field of view), and (6) open-source code, assets, and tutorials for all those components and other subsystems that tie them together.

In order to generate the dataset, we first collect data in an online and interactive manner by immersing human agents in the virtual indoor environment so that they can freely move, look, and interact with the scene (respecting robot constraints). In this stage, we use UnrealROX with UE4 to render our scene into an Oculus/HTC Vive Pro VR headset worn by a person equipped with motion controllers for hand movement. During this phase, we gather all the information that we would need to replay the whole sequence offline to collect data and generate annotations without lagging the rendering process.

This data collection process is performed by an actor in UE4 which ticks on every rendered frame and asynchronously dumps to a text file the $SE(3)$ pose (location and rotation) of every object and camera in the scene. It also dumps the full pose for each bone of the robot. This text file just contains a timestamp for each frame and the aforementioned raw information to have a minimal impact on performance. In fact, this process allows us to render at 80+ FPS thanks to asynchronous and threaded writes to files. After the whole sequence is recorded, the text file is converted to JSON for better interpretability. Figure ?? shows a diagram of this sequence recording procedure.

After that we use the playback component of UnrealROX to reproduce sequences frame by frame by setting the corresponding poses for the robot and all objects and cameras. Once the frame is exactly as it was when we recorded the sequence, we just generate all the needed data offline by requesting the appropriate information such as RGB images, depth maps, or segmentation masks through the interface.

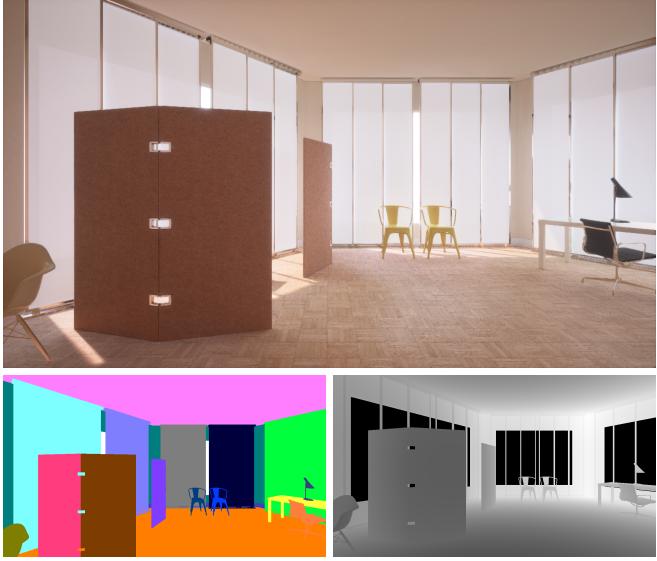


Fig. 6. RGB, instance segmentation masks, and depth map generated with UnrealROX for a synthetic scene rendered in Unreal Engine 4 (UE4).

Figure 6 shows some examples of raw data generated with UnrealROX.

F. Ground Truth Generation

After collecting RGB, depth, and instance segmentation masks for each frame of a sequence, we can use that data to generate annotations. In our dataset release, we only include that raw data from the sequence and certain configuration files generated by the client in order to be able to produce ground truth annotations offline and on demand. We decoupled the data collection and the ground truth generation processes for a simple reason: practicality. In this way, to use the dataset researchers only need to download the RGB, depth, instance segmentation masks, and the generator code to locally generate whichever annotation their problems require in the appropriate format instead of fetching the full bundle in a predefined one. The generator takes that raw data and additional information generated by the client (camera configuration, object classes, colors, and instance mapping) and outputs 2D/3D bounding boxes in VOC format, point clouds, 2D/3D class segmentation masks, and 3D instance segmentation masks (hand pose and camera pose information is embedded in the sequence recording). Figure 7 shows some examples of ground truth generation.

G. Content

Using this pipeline, we generated a dataset of 512 sequences recorded on 16 room layouts (some samples are shown in Figure 8) at +60 FPS with a duration that spans between one and five minutes. That means a total of approximately 8 million individual frames. For each one of those frames we provide the following data:

- 3D poses for the cameras, objects, and joints.
- RGB image @ 1920 × 1080 in JPEG.
- Depth map @ 1920 × 1080 in 16-bit PNG.

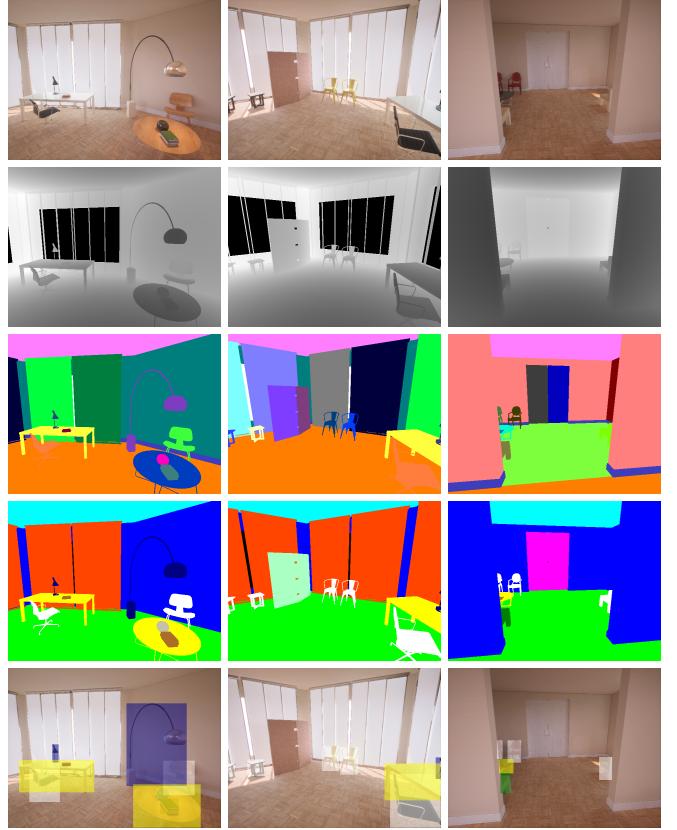


Fig. 7. Ground truth generation examples (from top to bottom: RGB, depth, instance masks, and bounding boxes).

- 2D instance mask @ 1920 × 1080 in 24-bit PNG.

And also annotations:

- 2D class mask @ 1920 × 1080 in 24-bits PNG.
- 2D/3D object instance oriented bounding boxes.
- 3D point cloud with RGB color.
- 3D instance/class mask.

This initial release of the dataset contains 32 detection classes and 39 semantic ones. These categories were selected from the most common and useful household goods in indoor environments for social robots. Note that semantic classes include structural elements such as walls that are not usually targets for detection that commonly focuses on relatively small and interactive objects. Table III shows both detection and semantic splits with their associated codes and colors.

All the tools, assets, and the dataset itself will be made available at <https://github.com/3dperceptionlab/therobotrix>.

IV. CONCLUSION

In this work, we presented The RobotriX, an extremely realistic suite of data and tools designed to boost progress in indoor robotic vision tasks with deep learning by: (1) creating the largest and most realistic synthetic dataset to date; (2) seamlessly integrating realistic robots and scenes within virtual reality to easily generate plausible and useful sequences with interactions; (3) providing video sequences in

TABLE III
CLASSES FOR SEMANTIC SEGMENTATION AND OBJECT DETECTION.

Type	0	1	2	3	4	5	6	7	8	9	10	11	12
Semantic Detection	void	wall	floor	ceiling	window	door	table	chair	lamp	sofa	cupboard	screen	robot
Type	13	14	15	16	17	18	19	20	21	22	23	24	25
Semantic Detection	frame	bed	fridge	whiteboard	book	bottle	plant	furniture	toilet	phone	bathtub	cup	mat
Type	26	27	28	29	30	31	32	33	34	35	36	37	38
Semantic Detection	mirror	sink	box	mouse	keyboard	bin	cushion	shelf	bag	curtain	kitchen_stuff	bath_stuff	prop
Type	mirror	sink	box	mouse	keyboard	bin	cushion	shelf	bag	-	kitchen_stuff	bath_stuff	prop

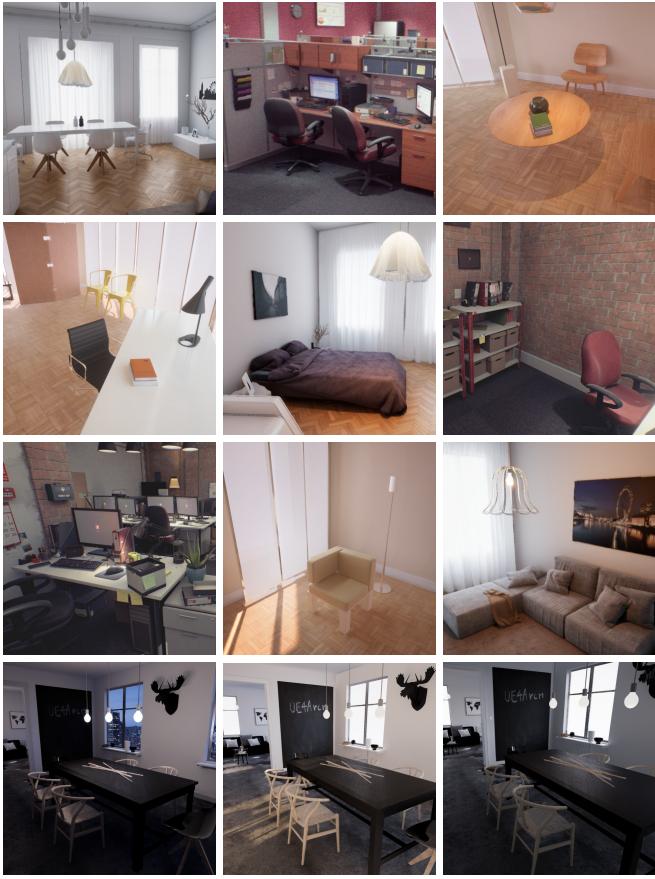


Fig. 8. Snapshots of photorealistic scenes in the dataset.

multiple data modalities with perfect ground truth for solving and pushing forward the state of the art of a wide variety of problems, e.g., object detection, semantic segmentation, depth estimation, object tracking, object pose estimation, visual grasping, and many more. By releasing this dataset, our methodology, and the whole pipeline for generating data, we hope to satisfy the ever-growing need for data of deep learning approaches with easily generated and extremely realistic synthetic sequences which facilitate the deployment of those systems in real-world scenarios.

As future works we plan on adding more complexity to the data and extend the range of problems that can benefit from it. For instance, we want to add non-rigid objects which can

be simulated with Unreal Engine 4 physics such as elastic bodies, fluids, or clothes for the robots to interact with. We also want to automatically generate semantic descriptions for each frame to provide ground truth for captioning and question answering. In addition, we also want to add simulated force sensors on robotic hands to provide annotations for more sophisticated grasping tasks.

At last, we would like to remark that The RobotriX is intended to adapt to individual needs (so that anyone can generate custom data and ground truth for their problems) and change over time by adding new sequences thanks to its modular design and its open-source approach.

ACKNOWLEDGMENT

This work has been funded by the Spanish Government TIN2016-76515-R grant for the COMBAHO project, supported with Feder funds. This work has also been supported by a Spanish national grant for PhD studies FPU15/04516 and by the University of Alicante project GRE16-19 and by the Valencian Government project GV/2018/022. Experiments were made possible by a generous hardware donation from NVIDIA.

REFERENCES

- [1] A. Handa, R. A. Newcombe, A. Angeli, and A. J. Davison, “Real-time camera tracking: When is high frame-rate best?” in *European Conference on Computer Vision*. Springer, 2012, pp. 222–235.
- [2] D. Held, S. Thrun, and S. Savarese, “Learning to track at 100 fps with deep regression networks,” in *European Conference on Computer Vision*. Springer, 2016, pp. 749–765.
- [3] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” *Computer Vision–ECCV 2012*, pp. 746–760, 2012.
- [4] S. Song, S. P. Lichtenberg, and J. Xiao, “Sun rgb-d: A rgbd scene understanding benchmark suite,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.
- [5] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese, “Joint 2D-3D-Semantic Data for Indoor Scene Understanding,” *ArXiv e-prints*, Feb. 2017.
- [6] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3D: Learning from RGB-D data in indoor environments,” *International Conference on 3D Vision (3DV)*, 2017.
- [7] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, “Semantic scene completion from a single depth image,” *arXiv preprint arXiv:1611.08974*, 2016.
- [8] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser, “Physically-based rendering for indoor scene understanding using convolutional neural networks,” *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [9] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, “Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2678–2687.
- [10] S. Brodeur, E. Perez, A. Anand, F. Golemo, L. Celotti, F. Strub, J. Rouat, H. Larochelle, and A. Courville, “Home: a household multimodal environment,” *arXiv preprint arXiv:1711.11017*, 2017.
- [11] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian, “Building Generalizable Agents with a Realistic and Rich 3D Environment,” *ArXiv e-prints*, Jan. 2018.
- [12] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, “Ai2-thor: An interactive 3d environment for visual ai,” *arXiv preprint arXiv:1712.05474*, 2017.
- [13] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, “Minos: Multimodal indoor simulator for navigation in complex environments,” *arXiv preprint arXiv:1712.03931*, 2017.
- [14] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *European Conf. on Computer Vision (ECCV)*, ser. Part IV, LNCS 7577, A. Fitzgibbon et al. (Eds.), Ed. Springer-Verlag, Oct. 2012, pp. 611–625.
- [15] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.
- [16] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, “Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” 2017.
- [17] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, “Learning from simulated and unsupervised images through adversarial training,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2242–2251.
- [18] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, “Looking beyond appearances: Synthetic training data for deep cnns in re-identification,” *Computer Vision and Image Understanding*, vol. 167, pp. 50 – 62, 2018.
- [19] J. Lin, X. Guo, J. Shao, C. Jiang, Y. Zhu, and S.-C. Zhu, “A virtual reality platform for dynamic human-scene interaction,” in *SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*. ACM, 2016, p. 11.
- [20] A. Mahendran, H. Bilen, J. F. Henriques, and A. Vedaldi, “Research-doom and cocodoom: Learning computer vision with games,” *arXiv preprint arXiv:1610.02431*, 2016.
- [21] C. Jiang, Y. Zhu, S. Qi, S. Huang, J. Lin, X. Guo, L.-F. Yu, D. Terzopoulos, and S.-C. Zhu, “Configurable, photorealistic image rendering and ground truth synthesis by sampling stochastic grammars representing indoor scenes,” *arXiv preprint arXiv:1704.00112*, 2017.
- [22] M. Mueller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem, “Ue4sim: A photo-realistic simulator for computer vision applications,” *arXiv preprint arXiv:1708.05869*, 2017.
- [23] Y. Zhang, W. Qiu, Q. Chen, X. Hu, and A. L. Yuille, “Unrealstereo: A synthetic dataset for analyzing stereo vision,” *CoRR*, 2016.
- [24] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, and R. Vasudevan, “Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?” *arXiv preprint arXiv:1610.01983*, 2016.
- [25] 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,” *arXiv preprint arXiv:1703.06907*, 2017.
- [26] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, “Clockwork convnets for video semantic segmentation,” in *Computer Vision–ECCV 2016 Workshops*. Springer, 2016, pp. 852–868.
- [27] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, “Deep image: Scaling up image recognition,” *arXiv preprint arXiv:1501.02876*, vol. 7, no. 8, 2015.
- [28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [29] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, “Scenenet: Understanding real world indoor scenes with synthetic data,” *arXiv preprint arXiv:1511.07041*, 2015.
- [30] P. Martinez-Gonzalez, S. Oprea, A. Garcia-Garcia, A. Jover-Alvarez, S. Orts-Escolano, and J. Garcia-Rodriguez, “Unrealrox: An extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation,” *arXiv preprint arXiv:1810.06936*, 2018.