Supplementary Material for
VRGym: A Virtual Testbed for Physical and Interactive AI

Xu Xie Hangxin Liu Zhenliang Zhang Yuxing Qiu Feng Gao Siyuan Qi Yixin Zhu Song-Chun Zhu
UCLA Center for Vision, Cognition, Learning, and Autonomy
Los Angeles, CA 90095
{xuxie,hx.liu,zlz,yxqiu,f.gao,syqi,yixin.zhu}@ucla.edu,sczhu@stat.ucla.edu

ACM Reference Format:
Xu Xie Hangxin Liu Zhenliang Zhang Yuxing Qiu Feng Gao Siyuan Qi Yixin Zhu Song-Chun Zhu. 2019. Supplementary Material for VRGym: A Virtual Testbed for Physical and Interactive AI. In *ACM Turing Celebration Conference - China (ACM TURC 2019) (ACM TURC 2019), May 17–19, 2019, Chengdu, China.* ACM, New York, NY, USA, 4 pages.
https://doi.org/10.1145/3321408.3322633

1 RELATED WORK

1.1 Passive Dataset

Large-scale labelled datasets play an important role in current development of artificial intelligence (AI) and machine learning, *e.g.*, ImageNet [7] and Microsoft COCO [27] have greatly facilitated the advancements in both object detection and classification. More recently, some datasets are beyond categorization tasks, *e.g.*, Visual Genome [23] focuses on learning the relationship. However, manually labeling dataset is proven to be tedious and error-prone, limiting both its quantity and accuracy.

Synthetic image datasets have recently been a source of training data for object detection and correspondence matching [8, 11, 12, 29, 30, 34, 59], single-view reconstruction [17], view-point estimation [49], human pose estimation [10, 31, 41, 45, 46, 52, 57, 60], depth prediction [48], pedestrian detection [16, 28, 32, 53], action recognition [37, 38, 40], semantic segmentation[39], scene understanding [5, 13, 14, 47], and in benchmark data sets [15, 18]. Previously, synthetic imagery, generated on the fly, online, had been used in visual surveillance [36] and active vision / sensorimotor control [50]. However, these datasets can only afford passive observation or very limited interactions, thereby difficult to generalize to scenarios where an AI agent can interact with a human.

1.2 Simulation Platform

Table 1 summarizes detailed comparisons against similar simulation platforms, showing the uniqueness of VRGym.

Robotics simulation platforms originated from the Robotics Operating System (ROS) (*e.g.*, Gazebo [21] and V-Rep [42]) have been playing an important role in robotics development. However, as these platforms focus on the robotics application, they rarely provide means for either the virtual reality (VR) integration or the human interactions.

Virtual training platforms including OpenAI Gym [3] and MuJoCo [51] are designed to evaluate and benchmark machine learning algorithms. Although these platforms allow easy-setup and fast training by providing interfaces with popular packages and games, they still lack sufficient levels of interactions, especially for physical agents.

Physics-based simulation platforms leverage the content developed by the game industry and incorporate sophisticated physics-based simulation. For instance, CARLA [9] is an open-source simulator for autonomous driving, whereas AirSim [44] provides a photorealistic rendering of outdoor scenes for drone navigation. They are, however, platforms for specific tasks, *e.g.*, vehicle or drone navigation. More general-purpose platforms are also available, including AI2THOR [22] and Gibson [55]. Specifically, AI2THOR provides detailed 3D indoor scenes where AI agents can navigate in the scenes and interact with objects to perform tasks; however, there is no human embodiment, and most of the interactions are symbolic-level. Although virtual agents in Gibson can receive a constant stream of visual observations and a human can be represented as a Mujoco humanoid, it still lacks a sufficient level of manipulations and interactions for the human embodiment.

1.3 Domain Adaptation

Although the presented work does not directly involve domain adaptation, this plays a vital role in learning from virtual environments, as the goal of using virtual training is to transfer the learned model and apply it to real-world scenarios. A review of existing work in

Table 1: Comparison with existing 3D virtual environments. Scale: Contains a large number of scenes. Physics: Supports physics-based simulation on agents and objects. Real: Provides a life-like rendering. Action: Object states can be changed by actions. Fine-grained: Enables fine-grained actions and simulates plausible object state changes. Human: Humanoid agents. Multi: Supports a multi-agent setting.

Environment	Scale	Physics	Real	Action	Fine-grained	Human	Multi
SUNCG [47]	✓						
Matterport3D [5]	✓						
Malmo [19]	✓			✓			✓
DeepMind Lab [2]							
VizDoom [20]							✓
MINOS [43]	✓		✓				
HoME [4]	✓	✓	✓				
Gibson [55]	✓	✓	✓			✓	
House3D [54]	✓	✓	✓				
AI2-THOR [22]		✓	✓	✓			
VirtualHome [33]		✓	✓	✓		✓	
VRGym	✓	✓	✓	✓	✓	✓	✓

ACM TURC 2019, May 17–19, 2019, Chengdu, China

ACM ISBN 978-1-4503-7158-2/19/05…$15.00
https://doi.org/10.1145/3321408.3322633

this area is beyond the scope of this paper; we refer the reader to a recent comprehensive survey [6].

1.4 VR for Cognitive Studies and Machine Learning

VR is capable of offering versatile settings for human training and testing, providing a convenient way for cognitive studies to quickly set up a specific scenario without building a costly physical apparatus. There are many successful cases; some recent work includes studying the deceptive behaviors [1], examining human physical judgments in abnormal environments [58], improving driving habits [24], designing the game level automatically [56], and training for earthquake [26].

VR is also a fast means to collect data and train for machine learning. Towards this goal, some researchers have built plugins for game engines, such as UETorch [25] and UnrealCV [35]. However, to date, such plugins only offer APIs to control game state and record data, requiring additional packages to train virtual agents.

In contrast, VRGym is capable of providing detailed logging of the data generated inside VR environment, as well as supporting training virtual agents directly for various tasks, subsuming the functions provided in previous virtual environments.

2 SYSTEM PERFORMANCE

VRGym runs in real-time (30fps) on a modern PC with an Intel 8700K CPU, a set of DDR4 memory totaling 64GB, and an EVGA GTX 1080 Ti GPU. Figure 1 shows the system performance running the physics-based simulation of complex manipulation using the human input devices, together with the software and hardware interface we develop. The results indicate that the VRGym is efficient on CPU and memory utilization. As the real-time physics-based simulation relies heavily on parallel computing, it requires relatively more GPU power. In general, VRGym can be supported by modern computers without requiring special setups or dependencies.

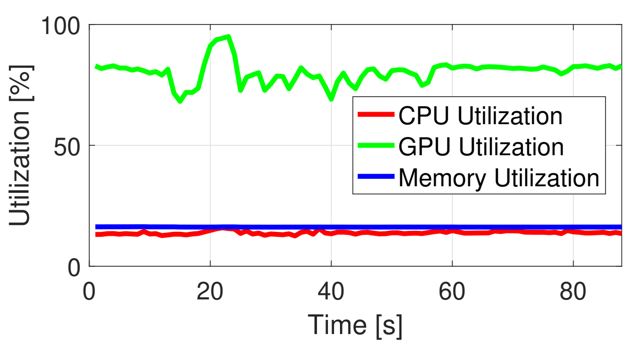

Figure 1: System performance: GPU (green), memory (blue), and CPU (red) utilization.

3 EVALUATION OF COMMUNICATION BANDWIDTH

A profile of the performance is shown in Figure 2, in which 20 packages are sent individually for each concurrent connection. The communication latency for the VRGym-ROS bridge increases from 0.04*sec* to 0.41*sec*. Linear regression is fitted to the mean of the latency $t9 = 2.9025, p = 0.01, r^2 = 0.9998$, indicating a strong linear trend with respect to the increase of the concurrent connections in the VRGym-ROS bridge.

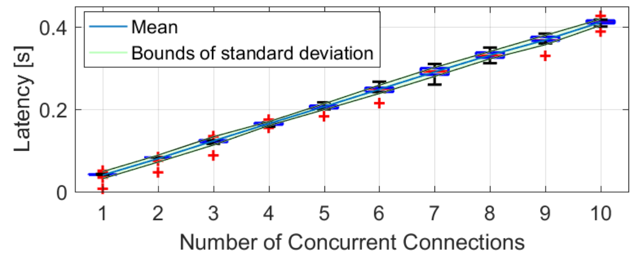

Figure 2: Evaluation of the latency in VRGym-ROS communication bridge. Each connection contains 20 packages, in which 512Kb data was sent. Linear regression is fitted to the mean of the latency $t9 = 2.9025, p = 0.01, r^2 = 0.9998$, indicating a strong linear trend with respect to the increase of the concurrent connections.

REFERENCES

[1] Carla Aravena, Mark Vo, Tao Gao, Takaaki Shiratori, and Lap-Fai Yu. 2017. Perception Meets Examination: Studying Deceptive Behaviors in VR. In *CogSci*.

[2] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. 2016. Deepmind lab. *arXiv preprint arXiv:1612.03801* (2016).

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym.

[4] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, and Aaron Courville. 2017. HoME: A household multimodal environment. *arXiv preprint arXiv:1711.11017* (2017).

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *International Conference on 3D Vision (3DV)*.

[6] Gabriela Csurka. 2017. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374* (2017).

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *CVPR*.

[9] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Annual Conference on Robot Learning (CoRL)*.

[10] Yu Du, Yongkang Wong, Yonghao Liu, Feilin Han, Yilin Gui, Zhen Wang, Mohan Kankanhalli, and Weidong Geng. 2016. Marker-less 3D human motion capture with monocular image sequence and height-maps. In *ECCV*.

[11] Sean Ryan Fanello, Cem Keskin, Shahram Izadi, Pushmeet Kohli, David Kim, David Sweeney, Antonio Criminisi, Jamie Shotton, Sing Bing Kang, and Tim Paek. 2014. Learning to be a depth camera for close-range human capture and interaction. *TOG* 33, 4 (2014), 86.

[12] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*.

[13] Ankur Handa, Viorica Pătrăucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. 2016. Understanding real world indoor scenes with synthetic data. In *CVPR*.

[14] Ankur Handa, Viorica Patraucean, Simon Stent, and Roberto Cipolla. 2016. SceneNet: an Annotated Model Generator for Indoor Scene Understanding. In *ICRA*.

[15] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. 2014. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *ICRA*.

[16] Hironori Hattori, Vishnu Naresh Boddeti, Kris M Kitani, and Takeo Kanade. 2015. Learning scene-specific pedestrian detectors without real data. In *CVPR*.

[17] Qixing Huang, Hai Wang, and Vladlen Koltun. 2015. Single-view reconstruction via joint analysis of image and shape collections. *TOG* 34, 4 (2015).

[18] Chenfanfu Jiang, Siyuan Qi, Yixin Zhu, Siyuan Huang, Jenny Lin, Lap-Fai Yu, Demetri Terzopoulos, and Song-Chun Zhu. 2018. Configurable 3D Scene Synthesis and 2D Image Rendering with Per-pixel Ground Truth Using Stochastic Grammars. *IJCV* 126, 9 (2018), 920–941.

[19] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The Malmo Platform for Artificial Intelligence Experimentation. In *IJCAI*.

[20] Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. 2016. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *Computational Intelligence and Games*.

[21] Nathan P Koenig and Andrew Howard. 2004. Design and use paradigms for Gazebo, an open-source multi-robot simulator.. In *IROS*.

[22] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An interactive 3d environment for visual AI.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 123, 1 (2017), 32–73.

[24] Yining Lang, Liang Wei, Fang Xu, Yibiao Zhao, and Lap-Fai Yu. 2018. Synthesizing Personalized Training Programs for Improving Driving Habits via Virtual Reality. In *Virtual Reality and 3D User Interfaces (VR)*.

[25] Adam Lerer, Sam Gross, and Rob Fergus. 2016. Learning physical intuition of block towers by example. In *ICML*.

[26] Changyang Li, Wei Liang, Chris Quigley, Yibiao Zhao, and Lap-Fai Yu. 2017. Earthquake safety training through virtual drills. *TVCG* 23, 4 (2017), 1275–1284.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

[28] Javier Marin, David Vázquez, David Gerónimo, and Antonio M López. 2010. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*.

[29] Yair Movshovitz-Attias, Takeo Kanade, and Yaser Sheikh. 2016. How useful is photo-realistic rendering for visual learning?. In *ECCV*.

[30] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. 2015. Learning deep object detectors from 3D models. In *CVPR*.

[31] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. 2012. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*.

[32] Leonid Pishchulin, Arjun Jain, Christian Wojek, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. 2011. Learning people detection models from few training samples. In *CVPR*.

[33] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. VirtualHome: Simulating Household Activities via Programs. In *CVPR*.

[34] Charles R Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. 2016. Volumetric and Multi-View CNNs for Object Classification on 3D Data. In *CVPR*.

[35] Weichao Qiu and Alan Yuille. 2016. Unrealcv: Connecting computer vision to unreal engine. In *ECCV Workshop*.

[36] Faisal Qureshi and Demetri Terzopoulos. 2008. Smart camera networks in virtual reality. *Proc. IEEE* 96, 10 (2008), 1640–1656.

[37] Hossein Rahmani and Ajmal Mian. 2015. Learning a non-linear knowledge transfer model for cross-view action recognition. In *CVPR*.

[38] Hossein Rahmani and Ajmal Mian. 2016. 3d action recognition from novel viewpoints. In *CVPR*.

[39] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. In *ECCV*.

[40] Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez. 2017. Procedural Generation of Videos to Train Deep Action Recognition Networks. In *CVPR*.

[41] Grégory Rogez and Cordelia Schmid. 2016. MoCap-guided data augmentation for 3D pose estimation in the wild. In *NIPS*.

[42] Eric Rohmer, Surya PN Singh, and Marc Freese. 2013. V-REP: A versatile and scalable robot simulation framework. In *IROS*.

[43] Manolis Savva, Angel X Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. 2017. MINOS: Multimodal Indoor Simulator for Navigation in Complex Environments. *arXiv preprint arXiv:1712.03931* (2017).

[44] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*. Springer, 621–635.

[45] Gregory Shakhnarovich, Paul Viola, and Trevor Darrell. 2003. Fast pose estimation with parameter-sensitive hashing. In *ICCV*.

[46] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (2013), 116–124.

[47] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. 2017. Semantic Scene Completion From a Single Depth Image. In *CVPR*.

[48] Hao Su, Qixing Huang, Niloy J Mitra, Yangyan Li, and Leonidas Guibas. 2014. Estimating image depth using shape collections. *TOG* 33, 4 (2014), 37.

[49] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. 2015. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *ICCV*.

[50] Demetri Terzopoulos and Tamer F Rabie. 1995. Animat vision: Active vision in artificial animals. In *ICCV*.

[51] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *IROS*.

[52] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from Synthetic Humans. In *CVPR*.

[53] David Vázquez, Antonio M Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. 2014. Virtual and real world adaptation for pedestrian detection. *PAMI* 36, 4 (2014), 797–809.

[54] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018. Building generalizable agents with a realistic and rich 3D environment. *arXiv preprint arXiv:1801.02209* (2018).

[55] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. 2018. Gibson Env: Real-World Perception for Embodied Agents. In *CVPR*.

[56] Biao Xie, Yongqi Zhang, Haikun Huang, Elisa Ogawa, Tongjian You, and Lap-Fai Yu. 2018. Exercise Intensity-Driven Level Design. *TVCG* 24, 4 (2018), 1661–1670.

[57] Hashim Yasin, Umar Iqbal, Björn Krüger, Andreas Weber, and Juergen Gall. 2016. A Dual-Source Approach for 3D Pose Estimation from a Single Image. In *CVPR*.
[58] Tian Ye, Siyuan Qi, James Kubricht, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. 2017. The Martian: Examining human physical judgments across virtual gravity fields. *TVCG* (2017).
[59] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. 2016. Learning Dense Correspondence via 3D-guided Cycle Consistency. In *CVPR*.
[60] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. 2016. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*.