ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

REMOTE EXPERIMENTS INVOLVING ARTIFICIAL INTELLIGENCE IN A 3D VIRTUAL ENVIRONMENT

Vijay Kumar Gumasa[1] and Dr. Manoj Eknath Patil[2]

[1]Research Scholar and [2]Research Guide, Department of Computer Science & Engineering, Faculty of Engineering & Technology, Dr. A. P. J. Abdul Kalam University Indore, MP, India

[1]vijay.gumasa@gmail.com and [2] mepatil@gmail.com

ABSTRACT

Recent advancements in Visual Artificial Intelligence (AI) research have led to significant progress, driven by the need to collect extensive data across diverse conditions and environments. However, the collection of such data is resource-intensive in terms of time and labor. Moreover, developing and testing visual AI algorithms, particularly those involving multi-sensory models, can incur substantial costs and pose real-world risks. To address these challenges, we introduce a 3D environment simulator equipped with a view synthesis module capable of generating highly realistic simulations and supporting adaptable configurations of multimodal sensors.

This simulator integrates cutting-edge techniques including adaptive view selection, depth refinement, and layered rendering to enhance the realism of generated imagery. The platform, known as PreSim, offers several benefits: (i) it facilitates the integration of multisensory models into a photo-realistic 3D environment, enabling perception and navigation within simulated scenes; (ii) leveraging an internal view synthesis module, PreSim streamlines the transition of algorithms from simulation to physical platforms, eliminating the need for domain adaptation; and (iii) it generates extensive datasets crucial for vision-based applications such as object pose and depth estimation.

Keywords: Simulation and Animation, RGB-D Perception, Sensor Fusion, Remote Experimentation, 3D Virtual Worlds

I. INTRODUCTION

Recent advancements in deep learning-based approaches have yielded significant progress in computer vision tasks like depth estimation [1] and 6D object pose estimation [2]. However, training and evaluating these data-driven methods require substantial amounts of labeled data, making collection and labeling a tedious and time-consuming process [2]. Simulated environments are emerging as a solution, offering vast quantities of annotated data for various AI tasks [2].

A key focus of environment simulators is achieving high-fidelity rendering of real-world scenes from diverse viewpoints. Open-source simulators exist that target this goal by manipulating scene parameters like geometry, lighting, texture, and 3D object representation [e.g., 3]. However, configuring these parameters can be labor-intensive [2]. Even with accurate modeling and parameter settings, simulated environments often struggle to capture the full richness and diversity of the real world [2]. This limitation, known as the "reality gap" [5], hinders the smooth transfer of algorithms developed and tested in simulation to real-world applications like obstacle avoidance, object recognition, and visual navigation [2].

Game engines have been leveraged to construct virtual environments due to their photorealistic rendering capabilities [4]. However, users are limited in creating custom environments with their own datasets due to the reliance on the game engine's detailed datasets [4]. Additionally, real-time rendering in game engines utilizes 3D graphics pipelines, where rendering time scales linearly with scene complexity (number of polygons) [6]. Dedicated hardware and architectures are often required for real-time performance [6].

Image-based rendering offers a solution, enabling real-time, realistic imagery without the limitations of 3D graphics pipelines [7]. This method utilizes a limited set of captured images to realistically visualize a 3D scene without complete reconstruction [7]. It has shown success in various environments [8]. Unlike traditional

rendering, runtime in image-based rendering is primarily dictated by the displayed image resolution, not scene complexity [7].

Leveraging image-based rendering, we introduce PreSim, a 3D photo-realistic environment simulator for AI algorithm training and testing. PreSim aims to bridge the reality gap by offering a vast collection of photorealistic Red Green Blue (RGB)-D views from arbitrary viewpoints, improving its suitability for vision-based applications. A key contribution of PreSim is its ability to provide a photorealistic 3D environment even in regions with inaccurate or missing 3D data. This ensures users still have access to precise multisensory model poses and color-depth image pairs from free viewpoints. The system incorporates a comprehensive visualizer offering real-time positions and paths of moving sensors, along with a global 3D map. Additionally, it integrates recorder components and a sequence controller for regulating sensor motion and capturing data crucial for AI development.

Our novel view synthesis module, built on image-based rendering principles, employs a combination of adaptive view selection, depth refinement, and layered 3D warping techniques. This combination aims to enhance the overall quality of synthesized images while minimizing rendering complexity.

II. LITERATURE SURVEY

C. Hong and colleagues [9] described the modern service industries as relying heavily on advanced technologies and the proficiency of service professionals to effectively conduct business operations in a globalized world. They also emphasize that service skills differ significantly from basic concepts or familiarity with science and technology. Without practical experience, individuals cannot receive effective training within a classroom. Service skills primarily involve soft skills, particularly those that require service staff to engage with applicable stakeholders, demonstrate an understanding of the service domain, address requests from multiple groups, evaluate potential solutions, prioritize tasks, and all agreed-upon final decisions.

S. Bronack and colleagues [11] employed a social constructivist framework to analyze the learning environment within the realm of 3D virtual worlds, emphasizing the significant disparities it possesses compared to web-based learning or traditional classroom environments. The authors argue that within the 3-dimensional (3D) world, students should be granted increased choices and support to create individual paths within this environment.

A. De Lucia and colleagues [7] present a virtual campus constructed as Second Life, incorporating four unique virtual spaces: a shared student campus, recreational areas, lecture rooms, and collaborative zones. Within a 3D multi-user virtual environment, the authors emphasize the importance of a user's connection to a learning community, along with their presence, awareness, and communication skills. A study was conducted with university students to assess the effectiveness of synchronous distance lectures using Second Life within the described learning environment. The findings of the study were highly positive.

D.C. Cliburn, J.L. Gross, et al. [8] employed a quasi-experimental design with pretest-posttest comparison groups approach to assess the impact of delivering a lecture in Second Life in contrast to conducting a lecture in the real world. The study demonstrated that individuals who attended the lecture in a physical setting outperformed participants who experienced a similar lecture within the virtual environment of Second Life during a terminal test quiz. In their comments, the researchers also highlighted the challenges faced by students, including difficulties in accessing the lecture material and the lack of restrictions on avatar behavior within the academic context.

P. Dev, et al [10] documented an extensive project that involved the development and evaluation of the computer-based simulator known as the Virtual Emergency Department. The objective of this project was to enable distance training for emergency medicine residency programs, emphasizing leadership and teamwork during trauma management. This aimed to successfully handle trauma without the need for practice with real patients.

L. Jarmon, et al [6] suggests that 3D virtual worlds have significant potential as suitable environments for experiential learning. They utilize a combination of research methods, including focus groups, surveys, journal content analysis, as well as virtual world snapshots and video, to systematically evaluate the instructional influence that Second Life has for facilitating experiential learning in interdisciplinary communication.

C. M. Itin, et.al [15] contend that experiential learning encompasses deriving meaning from direct experience and emphasizes the significance of the individual's learning process. This approach establishes a deeper connection with the learner by addressing their specific needs and desires on a personal level. Based on this definition, a narrative script is developed for the educational service offered within Second Life.

N. Koenig and A. Howard, et.al [14] are widely acknowledged for their utilization of advanced physics engines to render both indoor and outdoor environments. While Gazebo possesses a range of features, it has limitations when it comes to creating visually rich environments on a large scale and providing realistic imagery. It has not kept pace with the rapid progress made in the latest rendering techniques that enable photo-realistic rendering. Another category of methodologies utilizes game engines with the capability to render camera streams with photo-realistic quality.

M. Savva et al., [2] describe the utilization of the Magnum engine for the creation of photo-realistic virtual environments. They also introduce a modular library that facilitates the development of AI tasks, such as visual navigation, within this framework. Nevertheless, the richness of simulated environments is constrained by their strong dependence on the capabilities of the engines. On the contrary, environment simulator empowers individuals to construct customized environments using the datasets they possess.

R. Ortiz-Cayon, G. Drettakis, and A. Djelouah, et al. [5] employ a strategy of dividing the image into super pixels to maintain the boundaries of objects. They then individually project each super pixel onto the virtual perspective using a local shape-preserving warping technique, with the aim of improving the blending quality. However, the specified approach overlooks photo-consistency and continues to encounter challenges such as inaccurate occlusion edges and the flattening of silhouettes. There have been several works that have improved the quality of synthesized images by filling holes. However, these methods have a fixed number of input views, which can result in failure to fill holes when the selected views are irrelevant or redundant. To avoid such a scenario, an adaptive approach is utilized for selecting views.

Author(s)	Study Focus	Key Findings	Strengths	Limitations
C. Hong et al. [9]	Service skills in a globalized world	Service skills require practical experience and soft skills for stakeholder engagement, domain understanding, and collaborative decision-making.	Emphasizes the importance of practical skills beyond theory.	Not directly related to 3D virtual worlds for learning.
S. Bronack et al. [11]	Learning environment in 3D virtual worlds	3D virtual worlds offer a social constructivist learning environment with opportunities for student choice and individual learning paths.	Highlights learner agency and social aspects of 3D virtual learning.	Lacks details on specific learning outcomes or instructional strategies.
A. De Lucia et al. [7]	Virtual campus in Second Life	Positive student outcomes in a synchronous distance lecture using Second Life, emphasizing community, presence, and communication skills.	Demonstrates potential for virtual environments in distance learning.	Limited to a single study and specific platform (Second Life).
D.C. Cliburn et al. [8]	Lecture delivery in Second Life vs. real world	Lower performance in a Second Life lecture compared to a physical lecture. Highlights	Provides a comparative study on learning	Limited to a single lecture format and may not generalize

		technical challenges and lack of behavioral control in virtual environments.	effectiveness.	to other learning activities.
P. Dev et al. [10]	Virtual Emergency Department for training	Virtual Emergency Department simulator effectively trains leadership and teamwork in trauma management for medical residents.	Shows promise for virtual simulations in healthcare training.	Focuses on a specific training program and may not apply to all learning domains.
L. Jarmon et al. [6]	Second Life for experiential learning in interdisciplinary communication	Second Life has potential for experiential learning due to its immersive and interactive capabilities.	Emphasizes the potential for experiential learning in virtual environments.	Lacks a clear definition of "experiential learning" and specific learning outcomes.
C. M. Itin et al. [15]	Narrative scripts for learning in Second Life	Narrative scripts enhance the effectiveness of educational services in Second Life by personalizing the learning experience.	Highlights the importance of personalization in virtual learning environments.	Limited details on the narrative script development and its impact on learning.
N. Koenig et al. [14]	Rendering techniques for virtual environments	Advanced physics engines like Gazebo offer limitations in visual richness and photo-realism compared to game engines.	Provides insights into rendering techniques for virtual environments.	Not directly focused on 3D virtual worlds for learning.
M. Savva et al. [2]	Photo-realistic virtual environments with Magnum engine	Magnum engine enables photo-realistic virtual environments for AI tasks like visual navigation, but limitations exist due to engine dependence.	Demonstrates the potential for photo-realistic virtual environments.	Lacks focus on learning applications and user experience.
R. Ortiz-Cayon et al. [5]	View synthesis techniques for image quality	View synthesis techniques improve image quality but may face challenges with photo-consistency and hole filling.	Explores technical aspects of image quality in virtual environments.	Not directly related to learning applications in 3D virtual worlds.

III. METHODOLOGY

Fig.1 displays the architecture of our simulator. The system encompasses a multisensory model, a global visualizer, scene datasets, controllers, and a view synthesis module. The simulator developed utilizes the Robot Operating System (ROS), which is well-known for its modular design, facilitating effortless customization, upgrades, and reusability. In the virtual environment, the process begins by importing the point cloud generated through 3D reconstruction of the real scene within the ROS. Then, the point cloud alongside the camera poses of the input images is showcased using Rviz, a graphical 3D visualization tool customized for utilization within the ROS framework. Following that, the movement of the virtual camera within the virtual world and real-time estimation of its 6 Dimensional pose are accomplished by employing ROS. Utilizing the estimated pose as a reference, the system identifies the most closely matched pairs of color and depth images within a provided input dataset. Afterward, the chosen color and depth image pairs are employed to generate the synthesized view through the utilization of the view synthesis module. Simultaneously, the complete movement path of the mobile camera along with the generated color-and-depth image pairs is captured. Within the subsequent sections, detailed information is presented regarding the individual components of the simulator.

The aim is to construct a vision-based environment that is photo-realistic and allows free-viewpoint capabilities for tasks. Departing from previous techniques that build the entire virtual environment based on precisely reconstructed 3D geometry, the view synthesis module make use of utilizes a limited collection of RGB-D images for input. This module is capable of generating new color-and-depth image pairs with arbitrary viewpoints. The methodology incorporates innovative depth refinement and view selection procedures, which are subsequently succeeded by a rapid rendering process. The collaborative efforts of these components aim to enhance the overall quality of the synthesized images, concurrently reducing rendering complexity.

Achieving high-quality rendering requires precise alignment of object boundaries across color-and-depth image pairs, as well as accurate depth values. This is due to the fact that inaccurate depth values and misalignment frequently result in noticeable artifacts. In the offline preprocessing stage, the aim is to accomplish this specific objective by utilizing an algorithm for pixel-level refinement of depth across multiple views.

In addition to rectifying misalignment between color-and-depth image pairs and filling in holes, achieving high-quality synthesized images also depends on the careful selection of input views. Selecting redundant or incorrect views by considering distances or angles among them commonly leads to the blurring of images. To prevent such scenarios, the selection of input images is conducted meticulously, taking into account the angles, distances, and overlaps between two views.

Our proposal revolves around employing a technique called layered depth image-based rendering to generate new sets of color-and-depth image pairs. A fundamental element of image-based rendering is 3D warping, encompassing the projection of pixels from the source image plane to the global coordinate system. Subsequently, these pixels are reprojection to their new positions in a different image plane by utilizing camera intrinsic and extrinsic matrices. In cases where foreground and background entities occupy the same position during projection, it can result in the concealment or obstruction of the foreground objects by the background objects. This issue arises due to incorrect depth information or errors in the reprojection process. In order to address this issue, the depth map is partitioned into separate layers by utilizing the minimum and maximum depth values. For each individual layer, 3D warping is employed using matching color-and-depth image pairs to generate fresh images, followed by applying a median filter using a 3×3 window to complete any missing information within the generated image. Consequently, the newly generated images are combined in order to generate the final synthesized image. The method effectively addresses the issue of visibility by utilizing the capabilities of layered depths to represent concealed elements. A trade-off between speed and quality resulted in the determination that four layers are the optimal.

Upon the completion of the blending process, the synthesized image often contains gaps or holes resulting from the restricted quantity of input views utilized. In response to this challenge, we introduce an adaptive view selection approach, utilizing a flexible number of input images in order to efficiently address the gaps within the synthesized image. The process starts with the projection of a key image onto the virtual position, which is carefully chosen considering its distance, angle, and overlap with the virtual view. Subsequently, we identify and locate holes in the synthesized image. In the event that the size of the largest crater exceeds a predefined threshold (e.g., 0.04% of the entire image), an alternative input image is selected to fill the crater. This process is iterated as long as the size of the largest hole diminishes below the specified threshold or the maximum limit for the number of input views is reached.

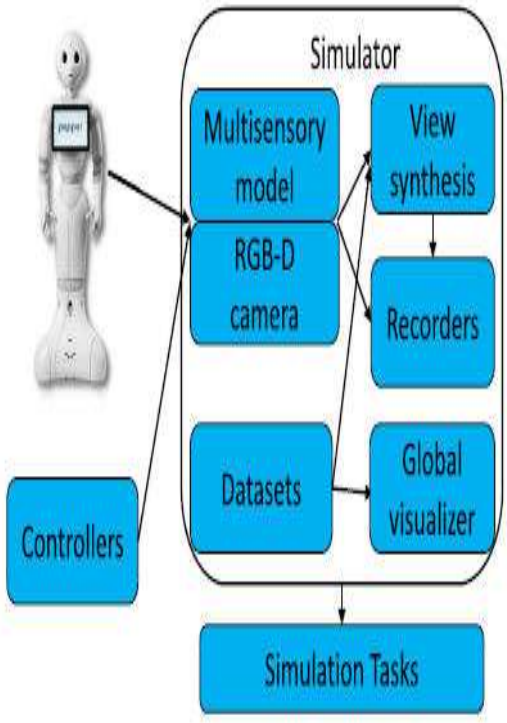

Fig.1: The architecture of simulator

Multisensory Models and Controllers: PreSim is purpose-built to address the challenge of transferring domains between simulation and the real world. Therefore, the multisensory model must consistently adhere to constraints imposed by space and physics, including considerations for gravity and collision.

Universal Robotic Description Formats (URDF) are employed for the description of multisensory models, including humanoid robots. Thus, it is possible to customize the model and its characteristics (e.g., varieties of sensors). For demonstration purposes, this utilize the Pepper robot, a social humanoid robot developed by Soft Bank.

To simplify the control complexity involved in the dynamic motions of the model, we offer a comprehensive range of practical controllers, such as navigation controllers and joint state. The joint state controller is utilized to regulate the movements of the model's joints, encompassing adjustments to the roll, pitch, and yaw angles. The navigation controller facilitates direct control of the model through the transmission of movement commands. Moreover, data recorders are provided, allowing for the storage of all the necessary data for learning-based approaches. Here's an example of a multisensory model and its trajectory.

IV. RESULT ANALYSIS

The evaluation of PreSim encompasses seven static datasets, which include three datasets we created ourselves (Table1, Table2, and Study room), four datasets (Playroom, Attic, Reading corner, Dorm) from and two dynamic datasets (Ballet and Break dancers). The seven static datasets consist of approximately 220 color-and-depth image pairs, encompassing objects such as black and texture-less items (e.g., writing boards and white walls), reflective objects (e.g., lights and bottles), and objects with small geometric features. In both dynamic datasets, there are sequences of 100 color-and-depth image pairs capturing individuals engaged in dancing activities. These sequences are captured using a set of eight static cameras arranged in an arc, each spaced 20 degrees apart.

In this approach, the ground truth image is selected at random from the initial captured dataset, representing a color image. Afterward, the remaining images are utilized to synthesize the selected image. It presents various instances of synthesized color images, displaying them alongside the corresponding ground truth images. The process of synthesizing a single image (1280 × 720) typically requires approximately 500–600 ms using a computer utilizing a 6-core Intel Core i7 8700 CPU operating at 3.19 GHz. While achieves a faster processing speed that compared to our approach, it is important to highlight that is dependent on a GPU, whereas our method functions independently of a GPU. It is evident that our proposed method successfully generates high-quality synthesized images.

Table.1: Perforamnce Analysis

Methods	Average PSNR Over 100 Images	
	Ballet	Breakdancers
VSRS[24]	30.23	31.17
Liu [25]	35.52	33.33
Dai [26]	32.55	31.77
Loghman [27]	33.36	31.64
Proposed	33.41	33.61

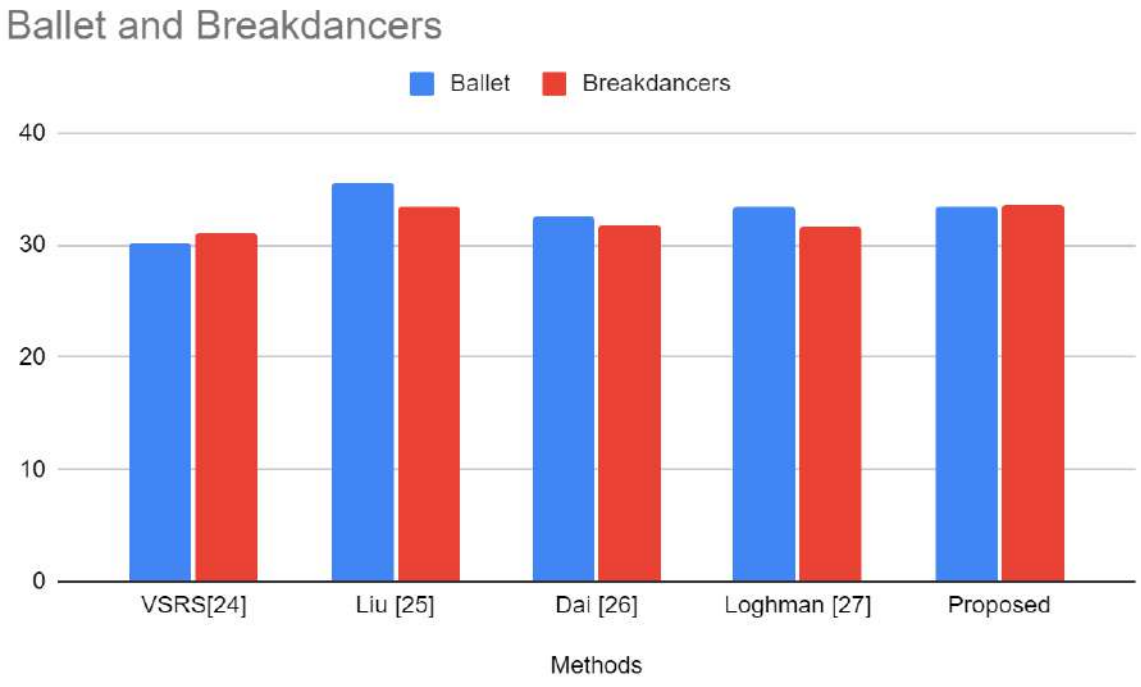

Fig.2: The PSNR comparison with layered 3Dwarping and Z-buffer on each frame.

V. CONCLUSION

Virtual environments and remote experimentation are pivotal tools that facilitate collaborative processes, offering unique insights into teaching collaboration and distributed learning across diverse applications. These technologies significantly enhance immersion by providing a genuine sense of presence and interaction. The exhibition aims to highlight the seamless integration of 3D remote experiments with virtual worlds, enriching the understanding of fundamental concepts in science and technology fields. The data generated from these experiments holds immense potential for training and validating data-driven approaches across various AI

applications, such as 6D object pose estimation and depth estimation. Through conducted experiments, our simulator demonstrates its capability to narrow the reality gap between real scenes and virtual environments. Consequently, vision-based algorithms developed within the simulation can be directly deployed on physical platforms without requiring extensive domain adaptation.

REFERENCES

[1] C. Wang, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3343–3352.

[2] M. Savva, "Habitat: A. platform for embodied ai research," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9339–9347

[3] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 239–248.

[4] P. Hedman, T. Ritschel, G. Drettakis, and G. Brostow, "Scalable inside-out image-based rendering," *ACM Trans. Gr.*, vol. 35, no. 6, pp. 1–11, 2016.

[5] R. Ortiz-Cayon, A. Djelouah, and G. Drettakis, "A bayesian approach for selective image-based rendering using superpixels," in *Proc. Int. Conf. 3D Vis.*, 2015, doi: 10.1109/3DV.2015.59.

[6] L. Jarmon, "Virtual world teaching, experiential learning, and assessment: An interdisciplinary communication course in Second Life," *Computers & Education*, vol. 53, no. 1, 2009, pp. 169-182.

[7] A. De Lucia, "Development and evaluation of a virtual campus on Second Life: The case of SecondDMI," *Computers & Education*, vol. 52, no. 1, 2009, pp. 220-233.

[8] D.C. Cliburn and J.L. Gross, "Second Life as a Medium for Lecturing in College Courses," *Proc. System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, 2009, pp. 1-8.

[9] C. Hong, "Service Design for 3D Virtual World Learning Applications," *Book Service Design for 3D Virtual World Learning Applications*, Series Service Design for 3D Virtual World Learning Applications, ed., Editor ed.^eds., IEEE Computer Society, 2008, pp.795-796

[10] P. Dev, "Virtual Worlds and Team Training," *Anesthesiology Clinics*, vol. 25, no. 2, 2007, pp. 321-336.

[11] S. Bronack, "Learning in the Zone: A social constructivist framework for distance education in a 3D virtual world," *Proc. Society for Information Technology & Teacher Education International Conference 2006*, AACE, 2006, pp. 268-275.

[12] M. Cavazza, "Causality and Virtual Reality Art," *Proc. 5th Conf. Creativity and Cognition*,ACM Press, 2005, pp. 4–12.

[13] J. Jacobson, "The CaveUT System: Immersive Entertainment Based on a Game Engine," *Proc. 2nd ACM SIGCHI Conf. Advances in Computer Entertainment Technology* (ACE 05), ACM Press, 2005, pp. 184–187.

[14] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, vol. 3, 2004, pp. 2149–2154.

[15] C. M.Itin, "Reasserting the Philosophy of Experiential Education as a Vehicle for Change in the 21st Century," *The Journal of Experiential Education*, vol. 22, no. 2, 1999, pp. 91-98.