3D Gaussian Splatting for Human-Robot Interaction

Shawn Bowser[1] and Stephanie M. Lukin[1]

***Abstract*— In order to assist a humans' ability to make decisions in uncertain and high-stakes scenarios, e.g., disaster relief, we aim to provide an interactive and "smart" visual model of an environment that a human can explore and query. We contribute a method for photorealistic 3D reconstruction of a scene from 2D images using improvements to 3D Gaussian Splatting (3DGS) methods. We showcase our process using a synthetic scene and showing a high level of fidelity between the ground truth synthetic scene and the reconstruction. We visualize the 3D reconstruction through a proof-of-concept web interface with robot ego-centric and exo-centric views, as well as semantic labels of objects within the scene, through which a human can interact. We discuss our ongoing design of one such human-robot collaborative task using this interface.**

I. INTRODUCTION

Robots have been increasingly deployed as assistants in disaster relief or robot rescue tasks with human partners, yet the hazardous nature of these environments often necessitates that human operators remain in safe yet remote locations while the robot navigates the dangerous areas [1]–[4]. This physical separation introduces unique challenges for effective human-robot interaction (HRI), including low-bandwidth communication and insufficient operator situational awareness. Under these circumstances, common ground must be established and maintained across the remote communication [5]. Shared visuals of the remote environment have been an effective tool in this, e.g., [6], [7], however, we posit that a "smart" visual model of the remote environment may provide novel ways for the human to query and explore in order to support their information gathering and decision making.

To design this "smart" visual model, we rely on a representation that 1) is underlying compatible with interactivity to support human-robot teaming, and 2) is an accurate reconstruction of a real-world environment. While many platforms, systems, and frameworks exist to develop virtual environments for interactive robots, e.g., [8]–[13], they may not effectively model the complex and unexpected conditions that may arise in disaster relief. Game engines or graphics suites may enable the creation of accurate real-world environments, but they are limited by hardware constraints to view these environments in real-time. Thus, we seek a simulation method that can rely on 3D reconstruction from images, either in real-world or virtual environments, that preserves the environment's fidelity and complexity.

In this paper, we leverage recent advancements in 3D Gaussian Splatting (3DGS) [15] for 3D scene reconstruction in order to create a semantically rich and interactive representation of an environment that can render novel views in real-time. 3DGS has proven effective in generating high-quality 3D scene reconstructions from 2D image sequences, a crucial capability for bridging the gap between simulation and real-world robot perception. Furthermore, the real-time rendering capabilities of 3DGS serves as an ideal backbone for our envisioned interactive user interface: a web-based system where the human operator can not only visualize the robot's surroundings in 3D but also actively interact with the scene, querying the robot about specific objects using natural language, and receiving text responses about the scene. This approach would not only enhance situational awareness and maintain common ground, but also facilitate more intuitive and efficient communication between the human and the robot, which may ultimately improve the effectiveness of disaster response operations. Models created with 3DGS do not need a complete representation of the underlying world, and instead infer novel views from a small collection of images from the environment. This has advantages over game engines or simulation platforms in terms of real-time rendering as a human explores the reconstructed scene.

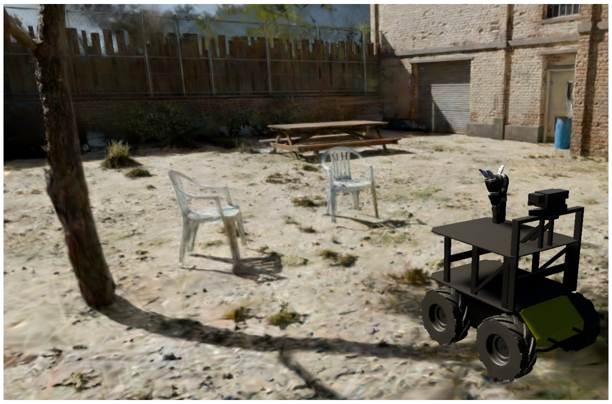


Fig. 1: A robot explores our outdoor 3D reconstructed environment in the human-robot interaction platform, RIVR [14]

The contributions of this paper are as follows:

- A method for photorealistic 3D reconstruction from images that enables a understanding through semantic segmentation and querying of an environment for both robotic systems and human operators.
- A proof-of-concept web interface with robot ego-centric and exo-centric views that can be used for testing human-robot interactions in 3D interactive scenes.

Figure 1 shows a robot in a 3D reconstructed environment following our method. To our knowledge, this is the first time the 3DGS approach for scene reconstruction has been

[1]DEVCOM Army Research Laboratory, 12015 Waterfront Drive, Playa Vista, 90094 stephanie.m.lukin.civ@army.mil

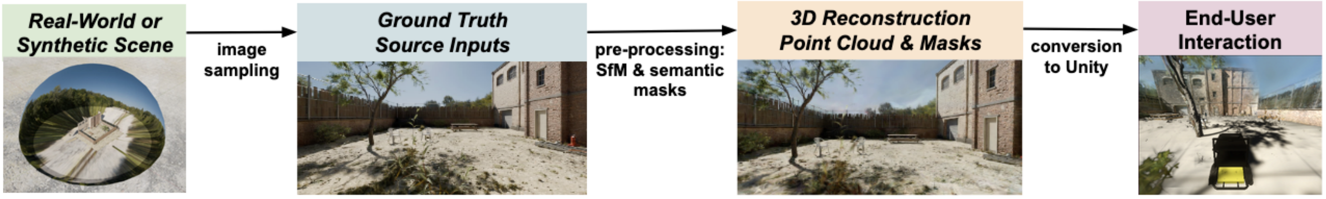

Real-World or Synthetic Scene
image sampling
Ground Truth Source Inputs
pre-processing: SfM & semantic masks
3D Reconstruction Point Cloud & Masks
conversion to Unity
End-User Interaction

Fig. 2: 3D Reconstruction Workflow

utilized for human-robot teaming with multiple views and semantic understanding.

In Section II we discuss related works that we build upon or take inspiration from or contrast in developing our approach. In Section III we describe our method for creating and optimizing 3D reconstructions and integrating them into a platform for human-robot teaming. We discuss how we envision this framework being used in an anomaly detection task as well as extend to others more broadly in Section IV, and we conclude in Section V by summarizing the feasibility and impactfulness of our proposed approach.

II. RELATED WORK

A. Sim-to-Real Transfer

The sim-to-real paradigm in robotics involves training behavior policies or learning models in simulated environments and transferring them to physical robots. Numerous sim-to-real datasets have been developed to facilitate the transfer of robotic skills from simulated environments to the real world [16]–[24]. These and other large-scale datasets are commonly used to evaluate autonomous robotic systems in tasks such as exploration, manipulation, and perception [25], [26]. However, an inherent sim-to-real gap persists due to the fundamental differences in visual appearance between simulated and real-world environments. To address this gap, the "real-to-sim" paradigm aims to inform and refine the simulation environment using real-world data [27], [28]. We adopt this approach using RGB images to fully reconstruct the geometry and appearance of 3D scenes.

B. 3D Scene Reconstruction

3D scene reconstruction is a fundamental task in HRI, enabling shared perception and understanding of an environment in three dimensions. Recent advancements in this area, including neural radiance fields (NeRFs) [29] and 3D Gaussian Splatting (3DGS) [15], have opened up new possibilities for creating immersive and interactive 3D representations of scenes. While both methods have demonstrated state-of-the-art results in appearance, the dependency on neural networks makes NeRFs resource-intensive to train and render. In contrast, 3D Gaussians can be rendered in real-time, are memory efficient, and can be directly embedded with language features [30]. This enables text-guided scene editing and scene object manipulation [31]–[34]. Additionally, 3DGS has been leveraged in recent works for autonomous navigation [35], [36] and can potentially be used for simultaneous localization and mapping (SLAM) [37]–[39]. We build off of 3DGS for its real-time visualization capabilities and discrete geometry representation, which we augment for potential real-time interactive applications (see [40] for a survey).

C. HRI Interfaces for Robot Operation

We focus our discussion of HRI as it pertains to sharing visuals for robot operation and collaboration. To this end, the vision of human-robot interactivity in this work is most similar to Walker et al. [6] who display robot-egocentric views (via onboard video streams) and robot-exocentric views (via remote environment reconstruction) for teleoperation. Their construction of the environment uses LiDAR-based point cloud to complete a complex inspection task in an unseen environment. Our approach similarly seeks to evaluate the performance of a human operator in remote control of a robot, but our 3D representation can handle highly detailed textures and geometry to be rendered using only an RGB sensor.

To allow a human to interact with our shared visual, we make use of the Robot Interaction in Virtual Reality (RIVR) platform [14]. RIVR supports environments in Unity with built-in compatibility with the Robot Operating System (ROS). It has been shown to be effective for a range of human-robot tasks including sim-to-real transfer and multi-modal tasks such as collaborative assembly of small-scale building exercises [41], using head-pose for object selection tasks [14], and using natural language instructions to navigate a robot [42]. We utilize the RIVR framework once our 3D representation has been created.

III. METHOD

Our work focuses on creating interactive 3D scenes from a small set of readily available RGB images. We demonstrate this process with a synthetic scene, then transfer the reconstruction to near-photoreal images in the sim-to-real transfer process and ensure that the resulting 3D scene representation is highly interpretable, as every 3D element can be directly traced back to the original 2D input (see Figure 2 for the workflow). Our reason for not using a real-world scene in this work is two-fold. First, we aim to have full control over the environment including the design of the space and placement of objects, in order to conduct a human-robot cooperative task in detecting anomalies and seeking objects of interest. This does not preclude the use of our approach on real-world scenes, however, as the 3DGS method was originally designed to create scenes from real-world images. The second reason to design the environment in simulation is to demonstrate how high-quality 3D recreations of synthetic

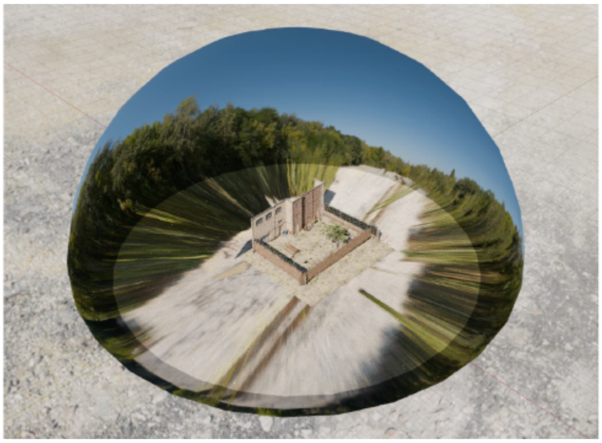


Fig. 3: Scene creation in Blender showing the synthetic environment

scenes can be effective in training and testing in human-robot teaming. We discussed these factors in more detail in Section IV.

In the remainder of this section, we describe our method for generating the synthetic scenes, reconstructing the 3D environment, and integrating it into a platform for interactive human-robot teaming.

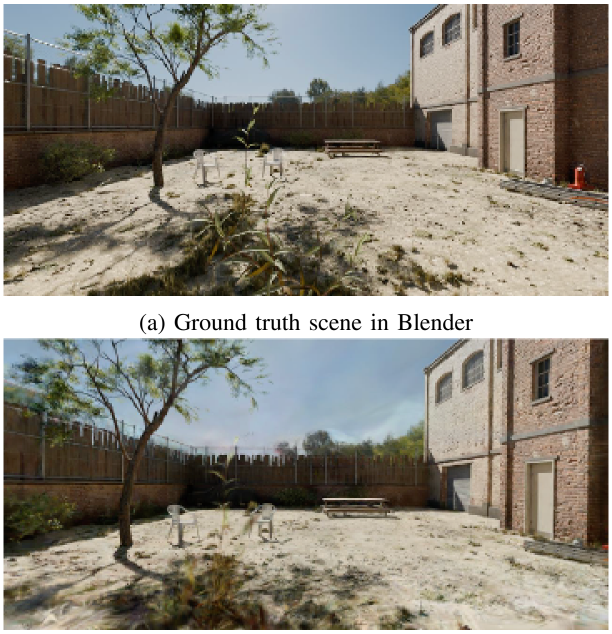
A. Scene Creation and Data Generation

Our scene layout is created with Blender [43], a free and open-source graphics tool, and consists of 3D assets from Polyhaven [44]. The navigable area spans approximately 374 sq. meters in an outdoor setting (see Figure 3 for a full dome-view of the scene). We distribute several objects of interest around the scene as targets for the anomaly detection task. These objects include potential hazards such as a chemical container, a fire extinguisher, and a box of electronics. We render 100 images sampled within the navigable area of the scene as the ground truth to use in our 3DGS reconstruction pipeline. Figure 4a is one of these ground truth images.

B. 3D Gaussian Splatting Reconstruction

Our 3D reconstruction pipeline leverages the 3DGS approach first introduced in Kerbl et al., [15]. We incorporate key optimizations from recent works to enhance its performance, efficiency, and semantic capabilities. We begin by processing the rendered images from the Blender scene using COLMAP [45], an open-source Structure-from-Motion (SfM) library, to obtain camera poses and a sparse 3D point cloud. This point cloud serves as the basis for our 3DGS representation. Following the approach of Girish et al. [46], we adopt a coarse-to-fine training strategy and quantized embeddings to significantly reduce per-point memory storage requirements, enabling faster training and rendering for real-time interaction. See references for full implementation details. Figure 4b shows near-photoreal images from our sim-to-real 3D reconstruction process from the same perspective as the ground truth in Figure 4a to show the visual fidelity.

Our reconstructed 3D scene is stored in point cloud format, with each 3D Gaussian represented by its mean position, color, opacity, and covariance matrix. To enable semantic understanding of the scene, we follow Cen et al. [47] and incorporate semantic feature fields obtained from Segment Anything Model (SAM) [48]. This extends our model without changing the base appearance of the 3D reconstruction, and can be easily rendered in Unity [49] or web browsers [50], [51].



(a) Ground truth scene in Blender

(b) 3D Gaussian Splat of Blender scene

Fig. 4: Warehouse yard in simulation showing the high preservation of visual fidelity of the 3D reconstruction (bottom) to the synthetic environment (top)

C. Integration

We integrate the Unity conversion of our scene into RIVR as the base environment for a simulated robot. A simulated Clearpath Husky, equipped with an onboard RGB camera, provides a first-person, ego-centric view for the operator to view from a web browser (shown in Figure 4b), as well as an exo-centric view (shown in Figure 1). The compatibility between our 3D reconstruction workflow and RIVR allows for many unique environments to be represented and tested under different human-robot experimental stimuli. Within RIVR, a user may view the environment via both the ego-centric and exo-centric views, as well as the semantic view of the scene. With RIVR's native support for ROS, the user may collaborate seamlessly with a robot through different methods of interactivity.

IV. DISCUSSION

We envision our framework for 3D scene recreation and interactivity as a versatile tool with dual applications: it can serve as an effective training platform for remote robot operations in disaster response scenarios, and it also holds the potential to be deployed as a near real-time decision support tool during active and ongoing incidents.

We consider both these applications in the context of a specific task: anomaly detection. While this task has been studied at the pixel-level, e.g., detecting a tear in a piece of fabric [52]–[54], it has only recently expanded to the scene-level where a human or robot seeks to visually analyze a room and identify anomalies, e.g., [55]–[57]. We posit that a human with the ability to actively explore the environment from multiple perspectives (i.e., robot ego-centric and exo-centric) as well as query the environment via natural language that can isolate and highlight objects (i.e., a "smart" environment that conveys crucial information from the semantic segmentation features) will excel at identifying out of place or hazardous objects within a large-scale environment.

In a training capacity, our platform can allow operators to familiarize themselves with the robot's capabilities and practice collaborative tasks in a safe, simulated space with a high level of realism to synthetic or real-world environments. The human's input can function as feedback to the robot's decision-making process. This interactive loop lays the foundation for future human-robot dialogue systems that can seamlessly integrate human expertise with robotic capabilities. Moreover, the immersive 3D environment and natural language interface provide a realistic and intuitive experience, enabling operators to develop the skills and strategies necessary for effective robot control.

Furthermore, our framework's real-time 3D reconstruction capabilities open up the possibility of utilizing it as a valuable tool during active disaster response efforts. By rapidly generating 3D representations of the affected environment from the robot's onboard camera feed, our platform may provide remote operators with crucial situational awareness, enabling them to make informed decisions and guide the robot's actions more effectively. This real-time feedback loop may significantly enhance the speed and accuracy of response efforts, potentially saving lives and minimizing damage.

V. CONCLUSION AND FUTURE WORK

This paper presents a method for 3D reconstruction from images using 3D Gaussian Splatting, and a pipeline for enabling human-robot interaction with these environments. We plan to conduct user studies to evaluate the effectiveness of our interface for remote robot operation on an anomaly detection task, placing particular focus on the utility of the semantic masks and the exo-centric views. We will investigate its potential for generating multi-modal grounded language datasets, focusing on metrics such as task completion time, accuracy, and subjective measures of visual quality and situational awareness for our anomaly detection task. To further enhance the real-time capabilities of our system, we will explore recent works that leverage 3DGS for visual SLAM [38], [39]. This would enable more responsive interactions between the operator and the robot, for example in highly dynamic environments.

We envision our platform as a versatile tool for accelerating research and development in HRI for complex, human-centeric domains such as disaster response. By providing operators with a more intuitive, immersive, and interactive interface, they may be empowered to communicate more effectively with robots and ultimately save more lives.

REFERENCES

[1] D. S. Drew, "Multi-agent systems for search and rescue applications," *Current Robotics Reports*, vol. 2, pp. 189–200, 2021.

[2] M. Chiou, G.-T. Epsimos, G. Nikolaou, P. Pappas, G. Petousakis, S. Mühl, and R. Stolkin, "Robot-assisted nuclear disaster response: Report and insights from a field exercise," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4545–4552.

[3] H. Chitikena, F. Sanfilippo, and S. Ma, "Robotics in search and rescue (sar) operations: An ethical and design perspective framework for response phase," *Applied Sciences*, vol. 13, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/3/1800

[4] K. Kanazawa, N. Sato, and Y. Morita, "Considerations on interaction with manipulator in virtual reality teleoperation interface for rescue robots," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023, pp. 386–391.

[5] H. H. Clark and S. E. Brennan, *Grounding in communication.* American Psychological Association, 1991.

[6] M. E. Walker, M. Gramopadhye, B. Ikeda, J. Burns, and D. Szafir, "The Cyber-Physical Control Room: A Mixed Reality Interface for Mobile Robot Teleoperation and Human-Robot Teaming," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 762–771.

[7] C. Bonial, S. M. Lukin, M. Abrams, A. Baker, L. Bonatelli, A. Foots, C. J. Hayes, C. Henry, T. Hudson, M. Marge, K. A. Pollard, R. Artstein, D. Traum, and C. Voss, "Human-Robot Dialogue Annotation for Multi-Modal Common Ground," *Language Resources and Evaluation*, To appear, 2024.

[8] M. Savva, A. X. Chang, X. Han, D. Gordon, L. Yu, and S. Song, "Virtualhome: A virtual environment for simulated human-robot interaction research," https://ai.stanford.edu/ alfredx/VirtualHome.html, 2017.

[9] C. Pinciroli, V. Trianni, R. O'Grady, G. Pini, A. Brutschy, M. Brambilla, N. Mathews, E. Ferrante, G. Di Caro, F. Ducatelle, and M. Birattari, "Argos multi-physics robot simulator," https://www.argos-sim.info/, 2012.

[10] C. Robotics, "Coppeliasim robot simulator," https://www.coppeliarobotics.com/, 2024.

[11] O. S. R. Foundation, "Gazebo simulator," https://gazebosim.org/, 2024.

[12] C. Ltd., "Webots robot simulator," https://cyberbotics.com/, 2024.

[13] J. Echeverria, N. Lassabe, A. Degroote, and S. Lemaignan, "Modular openrobots simulation engine," https://www.openrobots.org/morse/doc/stable/, 2011.

[14] P. Higgins, G. Y. Kebe, A. Berlier, K. Darvish, D. Engel, F. Ferraro, and C. Matuszek, "Towards making virtual human-robot interaction a reality," 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions, 2021.

[15] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–14, 2023.

[16] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar, and A. Farhadi, "RoboTHOR: An Open Simulation-to-Real Embodied AI Platform," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, K. Ehsani, J. Salvador, W. Han, E. Kolve, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5982–5994, 2022.

[18] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. De Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[19] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 912–10 922.

[20] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, “Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset,” *arXiv preprint arXiv:1809.00716*, 2018.

[21] M. Khanna, Y. Mao, H. Jiang, S. Haresh, B. Schacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang, and M. Savva, “Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation,” *arXiv preprint arXiv:2306.11290*, 2023.

[22] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao *et al.*, “3d-front: 3d furnished rooms with layouts and semantics,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 933–10 942.

[23] A. Garcia-Garcia, P. Martinez-Gonzalez, S. Oprea, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Jover-Alvarez, “The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6790–6797.

[24] B. Liu, J. Zhang, X. Zhang, W. Zhang, C. Yu, and Y. Zhou, “Furnishing your room by what you see: An end-to-end furniture set retrieval framework with rich annotated benchmark dataset,” *arXiv preprint arXiv:1911.09299*, 2019.

[25] 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,” *arXiv preprint arXiv:1709.06158*, 2017.

[26] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *Proc. Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[27] M. Torne, A. Simeonov, Z. Li, A. Chan, T. Chen, A. Gupta, and P. Agrawal, “Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation,” *Arxiv*, 2024.

[28] M. Chen, Q. Hu, Z. Yu, H. Thomas, A. Feng, Y. Hou, K. McCullough, F. Ren, and L. Soibelman, “Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset,” *arXiv preprint arXiv:2203.09065*, 2022.

[29] 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,” *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[30] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, “Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 676–21 685.

[31] M. Ye, M. Danelljan, F. Yu, and L. Ke, “Gaussian grouping: Segment and edit anything in 3d scenes,” *arXiv preprint arXiv:2312.00732*, 2023.

[32] J. Huang and H. Yu, “Point’n move: Interactive scene object manipulation on gaussian splatting radiance fields,” *arXiv preprint arXiv:2311.16737*, 2023.

[33] B. Dou, T. Zhang, Y. Ma, Z. Wang, and Z. Yuan, “Cosseggaussians: Compact and swift scene segmenting 3d gaussians,” *arXiv preprint arXiv:2401.05925*, 2024.

[34] Q. Gu, Z. Lv, D. Frost, S. Green, J. Straub, and C. Sweeney, “Egolifter: Open-world 3d segmentation for egocentric perception,” *arXiv preprint arXiv:2403.18118*, 2024.

[35] X. Lei, M. Wang, W. Zhou, and H. Li, “GaussNav: Gaussian Splatting for Visual Navigation,” *arXiv preprint arXiv:2403.11625*, 2024.

[36] G. Liu, W. Jiang, B. Lei, V. Pandey, K. Daniilidis, and N. Motee, “Beyond Uncertainty: Risk-Aware Active View Acquisition for Safe Robot Navigation and 3D Scene Understanding with FisherRF,” *arXiv preprint arXiv:2403.11396*, 2024.

[37] F. Tosi, Y. Zhang, Z. Gong, E. Sandström, S. Mattoccia, M. R. Oswald, and M. Poggi, “How nerfs and 3d gaussian splatting are reshaping slam: a survey,” *arXiv preprint arXiv:2402.13255*, vol. 4, 2024.

[38] M. Li, S. Liu, and H. Zhou, “Sgs-slam: Semantic gaussian splatting for neural dense slam,” *arXiv preprint arXiv:2402.03246*, 2024.

[39] S. Zhu, R. Qin, G. Wang, J. Liu, and H. Wang, “Semgauss-slam: Dense semantic gaussian splatting slam,” *arXiv preprint arXiv:2403.07494*, 2024.

[40] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao, “Recent advances in 3d gaussian splatting,” *Computational Visual Media*, pp. 1–30, 2024.

[41] P. Higgins, R. Barron, D. Engel, S. Lukin, and C. Matuszek, “A collaborative building task in vr vs. reality,” International Symposium on Experimental Robotics (ISER), 2023.

[42] S. Lukin, J. South, and S. Bowser, “CHRIS-Bot: A Robot for Dialogue and Scene Understanding of Anomalous Environments in Virtual Reality,” Tech. Rep. ARL-TR-9906, 2024.

[43] Blender Online Community, “Blender - a 3d modelling and rendering package,” Amsterdam, 2018. [Online]. Available: https://www.blender.org/

[44] “Poly haven,” https://polyhaven.com/, 2024.

[45] J. L. Schönberger, “Robust methods for accurate and efficient 3d modeling from unstructured imagery,” Ph.D. dissertation, ETH Zurich, 2018.

[46] S. Girish, K. Gupta, and A. Shrivastava, “Eagles: Efficient accelerated 3d gaussians with lightweight encodings,” *arXiv preprint arXiv:2312.04564*, 2023.

[47] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian, “Segment any 3d gaussians,” *arXiv preprint arXiv:2312.00860*, 2023.

[48] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[49] A. Pranckevičius, “UnityGaussianSplatting,” https://github.com/aras-p/UnityGaussianSplatting, 2023.

[50] antimatter15, “WebGL 3D Gaussian Splat Viewer,” https://github.com/antimatter15/splat, 2024.

[51] M. Kellogg, “GaussianSplats3D,” https://github.com/mkkellogg/GaussianSplats3D, 2024.

[52] X. Jiang, G. Xie, J. Wang, Y. Liu, C. Wang, F. Zheng, and Y. Jin, “A survey of visual sensory anomaly detection,” arXiv preprint arXiv:2202.07006, 2022.

[53] V. Zavrtanik, M. Kristan, and D. Skočaj, “Reconstruction by inpainting for visual anomaly detection,” *Pattern Recognition*, vol. 112, p. 107706, 2021.

[54] J. Li, X. Xu, L. Gao, Z. Wang, and J. Shao, “Cognitive visual anomaly detection with constrained latent representations for industrial inspection robot,” *Applied Soft Computing*, vol. 95, p. 106539, 2020.

[55] S. M. Lukin and R. Sharma, “Anomaly Detection with Visual Question Answering,” DEVCOM Army Research Laboratory, Tech. Rep. ARL-TR-9817, 2023.

[56] S. M. Lukin, R. Sharma, and M. Bellissimo, “Learning to Understand Anomalous Scenes from Human Interactions,” DEVCOM Army Research Laboratory, Tech. Rep. ARL-TR-9624, 2023.

[57] J. F. Mullen Jr, P. Goyal, R. Piramuthu, M. Johnston, D. Manocha, and R. Ghanadan, ““Don’t forget to put the milk back!” Dataset for Enabling Embodied Agents to Detect Anomalous Situations,” *arXiv preprint arXiv:2404.08827*, 2024.