Artifact Removal 3D Neural Reconstruction based Large-scale Digital Twin Construction for Industrial Robot Simulation

Ki Hoon Kwon, Suwoong Lee, and Byeong Hak Kim*

Abstract— **Digital twins are essential for robot simulation and real-time monitoring, but conventional graphics engine-based modeling methods suffer from limited realism and poor adaptability to dynamic environments, resulting in a significant Sim2Real gap. This paper presents a neural reconstruction based digital twin construction method that learns 3D spatial representations directly from multi-view 2D images to achieve photorealistic virtual environments. Using industrial site images captured with a 360° camera, dynamic object inpainting and clustering-based noise removal are applied to improve the quality of the initial 3D representation. The 3D Gaussian Splatting based neural reconstruction model is trained with a combined loss function incorporating Smooth L1, MS-SSIM, and LPIPS. To further enhance reconstruction efficiency and visual quality in large-scale scenes, we introduce a hard negative mining-based adaptive learning strategy and a diffusion model -based 3D artifact refinement module. Experimental results demonstrate high visual fidelity, achieving improved PSNR (26.69 dB) and SSIM (0.761) with a 10.2% lower standard deviation than the baseline 3D Gaussian Splatting, validating the feasibility of the proposed approach for constructing reliable digital twins for robot simulation.**

I. INTRODUCTION

Digital twin is a core technology of digital transformation that enables monitoring, simulation, prediction, and control by precisely replicating physical objects, spaces, and systems in virtual environments. In the industrial robotics field, digital twins enable virtual execution of robot placement, path planning, and task simulation in real environments, significantly reducing costs and risks. The consistency between virtual space information and the real environment is essential for advancing digital twin functionality, and simulation results become reliable only when such consistency is ensured.

However, current commercialized digital twin construction methods have several limitations. First, CAD/BIM-based approaches using design drawings provide accurate spatial information but struggle to reflect changes occurring on-site in real-time. Second, graphics engine-based manual modeling offers high flexibility but requires significant production costs and time while lacking realism. Third, photogrammetry-based approaches reconstruct 3D from camera-captured images providing high realism but lack detail and require substantial computational resources. This limited realism becomes a major cause of the domain gap (Sim2Real Gap) between real and virtual environments, causing performance degradation when robot algorithms trained in simulation are transferred to real environments.

With recent advances in computer vision and deep learning, neural reconstruction technology has garnered significant attention. This technology learns image data and 3D spatial information to generate realistic images from new viewpoints, enabling free viewpoint movement based on photorealistic imagery. Representative methods include Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3D GS) [1,2]. 3D Gaussian Splatting, in particular, uses explicit 3D Gaussian primitives enabling real-time rendering, trained with Structure from Motion (SfM) data to reconstruct 3D models by minimizing the difference between rendered and actual images through loss functions.

To overcome the domain gap limitations of conventional graphics-based robot simulation, recent research has actively explored applying neural reconstruction technology to robot simulation. SplatSim achieved zero-shot Sim2Real transfer using Gaussian Splatting, and Vid2Sim built realistic and interactive simulation environments from video [3,4]. These studies demonstrate that providing visual, spatial, and physical information similar to reality can minimize performance degradation when transferring simulation learning results to real environments.

Nevertheless, challenges remain in applying existing neural reconstruction techniques to large-scale industrial environments. Performance is significantly influenced by the precision of initial SfM results used for training, and noise and outliers in SfM data can persist as floaters in trained models. Additionally, in large-scale environments, the number of times (epochs) each image can contribute to Gaussians during training relatively decreases, increasing the likelihood of learning bias toward specific viewpoints or regions.

This paper proposes a large-scale digital twin construction method that overcomes these limitations through effective data preprocessing and improved learning algorithms. Using industrial site images captured with a 360° camera, we improve initial 3D model precision through dynamic object inpainting and clustering-based noise removal. During training, we develop a loss function

*This study has been conducted with the support of the Korea Institute of Industrial Technology as Intelligent Production System Technology Enabled by the Convergence of Multi-Domain Digital Twin and Connected AI(kitech EH-26-0002), Development of a Flex-Demo-Workshop for Small and Medium-Sized Factories(kitech EO-26-0005) and the Robot Industry Technology Development Project of the Korea Evaluation Institute of Industrial Technology (RS-2024-00422721, Development of safe and reliable XAI-based robotic workcell safety sensors and control modules) funded by the Ministry of Trade, Industry Energy (MOTIE, Korea).

Ki Hoon Kwon is with the Korea Institute of Industrial Technology, Daegu 42994, Republic of Korea (e-mail: kwon149@kitech.re.kr).

Suwoong Lee is with the Korea Institute of Industrial Technology, Daegu 42994, Republic of Korea (e-mail: lee@kitech.re.kr).

Byeong Hak Kim is with the Korea Institute of Industrial Technology, Daegu 42994, Republic of Korea (phone: 053-580-0181; fax: 053-580-0120; e-mail: bhkim81@kitech.re.kr).

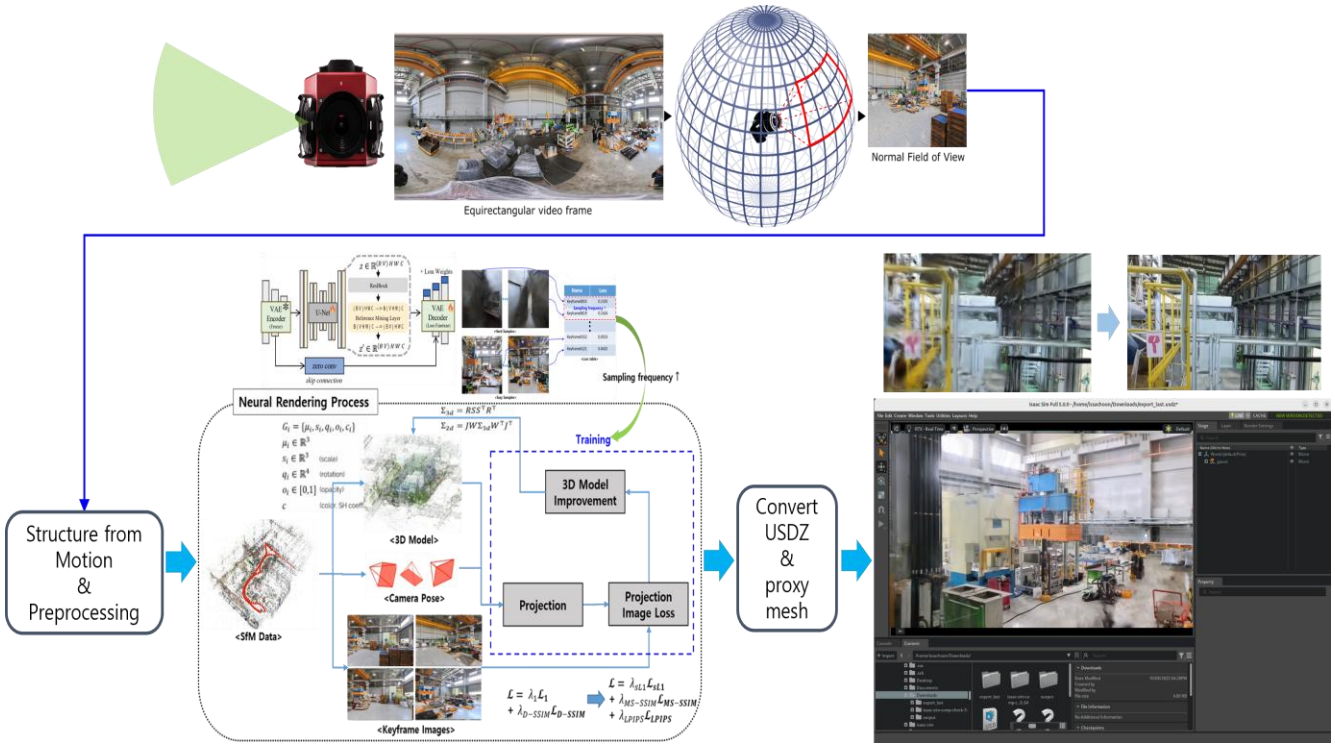

Figure 1. Overall architecture of the proposed method.

combining Smooth L1, Multi-Scale Structural Similarity (MS-SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), along with a hard negative mining-based adaptive learning algorithm and a diffusion-based 3D artifact refinement model to achieve efficient reconstruction and regional quality balance for large-scale spaces.

II. Proposed Method

The proposed system comprises three primary stages: SfM-based data preprocessing, 3D reconstruction via neural rendering, and simulation platform integration. Initially, multiple on-site images captured with a 360-degree camera serve as input, and initial 3D data is obtained through SfM with multi-view geometric consistency analysis. The SfM data containing initial 3D data, images, and camera pose information serves as input data for neural reconstruction training, and the quality of this SfM data significantly impacts final model quality. Subsequently, high-quality 3D virtual spaces are generated through proposed 3D GS model training based on preprocessed data. The model with learned 3D information provides high-quality 3D representation capable of real-time rendering and can generate novel view images not used in training. Finally, the generated 3D space model can be converted to USDZ format and proxy mesh for potential integration with robot simulation platforms such as NVIDIA Isaac Sim.

A. SfM Data Preprocessing

Effective data preprocessing techniques were applied to improve 3D reconstruction quality. This study developed and applied two key preprocessing techniques. For dynamic object removal, we developed a deep learning-based automatic detection and inpainting algorithm. Industrial sites contain various dynamic objects such as workers, vehicles, and mobile equipment, and including these objects in training data causes floaters and noise during neural reconstruction. To address this, we detect dynamic objects (people, vehicles, etc.) using an object detection model, then obtain precise instance masks through a foundation segmentation model, after which the masked regions are naturally restored with surrounding backgrounds using an inpainting pipeline [5], thereby pre-removing artifacts that could occur due to dynamic objects during training. For noise filtering, we applied a clustering-based technique to prevent quality degradation caused by noise and outliers in initial 3D data. By analyzing point cloud density and distribution information, we identify isolated point clusters separated from main structures. Through confidence-based and statistical analysis, point clouds determined to be outliers are removed, thereby improving initial 3D data quality.

B. Adaptive Learning for Large-scale 3D Reconstruction

To achieve immersive 3D virtual space generation, this study utilizes 3D GS, leveraging its explicit neural reconstruction capabilities to learn and reconstruct detailed 3D spatial information. 3D GS represents scenes with millions of 3D Gaussian primitives, where each Gaussian has parameters including position, scale, rotation, opacity, and color (represented by Spherical Harmonics coefficients). These parameters are optimized via gradient descent to minimize the difference between rendered and actual images. This study proposes improved neural reconstruction training methods to enhance large-scale space reconstruction efficiency and mitigate regional quality imbalance.

Advanced Loss Function: While Conventional 3D GS uses a combination of L1 loss and D-SSIM loss, this study designed an advanced loss function combining Smooth L1, MS-SSIM, and LPIPS as formulated in Equation (1)-(5).

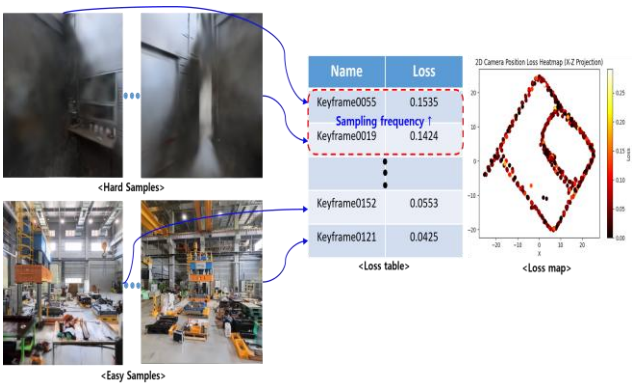

Figure 2. Example of loss table and loss map.

$$\mathcal{L}_{sL1} = \sum_{n=1}^{N} \begin{cases} \frac{(k_n - r_n)^2}{2\beta}, & if\ |k_n - r_n| < \beta \\ |k_n - r_n| - \frac{1}{2}\beta, & otherwise \end{cases} \quad (1)$$

$$MS\text{-}SSIM(x,y) = l_M^{\alpha}(x,y) \cdot \prod_{j=1}^{M} cs_j^{\beta_j}(x,y) \quad (2)$$

$$\mathcal{L}_{MS-SSIM} = 1 - MS\text{-}SSIM(k,r) \quad (3)$$

$$\mathcal{L}_{LPIPS} = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\widehat{k_{hw}^l} - \widehat{r_{hw}^l} \right) \right\|_2^2 \quad (4)$$

$$\mathcal{L}_{total} = \lambda_{sL1}\mathcal{L}_{sL1} + \lambda_{MS-SSIM}\mathcal{L}_{MS-SSIM} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} \quad (5)$$

Here, k and r denote the keyframe and rendered images respectively, $\beta$is a tuning parameter for Smooth L1, w_l represents feature weights in layer l, and λ denotes the respective weight for each loss component. Smooth L1 loss provides robustness to outliers in pixel-wise difference learning and contributes to training stability and detail preservation. MS-SSIM loss measures structural similarity at multiple scales, providing robustness to scale changes. LPIPS loss measures perceptual similarity based on features from pre-trained Convolutional neural network (CNN). LPIPS loss is effective for texture quality improvement, and in this study, it is added in the later training phase (iteration >= 80%) to focus on fine-grained texture improvement after basic structure learning is completed.

Hard Negative Mining-based Adaptive Learning: To address regional quality imbalance in large-scale environments, we developed a hard negative mining-based adaptive learning algorithm. In large-scale environments, the number of times each image can contribute to Gaussian optimization during training relatively decreases, leading to learning bias toward specific viewpoints or regions. To overcome this limitation, loss function values are calculated for each keyframe image used in training and sorted into a table. As shown in Figure 2, this loss table provides a quantitative measure of reconstruction difficulty for each region, where higher loss values indicate regions that require more learning attention. In the middle training phase (iteration 30%~70%), the learning frequency of the automatically selected top 30% data (hard samples) based on loss ranking is increased. Specifically, in every 10 iterations, one iteration randomly selects from hard samples with high loss values for training, thereby increasing learning opportunities for those regions. This hierarchical learning approach minimizes regional deviation and resolves quality imbalance. Additionally, the loss table analysis enables identification of regions requiring additional data capture or retraining, facilitating efficient data collection and model improvement.

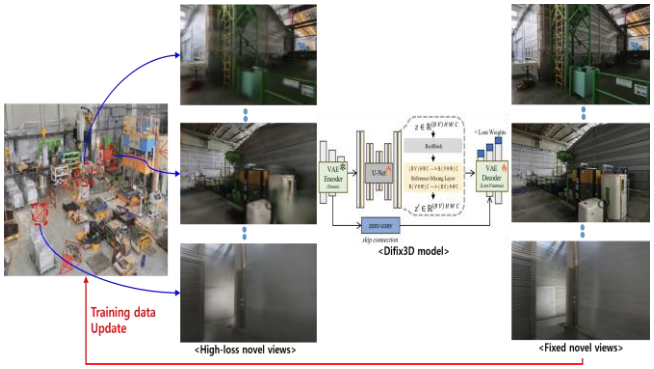

Figure 3. Artifact removal and learning data augmentation process.

Diffusion-based Artifact Removal: Figure 3 shows how the diffusion-based artifact removal model is applied as a data augmentation technique for hard sample regions during the later training phase (iteration ≥ 70%). Even after hard negative mining, some challenging regions may still exhibit artifacts such as blurring or structural distortions due to insufficient viewpoint coverage or complex geometry. To address this, novel view images are generated from camera poses interpolated around keyframe images with high loss values. These synthesized novel views often contain artifacts that degrade visual quality. The Difix3D model, a single-step diffusion model specifically designed for 3D reconstruction refinement, is then applied to remove these artifacts and improve the quality of generated images [6]. The improved novel view images serve as new spatial training data to further enhance quality in hard sample regions. This iterative refinement process enables the model to learn from cleaner, artifact-free representations, resulting in improved reconstruction quality even in previously challenging regions.

III. Experiments and Results

Experiments were conducted in a large-scale indoor experimental facility at the research institute. The Ladybug5+ (360-degree camera) was used for image capture, and NVIDIA RTX 4090 GPU was utilized for training. A total of

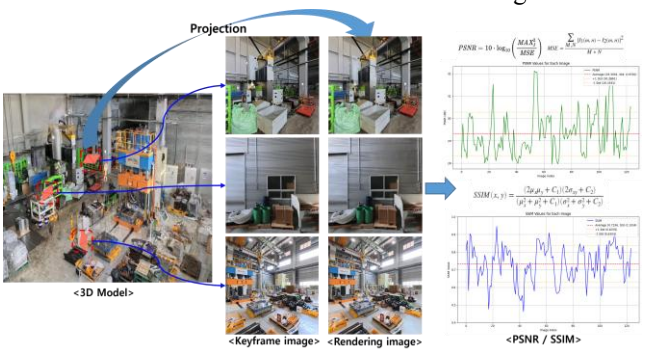

Figure 4. Rendering quality analysis results.

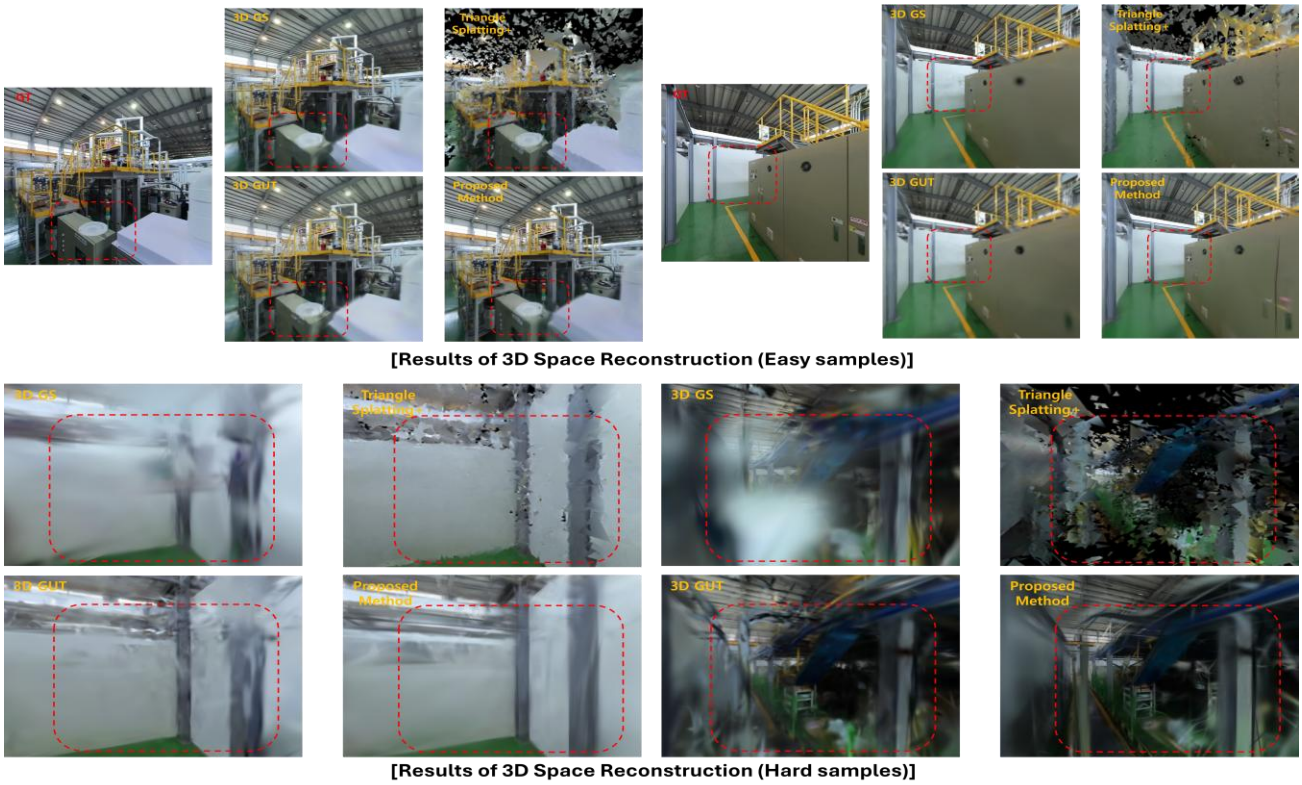

Figure 5. Qualitative comparison of different methods.

1,316 images were captured within an indoor experimental facility to construct the dataset. To evaluate the quality of generated 3D virtual spaces, the 3D model was projected and rendered at specific camera poses, and quality and similarity were measured through comparative analysis with corresponding keyframe images, as shown in Figure 4. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) were used as evaluation metrics, with mean and standard deviation presented together to evaluate regional quality balance.

Quantitative Evaluation Results: Table I shows the quantitative comparison results for the experimental building.

Comparison methods include baseline 3D Gaussian Splatting, Triangle Splatting+, and 3D Gaussian Unscented Transform (3D GUT)[7,8]. The proposed method achieved improvements of 0.64 dB in PSNR and 0.045 in SSIM compared to baseline 3D Gaussian Splatting. Notably, the standard deviation decreased from 1.09 to 0.979, confirming that regional quality imbalance has been mitigated. These results validate that the proposed method can be effectively applied to various large-scale industrial environments.

TABLE I. QUANTITATIVE COMPARISON RESULTS

Method	PSNR (mean/std)	SSIM (mean/std)
3D GS[2]	26.05 / 1.09	0.716 / 0.110
Triangle Splatting+[7]	20.66 / 1.18	0.593 / 0.151
3D GUT[8]	26.42 / 1.11	0.745 / 0.108
Proposed	**26.69 / 0.979**	**0.761 / 0.102**

Qualitative Evaluation Results: Figure 5 shows the results of the qualitative analysis, which further confirms the superiority of the proposed method. In easy sample regions with low loss values, all methods showed relatively good results, but differences between methods became clear in hard sample regions with high loss values. Baseline 3D GS exhibited severe blurring in hard sample regions, with structural detail loss observed. Triangle Splatting+ showed serious artifacts and holes, demonstrating limitations for industrial environment applications. 3D GUT showed improved results over baseline 3D GS but still exhibited quality degradation in some regions.

In contrast, the proposed method effectively preserved structural details even in hard sample regions while maintaining uniform quality throughout. Equipment structure edges and textures were clearly reconstructed, confirming the

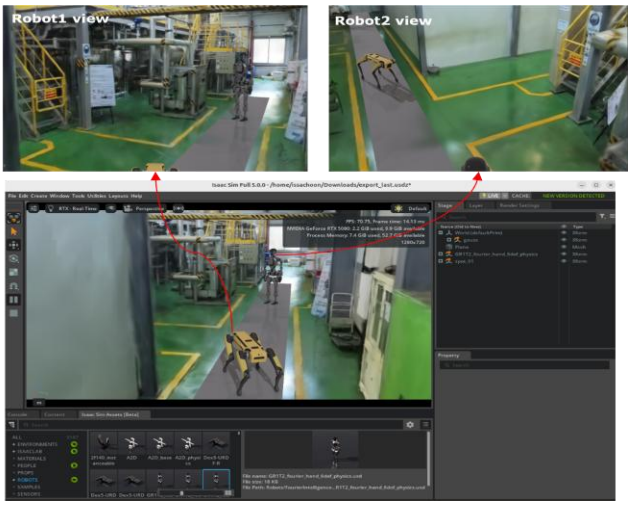

Figure 6. Application example of digital twin for robot simulation.

ability to satisfy environment recognition accuracy required for robot simulation. These results demonstrate that the proposed methods, including hard negative mining, the improved loss function, and the diffusion-based artifact refinement, effectively resolve quality imbalance issues in large-scale environments.

The generated high-quality 3D virtual space can be utilized as background and environment data for robot simulation, as shown in Figure 6. Unlike conventional graphics-based simulation environments, the proposed method provides photorealistic rendering images that closely resemble real-world scenes. This enables robot simulation and training in visually authentic environments, thereby mitigating the Sim2Real gap when deploying robots in actual industrial sites.

IV. Conclusion

This paper proposed a neural reconstruction-based large-scale 3D digital twin construction technology for industrial robot simulation. Using industrial site images captured with a 360-degree camera, we applied data preprocessing including dynamic object inpainting and clustering-based noise removal to improve initial 3D model precision. Furthermore, we achieved efficient reconstruction and quality balance for large-scale spaces through a loss function combining Smooth L1, MS-SSIM, and LPIPS, along with hard negative mining-based adaptive learning algorithm and diffusion-based artifact removal techniques. Experimental results showed that the proposed method achieved PSNR of 26.69 dB and SSIM of 0.761 at the experimental building, demonstrating improved performance over existing methods, with reduced standard deviation confirming mitigation of regional quality imbalance. The generated 3D virtual spaces demonstrated high visual similarity to real environments, confirming the feasibility of constructing reliable simulation environments where robots can visually interact almost identically to real environments. These results confirm that the proposed method shows promise for large-scale industrial environments, although the current validation is limited to a single indoor facility and the practical Sim2Real benefit for downstream robotic tasks remains to be quantitatively verified. Future work will focus on expanding to a digital twin-based smart factory management platform through integration with robot automation systems and process data. We will also continue research on Real2Sim/Sim2Real technology development to verify and improve the real-world transfer performance of robot algorithms trained in virtual environments.

References

[1] 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," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139:1–139:14, 2023.

[3] M. N. Qureshi, S. Garg, F. Yandun, D. Held, G. Kantor, and A. Silwal, "SplatSim: Zero-shot sim2real transfer of RGB manipulation policies using Gaussian splatting," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025.

[4] Z. Xie, Z. Liu, Z. Peng, W. Wu, and B. Zhou, "Vid2Sim: Realistic and interactive simulation from video for urban navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

[5] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen, "Inpaint anything: Segment anything meets image inpainting," *arXiv preprint arXiv:2304.06790*, 2023.

[6] J. Z. Wu, Y. Zhang, H. Turki, X. Ren, J. Gao, M. Z. Shou, S. Fidler, Z. Gojcic, and H. Ling, "DiFix3D+: Improving 3D reconstructions with single-step diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

[7] J. Held, R. Vandeghen, S. Son, D. Rebain, M. Gadelha, Y. Zhou, M. C. Lin, M. V. Droogenbroeck, and A. Tagliasacchi, "Triangle Splatting+: Differentiable rendering with opaque triangles," *arXiv preprint arXiv:2509.25122,* 2025.

[8] Q. Wu, J. M. Esturo, A. Mirzaei, N. Moenne-Loccoz, and Z. Gojcic, "3DGUT: Enabling distorted cameras and secondary rays in Gaussian splatting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR),* 2025.