NeRF-Con : Neural Radiance Fields for Automated Construction Progress Monitoring

Yuntae Jeon[1], Almo Senja Kulinan[1], Dai Quoc Tran[2], Minsoo Park[3] and Seunghee Park[4]

[1]Department of Global Smart City, Sungkyunkwan University, Suwon, South Korea
[2]Global Engineering Institute for Ultimate Society, Sungkyunkwan University, Suwon, South Korea
[3]Sungkyun AI Research Institute, Sungkyunkwan University, Suwon, South Korea
[4]School of Civil, Architectural Engineering and Landscape Architecture, Sungkyunkwan University, South Korea

jyt0131@g.skku.edu, almosenja14@skku.edu, daitran@skku.edu, pms5343@skku.edu, shparkpc@skku.edu

Abstract -

The monitoring of construction progress is crucial for ensuring project timelines, budget adherence, and quality control. Traditional methods often involve manual inspection, which is labor-intensive and prone to human error. We introduce NeRF-Con, an innovative approach utilizing Neural Radiance Fields (NeRF) to automate the process of construction progress monitoring. NeRF-Con can infer images that render the construction site with a level of quality comparable to reality by utilizing NeRF, which synthesizes novel views of complex scenes from a sparse set of images. Additionally, by employing a segmentation model, NeRF-Con can compare these rendered images with BIM to evaluate the progress of the work. This capability is achieved by training the model using handheld smartphone-captured video. This paper details a method for applying NeRF in real construction sites with data collection, pre-processing, and progress evaluation. In assessing the model's performance, comparisons are made with data from mobile-LiDAR, stand-LiDAR, and BIM. With this research, we suggest potential future studies in applying NeRF models to construction progress monitoring systems.

Keywords -

NeRF; 3D Computer Vision; Deep Learning; Segmentation; Construction Progress Monitoring

1 Introduction

In the field of construction, progress monitoring stands as a essential work ensuring timely and cost-effective project delivery. The advent of advanced AI and deep learning technologies has initiated a new era of innovation in this domain, enabling automated progress monitoring with remarkable accuracy and efficiency. In recent years, AI advancements utilizing computer vision, such as object detection and instance segmentation for construction object recognition, have been increasingly adopted, transforming traditional monitoring techniques with automated, data-driven approaches. Among these advancements, Neural Radiance Fields (NeRF) [1] have emerged as a innovative approach in the field of 3D data processing and visualization. This study introduces NeRF as a deep learning model that excels in synthesizing photo-realistic images by considering light and material properties, rendering images on novel views in construction sites or built environments that closely replicate real-life visuals. The integration of NeRF into construction progress monitoring marks a significant advancement, providing a method that not only improves visual comprehension but also greatly contributes to the automation and precision of tracking construction progress.

For automating construction progress monitoring, the integration of vision sensors and deep learning methods has drastically changed traditional approaches. Beginning with the use of traditional image processing skills like filtering, edge and corner detection to analyze site images [2], the approach has evolved to incorporate deep learning for object detection [3] and segmentation [4]. This advancement significantly improves the accuracy of construction progress assessments from 2D sensors by enabling more precise comparisons of site images with designs derived from Building Information Modeling (BIM).

Furthermore, the progression in construction monitoring has greatly benefited from the adoption of 3D scanning technologies like LiDAR [5], which have revolutionized the field by enabling comprehensive three-dimensional site captures. These methods allow for detailed and precise comparisons between ongoing construction and BIM designs. Advancements such as real-time 3D point cloud mapping with Simultaneous Localization and Mapping (SLAM) [6], further enhance geometry analysis in construction environments. Combining these cutting-edge 3D scanning techniques with AI and deep learning significantly improves the accuracy and efficiency of construction monitoring, setting a new standard in the industry.

While previous studies in automated construction progress monitoring have significantly utilized 2D and 3D sensing technologies for gathering building or construction site data, they commonly entail transforming scanned data into a 2D image with orthogonal view [4, 7, 8]. However,

challenges remain, such as: 1) Achieving efficiency and quality in rendered parallel 2D images. The creation of parallel 2D images from RGB cameras is a detailed, rule-based process requiring manual refinement. Moreover, while LiDAR or SLAM methods often lack the realistic appearance of actual images, resulting in lower quality renderings. 2) The cost and user-convenience of data acquisition. Methods such as SLAM, which utilize robotic or drone sensing, necessitate predefined operational paths. Compared to manual, hand-held capture, these methods are operationally more complex and constrained by environmental factors like limited pathways or airspace, reducing their feasibility in diverse construction environments.

To address the issues of existing 2D and 3D sensor-based methods in automated progress monitoring, we propose an approach utilizing NeRF. This approach utilizes deep learning to achieve a degree of realism in spatial rendering that significantly exceeds the capabilities of traditional methods. A key advantage of our methodology is the use of smartphone-captured video as input. Furthermore, our approach is not limited to rendering the site in 3D; it also generates orthogonal views, which can be directly compared with BIM for accurate construction process monitoring. Our NeRF-based method's ability to generate both realistic 3D renderings and orthogonal views establishes it as a versatile and effective solution for construction progress monitoring. We further enrich our research by testing and comparing various NeRF models—vanilla NeRF [1], Instant-NGP [9], and Nerfacto [10]. Utilizing the Nerfstudio [10] platform, we efficiently train and visualize our models. Our research includes the collection and analysis of data from two different indoor scenes and one outdoor scene, all derived from actual built environments. Our main contributions are:

- We utilize the concept of neural radiance fields (NeRF) to comprehend the 3D spatial information of construction sites and render images from novel views that closely resemble the actual environment.
- We demonstrate the use of a common smartphone camera, easily handheld and maneuvered, to capture videos in a user-friendly and uncomplicated manner. These videos are then used as the input for NeRF model training.
- We evaluate and apply the NeRF model in various built environments, including indoor and outdoor settings, specifically for the purpose of automated construction progress monitoring.

2 Background

2.1 Automated progress monitoring

Computer vision technology has increasingly been applied in automated construction monitoring. Initial approaches involved image processing techniques like edge detection and deep learning-based object segmentation to compare material edges with as-designed BIM [2, 3, 4]. The focus then shifted to LiDAR-based 3D scanning [5], providing detailed site comparisons with BIM, typically evaluated using Root Mean Square Error (RMSE). Advancements continued with SLAM [6], using moving robots capable of capturing diverse scenes, thereby enhancing segmentation and detection for more accurate progress tracking against as-designed BIM. Recently, Pal et al. [8] employed vanilla NeRF [1] to generate orthographic views of under-construction elements, performing semantic segmentation to monitor construction progress in comparison with BIM designs. In this paper, we utilize various NeRF models such as vanilla NeRF, Instant-NGP [9], and Nerfacto [10].

2.2 Neural Radiance Field (NeRF)

Neural Radiance Fields [1], or NeRF, represent a novel approach in the field of 3D scene reconstruction from 2D images. Traditionally, rendering realistic 3D objects involved the use of expensive 3D scanners or photogrammetry that transform images into voxel, point cloud, or mesh forms, NeRF introduces a novel approach in novel view synthesis. In recent, research in the field of 3D computer vision is largely centered around the use of NeRF. First, Vanilla NeRF [1], as the foundational model, utilizes an Multi-Layer Perceptron (MLP) with 8 linear layers, offering a distinctive approach to 3D scene representation. It processes 3D coordinates through positional encoding to enrich the input data, thereby enhancing the details captured in the scene. This architecture extracts density outputs and integrates ray viewing directions, allowing the final RGB output to dynamically reflect how the appearance of objects changes with the viewer's perspective. Building on this, Instant-NGP [9] innovates by encoding coordinates with HashMap and Linear Interpolation to significantly reduce computational load and accelerate training. This approach efficiently creates feature vectors from selected coordinates and auxiliary values, streamlining the MLP processing. Finally, Nerfacto [10] builds upon previous NeRF advancements by combining several techniques for real data capture of static scenes. It integrates camera pose refinement, per-image appearance conditioning, proposal sampling, scene contraction, and hash encoding.

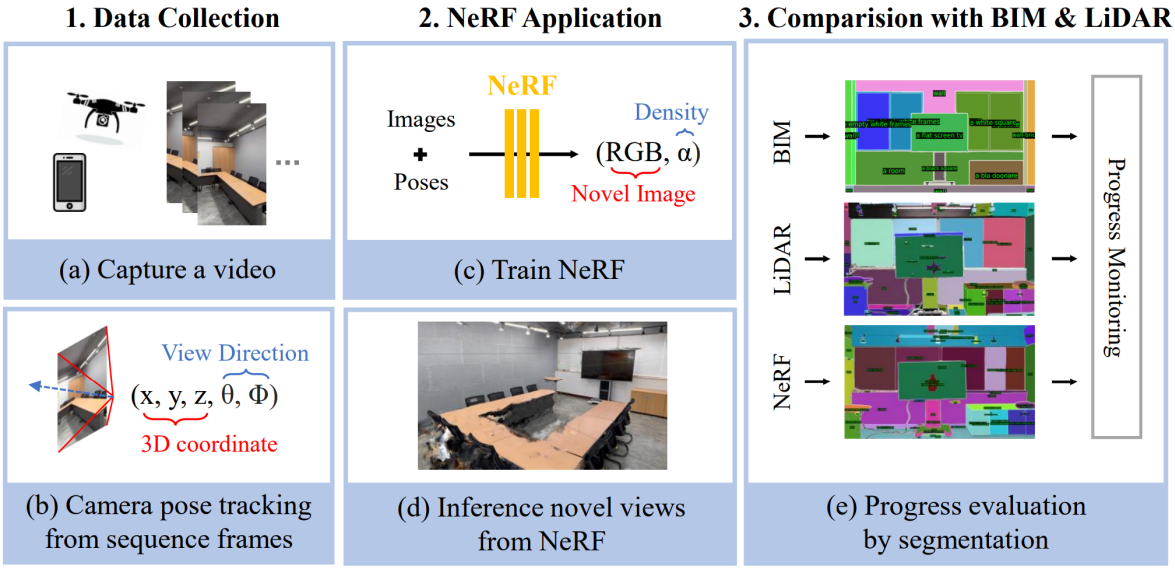


Figure 1. Overview of NeRF-Con pipeline for construction progress monitoring

3 Method

Fig. 1 shows our NeRF-based method for automated construction progress monitoring, starting with data collection. This model is then trained to accurately render photo-realistic 2D images of the site from novel perspectives, aiming for realistic visualizations of the 3D space. The final comparison stage involves aligning NeRF-generated images, potentially orthographic views, with BIM designs, employing instance segmentation for precise progress assessment. This method, leveraging NeRF's strengths in 3D spatial representation and image synthesis, offers a novel, accurate, and efficient approach to quantifying construction progress.

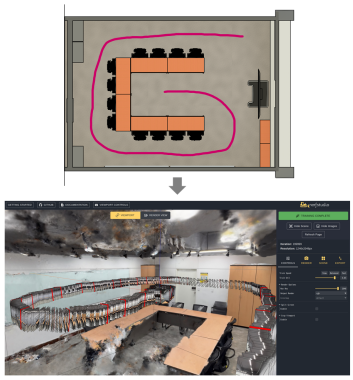


Figure 2. Visualization of the actual move path for capturing a video (above) and the Nerfacto model application on the Nerfstudio platform (below)

3.1 Data collection

For video capture in construction progress monitoring, various methods are employed, such as smartphones, drones, and robots, which can be hand-held or integrated into automated systems. We predominantly utilize the iPhone 15 Pro, chosen for its high accessibility and efficiency, and employ COLMAP [11], a structure-from-motion (SfM) technique, to extract camera poses from image sequences. This approach, as depicted in Fig. 2, involves using SfM, a photogrammetric method, to estimate three-dimensional structures from two-dimensional images. The process identifies key features across images and uses their relative motion to infer depth and structure, with a focus on the epipolar line, which indicates the trajectory of a point in one image across another, based on camera movement. COLMAP processes video frames to generate accurate 3D coordinates and view directions for the camera, constructing a 3D point cloud of the site and determining the camera's position and orientation for each frame. This method maintains the practicality and convenience of data collection, ensuring regular monitoring feasibility across various environments without specialized equipment. The detailed process guarantees a precise representation of the construction site, facilitating high-precision training of the NeRF model, aligning with our goals for efficient and comprehensive construction progress monitoring.

3.2 NeRF application

The fundamental concept of NeRF [1] involves sampling points in a 3D space along rays that emanate from

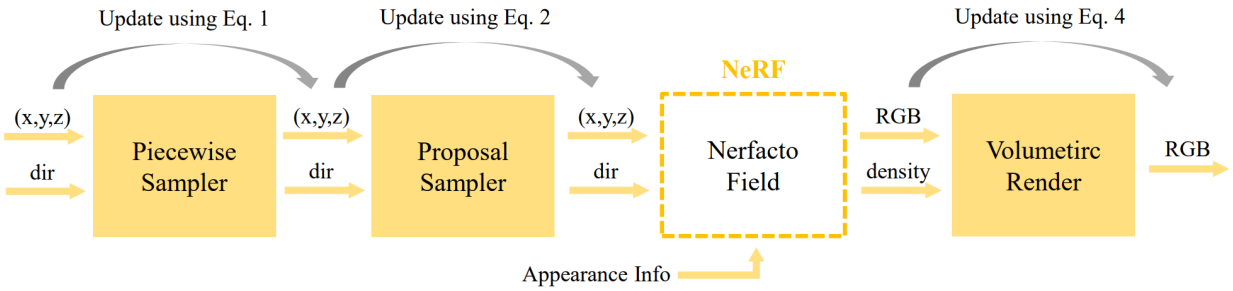


Figure 3. The architecture of the Nerfacto model

the camera's viewpoint. These sampled points are used to estimate both the color and the density at every location through the scene, which are then composited into a final image via volumetric rendering techniques. The input to a NeRF model typically includes the 3D coordinates of sample points, the direction of the viewing rays, and any appearance features that characterize the scene's properties, such as lighting or texture. The output is the rendered 2D image that approximates the real-world scene from the camera's perspective. Fig. 3 shows Nerfacto [10] model's pipeline for NeRF application at construction sites by creating photo-realistic images from captured videos.

3.2.1 Piecewise Sampler

The rendering pipeline begins with the Piecewise Sampler, selecting sample points along camera rays to evaluate the scene. It distributes half of the samples uniformly within a distance of 1 (unit distance) from the camera, ensuring a detailed sampling of nearby areas. The remaining samples are placed with increasing step sizes, effectively scaling the sampling frustums and allowing for a broader coverage that includes distant objects without compromising the sampling density for closer areas. This two-part approach can be expressed as:

$$d_i = \begin{cases} \frac{i}{N/2} & \text{if } i \leq \frac{N}{2}, \\ f(i) & \text{if } i > \frac{N}{2}, \end{cases} \tag{1}$$

where i is the index value of the samples, d_i is the distance from the camera, N is the total sample count, and $f(i)$, a monotonically increasing function, adjusts the samples based on conical frustum.

3.2.2 Proposal Sampler

After the initial sampling phase, the Proposal Sampler, utilizing two density functions, refines the sample locations. Its primary goal is to maximize sampling around surface boundaries, which are crucial for the scene's visual accuracy. These density functions, constituting the Density Field, guide the sampling process. Each density function in the Proposal Sampler is an MLP that receives 3D coordinates as input and is combined with hash encoding. This structure is designed to provide a coarse representation of density, which is crucial for efficient sampling. The density function can be expressed as:

$$\rho(\mathbf{x}) = \Theta_{\text{density}}(\phi(\mathbf{x})) \tag{2}$$

where $\mathbf{x}$ is a spatial coordinate, ϕ is a hash encoder [9], Θ_{density} is MLP for density and $\rho(\mathbf{x})$ is the estimated density at that location. The hash encoding transforms the 3D coordinates into a suitable format for the MLP, enabling it to compute the density. The two density fields in the Proposal Sampler work together to concentrate sample points around significant areas like surface boundaries. The design of these density fields focuses on capturing only a coarse representation of scene density. This approach is sufficient for guiding the sampling process, ensuring that the model concentrates computational resources on the most important aspects of the scene without being burdened by the intricacies of high-frequency details.

3.2.3 Nerfacto Field

The Nerfacto Field is an integral component of the rendering pipeline that takes as input the 3D coordinates $\mathbf{x}$, the view direction $\mathbf{d}$, and the appearance features $\mathbf{f}$, and outputs both the color C and the density $\rho(\mathbf{x})$ at the given spatial location. For extracting the RGB color, the Nerfacto Field employs a neural network function which can be expressed as:

$$C(\mathbf{x}, \mathbf{d}) = \Theta_{\text{RGB}}(\phi(\mathbf{x}), SH(\mathbf{d}), \mathbf{f}) \tag{3}$$

where Θ_{RGB} is MLP for density, SH is the spherical harmonic encoding of the view direction, and the appearance features $\mathbf{f}$ capture the variations in scene appearance such as lighting and material properties. The density $\rho(\mathbf{x})$ is inferred using the same equation to Eq. 2.

3.2.4 Volumetric Render

The last stage in the pipeline is the Volumetric Render, which integrates the density and color information along the rays to form the final rendered image. This integration can be described by the following equation:

$$\mathrm{RGB}_{\mathrm{final}} = \int \rho(\mathbf{x}) \cdot C(\mathbf{x}, \mathbf{d})\, d\mathbf{x}, \tag{4}$$

where the integration is performed along the ray path, accumulating the product of density and color to yield the final pixel color value. The rendered color $\mathrm{RGB}_{\mathrm{final}}$ is then compared to the ground truth image's RGB values, using the L-2 distance as a loss function during the training process. This loss function quantifies the difference between the rendered image and the actual image, guiding the optimization of the network parameters to minimize these discrepancies.

We delve into the application of NeRF for creating 2D novel images. These orthogonal projected images offer a distinctive view of construction sites. Utilizing the NeRF model, we efficiently segment building elements through the semantic-segment-anything [12]. This segmented output is then compared with 2D plane images derived from BIM model, which similarly employ orthogonal projections. By evaluating the segmented outcomes from the NeRF model against those from BIM, we are able to not only gauge construction progress with great precision but also visually confirm the consistency with the original design. This approach provides a layered insight into project development, facilitating a thorough comparison between what was planned and what is being constructed.

4 Experiments

4.1 Dataset

Our study involved three experiments - two indoors and one outdoors - using an iPhone 15 Pro for data collection. In the first indoor experiment, a 90-second site video was captured for NeRF model training, complemented by a 120-second mobile LiDAR (iPhone 15 Pro) scan and a 210-second FARO LiDAR scan, the latter offering higher accuracy but at a significantly higher cost (50x expensive) than mobile LiDAR. The scanning time differences between mobile LiDAR and FARO LiDAR are due to their operational designs. Mobile LiDAR, a handheld device, necessitates manual navigation for comprehensive site imaging, conversely, FARO LiDAR, a stationary system, automates image capture from all directions.

The second indoor experiment used a 120-second video capture and the last outdoor experiment used a 20-second video. Notably, in each experiment, we downsampled the video frames to one-third of the total frames for both training and testing, dividing the data in a 0.9 to 0.1 training-to-testing ratio. This methodology created a diverse dataset, integrating various technologies for a comprehensive assessment of our NeRF-based monitoring system.

4.2 Implementation details & Metrics

We utilize the Nerfstudio [10] platform for train and visualization, and our experiments involved three NeRF models – Nerfacto, instantNGP, and vanilla NeRF – to compare their performance. Common settings across these models included 200k iterations and 4096 for train/test number of rays per batch. For Nerfacto and instantNGP, the optimizer used was Adam with a learning rate of 0.01. In contrast, vanilla NeRF utilized the RAdam optimizer, featuring a lower learning rate of 0.0005. These models were trained on an NVIDIA RTX4090 GPU, using PyTorch version 2.0.1 and CUDA 11.8, ensuring efficient computation and model optimization.

To evaluate the performance of our models, we employed three key metrics: PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure), and LPIPS (Learned Perceptual Image Patch Similarity) [13].

- **PSNR:** Expressed in the logarithmic decibel scale, with values typically ranging from 20 to 30. Higher values indicate better image quality.
- **SSIM:** Values range between -1 and 1. A value of 1 indicates perfect similarity between the test image and the reference image. SSIM assesses visual impacts based on luminance, contrast, and structure, aligning more with human visual perception than PSNR.
- **LPIPS:** Scores typically range from 0 to 1, where a lower score indicates greater perceptual similarity between compared images. Unlike PSNR and SSIM, LPIPS leverages deep learning models to better approximate human visual perception.

4.3 Results

In our research, we conducted a comparison using two different LiDAR sensors with NeRF-based approaches. One of the LiDAR sensors is a mobile LiDAR embedded in the iPhone 15 Pro, utilizing Pix4Dcatch for analysis. The other is a stationary Faro LiDAR, known for its exceptional precision and high cost. In contrast, for our NeRF-based approaches, we used a smartphone or drone equipped with only a RGB camera. Thus, we experimented with three different NeRF models' rendering image quality (Tab. 1) related to creating the parallel view images and performed an additional comparison between NeRF, stable LiDAR, and BIM with SAM [12] (Fig. 5). We also visualized the infeasible result from the mobile LiDAR (Fig. 6).

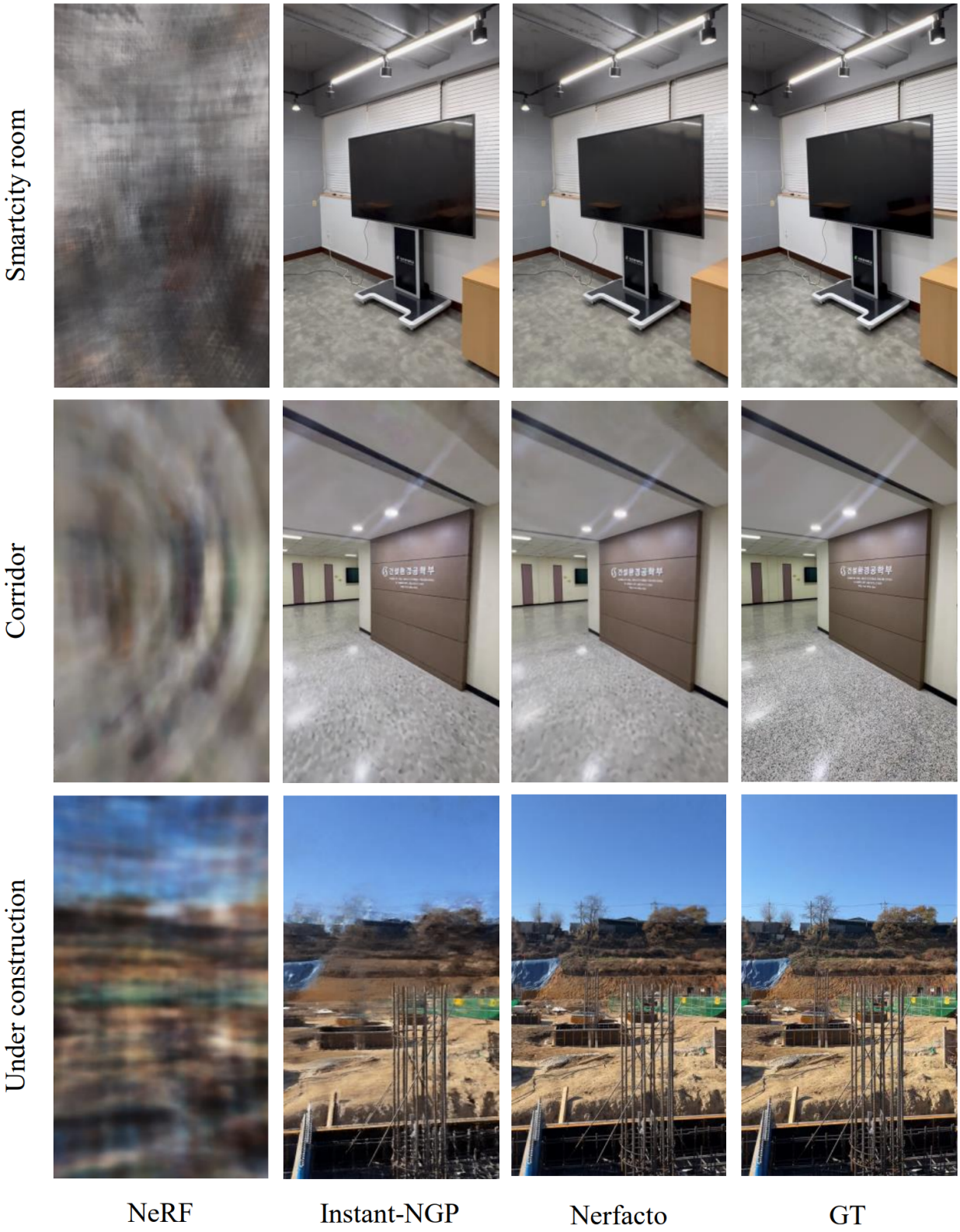
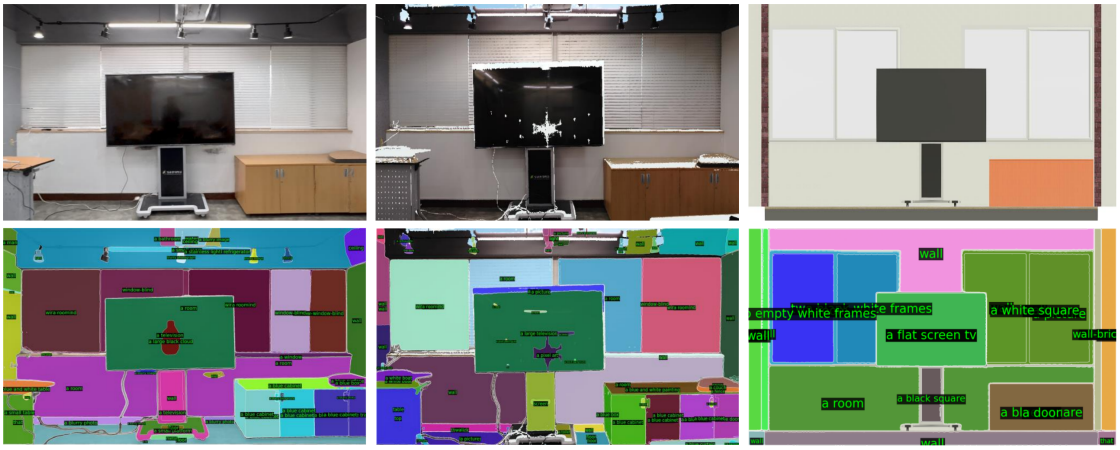


Figure 4. Qualitative comparison of three NeRF models on three different scenes



NeRF LiDAR BIM

Figure 5. Visualization of semantic segmented results on 2D orthogonal images from three different sources

Method	PSNR ↑	SSIM ↑	LPIPS ↓
exp#1 : smartcity room (indoor)			
NeRF	10.8	0.68	0.64
Instant-NGP	30.7	0.91	0.20
Nerfacto	**31.0**	**0.92**	**0.19**
exp#2 : corridor (indoor)			
NeRF	14.7	0.63	0.71
Instant-NGP	**24.8**	**0.74**	**0.35**
Nerfacto	24.4	0.73	0.37
exp#3 : under construction (outdoor)			
NeRF	11.4	0.43	0.93
Instant-NGP	16.8	0.56	0.38
Nerfacto	**19.1**	**0.57**	**0.22**

Table 1. Quantitative comparison of three NeRF models on three different scenes

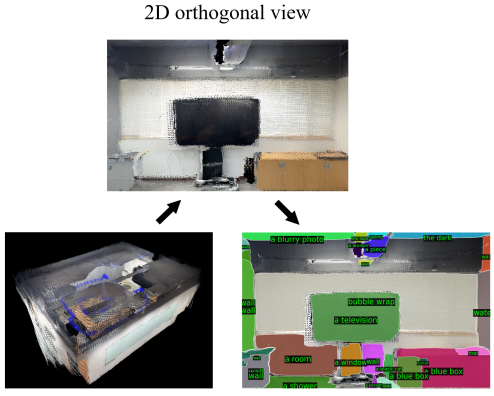


Figure 6. Visualization of semantic segmented results on 2D orthogonal images from mobile LiDAR

We trained three distinct NeRF models - vanilla NeRF, Instant-NGP, and Nerfacto - in diverse environments: a smartcity room (indoor), a corridor (indoor), and an under construction site (outdoor). The results, detailed in Tab. 1, exhibit a notable trend as the spatial scale increases from a confined room to a more expansive corridor and then to an open outdoor space, there's a discernible decrease in model accuracy, as reflected by metrics such as PSNR, SSIM, and LPIPS. This pattern suggests that the complexity and size of the environment negatively impact the rendering quality of these models. In particular, the outdoor scene (exp#3) highlighted the strengths of the Nerfacto model. It achieved a PSNR of 19.1 and an LPIPS of 0.22, surpassing the Instant-NGP model, which managed a PSNR of 16.8 and an LPIPS of 0.38. Furthermore, the qualitative visual results in Fig. 4 corroborate this finding, showing that in the construction site scene of exp#3, Nerfacto outperforms Instant-NGP, providing relatively superior visual quality.

Fig. 5 presents a comparison of parallel view images from different sources. The NeRF image was obtained from the Nerfacto model in exp#1, alongside 2D views acquired from stand-LiDAR and BIM models. These images were further processed using the SAM model. This comparison highlights the practical utility of applying segmentation to NeRF-generated images, considering the higher cost and lower usability of stable LiDAR sensor. We also tested scanned images from mobile LiDAR shown in Fig. 6, but the resolution of images acquired from mobile-LiDAR are inferior compared to stand LiDAR or NeRF. Therefore, we can use smartphones to easily capture visual information at construction sites and then utilize NeRF and SAM models to visualize the level of progress.

5 Conclusion

In conclusion, this study has successfully demonstrated the efficacy of Neural Radiance Fields (NeRF) in automating construction progress monitoring, marking a significant leap over traditional methods. By leveraging NeRFCon, we have shown that it is possible to generate photorealistic, 3D rendered images from simple smartphone-captured videos, offering a more efficient, accurate, and cost-effective solution compared to existing 2D and 3D sensor-based methods. The application of NeRF in various environments - small room, corridor, construction site - proves the robustness and versatility. The integration with segmentation models to compare these renders with BIM designs, ensuring more precise and automated monitoring of construction progress. In conclusion, our proposed methods, NeRF-based rendering and SAM-based comparison with BIM, can enable more efficient project planning and facilitate communication among construction site stakeholders.

In future work, we aim to address two main challenges: the decline in NeRF model's rendering accuracy with increased spatial scale, especially outdoors, and the current reliance on only qualitative SAM result images for progress monitoring. Our focus will be on optimizing NeRF's application for large outdoor sites and developing quantitative assessment methods, such as completion percentages, to enhance automated progress monitoring.

Acknowledgment

This research was supported by a grant [2022-MOIS38-002 (RS-2022-ND630021)] from the Ministry of Interior and Safety (MOIS)'s project, a grant from the Korean Government (MSIT) to the NRF [RS-2023-00250166] and Korea Ministry of Land, Infrastructure and Transport(MOLIT) as Innovative Talent Education Program for Smart City.

References

[1] Srinivasan P. P. Tancik M. Barron J. T. Ramamoorthi R. Ng R. Mildenhall, B. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. doi:https://doi.org/10.1145/3503250.

[2] Kim B. Kim H. Kim, C. 4d cad model updating using image processing-based construction progress monitoring. *Automation in Construction*, 35:44–52, 2013. doi:https://doi.org/10.1016/j.autcon.2013.03.005.

[3] Bienvenido-Huertas D. Carretero-Ayuso M. J. Della Torre S. Marín-García, D. Deep learning model for automated detection of efflorescence and its possible treatment in images of brick facades. *Automation in Construction*, 145:104658, 2023. doi:https://doi.org/10.1016/j.autcon.2022.104658.

[4] Lee-S. Ying, H. Q. A mask r-cnn based approach to automatically construct as-is ifc bim objects from digital images. In *ISARC*, pages 764–771, 2019. doi:https://doi.org/10.22260/ISARC2019/0103.

[5] Turkan Y. Puri, N. Bridge construction progress monitoring using lidar and 4d design models. *Automation in Construction*, 109:102961, 2020. doi:https://doi.org/10.1016/j.autcon.2019.102961.

[6] Chen J.-Cho Y. K. Kim, P. Slam-driven robotic mapping and registration of 3d point clouds. *Automation in Construction*, 89:38–48, 2018. doi:https://doi.org/10.1016/j.autcon.2018.01.009.

[7] Bosché F.-Lu C. X.-Wilson L. Li, J. Occlusion-free orthophoto generation for building roofs using uav photogrammetric reconstruction and digital twin data. In *ISARC*, pages 371–378, 2023. doi:https://doi.org/10.22260/ISARC2023/0051.

[8] Lin J. J. Hsieh S. H. Golparvar-Fard M. Pal, A. Activity-level construction progress monitoring through semantic segmentation of 3d-informed orthographic images. *Automation in Construction*, 157:105157, 2024. doi:https://doi.org/10.1016/j.autcon.2023.105157.

[9] Evans A. Schied C. Keller A. Müller, T. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. doi:https://doi.org/10.1145/3528223.3530127.

[10] Weber E. Ng E. Li R. Yi-B. Wang T. ... Kanazawa A. Tancik, M. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH*, pages 1–12, 2023. doi:https://doi.org/10.1145/3588432.3591516.

[11] Frahm J. M. Schonberger, J. L. Structure-from-motion revisited. In *CVPR*, 2016. doi:https://doi.org/10.1109/CVPR.2016.445.

[12] Arfeto B. E. Zhang C. Shin H. Tariq, S. Segment anything meets semantic communication. *arXiv preprint*, 2023. doi:https://doi.org/10.48550/arXiv.2304.02643.

[13] Isola P. Efros A. A. Shechtman E.-Wang O. Zhang, R. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. doi:https://doi.org/10.1109/CVPR.2018.00068.