Scaffold Component Recognition and Joint-Level Understanding using 3D Gaussian Splatting and Vision-Language Models

Boyu Wang[1] and Borja García de Soto[1]

[1]S.M.A.R.T. Construction Research Group, Division of Engineering, New York University Abu Dhabi (NYUAD), Experimental Research Building, Saadiyat Island, P.O. Box 129188, Abu Dhabi, United Arab Emirates
bw3005@nyu.edu, garcia.de.soto@nyu.edu

Abstract

Scaffold inspection is critical for ensuring construction safety. However, scaffolds are typically dense, multilayered, and continuous structures, making them difficult to analyze automatically using traditional geometric methods. In this paper, a framework for automated scaffold component recognition and joint-level understanding is proposed by leveraging Vision–Language Models (VLMs) and 3D Gaussian Splatting (3DGS) to interpret complex scaffold assemblies. Domain knowledge is incorporated to decompose the structure using 3DGS, where photorealistic rendering is exploited to generate clear visualizations for semantic interpretation by pretrained VLMs. The semantic understanding of scaffold components lays a crucial foundation for future compliance-oriented analysis by enabling the precise identification and spatial reasoning of code-relevant and design-relevant scaffold components. This approach is expected to provide a reference for analyzing other continuous and complex structures in construction scenarios.

Keywords –

Scaffold; Vision-Language Model (VLM); 3D Gaussian Splatting (3DGS); Component Recognition

1 Introduction

Construction safety remains a critical concern worldwide, with falls from height consistently ranking among the leading causes of construction-related fatalities [1]. Temporary structures such as scaffolds are major contributors to these incidents, as their frequent erection and dismantling cycles significantly increase the risk of structural failure [2]. To mitigate these risks, safety regulations require that scaffolds be inspected regularly for compliance with both construction codes and design specifications throughout their lifecycle.

As shown in Figure 1, scaffold inspection tasks can be broadly categorized into two parts. The first is code compliance checking, which verifies whether key components meet regulatory requirements in terms of installation and dimensions. For example, whether base plates are properly placed, whether the vertical distance between adjacent guard rails is within the 600 mm limit, and whether toe board heights exceed the minimum of 90 mm, as mandated by standard scaffolding safety regulations [3]. The second is design compliance checking, which ensures that the scaffold has been assembled exactly as planned including checking for missing elements such as bracings, and verifying whether couplers match the design specifications.

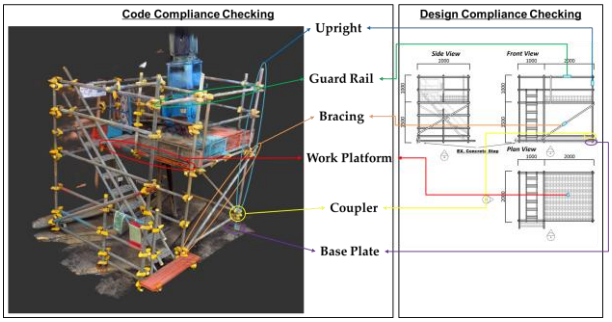

Figure 1. Conceptual workflow of scaffold component classification and compliance checking against code and design specifications. (Created by the authors based on the information provided by onsite engineers).

However, current inspection practices are still predominantly manual, relying heavily on the experience of safety managers and site supervisors. These methods are time-consuming, labor-intensive, and prone to human error, particularly in dense, multilayered scaffold systems. Although recent research has explored the use of computer vision for scaffold analysis, most existing approaches are limited to the geometric interpretation of textureless 3D point clouds [4-7]. These methods often

lack the photorealistic fidelity and semantic richness needed for detailed component recognition and are difficult to scale for complex real-world scenarios.

To address these limitations, this paper presents a framework for automated as-built scaffold understanding by integrating 3D Gaussian Splatting (3DGS) with Vision-Language Models (VLMs). Specifically, unlike sparse and textureless point clouds, 3DGS reconstructs 3D scenes with high photorealistic fidelity, preserving both accurate geometry and rich visual appearance (color, texture) to provide a high-quality foundation for downstream analysis [8-10]. Concurrently, VLMs solve the generalizability limitations of traditional deep learning methods by leveraging the photorealistic 2D renderings generated by 3DGS to perform flexible, prompt-based semantic segmentation without the need for extensive task-specific training [11-13]. This bypasses the need for extensive task-specific 3D training data and enables human-interpretable component recognition. Together, this integration serves as a robust first step toward enabling automated component recognition and compliance checking workflows.

To operationalize this integrated framework, the proposed pipeline follows a systematic workflow. Starting from site-captured point clouds and images, a photorealistic 3DGS model is constructed to represent the as-built scaffold with high visual and geometric fidelity. The structure is then decomposed into semantically meaningful subcomponents via 3DGS-based editing. Pretrained VLMs are applied to these substructures to extract and interpret semantic information, enabling fine-grained component segmentation and classification. Furthermore, critical joints can be examined in detail by adjusting viewpoints and zooming within the 3DGS model, supporting close-up inspection and providing the necessary information for compliance verification.

2 Literature Review

2.1 Scaffold Inspection

Previous research on scaffold inspection has primarily focused on two major directions. The first line of work aims to monitor scaffold deformation, which is crucial for early detection of structural instability. Two main approaches have been explored: sensor-based methods, which use strain gauges or other physical sensors to measure deformation directly [14,15], and vision-based methods, which infer deformation by analysing pixel-level changes across image sequences [16,17]. While the latter can generate heatmaps to visualize deformation trends, they often lack fine-grained, component-level information.

The second line of research focuses on semantic segmentation of scaffolds into discrete components such as uprights, guardrails, bracing, work platforms, stairs, and toe boards. These methods typically operate on 3D point cloud data, attempting to segment the scaffold into meaningful elements [4,7,18]. However, they often rely heavily on manually defined geometric rules, which are sensitive to noise and make the process unstable and difficult to generalize for large-scale scaffold inspection. While some deep learning–based methods have been proposed to improve robustness and generality, they usually require extensive annotated training data. Moreover, due to the black-box nature of deep models, it is often difficult to diagnose or fine-tune the system when performance is suboptimal in complex real-world conditions.

2.2 3D Gaussian Splatting

To remove the ambiguity of point cloud analysis–based component extraction, 3DGS provides a highly suitable solution. 3DGS is an emerging technique for real-time radiance field rendering, offering a compelling alternative to traditional mesh or point cloud based scene representations [9]. 3DGS models a scene as a collection of anisotropic 3D Gaussians, each parameterized by position, scale, orientation, color, and opacity. These Gaussians are efficiently splatted onto the image plane to generate photorealistic renderings from arbitrary viewpoints at interactive frame rates.

Recent studies have demonstrated the capability of 3DGS to capture fine surface details and complex lighting with high fidelity, which is especially valuable for detailed visual inspection in unstructured and cluttered environments. Compared to sparse and noisy point clouds, 3DGS produces a dense, view-dependent scene representation that preserves both geometric and appearance cues. This enables users to zoom into specific components, change viewpoints, and visually isolate substructures, which is essential for accurate and interpretable inspection workflows. Although 3DGS has been primarily applied in computer graphics and immersive rendering, its potential for component-level analysis in the Architecture, Engineering, and Construction (AEC) industry has not been fully explored. This paper leverages 3DGS to enhance semantic decomposition and inspection of as-built scaffold structures.

2.3 Vision-Language Model

While 3DGS enables photorealistic visualization of as-built scenes, VLMs provide potential solution to the semantic understanding of image content. VLMs have recently emerged as powerful tools for multimodal reasoning by jointly learning from both visual and textual data. These models, often pre-trained on large-scale

image text pairs, are capable of associating natural language descriptions with visual content, enabling tasks such as image captioning, visual question answering, and semantic segmentation through textual prompts. Recent advances, including CLIP [19], BLIP [20], and GPT-4V [21], have demonstrated strong generalization across domains, even with zero-shot or few-shot inputs. Such capabilities make VLMs especially promising for domains with limited labeled data or where human-like semantic understanding is needed [22-24].

In the context of construction automation, VLMs offer a new paradigm for interpreting complex scenes in a flexible and human-interpretable way. Unlike traditional vision models that require task-specific training datasets, VLMs can be prompted using natural language queries to identify or describe components, attributes, or spatial relationships. Despite their success in general vision tasks, the application of VLMs in 3D construction environments remains underexplored, and this paper aims to bridge that gap through an integration with 3DGS.

3 Methodology

This paper presents a method for as-built scaffold understanding by integrating 3DGS with VLMs, providing a foundation for downstream compliance inspection tasks. As illustrated in Figure 2, the proposed workflow consists of five main stages: (1) Mobile laser scanning is first employed to capture high-resolution point clouds and images of the as-built scaffold. (2) These multi-modal data are fused to generate a photorealistic 3DGS scene representation that preserves both geometric structure and visual appearance. (3) A 3DGS-based structure decomposition process is then applied to isolate scaffold substructures based on spatial layout. (4) Each substructure is further segmented through VLM-based semantic analysis, where pretrained VLMs interpret rendered 3DGS views to assign component-level labels (e.g., upright, bracing, guardrail). (5) Finally, a component-wise graph is constructed to encode the topological relationships among scaffold elements, enabling joint-level analysis that is relevant for code compliance and design conformance.

.

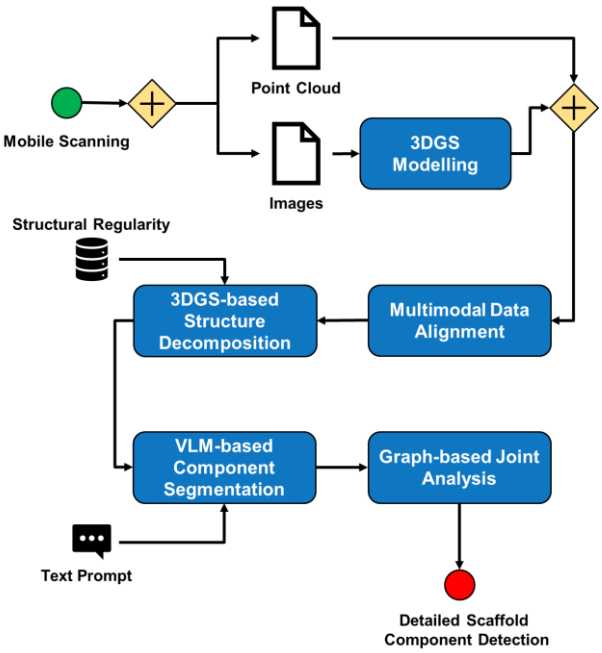

Figure 2. Pipeline for 3DGS-based, VLM-enabled understanding of as-built scaffold data.

3.1 Mobile Scanning

To capture as-built conditions of scaffold structures, a mobile scanning system was employed, as shown in Figure 3. The data acquisition platform integrates a LiDAR sensor and two high-resolution RGB cameras, enabling synchronous collection of both dense point clouds and photorealistic images. The scanning process is conducted manually by an operator walking around the scaffold with the handheld device, ensuring flexible access to complex and confined construction environments. The system adopts Fast-LIVO2 [25] as its Simultaneous Localization and Mapping (SLAM) backend, which provides robust real-time pose estimation and trajectory optimization. This allows for accurate spatial registration of the captured LiDAR frames and camera images, resulting in a globally consistent 3D representation of the scene. The acquired data serve as the foundation for 3DGS and subsequent semantic analysis in the pipeline.

Figure 3. Mobile scanning for data collection

3.2 3DGS Modelling

To obtain a high-fidelity and interpretable representation of the as-built scaffold structure, 3DGS is employed as the core modeling technique. Rather than relying on explicit mesh reconstruction or discrete point clouds, the scene was modeled as a radiance field parameterized by a set of volumetric 3D Gaussians. This representation enables photorealistic rendering from arbitrary viewpoints and provides strong support for visual inspection and downstream analysis.

The modeling process was initialized using a sparse Structure-from-Motion (SfM) reconstruction, from which camera parameters and initial 3D point distributions were estimated to guide the placement of Gaussians. An iterative optimization was then performed to refine each Gaussian's parameters including position, scale, opacity, and appearance (encoded via spherical harmonics) by minimizing a photometric loss between rendered and observed images. A visibility-aware renderer was used to simulate training views and guide updates. To ensure compactness and sufficient coverage, a density control mechanism was applied, dynamically adding or pruning Gaussians based on gradient feedback.

As shown in the left part of Figure 4, this process transforms a large set of unstructured images into a coherent 3D model that captures both photorealistic appearance and explicit spatial structure. Unlike traditional image-based methods, the resulting 3DGS model encodes geometric relationships in 3D space, allowing the scene to be interpreted as a spatially organized structure. This spatial coherence provides a robust foundation for structural decomposition, where scaffold components can be systematically separated based on their relative positions, orientations, and connectivity, enabling higher-level reasoning about component grouping and configuration in a data-driven yet interpretable manner.

3.3 Multimodal Data Alignment

Since the 3DGS reconstruction does not preserve an accurate global scale, SLAM-based point clouds were incorporated to leverage their precise geometric information. To integrate the SLAM point cloud with the 3DGS model, a registration step was performed. As the two representations exist in different coordinate frames, a two-stage alignment process was adopted. First, a coarse transformation was estimated using a three-point alignment based on corresponding anchor points identified in both datasets. The alignment was then refined using the Iterative Closest Point (ICP) algorithm to minimize geometric discrepancies. As a result, the photorealistic 3DGS model was accurately aligned with the globally referenced SLAM point cloud, as shown in Figure 4, enabling consistent spatial reasoning across subsequent processing stages.

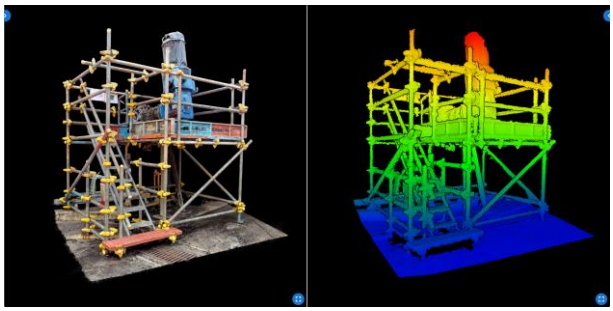

Figure 4. Geometrically aligned 3DGS (left) and point clouds (right).

3.4 3DGS-based Structure Decomposition

To simplify the analysis of complex scaffold assemblies, structure decomposition was performed directly on the 3DGS model. Based on the global coordinate system established during registration, the scaffold was segmented into meaningful substructures according to their geometric locations and spatial hierarchy. Specifically, spatial filters were applied along the X, Y, and Z axes to crop out regions corresponding to the front face, side frames, and intermediate bays. This axis-aligned cropping strategy leverages the inherent regularity of scaffold structures, which are typically aligned with global Cartesian axes. It enables clean separation of components while preserving substructure integrity, offering a simple and interpretable approach suitable for early-stage inspection workflows. By iteratively applying these geometric cropping operations, the continuous and highly layered scaffold structure was decomposed into a set of clear, modular components, each representing a distinct structural unit. As illustrated

in the left part of Figure 5, the front face has been isolated, enabling precise VLM-based analysis and component segmentation in subsequent stages.

Figure 5. 3DGS-based structure decomposition and VLM-enabled component recognition. The left image shows the cropped frontal view of the scaffold's 3DGS model; the right image illustrates the segmentation results using the text prompt “bar”.

3.5 VLM-Based Component Segmentation

To achieve semantic understanding of scaffold components, a vision-language model (VLM) was employed to segment 3DGS-rendered views based on natural language prompts. Specifically, SAM3 [12] is adopted, a recent model designed for promptable concept segmentation, which supports segmentation of all instances matching a concept prompt across images and videos. SAM 3 accepts short noun phrases or image exemplars as prompts and returns corresponding segmentation masks along with object identities. Prompt-based segmentation was adopted to minimize manual labeling and avoid rigid class definitions, enabling flexible and interpretable semantic analysis. In the proposed pipeline, one rendering of the 3DGS model were generated for each substructure from calibrated camera poses, and each image was prompted with component-level concepts such as “bar”, “stair”, and “board”. The SAM 3 model uses a DETR-based detector and a mask decoder, conditioned on both text and visual prompts, to produce high-quality masks. A presence token is used to decouple recognition and localization, improving segmentation accuracy in cluttered scenes. The segmentation results for the front side, obtained using the text prompt “bar,” are presented in Figure 5.

The resulting 2D masks were projected back to the 3D space by leveraging the 2D–3D correspondences inherent in the 3DGS rendering process. Specifically, each 2D pixel in the segmentation mask was mapped to its nearest 3D point clouds by reprojecting the camera ray and using visibility-aware associations. In this way, the semantic labels predicted by the VLM were transferred to the corresponding subset of 3D point clouds, effectively labeling the scaffold model at the point level. By aggregating segmentation results across multiple views, a consistent 3D semantic segmentation was obtained as shown in the left part of Figure 6, forming the basis for subsequent component graph construction and compliance analysis.

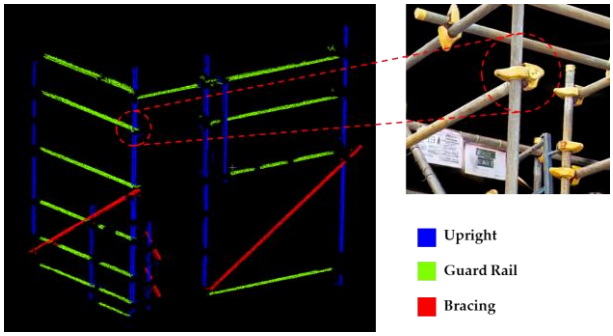

Figure 6. Scaffold component segmentation and corresponding as-built details. Left: segmented scaffold structure showing uprights (blue), guard rails (green), and bracings (red). Right: 3DGS of selected joints located at the intersection between a guard rail and an upright.

3.6 Graph-based Joint Analysis

Following instance segmentation, each scaffold component is represented as a labeled subset of 3D points corresponding to structural elements such as uprights, guard rails, and bracings. To capture the structural connectivity among these components, a spatial graph is constructed based on proximity-based interaction. Specifically, pairwise distances between segmented components are computed, and potential joints are inferred wherever two elements fall within a predefined spatial threshold. These candidate connection points serve as the edges in a component-level graph, while the scaffold members themselves define the nodes, forming a graph-based abstraction of the scaffold topology.

Once the graph is established, the precise 3D positions of joint candidates are localized using the high-fidelity 3DGS model. Since 3DGS preserves photorealistic appearance and scene geometry, rendered views of the detected joints allow for visual verification and detailed analysis of couplers and base plates. As illustrated in Figure 6, this enables the system to infer the presence of joints from geometry, providing detailed visual cues that assist in inspecting their physical condition and evaluating compliance with structural regulations. By combining geometric reasoning with appearance-based cues, this graph-based framework supports joint-level inspection in a scalable and interpretable manner. While a full-scale quantitative evaluation of the spatial graph and automated compliance checking are reserved for future work, this graph-based

abstraction serves as a crucial topological bridge. Currently, it is utilized to localize candidate joints, directly guiding the targeted detection and visual inspection of critical connection elements like couplers (as evaluated in Section 4). Thus, this joint-level understanding provides a necessary foundation for downstream analyses.

4 Experiments and Results

To evaluate the effectiveness of the proposed pipeline, experiments were conducted on real-world scaffold scenes captured via mobile laser scanning. The evaluation focused on the object detection accuracy of individual scaffold components, assessed by comparing the predicted labels to manual annotations. The tested scene serves as a representative real-world case intended to validate the feasibility of the proposed pipeline, rather than to provide a comprehensive performance benchmark. Therefore, the current findings should be considered preliminary. Broader validation across multiple sites and scaffold configurations is left for future work. Table 1 summarizes the detection performance across five key component categories (uprights, guard rails, braces, couplers, and base plates) using standard evaluation metrics: true positives (TP), false positives (FP), false negatives (FN), precision (Pre), recall (Rec), and F1-score (F1).

Table 1 Results of scaffold component detection

Component Type	TP	FP	FN	Rec.	Pre.	F1
upright	7	0	0	1.00	1.00	1.00
guard rail	11	0	0	1.00	1.00	1.00
bracing	4	0	0	1.00	1.00	1.00
coupler	28	0	2	1.00	0.93	0.96
base plate	4	0	0	1.00	1.00	1.00

A total of 56 components were annotated and used as ground truth. The results show that the proposed method achieves high precision across the evaluated categories in this specific scene, with nearly zero false negatives recorded among the 56 annotated components. Notably, the detection of couplers (achieving 28 true positives) and the joint-level visualization shown in Figure 6 (Right) serve as a validation of the graph-based joint analysis proposed in Section 3.6. By successfully localizing the intersections among linear structural elements through the spatial graph, the system was able to accurately isolate and identify the corresponding couplers. This demonstrates the practical utility of the graph-based abstraction in supporting detailed component inspection, even though a few false negatives occurred due to occlusions and limited visibility in the scanned data. While these initial results demonstrate the fundamental feasibility of the proposed 3DGS and VLM-based pipeline for component recognition, we acknowledge that this small-scale evaluation does not fully capture the complexities of diverse construction sites.

Although a direct quantitative comparison is challenging due to the absence of standardized public datasets for scaffold inspection, contrasting the proposed approach with existing literature underscores its novelty and practical feasibility. Traditional methods predominantly rely on fully supervised 3D point cloud segmentation networks (e.g., PointNet or RandLA-Net). In the broader computer vision domain, these general frameworks typically achieve mean Intersection over Union (mIoU) of around 80% on standard point cloud benchmarks [26]. Specifically for scaffolding, recent studies report mIoU scores ranging from 80% to 87% [4]. However, these approaches suffer from two major limitations: first, they are restricted to semantic segmentation and cannot provide the object-level (instance) representations essential for detailed compliance checking; second, they necessitate hundreds of manually annotated 3D frames for training. In contrast, our proposed 3DGS and VLM-based framework achieves highly competitive object-level recognition, yielding higher recall and precision on major components in our test case without requiring any task-specific 3D training data. The core novelty lies in synergizing the high-fidelity rendering of 3DGS with the zero-shot, open-vocabulary capabilities of VLMs. This paradigm fundamentally bypasses the data preparation bottleneck inherent in traditional point cloud methods, demonstrating strong potential for rapid, scalable deployment in dynamic construction environments.

Overall, the integration of 3DGS modeling and VLM-based segmentation enables high-quality detection for the tested component categories, demonstrating promising potential for automated inspection workflows.

5 Conclusion

This paper presents an automated pipeline for scaffold component recognition and joint-level understanding by integrating mobile laser scanning, 3DGS, and VLMs. Through precise 3D reconstruction and prompt-based segmentation, the system effectively identifies and labels structural elements such as uprights, guard rails, and couplers. A graph-based analysis further enables joint detection and spatial reasoning. Experimental results demonstrate high accuracy in component detection, with zero false positives and minimal false negatives.

While the current evaluation is based on a single representative scene with a limited number of components, the results confirm the initial feasibility of the approach as a proof-of-concept. However, several

critical limitations remain. The small dataset size does not sufficiently demonstrate the framework's generalizability across diverse scaffold configurations. Furthermore, the system's robustness in highly cluttered construction environments, where severe occlusions, varying lighting conditions, and dense overlapping structures are common, requires further investigation. These aspects, along with sensitivity to scan quality and prompt phrasing, will be addressed in future work through large-scale data collection and comprehensive benchmarking across multiple complex sites.

Overall, although a complete automated compliance checking system remains a subject for future work, the proposed approach successfully lays the essential foundation for such systems by providing reliable component recognition and joint-level topological understanding capabilities.

Acknowledgement

This research was partially supported by the Center for Sand Hazards and Opportunities for Resilience, Energy, and Sustainability (SHORES), funded by Tamkeen under the NYUAD Research Institute Award CG013.

References

[1] S. Tan, Speech at SCAL Construction Safety, Health and Security Seminar, (25 March 2026) (2017)

[2] S.M. Whitaker, R.J. Graves, M. James, P. McCann, Safety with access scaffolds: Development of a prototype decision aid based on accident analysis, Journal of Safety Research, 34(3) (2003), pp. 249-261

[3] Workplace Safety and Health (Scaffolds) Regulations, Ministry of Manpower, Singapore, (2011)

[4] J. Kim, J. Kim, N. Koo, H. Kim, Automating scaffold safety inspections using semantic analysis of 3D point clouds, Automation in Construction, 166 (2024), pp. 105603

[5] J. Kim, J. Kim, Y. Kim, H. Kim, 3D reconstruction of large-scale scaffolds with synthetic data generation and an upsampling adversarial network, Automation in Construction, 156 (2023), pp. 105108

[6] J. Kim, D. Chung, Y. Kim, H. Kim, Deep learning-based 3D reconstruction of scaffolds using a robot dog, Automation in Construction, 134 (2022), pp. 104092

[7] Q. Wang, Automatic checks from 3D point cloud data for safety regulation compliance for scaffold work platforms, Automation in Construction, 104 (2019), pp. 38-51

[8] F. Luan, S. Sun, H. Zhang, Y. Yin, K. Wang, J. Yang, 3D Gaussian splatting technologies and extensions: A review, Neurocomputing, (2025), pp. 131629

[9] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3D Gaussian splatting for real-time radiance field rendering, ACM Transactions on Graphics, 42(4) (2023), pp. 139:131-139:114

[10] T. Lu, M. Yu, L. Xu, Y. Xiangli, L. Wang, D. Lin, B. Dai, Scaffold-gs: Structured 3d gaussians for view-adaptive rendering, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 20654-20664

[11] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, Grounding DINO: Marrying dino with grounded pre-training for open-set object detection, Arxiv Preprint, (2023)

[12] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K.V. Alwala, H. Khedr, A. Huang, Sam 3: Segment anything with concepts, arXiv preprint arXiv:.16719, (2025)

[13] L.H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, Grounded language-image pre-training, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10965-10975

[14] S. Sakhakarmi, J. Park, C. Cho, Enhanced machine learning classification accuracy for scaffolding safety using increased features, Journal of Construction Engineering Management, 145(2) (2019), pp. 04018133

[15] C. Cho, K. Kim, J. Park, Y.K. Cho, Data-driven monitoring system for preventing the collapse of scaffolding structures, Journal of Construction Engineering Management, 144(8) (2018), pp. 04018077

[16] Y. Jung, H. Oh, M.M. Jeong, An approach to automated detection of structural failure using chronological image analysis in temporary structures, International Journal of Construction Management, 19(2) (2019), pp. 178-185

[17] Y. Feng, F. Dai, H.-H. Zhu, Evaluation of feature-and pixel-based methods for deflection measurements in temporary structure monitoring, Journal of Civil Structural Health Monitoring, 5(5) (2015), pp. 615-628

[18] R. Luo, Z. Zhou, X. Chu, W. Ma, J. Meng, 3D deformation monitoring method for temporary structures based on multi-thread LiDAR point cloud, Measurement, 200 (2022), pp. 111545

[19] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, Learning transferable visual models from natural language supervision, International Conference on Machine Learning, 2021, pp. 8748-8763

[20] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, International Conference on Machine Learning, 2023, pp. 19730-19742

[21] OpenAI, GPT-4 technical report, (2023)

[22] B. Wang, Z. Chen, M. Li, Q. Wang, C. Yin, J.C. Cheng, Omni-Scan2BIM: A ready-to-use Scan2BIM approach based on vision foundation models for MEP scenes, Automation in Construction, 162 (2024), pp. 105384

[23] B. Wang, F. Lin, M. Li, Z. Liang, Z. Chen, M. Wang, J.C. Cheng, Informative As-Built Modeling as a Foundation for Digital Twins Based on Fine-Grained Object Recognition and Object-Aware Scan-vs-BIM for MEP Scenes, Advanced Engineering Informatics, 65 (2025), pp. 103382

[24] B. Wang, F. Lin, M. Li, Z. Liang, H. Yue, Q. Wang, J.C. Cheng, Aligning as-built and as-designed: Local point cloud to BIM registration via hybrid visibility map encoding for construction digital twins, Automation in Construction, 180 (2025), pp. 106551

[25] C. Zheng, W. Xu, Z. Zou, T. Hua, C. Yuan, D. He, B. Zhou, Z. Liu, J. Lin, F. Zhu, Fast-livo2: Fast, direct lidar-inertial-visual odometry, IEEE Transactions on Robotics, (2024),

[26] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, H. Zhao, Point transformer v3: Simpler faster stronger, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4840-4851