Evaluation of Mapping Computer Vision Segmentation from Reality Capture to Schedule Activities for Construction Monitoring in the Absence of Detailed BIM

Juan D. Núñez-Morales[1,3,†] **and Yoonhwa Jung**[1,3,†] **and Mani Golparvar-Fard**[2,3]

[1]Department of Civil and Environmental Eng., and Computer Science
[2]Department of Civil and Environmental Eng., Computer Science, and Technology Entrepreneurship
[3]University of Illinois at Urbana-Champaign, USA
[†] These authors contributed equally.

{jdnunez2, yoonhwa2, mgolpar}@illinois.edu

Abstract -

Over the past few years, research has focused on leveraging computer vision in construction progress monitoring, particularly in comparing construction photologs to Building Information Modeling (BIM), with or without schedule data. The practical application of these techniques and a large number of startups that have brought hyper AI and human-in-the-loop services around progress monitoring have revealed several gaps: 1) Current BIM-driven projects do not have model disciplines at the right level of maturity and Level of Development; 2) definitions of states of work-in-progress that are detectable from images are not formalized; 3) poor schedule quality and lack of frequent progress update challenges the incorporation of detailed 4D BIM for progress tracking. Such gaps are addressed in this work by exploring the requirements for mapping modern computer vision techniques for object segmentation with construction schedule activities to automate progress monitoring applications using computer vision without BIM as a baseline. The approach utilizes reality mapping practices to offer time machines for construction progress, organizing photologs over space and time. Additionally, this work shows how Large Language Models can structure schedule activity descriptions around <Uniformat Object Classification, Location>, focusing on how vision and language models can be trained separately with limited annotated data. ASTM Uniformat classification is utilized to map triangulated object segments from images to color-coded 3D point clouds aligned with schedule activities without the need for image and language feature alignments. Exemplary results on tied new transformer-based models with few-shot learning are shown, and the requirements for full-scale implementation are discussed.

Keywords -

Automated Progress Monitoring; Artificial Intelligence; Computer Vision; Natural Language Processing

1 Introduction

The field of construction progress monitoring has witnessed significant advancements in recent years, primarily driven by the integration of computer vision techniques and BIM. Specifically, leveraging devices such as LiDAR (Light Detection and Ranging), 360-degree cameras, and drones, a prevalent approach involves comparing generated as-built point clouds with the as-planned BIM model to assess construction progress by identifying deviations or discrepancies [1]. The integration between computer vision and BIM has enabled stakeholders to gain comprehensive insights into the construction process, facilitating improved decision-making, resource allocation, and offering real-time data analysis and project visualization [2].

However, despite these advancements, several critical gaps persist, posing challenges to the effective implementation of computer vision-based progress monitoring applications:

- *Lack of BIM model disciplines at the right level of maturity and Level of Development (LoD)*: The current computer vision-based progress monitoring relies heavily on the quality and completeness of BIM models. Insufficient BIM LoD and maturity in model disciplines, particularly around work zones and ASTM Codes, create discrepancies when attempting to align as-built point clouds with the as-planned BIM model [3, 4].
- *The absence of formalized definitions for states of work-in-progress detectable from images*: The lack of formalized definitions for work-in-progress states detectable from images poses challenges, especially in visually complex indoor environments with occlusion complexities like wall layers. [5, 6]. Establishing clear and universally accepted definitions for construction states is crucial for enhancing the accuracy and comparability of progress assessments in diverse project environments.
- *Poor schedule quality and lack of frequent progress*

update: The effectiveness of 4D BIM, integrating the temporal dimension, relies heavily on schedule quality and BIM LoD. However, the suboptimal schedule quality, varying schedule granularity, and infrequent progress updates impede the alignment of real-time progress with the as-planned 4D BIM model [7].

This paper addresses these gaps by investigating the requirements for mapping modern computer vision techniques, specifically object segmentation, to construction schedule activities. The focus is on automating progress monitoring applications using computer vision as a standalone tool without relying on BIM as a baseline. To achieve this, this work proposes leveraging reality mapping practices that effectively organize photologs over space and time, providing time machines for construction progress.

Furthermore, this work explores how Large Language Models (LLMs) can play a pivotal role in structuring schedule activity descriptions, emphasizing key elements such as Action, Location with an emphasis on utilizing the ASTM Uniformat classification to map triangulated object segments from images to space. Exemplary results from a novel transformer model, coupled with few-shot learning, are presented, and the paper concludes by discussing the requirements for the full-scale implementation of these proposed methodologies. Identified gaps and specific challenges and solutions are covered in detail in the following sections.

2 Literature Review

This section reviews the current state-of-the-art in construction for the creation of semantically rich point clouds, and how both visual and text feature information have been automatically aligned in different fronts of computer vision. Additionally, this section covers the main strategies widely utilized for mapping progress against schedules, specifically focusing on the challenges presented when aligning progress information against construction schedule documents. Lastly, BIM-based progress extraction methods are explored and further discussed as a function of their advantages and challenges.

2.1 Point cloud generation and segmentation

Point clouds, obtained via 3D laser scanning or similar technologies, consist of data points in a three-dimensional coordinate system. LiDAR, using laser beams, accurately generates detailed 3D representations of structures and environments, capturing geometry and spatial information at construction sites. Photogrammetry, another method, extracts 3D details from 2D images taken by cameras or drones. The process involves estimating camera parameters and correlating pixels between cameras to triangulate depth information [8, 9].

After creating a 3D point cloud, machine learning algorithms are used to classify individual points into categories such as building components and Mechanical-Electrical-Plumbing (MEP) systems[10]. However, existing detection systems often depend on complete object visualizations, making them sensitive to occlusions and capture completeness. In the context of photogrammetry-based point cloud reconstruction, classification to correlate 2D information with the 3D point cloud is achieved through numerous novel image-based classification and segmentation techniques, only using 2D features.

For example, [11] employs traditional vision-based algorithms using shape and color information to infer indoor partition wall progress. Others adopt deep-learning approaches like MaskRCNN [12], YOLO [13], and Transformer-based Swin [14] to directly detect construction elements from 2D images [15, 16]. Despite some photogrammetry approaches mapping camera pixels to real-world coordinates, the projection of semantic information from 2D inputs onto point clouds remains underexplored [16].

2.2 Image and Text Information Mapping

In the realm of Computer Science, significant progress has been made in mapping information between images and text as multimodal learning. Notable developments include pre-trained vision-language models like CLIP [17] and mPLUG [18], which demonstrate the capability to understand and associate textual descriptions with visual content. However, construction projects involve unique terminologies, specialized jargon, and context-specific information that may not be adequately addressed by generic image-text mapping models. Addressing these limitations is essential for creating effective tools that can assist professionals in the construction industry, providing them with streamlined access to information and enhancing communication between textual project descriptions and visual project elements.

2.3 Progress Against Schedules in Construction

Proactive construction workflows rely on critical schedule updates. Researchers have explored automated methods for optimizing schedules based on underlying reasons of sequencing logic [19], physical building component relationships [20, 21], aligning different levels of schedules [22], and ensuring consistency with BIM, schedule, and payment applications via ASTM Uniformat classifications [7]. These approaches often involve natural language processing and machine learning algorithms. However, the usefulness of the ASTM Uniformat II to properly report

visual progress information against construction schedules is often ignored and poorly explored.

2.4 BIM-based Progress Monitoring Applications

Traditionally, progress monitoring has been based on comparing documented reality against plans, and through the standardization of the usage of BIM, numerous research have explored this data format for calculating construction progress. This is seen in BIM methods to drive Earned Value Analysis (EVA) [23] through geometry and time-based heuristics. Similarly, other works have relied on geometry information from BIM models to compute progress when compared against photogrammetry-based point clouds and proximity-based heuristics [3, 4], eventually incorporating 4D BIM [24] to determine the presence of built elements over time. More advanced implementations eventually considered detecting semantics from images to increase the correct detection of construction elements against BIM elements [25, 26].

While these lines of work pioneered one of the great avenues of progress monitoring using BIM, they were exposed to dependencies on the BIM LOD, leaving a wide gap to properly comparing reality against plan given the lack of 4D adoption that could connect progress against construction documents and schedules or the incorrect comparison between visible elements without semantics in a point cloud cluster and their intersection against BIM elements.

3 Method

The proposed methodology addresses one of the applications proposed in [27]. This method consists of a four-step approach model ensemble to create connections between vision and language features. At a high level, this is achieved by aligning orthographic projections of semantically rich photogrammetry point clouds against quantity-take-off (QTO) construction drawings to detect completed quantities of each classification of construction object classes based on the ASTM Uniformat II. Simultaneously, this method leverages corresponding logical sequences from a construction schedule to determine the completeness constraints of overlapping construction classes as part of the sequence logic, constrained to a specific location, generally denoted in a construction drawing. Such imposed logical constraint allows for the correction of observed quantities in segmented point clouds due to occlusion and point cloud completeness issues, which are present in most photogrammetry-based point cloud reconstructions. Figure 1 presents a graphical representation of the description presented above.

The first step leverages modern computer vision Structure-from-Motion (SfM) [8] and Multi-View Stereo (MVS) [9] algorithms to create unstructured three-dimensional as-built representations of the construction environment from video capture frames or cameras C_i. At this stage, and specifically during the depth estimation for each pixel P_j of a registered and localized camera C_i, mappings between pixel and three-dimensional point coordinates are captured using camera matrix transformations M_i for each camera C_i. For a more in-depth understanding of the employed SfM with MVS approaches, readers are recommended to read the work shown in [28].

In parallel, using the image inputs from the first step, the second step leverages a few-shot trained Swin [14] transformer model, pretrained with synthetic data to create per-pixel semantic classes or segmentations. The model training step considers a class structure based on the ASTM Uniformat classification of construction objects to detect partial construction of different construction elements. A semantic label class L_k is stored for each pixel P_j in camera C_i, and added to a general dictionary D containing tuples $D = [C_i, Pj, L_k, point_{x,y,z}]$. Such a dictionary is used to create a semantically rich segmented point cloud based on ASTM Uniformat II classes for each visible object. By leveraging camera vectors and orientations, an orthographic projection O is automatically created and overlayed against corresponding IFC drawings with QTOs using a three-point aligning process.

Concurrently, during the third step, corresponding schedules are parsed and classified Uniformat Level 2 & 3 instances using UniformatBridge [7], built on the pretrained BERT model. At this stage, for each activity line item Act_m, similar ASTM Uniformat classification labels L_k as assigned, together with location-based constraints Loc_n based on the schedule text usign PoAT [29]. At this stage, using the schedule activity relationships, sequential constraints are extracted for each detected Uniformat classification (i.e., $[Floor > Wall > Windows > Ceiling]$). These extracted sequences provide the logical constraints in which overlapping detected activity class orthographic projections from a segmented point cloud.

The fourth step attempts to create a polygon-based completion coverage to extract current object quantities. For each polygon Pol_u from a QTO drawing containing construction class label L_k, its completeness is evaluated using the overlapping segmented pixel class from the aligned orthographic point cloud projection as a function of the percent complete of an object and its corresponding activity, as shown in equations 1 and 2, respectively:

$$\%Complete_{Obj_m} = \frac{Drawing_{Pol_u} \cap PointCloud_{Pol_u}}{|Drawing_{Pol_u,L_k} = PointCloud_{Pol_u,L_k}} \quad (1)$$

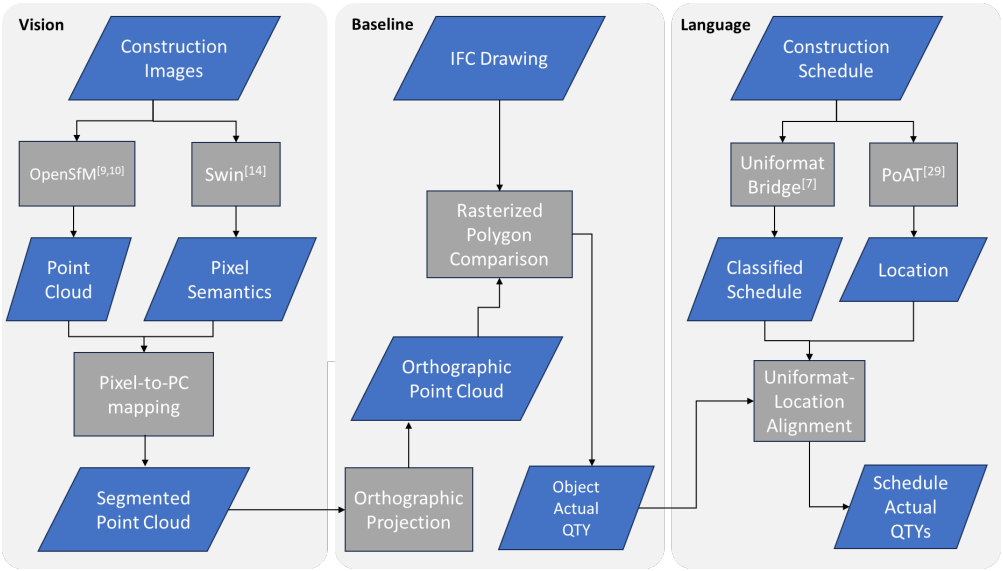

Figure 1. The proposed methodology of this work for automating progress monitoring using Images, IFC Drawings with QTOs, and construction schedules.

$$\%Complete_{Act_m} = \%Complete_{Obj_m} \in Loc_n \quad |Act_{m,Loc_n} = Drawing_{Pol_u,L_k} \tag{2}$$

Where $\%Complete_{Obj_m}$ is the estimated percentage completion of a construction object, $Drawing_{Pol_u}$ is the IFC Drawing polygon object, and $\%Complete_{Act_m}$ is the estimated percentage completion of an activity Act_m, upon constraining quantities to a location polygon Loc_n provided by the activity classification step, and extracting locations from the IFC Drawing.

As part of the logical constraint imposed by the extracted sequence in the second step of this method, quantities of initial objects in such sequence are corrected to reflect their completion imposed by constructability constraints of subsequent dependent objects (i.e., floor slabs quantities are corrected as a function of the detected ceiling quantities), as defined in the expression of equation 3, and the rules imposed by equation 4:

$$Seq = Obj_a, Obj_b, Obj_c \tag{3}$$

$$\%Complete_{Obj_a} = 100\% | \%Complete_{Obj_b} > \%Complete_{Obj_a} \tag{4}$$

4 Data and Experiment Settings

The experiment of this work evaluates a real-world scenario in the construction of a high-rise hotel building. It considers the utilization of an IFC drawing with computed quantities, a construction schedule, and a set of images encompassing five different ASTM Uniformat II classifications corresponding to floors, partition walls, windows, ceilings, and pipe installation.

For the creation of dense reality point clouds, this work makes use of the OpenSfM library [30] and a set of 120 images corresponding to a room area of the construction to evaluate, as shown in Figure 2. The available IFC drawings are manually scaled and annotated based on the Uniformat II object classifications and stored as a rasterized document, as shown in Figure 3.

Moreover, the detection of construction elements makes use of a Swin model, pretrained with synthetically created construction scene images, following the training strategies from work presented in [31]. For visualization purposes, the projected RGB values of the point cloud semantics are set to differ from those from the IFC drawing.

Given the limited visibility of information when comparing the orthographic projection of a point cloud against the rasterized IFC drawing, different elevations are considered to detect and measure objects properly. Specifically, three elevation values – each at one-third (1/3) incremental of the total height – are selected to extract key orthographic projections from the point cloud to compare and estimate detected quantities. To account for elements that may be detected but outside of the designated point cloud orthographic projection height slice, a sampling threshold of one-sixth (1/6) in the vertical direction is utilized, as shown in Figure 4.

To extract schedule sequential information, a pretrained

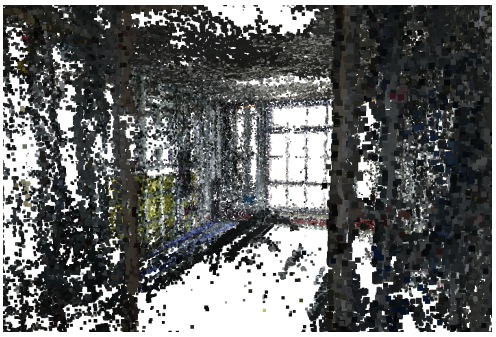

Figure 2. Sample view of the point cloud reconstruction results. As made evident by using photogrammetry in construction environments, common challenges include point cloud completion due to homogeneous surfaces and object occlusions due to non-construction-related elements. The current reconstructed elements in this view correspond to wall aluminum framing, windows, gypsum board on ceilings, and concrete floors.

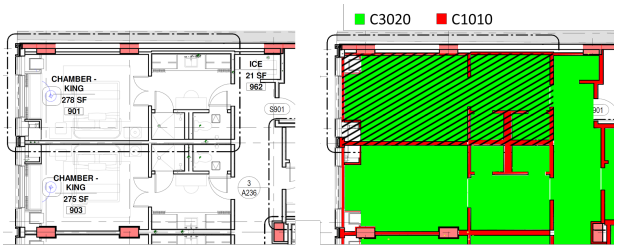

Figure 3. An a) Issued for Construction (IFC) drawing, along with b) polygon-based annotations for computing construction object quantities (QTYs). The scope of the experiment focuses on room areas marked with a hash pattern.

BERT model following the work proposed in [7] is used. A total of 1,700 construction activities are parsed and classified as a function of levels two and three of the ASTM Uniformat II. Specifically for this experiment, the scope of activities is focused on activities present in the room shown in Figure 3, focusing on the scope of structural and interior work. In addition to such classification, and to account for the mapping between a visually detected object and its correct activity line item, location information is extracted following the work presented in [29] and compared against the matched segmented point cloud segments, whose location information is extracted from the overlapping annotated IFC drawing. Additionally, Uniformat II sequences are created based on the extracted activity relationships present in the schedule and stored as separate recipes.

Lastly, the automatically detected progress quantities are evaluated against ground-truth actual progress quantities, estimated from the project visual documentation and available daily construction reports. The accuracy of the detected progress quantities using the proposed method is evaluated using the Mean Average Percentage Error (MAPE) and reported against each schedule line item.

5 Result and Discussion

The following section focuses on the results pertaining to the computation of actual quantities based on the comparison between actual and planned orthographic projections, the mapping accuracy against schedule line items via using Uniformat and Location information, and the ability to correct detected quantities as a function of sequential information from a schedule's Uniformat classifications and relationships.

As evident in Figure 5, the sparsity of the mesh in a point cloud may create inaccurate estimates of progress values. Table 1 showcases how each detection is assigned to the corresponding activity Uniformat code and location, the results of comparing the estimated quantities based on comparing a segmented mesh against the IFC drawing, and the corrected completion estimates for each entry.

This case study shows the success of utilizing the ASTM Uniformat II as the bridge to align schedules against detected vision information. The contrast provided between the results in Figure 5 and Table 1 shows how elements with large pixel coverage (such as walls, floors, and ceilings) contain more pixel-level information that allows for dense point cloud reconstruction, improved feature segmentation, and better heuristically-controlled calculations due to corresponding to middle-sequence tasks, which can make use of a predecessor and successor thresholds for progress-based dependencies against other completed or to-be-completed elements.

In contrast, having significant occlusions denotes how the precision of the reality capture is less significant than the strategies utilized for correcting or assuming completed quantities. Such a case becomes especially challenging for slender objects such as MEP components, thus explaining the higher degrees of error in the experiment. This is mainly attributed to the lower number of pixels available for reconstructions and feature segmentation. With fewer pixels, a reconstruction engine may produce non-dense point clouds, decreasing the degree of label projection against the point cloud. Moreover, fewer pixels decrease the success rate of correctly classifying pixel features against predetermined classes. Lastly, due to MEP components being at the very last step of a sequence, the rule-based checking of the quantities becomes less impactful compared to the middle steps of a sequence.

Nevertheless, occlusions are a phenomenon that would similarly impact LiDAR-based scanning strategies. Still,

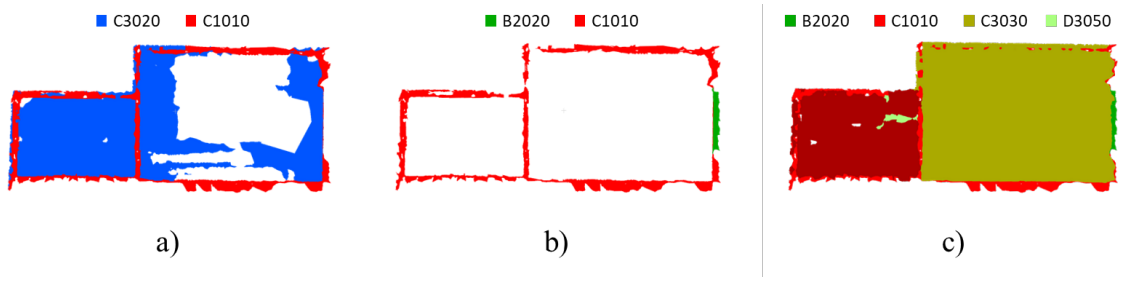

Figure 4. Pixel-based mapped against the orthographic representation of a point cloud at three different heights, where case a) shows the bottom projection, b) shows the middle height projection, and c) the top projection. These projections are scaled to the mesh representation of the resulting point cloud.

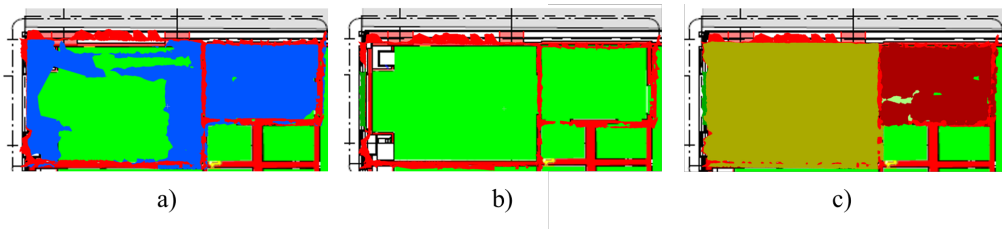

Figure 5. Initial orthographic overlay between the segmented point cloud and an IFC drawing with annotated quantities at three different heights. Different coverage is observed for each evaluated height, showcasing the need for heuristics based on Uniformat II sequences.

Table 1. Reported quantities for each construction activity. Classification results are based on the Uniformat Level 3, and Location information is automatically extracted from the Activity Name. Actual quantities (QTY) correspond to the actual quantity takeoff for completed elements from the IFC drawing (ground truth), while the Detected QTY comes from the area completeness ratio of the segmentation overlay against IFC drawings. The Corrected QTY is determined based on the extracted sequence from the classified activities and their relationships, and is compared against the ground truth to compute their mean average percentage error (MAPE).

Activity Name	Uniformat Code	Location	Actual QTY	Percentage Coverage (%)	Detected QTY	Corrected QTY	MAPE (%)
5th Floor - Wall Covering	C3010	Level 5	0 SQFT	0	0	0	0
5th Floor - Install Flooring	C3020	Level 5	182.6 SQFT	45.8	83.6 SQFT	179.6 SQFT	1.6
N-E 5th Floor - Hang Gyp at Ceiling - NE	C3030	Level 5 Zone NE	191.5 SQFT	93.8	179.6 SQFT	179.6 SQFT	6.2
5th Floor - Frame Walls	C1010	Level 5	579.4 SQFT	100	579.4 SQFT	579.4 SQFT	0
Exterior Skin - Window Installation Complete - 5th Floor	B2020	Level 5	1 EA	100	1 EA	1 EA	0
5th Floor - Plumbing Trim - Shower	D2010	Level 5	0 LF	0	0	0	0
5th Floor - Start Final MEP Wall/Ceiling Rough-In	D3050	Level 5	10.6 LF	4.6	0.5 LF	0.5 LF	95.3

using language models to extract sequential information encoded through the ASTM Uniformat II allows for the correction of certain construction elements that poor point cloud reconstructions may impact.

6 Conclusion

This work presented an application of automated progress monitoring, combining vision and language outputs only (i.e., without the use of BIM), and a practical way forward to creating automated progress monitoring applications, leveraging modern deep learning and computer vision algorithms. The proposed method utilizes novel transformer-based architectures such as Swin and BERT, each trained to classify detected construction objects based on the ASTM Uniformat II of Construction Objects from reality capture images and construction schedules. Additionally, the proposed method introduces an approach and heuristic to calculating the actual progress of construction. This approach is done by leveraging sequential and location-based information extracted from classified activity line items. Additionally, it imposes constraints on computed quantities from comparing per-pixel construction semantics projected on orthographic point cloud representations against IFC drawings with annotated QTOs.

The utilization of schedule-based-heuristics and sequential constraints addresses two types of limitations from the utilization of photogrammetry for the detection of progress: 1) the miscalculation of percent completed due to occluded construction objects in typical

scenes, and 2) the incomplete computed progress from low-density reconstructions showcasing incomplete objects. This method is validated by a real-world case study, which computes progress quantities for a high-rise hotel as a function of provided reality captures, IFC drawing QTOs, and construction schedules. In this study, the detected pixel segmentations are projected against point cloud reconstructions and compared against drawing QTOs to compute actual progress. Such progress is corrected based on the automatically extracted schedule sequences and locations.

Given the current state of BIM standardization across the industry, the applicability of this method in real-world scenarios may fill the gap that low LOD in BIM models create by bringing an alternative to progress monitoring based on globalized practices when working with IFC drawings and QTOs. Future steps in this line of work will evaluate the validity of the established heuristics for other ASTM Construction Object categories.

References

[1] Varun Kumar Reja, Koshy Varghese, and Quang Phuc Ha. Computer vision-based construction progress monitoring. *Automation in Construction*, 138:104245, 2022. ISSN 0926-5805. doi:10.1016/j.autcon.2022.104245.

[2] M.Q. Huang, J. Ninić, and Q.B. Zhang. Bim, machine learning and computer vision techniques in underground construction: Current status and future perspectives. *Tunnelling and Underground Space Technology*, 108:103677, 2021. ISSN 0886-7798. doi:10.1016/j.tust.2020.103677.

[3] Sebastian Tuttas, Alex Braun, Andre Borrmann, and Uwe Stilla. Comparision of photogrammetric point clouds with bim building elements for construction progress monitoring. volume 1, pages 341–345, 08 2014. doi:10.5194/isprsarchives-XL-3-341-2014.

[4] Seungho Kim, Sangyong Kim, and dong-eun Lee. 3d point cloud and bim-based reconstruction for evaluation of project by as-planned and as-built. *Remote Sensing*, 12:1457, 05 2020. doi:10.3390/rs12091457.

[5] Biyanka Ekanayake, Johnny Kwok-Wai Wong, Alireza Ahmadian Fard Fini, and Peter Smith. Computer vision-based interior construction progress monitoring: A literature review and future research directions. *Automation in Construction*, 127:103705, 2021. ISSN 0926-5805. doi:10.1016/j.autcon.2021.103705.

[6] Hélène Macher, Tania Landes, and Pierre Grussenmeyer. From point clouds to building information models: 3d semi-automatic reconstruction of indoors of existing buildings. *Applied Sciences*, 7(10):1030, 2017. doi:10.3390/app7101030.

[7] Yoonhwa Jung, Julia Hockenmaier, and Mani Golparvar-Fard. Transformer language model for mapping construction schedule activities to uniformat categories. *Automation in Construction*, 157:105183, 2024. ISSN 0926-5805. doi:10.1016/j.autcon.2023.105183.

[8] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. doi:10.1109/CVPR.2016.445.

[9] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Multi-view photometric stereo revisited. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3126–3135, 2023. doi:10.48550/arXiv.2210.07670.

[10] Qian Wang, Yi Tan, and Zhongya Mei. Computational methods of acquisition and processing of 3d point cloud data for construction applications. *Archives of computational methods in engineering*, 27:479–499, 2020. doi:10.1007/s11831-019-09320-4.

[11] Hesam Hamledari, Brenda McCabe, and Shakiba Davari. Automated computer vision-based detection of components of under-construction indoor partitions. *Automation in Construction*, 74:78–94, 2017. ISSN 0926-5805. doi:10.1016/j.autcon.2016.11.009.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi:10.48550/arXiv.1703.06870.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. doi:10.48550/arXiv.1506.02640.

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Pro-*

ceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021. doi:10.48550/arXiv.2103.14030.

[15] Nipun D. Nath and Amir H. Behzadan. Deep convolutional networks for construction object detection under different visual conditions. *Frontiers in Built Environment*, 6, 2020. ISSN 2297-3362. doi:10.3389/fbuil.2020.00097.

[16] Wei Wei, Yujie Lu, Yijun Lin, Ruihan Bai, Yichong Zhang, Haisong Wang, and Peixian Li. Augmenting progress monitoring in soil-foundation construction utilizing solov2-based instance segmentation and visual bim representation. *Automation in Construction*, 155:105048, 2023. ISSN 0926-5805. doi:10.1016/j.autcon.2023.105048.

[17] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734:pp. 1–10, 2021. doi:10.48550/arXiv.2111.09734.

[18] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mPLUG:: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. doi:10.48550/arXiv.2205.12005.

[19] Bonsang Koo, Martin Fischer, and John Kunz. A formal identification and re-sequencing process for developing sequencing alternatives in cpm schedules. *Automation in Construction*, 17(1):75–89, 2007. ISSN 0926-5805. doi:10.1016/j.autcon.2007.03.005.

[20] Yibrah Weldemihret Weldu. *Automated Generation and Visualization of Initial Construction Schedules from Building Information Models*. PhD thesis, Louisiana State University, 2016. URL `https://digitalcommons.lsu.edu/gradschool_dissertations/175`.

[21] Zhiliang Ma, Songyang Li, Yong Wang, and Zhenqing Yang. Component-level construction schedule optimization for hybrid concrete structures. *Automation in Construction*, 125:103607, 2021. ISSN 0926-5805. doi:10.1016/j.autcon.2021.103607.

[22] Fouad Amer, Yoonhwa Jung, and Mani Golparvar-Fard. Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction. *Automation in Construction*, 132: 103929, 2021. doi:10.1016/j.autcon.2021.103929.

[23] Yelda Turkan, Frédéric Bosché, Carl Haas, and Ralph Haas. Toward automated earned value tracking using 3d imaging tools. *Journal of Construction Engineering and Management*, 139:423–433, 01 2012. doi:10.1061/(ASCE)CO.1943-7862.0000629.

[24] Sebastian Tuttas, Alex Braun, Andre Borrmann, and Uwe Stilla. Acquisition and consecutive registration of photogrammetric point clouds for construction progress monitoring using a 4d bim. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 85:3–15, 02 2017. doi:10.1007/s41064-016-0002-z.

[25] Christopher Kropp, Christian Koch, and Markus König. Interior construction state recognition with 4d bim registered image sequences. *Automation in Construction*, 86:11–32, 02 2018. doi:10.1016/j.autcon.2017.10.027.

[26] Alex Braun, Sebastian Tuttas, Andre Borrmann, and Uwe Stilla. Improving progress monitoring by fusing point clouds, semantic data and computer vision. *Automation in Construction*, 116, 08 2020. doi:10.1016/j.autcon.2020.103210.

[27] Juan D. Nunez-Morales, Yoonhwa Jung, and Mani Golparvar-Fard. Bi-directional image-to-text mapping for nlp-based schedule generation and computer vision progress monitoring. In *Construction Research Congress 2024*, 2024. (In press).

[28] Moch Kholil, I Ismanto, and M Fu'ad. 3d reconstruction using structure from motion (sfm) algorithm and multi view stereo (mvs) based on computer vision. *IOP Conference Series: Materials Science and Engineering*, 1073:012066, 02 2021. doi:10.1088/1757-899X/1073/1/012066.

[29] Fouad Amer and Mani Golparvar-Fard. Automatic understanding of construction schedules: Part-of-activity tagging. In *EC3 Conference 2019*, volume 1, pages 190–197, 2019. doi:10.35490/EC3.2019.196.

[30] Pablo Fernández Alcantarilla, Jesús Nuevo, and Adrien Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *British Machine Vision Conference*, 2013. URL `https://api.semanticscholar.org/CorpusID:8488231`.

[31] Juan D. Núñez-Morales, Shun-Hsiang Hsu, and Mani Golparvar-Fard. *Synthetic Image Generation for Training 2D Segmentation Models at Scale for Computer Vision Progress Monitoring in Construction*, pages 273–281. doi:10.1061/9780784485224.034.