Integration of Vision Platforms: Generating a Unified 3D Digital Map for Real-time Construction Site Monitoring

J. Shin[1], P. Kim[1], J. Choi[1], Y. Cho[1], G. Lee[2], S. Chi[1]

[1]Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea
[2]Institute of Construction and Environmental Engineering (ICEE), Republic of Korea
wpals113@snu.ac.kr, leophilgu@snu.ac.kr, cjw5428@snu.ac.kr, cyh_i@snu.ac.kr, lgt0427@snu.ac.kr, shchi@snu.ac.kr

Abstract –

Various vision sensor-based platforms generate critical information for real-time construction site monitoring, but their independent applications limit the effective use of this data. This study proposes a framework that integrates three representative vision platforms—Unmanned Aerial Vehicle (UAV), Terrestrial Laser Scanner (TLS), and Closed-Circuit Television (CCTV)—to create a unified digital map. The framework operates as follows: UAVs capture aerial images for overall 3D digital map reconstruction, TLS refines vertical surfaces and structural details through map alignment, and CCTV cameras enable robust real-time 3D object tracking using ray marching and multi-view fusion, resulting in a dynamic digital map. Experimental results demonstrated that the framework enhances resolution of a digital map and improves dynamic object tracking compared to using individual platforms in isolation.

Keywords –

Unified Map, Registration, 3D Reconstruction, Object Tracking, UAV, TLS, CCTV

1 Introduction

In modern construction sites, various vision platforms such as Unmanned Aerial Vehicles (UAVs), Terrestrial Laser Scanners (TLS), and Closed-Circuit Television (CCTV) are actively utilized to enhance site monitoring by generating critical site information. However, Each platform has its strengths and limitations. UAVs, for example, offer high mobility and cost-effective coverage of large areas, making them widely used for 3D digital map generation in construction sites [1]. However, their high-altitude operations and the perpendicular orientation of cameras to the ground, commonly used in UAV surveys, result in low spatial resolution and a lack of information about vertical surfaces in their digital maps [2]. Furthermore, UAVs cannot capture the entire environment at once, which prevents representing real-time movements of dynamic objects on the digital map.

TLS, on the other hand, uses LiDAR and cameras for long-duration scans from the ground, producing high-resolution 3D digital models, particularly for vertical surfaces. While expensive(e.g., the FARO FOCUS S350 costs over $14,500), its benefits make it widely used in construction for applications such as detecting concrete cracks [3] and comparing as-is models with design plans [4]. Recent advancements in Scan-to-BIM technology [5] enable automatic BIM model generation from precise survey data, further driving TLS adoption [6]. However, its ground-level perspective limits the capture of horizontal surfaces like rooftops, and the extensive time and labor required for data collection make it impractical for large-scale mapping or dynamic object analysis [7].

CCTV systems offer a cost-effective solution for long-term, real-time monitoring, capturing sequential data with low operational costs [8]. Rising safety concerns over construction sites accidents have driven their widespread adoption [9]. They are particularly useful for tracking dynamic objects like heavy equipment productivity [10]. However, their limited field of view and issues with occlusion make it challenging to continuously track and monitor dynamic objects, such as materials and workers, in complex and active construction sites [11]. CCTVs, additionally, struggle to generate 3D spatial information about a scene because they are typically installed in scattered locations with fixed monocular cameras.

To summarize, while various vision platforms have been widely used in construction sites, their inherent limitations restrict the usability of the generated information when used independently. To produce comprehensive site information, it is essential to integrate the independent data from each platform into a unified system. UAVs can rapidly capture overall site data without being constrained by site complexity, while TLS can complement this by generating high-resolution models of critical structures and vertical surfaces that require greater detail. Meanwhile, CCTV systems can track dynamic objects, such as workers and equipment.

By integrating data from these platforms, a more complete and accurate digital models of the construction site can be achieved. To this end, this paper proposes a framework for the integration of information from each vision platform into a unified 3D digital map of the site, leading to more comprehensive construction site monitoring in real-time.

2 Literature Review

2.1 Integration of UAV and TLS Data

Many previous studies on aligning UAV and TLS data utilize the Iterative Closest Point (ICP). ICP is an algorithm that iteratively aligns two datasets by finding the closest point for each point and calculating the rotation and translation between them. However, it is highly sensitive to the quality of the initial alignment [12]. The previous studies often perform initial alignment using GPS data [13] or manual alignment via software [14], followed by fine alignment using ICP.

UAV-generated 3D digital maps, created from aerial imagery, and TLS-generated 3D digital models, created from ground-based LiDAR, differ significantly from the distribution and density characteristics of their points. These differences in point distribution can negatively impact alignment accuracy [15]. Moreover, construction sites are typically large-scale environments, resulting in digital maps with a high number of points. This makes it challenging to apply computationally intensive algorithms due to their high computational costs.

In response, this study aims to resolve the differences in point distribution and density characteristics, perform reliable automatic initial alignment, and maximize the performance of ICP fine alignment.

2.2 Camera-based 3D Object Tracking

Camera-based 3D object tracking can be broadly categorized into multi-view geometry-based methods and monocular methods. Multi-view geometry involves installing multiple cameras with overlapping fields of view and estimating relative spatial relationships through feature points to infer 3D coordinates. This approach, however, requires cameras to be installed with similar viewpoints and overlapping coverage, which is challenging to achieve in typical construction sites.

On the other hand, monocular methods primarily rely on homography or depth estimation models. Homography represents the transformation between points on two planes. Assuming the ground is a plane, a homography relationship can be established between the ground plane in the 3D real-world coordinate system and the 2D image plane. The relationship can be calculated if at least four corresponding point pairs are identified. However, since homography is limited to flat planes like the ground, accurately incorporating dynamic information captured by CCTV into a 3D digital map becomes challenging in more complex scenarios involving uneven or non-planer surfaces.

MiDaS [16], a state-of-the-art single-image depth estimation model, offers an alternative approach. MiDaS is trained to generate depth maps from 2D images labeled with depth information. However, monocular depth estimation models can experience significant performance degradation when applied to scenes that differ substantially from those seen during training.

3 Methodology

This study is conducted in three stages: overall 3D map reconstruction, map refinement with registration, and object tracking in a 3D map, as shown in Figure 1.

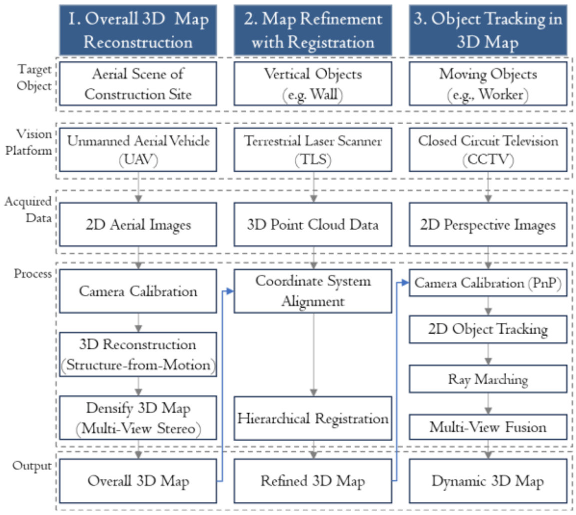

Figure 1. Framework of Vision Platforms Integration

The first stage involves generating an overall 3D map. A UAV is deployed to capture sequential 2D aerial images, and Structure-from-Motion (SfM) [17] is applied to reconstruct a sparse 3D map of the entire site. Subsequently, Patch-based Multi-View Stereo (PMVS) [18] creates a densified overall 3D map.

The second stage focuses on refining the generated 3D digital map. Aerial images alone have limitations in representing high-resolution details and generating 3D models of structures with vertical surfaces, such as walls. To address these limitations, high-resolution 3D point cloud data is collected using a TLS. The UAV-derived 3D digital map and the TLS-based 3D digital models are first coarsely aligned to a unified coordinate system. Subsequently, fine alignment is performed using hierarchical registration. The integration of UAV and TLS data enhances the level of detail and improves the representation of vertical surfaces in the initial map.

The third stage involves tracking objects and incorporating this information into the 3D map. For CCTV footage, objects are tracked, and their positions in the 3D map are estimated using ray marching. Multi-view fusion is then applied to ensure robustness against tracking failures and maintain the consistency of object information across multiple CCTVs. By integrating these three stages, this study generates a unified 3D digital map that not only provides refined 3D representations but also incorporates dynamic object information.

3.1 Overall 3D Map Reconstruction

The process of generating an overall 3D digital map using UAV involves the following steps: camera calibration is first conducted to estimate the camera matrix, which defines the transformation between 2D image points and their corresponding 3D coordinates. The camera matrix is composed of intrinsic and extrinsic parameters. Intrinsic parameters, such as focal length and principal point, describe the internal characteristics of the camera. Extrinsic parameters define the camera's position and orientation relative to the world coordinate system. The intrinsic parameters are typically determined using a chessboard pattern, while the extrinsic parameters are estimated using Structure from Motion (SfM). SfM is a photogrammetry-based technique that iteratively computes camera poses and generates a sparse 3D map from sequential images, optimizing both pose estimation and mapping accuracy throughout the process. Subsequently, Patch-based Multi-View Stereo (PMVS) is employed to identify overlapping patches between consecutive images and reconstruct them in 3D, thereby densifying the sparse map.

3.2 Map Refinement with Registration

The TLS captures high-resolution 3D data of the surrounding area using LiDAR technology. The 3D data generated by UAV and TLS are represented in different coordinate systems, each with distinct scales and origins. To unify these coordinate systems, GPS data is used to restore the scale of each 3D map to real-world dimensions and align their coordinate system references. Since GPS inherently has an error margin of several meters, an additional hierarchical registration process is required to achieve accurate integration.

3.2.1 Coordinate System Alignment with GPS

Geo-registration converts the overall 3D digital map from the SfM coordinate system into the Earth-Centered Earth-Fixed (ECEF) coordinate system, a global 3D Cartesian coordinate system centered on the Earth. This transformation is essential for accurately positioning and orienting objects on a global scale. The process uses Random Sample Consensus (RANSAC) to determine the optimal similarity transformation between the two coordinate systems.

In contrast, TLS directly captures 3D points via LiDAR, generating maps based on GPS data. To unify digital maps, the TLS-generated 3D map is transformed into the ECEF coordinate system, ensuring coarse alignment with the UAV digital map.

3.2.2 Hierarchical Registration of 3D Digital Maps

Hierarchical registration is used to efficiently align large-scale UAV and TLS point clouds by breaking the process into two stages, as shown in figure 2. For initial alignment, a Fast Point Feature Histogram (FPFH) analyzes the relationships between neighboring points in the point cloud data to extract feature points that encapsulate local geometric information. These features are matched using a RANSAC algorithm to establish an initial transformation. While FPFH provides reliable initial alignment, it incurs high computational costs when applied to large point cloud datasets. To overcome this challenge, both 3D digital maps are down-sampled with voxelization, where the points within each cubic voxel are represented by their average position. This process reduces computational complexity while mitigating differences in point distribution.

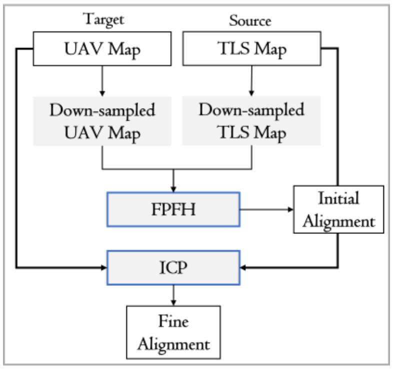

Figure 2. Hierarchical Registration Process

Following this, the ICP algorithm refines the alignment by identifying the closest corresponding points and updating the transformation matrix. Since finding the nearest points is a relatively simple operation and FPFH ensures a reliable initial alignment, ICP is performed on the entire point cloud. By combining FPFH for initial alignment and ICP for fine refinement, this hierarchical approach achieves efficient and accurate automatic alignment of UAV and TLS maps without succumbing to computational overhead or overfitting from differences in point distribution.

3.3 Object Tracking in a 3D Map

The proposed method for object tracking in a 3D map integrates information about dynamic objects into the map for enhanced situational awareness. The process begins with camera calibration, during which the intrinsic and extrinsic parameters of each camera are estimated. This calibration establishes the relative position and orientation of the CCTV cameras within the registered refined digital map. Subsequently, YOLO-based 2D object detection is applied, identifying objects and determining their positions as 2D pixel coordinates x_{object}. These coordinates are then back-projected into the 3D space using ray marching to estimate the objects' 3D positions X_{object}. To improve the accuracy and consistency of tracking, multi-view fusion is employed, leveraging data from multiple perspectives.

3.3.1 Camera Calibration for CCTV Systems

For CCTV cameras, the extrinsic parameters are determined using the Perspective-n-Points (PnP) [19] algorithm. When four matching pairs between the 2D image and the 3D map are specified, the PnP algorithm calculates the camera's rotation and position. Further optimization is conducted using the Levenberg-Marquardt [20] to refine these estimates. Since the extrinsic parameters of a CCTV camera typically remain fixed after installation, calibration only needs to be performed once during the initial setup of the camera. This ensures that the camera's position and orientation are accurately aligned with the 3D digital map.

3.3.2 2D Object Tracking with Pose Tracker

2D object tracking is conducted using the BoT-SORT [21], a multi-object tracking method that performs object tracking without additional training. BoT-SORT can also extract key points of the detected objects. As illustrated in Figure 3, for a worker object, the 2D point of the object is defined as the center point (marked as a black cross) between the coordinates of the left foot and right foot (marked as red dots).

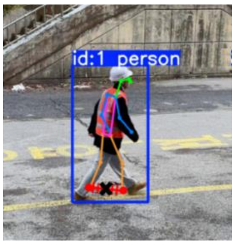

Figure 3. Definition of the worker's 2D point

3.3.3 Back Projection with Ray Marching

To represent the 2D point of an object into the 3D digital map, back projection is performed. Back projection is the process of reconstructing a 3D world point from a 2D image point.

$$\lambda P^{-1}\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix} \tag{1}$$

In Equation 1, P is a 3×4 camera matrix, u and v are 2D image points, and X, Y, and Z are 3D world points. λ is a scale parameter that represents the depth from the 2D image point to the corresponding 3D world point. The camera matrix P is obtained through camera calibration, leaving the depth parameter λ as the only unknown variable. In essence, back projection is the process of determining this unknown depth.

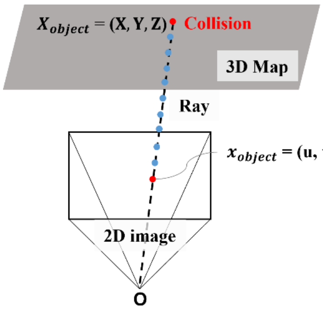

Figure 4. Ray Marching Method

Back projection can be achieved using ray marching. As shown in figure 4, ray marching is a technique that casts a ray from a 2D point x_{object} by sampling 3D points along the ray and searches for the point X_{object} where the ray collides with a surface in 3D space. Because this method directly interacts with the digital map, it requires a high-quality map that is noise-free, high-resolution , and free of voids. Fortunately, the registration of UAV and TLS data eliminates noise, enhancing both the resolution and completeness of the digital map. Ray marching involves sampling candidate 3D points along a ray extending from the camera's origin toward the corresponding 2D image point.

$$r(t) = O + tD \tag{2}$$

Equation 2 reformulates Equation 1 from the perspective of a ray, where O represents the camera origin and D is the direction vector of the ray. Both O and D are derived from the camera matrix P. The parameter t denotes a positive real number depth value, and by incrementing t, the 3D position $r(t)$ of a sample point along the ray is calculated. If the shortest distance between $r(t)$ and the points in the digital map falls below a predefined threshold, the corresponding point in the digital map is determined to be the object's position.

3.3.4 Multi-View Fusion

The 3D position obtained in the previous steps is managed by assigning a unique track ID for each camera. A single camera, however, loses the ability to track the object if it moves out of the camera's field of view. When the object reappears, a new track ID is assigned, resulting in inconsistency between multiple cameras. Multi-view fusion overcomes this inconsistency in track IDs and enables a robust object tracking system by integrating data from multiple cameras into a unified 3D digital map.

The detailed algorithm for multi-view fusion is explained with an example in Figure 5. At T=3, when a new track ID appears in a camera, the system queries the database to check whether any object is being tracked near the 3D position of the newly detected object. If no such object exists, a new object ID is assigned, and the database begins tracking this object. Subsequently, at T=40, if the same object appears in another camera, track matching is performed to associate the track ID from the camera with the object ID in the database. Once matched, the object's position can be obtained from both CAM0 and CAM1. This ensures that even if a tracking failure occurs in one camera, such as CAM1 at T=40, the object can still be tracked using CAM0.

If, at T=50, all cameras fail to track the object, the system predicts the next position based on the average position change over the previous five timestamps stored in the database. This prediction enables continued tracking, even during temporary tracking failures. Finally, at T=109, if different track IDs are assigned to the same object in other cameras, track matching ensures that the same object ID is used. However, duplicated 3D positions estimated by distinct cameras can be different due to errors in object detection. To address this, the system calculates the tracking error for each 3D position and selects the point with the lower error, reducing redundancy and improving accuracy. The error in 2D object tracking propagates into the 3D position estimation error through back projection. Since the relationship between 2D and 3D points involves a nonlinear transformation, this propagation is described by the Jacobian matrix J.

$$cov(X_w) = J \times cov(x_{pixel}) \times J^T \tag{3}$$

In Equation 3, $cov(X_w)$ represents the covariance of the 3D point, indicating the 3D object tracking error, while $cov(x_{pixel})$ represents the covariance of the 2D point, characterizing the error in 2D object tracking. This systematic error propagation ensures that the system accounts for inaccuracies in both 2D and 3D tracking, leading to more precise object localization.

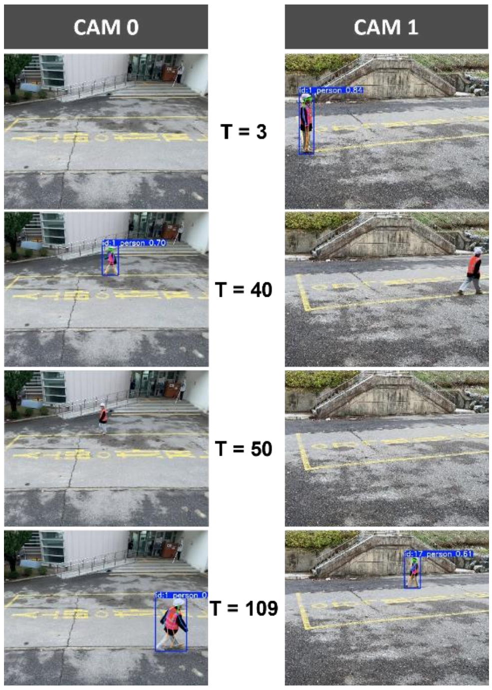

Figure 5. Example of Multi-view Fusion

4 Experiments

The experiments took place in a lane-marked area at Seoul National University. To simulate a multi-view CCTV system, two iPhone 12s were installed with overlapping fields of view, and their resolution was reduced to 360×640 to match typical CCTV settings. A DJI Mavic 3E UAV and a FARO Focus M Laser Scanner (TLS) were used for 3D map generation. The study consisted of two experiments.

- UAV and TLS Map Alignment

This experiment assessed the error in digital map alignment and the improvement in resolution for vertical surfaces. The evaluation metric for alignment was the Inlier Root Mean Square Error (Inlier RMSE) in meters. The metric was calculated by finding the nearest UAV-based neighbor for each transformed TLS point and calculating the RMSE. Points exceeding a set distance threshold were classified as outliers and excluded, ensuring a more precise alignment assessment. A voxelization size of 0.5m was used, with inlier thresholds of 0.5m, 1.0m, and 1.5m.

- Multi-View Fusion-based CCTV Object Tracking

This experiment evaluated the proposed ray marching-based object tracking method against baseline approaches using homography and the MiDaS depth estimation model. Since MiDaS generates only relative depth maps, four reference points were selected to obtain true depth values. A cubic spline interpolation was then applied to convert the relative depth map into an absolute depth map for the analysis. Moreover, the effectiveness of multi-view fusion was validated. Object tracking performance was assessed using the mean and RMSE of the Absolute Trajectory Error (ATE). The mean ATE is a standard measure of position estimation accuracy, while RMSE is more sensitive to large errors, providing a comprehensive evaluation of tracking.

The ground truth trajectory for evaluation was defined as the yellow rectangle shown in Figure 6. The ATE was calculated as the distance between the estimated 3D points and the closest ground truth points.

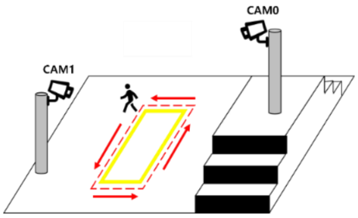

Figure 6. Testbed for Object Tracking

4.1 Evaluation of Detailed 3D Map

To validate the alignment accuracy, an ablation study, as shown in Table 1, compared the proposed hierarchical method against individual alignment algorithms. The best RMSE of 0.101m at a 0.5m inlier threshold was achieved using $FPFH_{down}$ with downsampling for initial alignment, followed by ICP. This highlights their complementary strengths: FPFH captures global geometric features for coarse alignment, while ICP refines precision during fine alignment.

The UAV-based digital map covers $150 \times 200m^2$ with 10.8 million points, while the TLS model spans $40 \times 76m^2$ with 74.2 million points. Table 1 analyzes runtime efficiency. Without downsampling, ICP successfully aligned in 4:36 minutes, whereas FPFH failed due to high computational costs. By applying 0.5m voxelization-based downsampling, FPFH completed in 3:12 minutes, approximately reducing UAV points to 207,000 (52× reduction) and TLS points to 64,000 (1,188× reduction). Since FPFH iterates over all points, downsampling reduced computational costs by over 1,000×, confirming its necessity.

Table 1 Inlier RMSE with Varying Inlier Thresholds and Runtime of Registration

	0.5m	$1.0m$	$1.5m$	Time(m)
ICP	0.300	0.610	0.920	4:36
$\text{FPFH}_{\text{down}}$	0.124	0.187	0.253	3:12
Proposed	**0.101**	**0.172**	**0.238**	7:48

Figure 7 illustrates the benefits of the proposed method. In the first row, vertical surfaces missing in the UAV digital map were completed through alignment, showing the ability to enhance map completeness. In the second row, the resolution of 3D models for structures of interest was significantly improved, increasing their utility for applications requiring high-quality 3D digital models, such as construction error analysis. In representing the wall structure, the UAV-based digital map had an average density of 124 pts/m^2. In contrast, the aligned digital map achieved an average density of 11,066 pts/m^2, almost 100-fold density increase. This substantial improvement demonstrates the enhanced capability of the proposed method to generate refined 3D representations of structures.

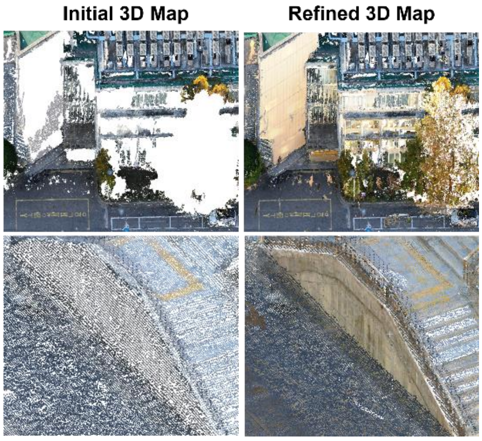

Figure 7. Qualitative Result of Registration

4.2 Evaluation of Object Tracking

Figure 8 illustrates the qualitative results of object tracking. The color gradient represents the progression of timestamps, with red indicating earlier timestamps and yellow indicating later ones. Ray marching exhibited a few errors near the bottom edge, while homography showed some errors along the top edge. Both the ray marching and homography methods, however, generally produced object trajectories that were closely aligned with the yellow rectangular ground truth path. In contrast, the depth estimation model demonstrated a trajectory that was relatively more deviated from the ground truth.

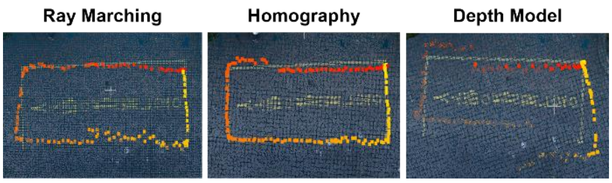

Figure 8. Qualitative Result of Object Tracking

As shown in Table 2, all methods achieved a mean ATE within 0.6m, indicating estimation errors in the tens-of-centimeters range. Among them, ray marching demonstrated the highest accuracy. Homography showed larger errors, likely due to its assumption of a perfectly planar ground surface, which does not fully reflect real-world conditions. Ray marching, by directly interacting with the refined digital map, provides more precise estimation, leading to lower errors. Meanwhile, the depth model showed relatively higher estimation errors, highlighting the limitations of single-camera-based depth estimation models, particularly when no elaborate finetuning is performed for the specific site.

Table 2 ATE (meters) for 3D Object Tracking Results

	Ray Marching	Homography	Depth Model
Mean	**0.204**	0.367	0.595
RMSE	**0.285**	0.382	0.717

Figure 9 presents the results of multi-view fusion. Blue dots represent object positions from CAM0, red dots correspond to those from CAM1, and green dots indicate predicted positions due to tracking failure. Before applying multi-view fusion, overlapping areas contained redundant red and blue dots. After fusion, these redundancies were eliminated by selecting the point with the lower error. Additionally, green dots were appropriately placed to compensate tracking failures, ensuring continuity in the object's trajectory. In summary, multi-view fusion effectively mitigates camera tracking failures, enabling consistent and robust object tracking.

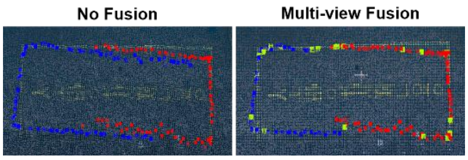

Figure 9. Validation Result of Multi-view Fusion

5 Limitations and Future Study

The experiment in this study was conducted in a controlled environment to accurately validate and compare the proposed methodology. However, real-world construction sites involve multiple objects and frequent occlusions, posing additional challenges in object tracking. As illustrated in Figure 10, when large objects obstruct the camera's view, ray marching may collide with obstacles before reaching the ground, resulting in inaccurate positioning. Furthermore, when multiple objects move in close proximity, mutual occlusion disrupts 2D tracking. While multi-view fusion helps mitigate tracking failures when objects are spaced apart, it becomes less effective in cases of close overlap, as shown in the right image of Figure 10, leading to track matching failures.

To address these challenges and enhance scalability, further enhancements are needed, such as integrating a depth estimation model for robust and seamless object tracking, along with further validation across diverse construction site scenarios.

Figure 10. Challenging Cases in Real-world Construction Sites

6 Conclusion

This study integrated commonly used vision platforms in construction sites—UAVs, TLS, and CCTV—to address their individual limitations. The proposed framework generated a unified 3D digital map, enabling seamless information sharing across platforms. A UAV provided an overall 3D digital map, while an alignment algorithm incorporating TLS enhanced resolution and vertical details. The refined map also served as a foundation for a robust CCTV-based 3D object tracking system, integrating dynamic object information into the 3D map.

Specifically, this study proposed a method to automatically align UAV and TLS maps using GPS, FPFH, and ICP in a step-by-step manner, effectively reducing computational costs for large-scale datasets and mitigating characteristic differences between heterogeneous data. In addition, the use of the ray marching technique and multi-view fusion enabled interaction with the refined 3D digital map, improving the robustness of CCTV-based 3D object tracking and ensuring consistency across cameras. The first experimental result demonstrated that the combination of FPFH and ICP significantly improved alignment accuracy compared to single alignment algorithms. The result of another experiment showed that the ray marching method achieved the highest accuracy in 3D object tracking.

In conclusion, this study highlights the significance of integrating vision platforms through a unified 3D digital map, improving the usability of site monitoring data. Nevertheless, challenges remain for the future research. A more robust object tracking system is needed to handle occlusions in complex construction environments, and further validation is required through real-world construction sites testing.

Acknowledgment

This research was conducted with the support of the

"National R&D Project for Smart Construction Technology (RS-2020-KA156875)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land Infrastructure and Transport. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00241758).

References

[1] Z. Ma and S. Liu. A review of 3D reconstruction techniques in civil engineering and their applications. Advanced Engineering Informatics, 37:163–174, 2018.

[2] B. Draeyer and C. Strecha. White paper: How accurate are UAV surveying methods. Pix4D White Paper, 4, 2014.

[3] M. Zhong, L. Sui, Z. Wang, and D. Hu. Pavement crack detection from mobile laser scanning point clouds using a time grid. Sensors (Switzerland), 20(15):1–20, 2020.

[4] E. B. Anil, P. Tang, B. Akinci, and D. Huber. Deviation analysis method for the assessment of the quality of the as-is building information models generated from point cloud data. Automation in Construction, 35:507–516, 2013.

[5] Q. Wang, J. Guo, and M.-K. Kim. An application-oriented Scan-to-BIM framework. Remote Sensing, 11(3):365, 2019.

[6] C. Wu, Y. Yuan, Y. Tang, and B. Tian. Application of terrestrial laser scanning (TLS) in the architecture, engineering and construction (AEC) industry. Sensors, 22(1):265, 2022.

[7] X. M. Garcia-Cruz, O. Y. Sergiyenko, V. Tyrsa, M. Rivas-Lopez, D. Hernandez-Balbuena, J. C. Rodriguez-Quiñonez, and P. Mercorelli. Optimization of 3D laser scanning speed by use of combined variable step. Optics and Lasers in Engineering, 54:141–151, 2014.

[8] X. Yang, et al. Computer-aided optimization of surveillance cameras placement on construction sites. Computer-Aided Civil and Infrastructure Engineering, 33(12):1110–1126, 2018.

[9] J. Seong, H.-S. Kim, and H.-J. Jung. The detection system for a danger state of a collision between construction equipment and workers using fixed CCTV on construction sites. Sensors, 23(20):8371, 2023.

[10] J. Kim and S. Chi. Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. Automation in Construction, 104:255–264, 2019.

[11] Q. Wang, H. Liu, W. Peng, C. Tian, and C. Li. A vision-based approach for detecting occluded objects in construction sites. Neural Computing and Applications, 36(18):10825–10837, 2024.

[12] S. Rusinkiewicz and M. Levoy. Efficient variants of the ICP algorithm. Proceedings third international conference on 3-D digital imaging and modelling, 145-152, IEEE, 2001.

[13] X. Tong, X. Liu, P. Chen, S. Liu, K. Luan, L. Li, S. Liu, X. Liu, H. Xie, Y. Jin, and Z. Hong. Integration of UAV-based photogrammetry and terrestrial laser scanning for the three-dimensional mapping and monitoring of open-pit mine areas. Remote Sensing, 7(6), 6635–6662, 2015.

[14] I. Aicardi, P. Dabove, A. M. Lingua, and M. Piras. Integration between TLS and UAV photogrammetry techniques for forestry applications. IForest, 10(1), 41–47, 2017.

[15] J. Šašak, M. Gallay, J. Kaňuk, J. Hofierka, and J. Minár. Combined use of terrestrial laser scanning and UAV photogrammetry in mapping alpine terrain. Remote Sensing, 11(18), 2019.

[16] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3):1623–1637, 2020.

[17] J. L. Schonberger and J. M. Frahm. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4104–4113, 2016.

[18] S. Yao, H. AliAkbarpour, G. Seetharaman, and K. Palaniappan. 3D patch-based multi-view stereo for high-resolution imagery. In Proceedings of Geospatial Informatics, Motion Imagery, and Network Analytics VIII, 2018.

[19] J. A. Hesch and S. I. Roumeliotis. A direct least-squares (DLS) method for PnP. In Proceedings of the International Conference on Computer Vision, pages 383–390, 2011.

[20] A. Ranganathan. The Levenberg-Marquardt algorithm. Tutorial on LM Algorithm, 11(1):101–110, 2004.

[21] N. Aharon, R. Orfaig, and B. Z. Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv Preprint arXiv:2206.14651, 2022.