Structural Features as Fiducial Marker Replacements for Multi-UAV Camera Pose Estimation in Construction Sites

Ayenew Yihune Demeke, Youjin Jang, and Inbae Jeong

Abstract— **Using multiple unmanned aerial vehicles (UAVs) for 3D monitoring of workers, equipment, and activities on construction sites requires the pose of every camera to be estimated in a shared metric frame. Fiducial markers provide such a reference, although markers observable from aerial viewpoints can be impractically large and require installation and maintenance. This paper investigates structural features with known 3D locations as marker replacements. Structural landmarks are detected in UAV images, associated with their corresponding model locations from their spatial layout, and used to estimate camera poses in the shared project frame. A four-UAV feasibility study was conducted in Webots to examine exterior slab corners, stair-opening corners, and column tops as structural references with distinct spatial configurations. Their respective median translation errors were 0.185 m, 1.192 m, and 0.163 m, while the corresponding rotation errors were 0.615°, 2.664°, and 0.470°. The slab corners and column tops provided lower median pose errors than the 1.0 m ArUco baseline, while all three structural references were substantially more accurate than the 0.5 m baseline. These findings establish the feasibility of using model-referenced structural features as marker-free metric anchors for multi-UAV camera localization, provided that their 3D locations are available.**

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) are routinely used on construction sites for surveying, progress documentation, and inspection because they enable rapid, repeated, and site-wide data collection while reducing the need for manual access. These capabilities also make them promising platforms for real-time site awareness and safety monitoring through the 3D tracking of workers, equipment, and site assets. Such information could support activity analysis, ergonomic assessment, and proximity warnings. Because monocular observations are affected by depth ambiguity, self-occlusion, and occlusion by surrounding site elements, reliable 3D pose estimation generally requires simultaneous observations from multiple cameras [1], [2]. The pose of each camera must therefore be continuously estimated in a shared coordinate frame as the UAVs move.

Existing pose-estimation approaches do not fully meet this requirement under construction-site conditions. Visual-inertial odometry tracks platform motion but accumulates drift and expresses pose in a platform-specific frame [3], [4]. Onboard sensors such as GNSS and inertial sensors can provide estimates of UAV position, attitude, and camera orientation; however, these estimates may not provide sufficient accuracy for precise camera localization, and GNSS accuracy may further deteriorate near structures because of occlusion and multipath [4]. An external visual reference can therefore provide additional geometric constraints needed to refine these onboard estimates and accurately anchor the cameras in a shared metric frame.

Fiducial markers provide such a reference by allowing camera pose to be recovered from a coded pattern of known size [5]. However, markers visible from UAV altitudes may need to be impractically large and must be installed, surveyed, and maintained as the site changes. They can also be obscured by dust, fresh concrete, or stored materials. Ground control points face similar limitations and are primarily intended for periodic surveys rather than continuous observation on cluttered construction sites from dynamically changing viewpoints.

Marker-free approaches offer two principal alternatives. The first estimates camera poses from the body joints of observed people, as demonstrated for UAV teams [6], [7] and ground cameras [8]–[10]. Because a person's dimensions are unknown, these methods recover scale only approximately [2] and produce a reference frame tied to the person, which becomes unavailable when no person is visible. The second estimates camera poses from anonymous scene features, such as SIFT or ORB, within structure-from-motion (SfM) or simultaneous localization and mapping (SLAM) pipelines, including multi-UAV systems [3], [11]. These approaches are effective when viewpoint changes between images are limited, but traditional feature descriptors become less reliable under the large viewpoint differences that can occur between spatially distributed cameras [12]. They also provide no fixed metric scale by themselves and depend on natural background texture, limiting their reliability on repetitive, dusty, or weakly textured construction surfaces [11].

Construction sites offer a distinct advantage because many site features have predefined geometry and semantics available from a building information model (BIM) or catalog. Structural elements such as columns, walls, slabs, and openings can provide site-based visual references, while standardized temporary elements such as formwork panels may offer additional references. Together, they form a distributed set of known geometries created by the construction process without separate installation or maintenance. Model-specified features have previously supported localization in construction applications. A markerless BIM registration for indoor augmented-reality inspection using an RGB-D camera was proposed in [13]. A BIM-derived object-landmark dictionary, RGB-D and inertial measurements, and pose-graph optimization were used for indoor relocalization in [14].

The authors would like to acknowledge financial support from ND EPSCoR for this research.

A. Y. Demeke is with the Department of Civil, Construction and Environmental Engineering, North Dakota State University, Fargo, ND 58108 USA (e-mail: ayenew.demeke@ndsu.edu).

Y. Jang is with the Department of Civil, Construction and Environmental Engineering, North Dakota State University, Fargo, ND 58108 USA (e-mail: y.jang@ndsu.edu).

I. Jeong is with the Department of Mechanical Engineering, North Dakota State University, Fargo, ND 58108 USA (corresponding author; e-mail: inbae.jeong@ndsu.edu).

While prior work has established the value of model-referenced scene geometry for localization, this paper examines a distinct setting in which structural landmarks available at different construction stages are detected in aerial RGB images, identified from their spatial layout, and used to estimate multiple camera poses in a common project frame. It makes two contributions. First, it presents a method that detects structural landmarks, associates them with their model elements using their spatial layout and an approximate onboard pose, and estimates each camera pose in the shared project frame. Second, it evaluates three structural landmark types in a four-UAV Webots feasibility study and compares their pose accuracy with large ArUco markers observed from the same viewpoints.

II. Methodology

The proposed approach uses structural features whose geometry and location are available from BIM as visual references for estimating UAV camera poses in a shared construction-site coordinate frame. The method consists of three stages: selecting suitable structural features, detecting landmarks on those features and associating them with BIM elements, and estimating the camera pose from the resulting 2D–3D correspondences.

A. Candidate Structural Features

The usefulness of a fiducial marker arises mainly from its known dimensions and unique identity. Its known dimensions provide metric scale, while its identity distinguishes one marker from another and indicates which marker has been observed. A construction feature intended to replace a marker must provide corresponding geometric and identity information. It must also remain fixed so that its position in the project coordinate frame remains valid. If it moves, its updated position must be known. Its construction-stage availability determines when and for how long it can serve as a reference, while its aerial visibility determines whether it can be reliably detected by the UAVs.

Based on these considerations, this study investigates column tops, exterior slab corners, and stair-opening corners as potential marker replacements. Their geometry and positions are specified in the BIM, and each remains fixed after the corresponding structural element is constructed. These features become available at different construction stages and provide reference configurations with different spatial extents and distributions. Reference features are selected for each stage, allowing the reference set to adapt as previously exposed features become concealed.

Determining the identity of each observed feature presents a separate challenge. Fiducial markers distinguish individual instances through unique binary codes, whereas structural elements of the same type often have similar appearances. A column or opening may therefore be readily detected without being distinguishable from other instances in the image. Feature descriptors such as SIFT offer one possible solution, but their reliability decreases under large viewpoint differences, as discussed in Section I. The proposed approach instead infers identity from the overall spatial layout of the detections relative to the BIM. A vision model first detects anonymous structural landmarks. These landmarks are then associated with model landmarks projected using an approximate onboard pose obtained from GNSS and attitude estimates. The resulting 2D–3D correspondences are used to estimate the corrected camera pose in the shared project frame.

B. Structural Landmark Detection

After a structural feature type is selected, each UAV image is processed to locate its visible instances as 2D landmarks. Each landmark is defined consistently with its model geometry. Column tops are represented by the centers of their exposed upper surfaces, while slab and stair-opening landmarks are represented by their corners. This consistency ensures that each detected image point can later be paired with a geometrically equivalent 3D location in the BIM.

Detection is performed at the feature-type level rather than the individual-element level. The detector locates column tops, slab corners, or stair-opening corners without determining which specific element has been observed. Individual identities are instead established from the spatial layout during the subsequent association stage. A keypoint detector trained on construction imagery could perform this feature-type detection in a real-time implementation.

C. BIM-Based Landmark Association

The detected landmarks identify the visible structural feature type but do not indicate which individual structural elements produced them. Their identities are determined by evaluating the possible assignments of the detected landmarks against the known spatial arrangement of the corresponding features in the building model. For each UAV image, the detector returns N image landmarks $\mathbf{u_i} = [u_i, v_i]^T$, where $i = 1, \dots, N$ identifies a detected landmark within the image. The BIM provides M candidate structural landmarks $\mathbf{X_j} = \left[X_j, Y_j, Z_j\right]^T$, where $j = 1, \dots, M$ identifies a 3D landmark expressed in the shared BIM coordinate frame.

An approximate camera pose is obtained from the onboard sensor position and attitude estimates. The candidate BIM landmarks are projected into the image using this pose to predict their spatial arrangement. The projection of BIM landmark $\mathbf{X_j}$ is expressed using the standard pinhole-camera model as

$$\hat{\mathbf{u}}_\mathbf{j} = \pi\left(\mathbf{K}\big(\tilde{\mathbf{R}}\mathbf{X_j} + \tilde{\mathbf{t}}\big)\right) \tag{1}$$

where $\mathbf{K}$ is the camera-intrinsic matrix, $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{t}}$ are the approximate camera rotation and translation, and $\hat{\mathbf{u}}_\mathbf{j}$ is the predicted pixel location of BIM landmark $\mathbf{X_j}$. The function $\pi(\cdot)$ converts the projected 3D point into pixel coordinates through perspective division.

Candidate associations are evaluated according to the agreement between the detected and projected landmark layouts. Each candidate association is represented by σ, where $\sigma(i)$ identifies the BIM landmark corresponding to detection $\mathbf{u}_\mathrm{i}$. The set $\mathcal{S}$ represents all feasible one-to-one associations, while σ represents one candidate association within that set. For N detected landmarks and M candidate model landmarks, the association space contains $M!/(M-N)!$ possible one-to-one associations when $N \leq M$ and reduces to $N!$ when the two sets have equal size. The minimum-cost association is obtained using a modified linear assignment algorithm [15], without explicitly enumerating these possibilities. The cost of each association is calculated as the sum of the squared pixel distances between the detected landmarks and their corresponding model projections. The association with the minimum total cost is selected as

$$\sigma^* = \arg\min_{\sigma \in \mathcal{S}} \sum_{i=1}^{N} \left\| \mathbf{u_i} - \hat{\mathbf{u}}_{\boldsymbol{\sigma}(\mathbf{i})} \right\|^2 \qquad (2)$$

where σ^* denotes the selected association. Evaluating the complete landmark layout allows repeated structural elements to be distinguished from their spatial arrangement, and the resulting 2D–3D correspondences are used to estimate the corrected camera pose.

D. Camera Pose Estimation

After association, the resulting 2D–3D correspondences are used to refine the camera pose. The approximate onboard pose is used only during landmark association. The final rotation and translation are estimated by minimizing the reprojection error of the established correspondences.

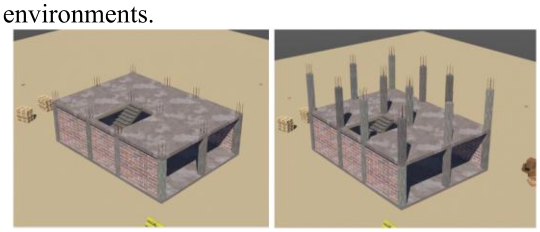
$$(\mathbf{R}^*, \mathbf{t}^*) = \arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} \left\| \mathbf{u_i} - \pi\left(\mathbf{K}\left(\mathbf{R}\mathbf{X}_{\boldsymbol{\sigma}^*(\mathbf{i})} + \mathbf{t}\right)\right) \right\|^2 \qquad (3)$$

where $\mathbf{X}_{\boldsymbol{\sigma}^*(\mathbf{i})}$ is the known 3D model landmark associated with image landmark $\mathbf{u_i}$, and $\mathbf{R}^*$ and $\mathbf{t}^*$ are the rotation and translation that minimize the total reprojection error. These parameters transform points from the shared project coordinate frame to the camera coordinate frame. The corresponding camera position in the project frame is obtained as $\mathbf{C}^* = -\mathbf{R}^{*\mathbf{T}}\mathbf{t}^*$, while the camera orientation in that frame is $\mathbf{R}^{*\mathbf{T}}$. The planar configurations use Infinitesimal Plane-Based Pose Estimation (IPPE) [16], retaining the positive-depth candidate with the lowest reprojection RMSE and using the approximate camera position when the preceding criteria do not yield a unique solution.

III. Simulation Study

The feasibility study was conducted in Webots [17], which provided a controlled multi-UAV environment and ground-truth structural-landmark coordinates and camera poses. The simulated reinforced-concrete building had a 3×4 column grid, with 4 m spans in one direction and spans of 4 m, 3 m, and 4 m in the other. Randomized, nonrepeating concrete textures improved visual realism and avoided artificially identical surfaces.

Two environments represented different construction stages. Before the upper-storey columns were cast, exterior slab and stair-opening corners served as landmarks. After casting, the exposed column tops were used. Fig. 1 shows both environments.



a b

Fig. 1. Simulation environment. (a) Slab and stair-opening landmarks. (b) Column-top landmarks.

The evaluated sets comprised four exterior slab corners, four corners of a 2.6 m by 3.6 m stair opening, and 12 column tops. Four DJI Mavic 2 Pro UAVs observed the site from its four sides, each capturing ten viewpoints and producing 40 images per landmark type. The cameras had a resolution of 1280 by 720 pixels, with intrinsic parameters $f_x = f_y =$ 1108.5 pixels, $c_x = 640$ pixels, and $c_y = 360$ pixels. Each UAV camera was localized independently using the same model-referenced landmark map, and no inter-UAV measurements or cooperative pose optimization were used. The resulting poses were nevertheless expressed in a common project frame, enabling subsequent multi-view use.

Approximate onboard poses were simulated by perturbing ground-truth positions by 3 m along each axis and attitudes by 2° in roll and pitch and 5° in yaw, with signs alternated across UAVs and viewpoints to reflect practical small-UAV pose uncertainty [18] and potential GNSS uncertainty near structures. Structural landmarks were detected using the vision-capable GPT-5.6 Sol model through the OpenAI API with low reasoning effort and original image detail. The model was used only for initial anonymous landmark detection, allowing evaluation of the subsequent association and pose-estimation stages.

The baseline used eight DICT_4X4_50 ArUco markers placed approximately 3 m from the building, with two on each side. Side lengths of 0.5 and 1.0 m were evaluated from the same viewpoints, and pose estimates from multiple detected markers were averaged.

IV. Results and Discussion

Table I summarizes structural-landmark detection and association performance, where association is the percentage of detected image landmarks matched to the correct BIM landmarks. All three feature types were detected reliably, with slab and opening corners achieving perfect association, while column tops showed slightly lower association accuracy and higher latency.

TABLE I. Structural-landmark detection and association results.

Landmark type	**Detection F1 (%)**	**Association (%)**	**Latency (s)**
Slab corners	99.10	100.00	13.28
Opening corners	96.88	100.00	13.91
Column tops	95.65	94.21	27.13

Fig. 2 illustrates representative landmark associations and final-pose reprojections. The final reprojections align closely with the observed slab corners and column tops, while the more compact stair-opening layout provides less geometric support for pose estimation.

Table II shows that column tops achieved the lowest median pose errors, while slab corners produced similarly low medians and tighter P90 errors. The stair-opening corners were less accurate, indicating that landmark spatial distribution is important in addition to detection and association quality. The 0.5 m ArUco markers were evaluated first; however, their small size made corner detection unreliable at the tested viewing distances, leading to the additional evaluation of the 1.0 m markers as a more representative baseline. Compared with ArUco, slab corners and column tops outperformed the 1.0 m markers in median translation and rotation error, while all three structural references substantially outperformed the 0.5 m markers.

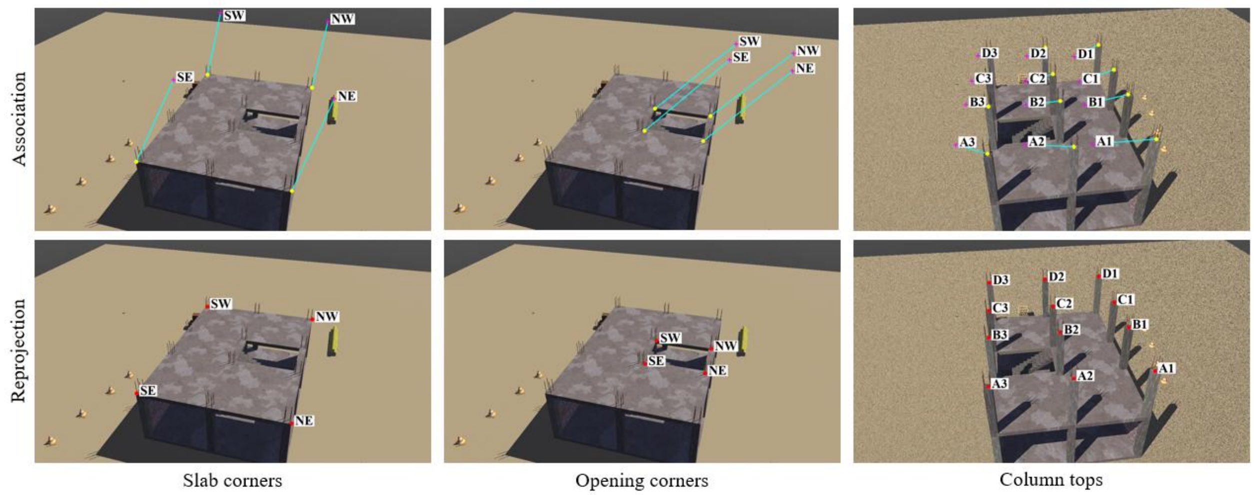


Fig. 2. Landmark associations and final-pose reprojections.

TABLE II. Camera-pose accuracy relative to ground truth.

Method (successful poses*)	**Translation error, median / P90 (m)**	**Rotation error, median / P90 (°)**
Slab corners (39/40)	0.185 / 0.363	0.615 / 1.142
Opening corners (39/40)	1.192 / 2.607	2.664 / 6.220
Column tops (40/40)	0.163 / 3.323	0.470 / 11.503
ArUco, 0.5 m (39/40)	14.929 / 35.958	35.862 / 93.235
ArUco, 1.0 m (38/40)	0.658 / 1.533	0.698 / 1.631

* Successful poses are valid estimates obtained from at least four structural-feature correspondences or at least one ArUco marker.

V. Conclusion

This paper investigated model-referenced structural features as marker replacements for multi-UAV camera localization on construction sites. The proposed method detects structural landmarks, determines their model correspondences from the observed spatial layout, and estimates camera poses in a shared project frame. In the Webots study, column tops and exterior slab corners achieved median translation errors of 0.163 m and 0.185 m and rotation errors of 0.470° and 0.615°, respectively, with lower median pose errors than the 1.0 m ArUco baseline. The stair-opening corners were less accurate because of their limited spatial extent. The vision model used for landmark detection may be suitable for applications requiring only periodic pose updates. Continuous real-time operation, however, will require a dedicated detector trained to identify the selected structural landmarks. The current evaluation also does not examine robustness to construction tolerances or the varied occlusions encountered on active sites. Future work will develop such a detector and validate the complete method under these conditions on real construction sites.

References

[1] B. Jiang, L. Hu, and S. Xia, “Probabilistic triangulation for uncalibrated multi-view 3D human pose estimation,” in Proc. IEEE/CVF ICCV, 2023.

[2] L. Müller, H. Choi, A. Zhang, B. Yi, J. Malik, and A. Kanazawa, "Reconstructing people, places, and cameras," in Proc. IEEE/CVF CVPR, 2025.

[3] L. Heng, G. H. Lee, and M. Pollefeys, "Self-calibration and visual SLAM with a multi-camera system on a micro aerial vehicle," Auton. Robots, vol. 39, no. 3, pp. 259–277, 2015.

[4] J. Hartzer and S. Saripalli, "Online multi camera-IMU calibration," in Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics (SSRR), Sevilla, Spain, 2022, pp. 360–365.

[5] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," Pattern Recognition, vol. 47, no. 6, pp. 2280–2292, 2014.

[6] N. Saini, E. Price, R. Tallamraju, R. Enficiaud, R. Ludwig, I. Martinovic, A. Ahmad, and M. J. Black, "Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles," in Proc. IEEE/CVF ICCV, 2019.

[7] N. Saini, E. Bonetto, E. Price, A. Ahmad, and M. J. Black, "AirPose: Multi-view fusion network for aerial 3D human pose and shape estimation," IEEE Robot. Autom. Lett., vol. 7, no. 2, pp. 4805–4812, 2022.

[8] K. Takahashi, D. Mikami, M. Isogawa, and H. Kimata, "Human pose as calibration pattern: 3D human pose estimation with multiple unsynchronized and uncalibrated cameras," in Proc. IEEE/CVF CVPR Workshops, 2018.

[9] S.-E. Lee, K. Shibata, S. Nonaka, S. Nobuhara, and K. Nishino, "Extrinsic camera calibration from a moving person," IEEE Robot. Autom. Lett., vol. 7, no. 4, pp. 10344–10351, 2022.

[10] B. Huang, Y. Shu, T. Zhang, and Y. Wang, "Dynamic multi-person mesh recovery from uncalibrated multi-view cameras," in Proc. Int. Conf. 3D Vision (3DV), 2021.

[11] S. Zhuge, Y. He, X. Xu, S. Gan, C. Li, B. Lin, X. Yang, and X. Zhang, "Markerless motion capture for humans through a multi-UAV system," IEEE Trans. Instrum. Meas., vol. 73, 2024.

[12] U. S. Parihar, A. Gujarathi, K. Mehta, S. Tourani, S. Garg, M. Milford, and K. M. Krishna, "RoRD: Rotation-robust descriptors and orthographic views for local feature matching," in Proc. IEEE/RSJ IROS, 2021, pp. 1593–1600.

[13] M. Kopsida and I. Brilakis, "Markerless BIM registration for mobile augmented reality based inspection," in Proc. ICCCBE, 2016.

[14] J. C. P. Cheng, C. Song, X. Zhang, and Z. Chen, "Pose graph relocalization with deep object detection and BIM-supported object landmark dictionary," J. Comput. Civ. Eng. (ASCE), vol. 37, no. 5, 2023.

[15] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," Computing, vol. 38, no. 4, pp. 325–340, 1987.

[16] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," Int. J. Comput. Vis., vol. 109, no. 3, pp. 252–286, 2014.

[17] O. Michel, "Webots: Professional mobile robot simulation," Int. J. Adv. Robotic Syst., vol. 1, no. 1, pp. 39–42, 2004.

[18] A. R. Vetrella, G. Fasano, and D. Accardo, “Attitude estimation for cooperating UAVs based on tight integration of GNSS and vision measurements,” Aerosp. Sci. Technol., vol. 84, pp. 966–979, 2019.