OPEN ACCESS

Artificial Intelligence ARTICLE

The Innovation Informatics

A fully autonomous UAV system for process-aware construction monitoring via few-shot learning

Yiquan Zou,[1] Tianxiang Liang,[2,*] Zhendong Xu,[3] Mudan Wang,[4] Ioannis Brilakis,[4] and Biao Xiong[3,5,*]

*Correspondence: liangliangtovo@gmail.com (T.L.); b.xiong@whut.edu.cn (B.X.)

Received: November 10, 2025; Accepted: October 30, 2026; Published Online: October 30, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100073

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- A new paradigm for construction progress sensing driven by few-shot learning.
- Construction progress monitoring still relies heavily on manual inspection and site observations.
- We develop a fully autonomous UAV system for process-aware monitoring using few-shot learning.
- The ViT-based framework integrates viewpoint classification, ROI detection, and process recognition.
- Real-world deployment achieves accurate monitoring and enables data-driven construction progress management.

ARTICLE Artificial Intelligence

OPEN ACCESS

A fully autonomous UAV system for process-aware construction monitoring via few-shot learning

Yiquan Zou,[1] Tianxiang Liang,[2,*] Zhendong Xu,[3] Mudan Wang,[4] Ioannis Brilakis,[4] and Biao Xiong[3,5,*]

[1]School of Civil Engineering, Architecture and The Environment, Hubei University of Technology, Wuhan 430068, China
[2]School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[3]China Construction Third Engineering Bureau Technology Innovation Industry Development Co., Ltd., Wuhan 430050, China
[4]Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom
[5]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China
*Correspondence: liangliangtovo@gmail.com (T.L.); b.xiong@whut.edu.cn (B.X.)

Received: November 10, 2025; Accepted: October 30, 2026; Published Online: October 30, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100073

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Zou Y., Liang T., Xu Z., et al. (2026). A fully autonomous UAV system for process-aware construction monitoring via few-shot learning. *The Innovation Informatics* **2**:100073.

Construction progress monitoring is critical for schedule control and project management, yet reliable site perception remains challenging due to dynamic environments, limited annotated data, and the high cost of manual inspection. Conventional rule-based or geometry-driven methods often lack robustness under complex site conditions, while fully supervised deep learning approaches require large labeled datasets that are rarely available in practice. To address this challenge, we develop a fully autonomous UAV system for process-aware construction monitoring via few-shot learning. The system integrates a dock-enabled UAV platform that autonomously performs takeoff, landing, charging, and data transmission, enabling persistent and repeatable aerial observation without on-site human intervention. UAV imagery is analyzed through a multi-stage recognition framework that combines camera viewpoint classification, region-of-interest detection, and process recognition within a Vision Transformer–Prototypical Network architecture. A daily multi-view fusion strategy further aggregates complementary observations from multiple viewpoints to produce stable progress estimates. The system is deployed on a 32-story residential project and evaluated using a five-month multi-view UAV dataset. With only limited annotations from early construction stages, the few-shot framework generalizes effectively to unseen floors, achieving 95.62% daily-level accuracy and a 93.83% F1-score after multi-view fusion across four representative workflows. During deployment, comparison of actual progress with the baseline schedule indicated a reduction of 21 working days in schedule delay, equivalent to approximately 14% of the planned project duration.

INTRODUCTION

Construction progress monitoring is central to schedule control, risk mitigation, and site management. Yet, on many high-rise projects, managers still rely on manual photo collection, chat groups, and subjective visual inspection to infer which construction processes are underway on each floor. Here, a construction process denotes a predefined construction operation performed within a specific floor or work zone, rather than a generic site activity, a broad construction stage, or a percentage of project completion. Process-level recognition classifies UAV observations into these categories, whereas process-aware monitoring integrates the results across space and time against the planned construction sequence. Linked with project schedules, these results can support automated progress updates, deviation detection, delay warnings, resource planning, and traceable site records for quality verification. Because construction processes are the fundamental units of progress and quality management, their accurate recognition is essential for reliable status perception, early risk identification, and evidence-based decision-making.[1] However, manual interpretation is labor-intensive, inconsistent across inspectors, and difficult to scale to the high-frequency, multi-level monitoring demanded by modern projects.[2] This gap between the complexity of contemporary construction and the limited observability provided by current practice has become a major barrier to digital and intelligent site management.

Unmanned aerial vehicles (UAVs) offer a promising route to improve this situation. Owing to their mobility, affordability, and broad coverage, UAVs can efficiently capture rich visual information from complex outdoor environments and tall structures.[3,4] Combined with advances in computer vision, they have supported automated inspection, safety monitoring, productivity assessment, and infrastructure documentation.[5,6,7] Multi-camera and video-based systems have further improved activity analysis,[8,9] while the integration of BIM and point cloud data has enabled more structured, model-based monitoring.[10,11] Despite this progress, most existing approaches remain focused on object detection, instance segmentation, or coarse progress comparison.[12,13] As a result, they often fail to provide the process-level understanding needed for accurate and timely progress assessment.[14,15] The key practical question therefore remains insufficiently addressed: what construction process is occurring at a given location, and how does it compare with the planned sequence?

Answering this question directly from UAV imagery is challenging. Construction sites are highly dynamic: viewpoints drift with flight deviations, floor geometry evolves, temporary facilities and materials move, and illumination changes from day to day.[16,17] Even the same nominal process can appear visually different across projects because of variations in construction methods, component layouts, scaffold configurations, and surrounding context.[5,17] This strong scene dependency limits the transferability of conventional deep learning models trained on a single project or site. At the same time, process-level annotation requires expert judgment and is expensive to obtain at scale.[2,18] Consequently, many existing monitoring systems rely on proxy indicators, such as equipment activity[19,20] or material presence, or are tuned to a single project and remain difficult to generalize.[21–23]

Few-shot learning (FSL) provides a natural way to address these challenges.[24] By learning representations that can recognize new categories from only a few labeled examples, FSL is well suited to settings characterized by annotation scarcity and high visual variability.[25,26] Prototype-based methods such as Prototypical Networks[27,28] are especially attractive because they are simple, interpretable, and effective under limited supervision.[29,30] Although few-shot approaches have shown promise in remote sensing, industrial monitoring, and several construction tasks such as hazard identification and irregular opening detection,[31,32] their use in process-level construction monitoring remains largely unexplored.[33,34] More importantly, existing studies typically rely on manually collected imagery and do not connect autonomous data acquisition with process recognition and schedule-aware progress analytics in a unified system.

To address this gap, we develop a fully autonomous UAV system for process-aware construction monitoring via few-shot learning. The system combines a dock-enabled UAV platform with a multi-stage few-shot visual intelligence framework for process-level recognition. The UAV autonomously executes daily missions, performs takeoff and landing, recharges, and uploads imagery without on-site human intervention. Building on this sensing infrastructure, we design a recognition framework that explicitly integrates three complementary elements of the monitoring task: camera viewpoint classification as a spatial prior, region-of-interest detection to suppress irrelevant background, and process recognition to infer ongoing construction activities under few-shot conditions. A daily multi-view fusion strategy then aggregates predictions from multiple viewpoints to generate stable and consistent progress estimates for each level.

We validate the system through a five-month deployment on an active 32-story residential project in Guangzhou, China. Using only limited annotations

from early floors, the proposed framework generalizes to unseen levels with evolving visual configurations, achieving **95.62% daily-level accuracy** and **93.83% F1-score** across four representative construction workflows. Integrated with the autonomous dock, the system produces continuous data-driven progress reports that helped recover **21 working days** and reduce delay-related losses by approximately **575,000 RMB**. These results demonstrate that combining autonomous UAV sensing with few-shot visual intelligence offers a practical and scalable pathway toward continuous, minimally supervised construction progress monitoring.

The main contributions of this work are as follows:

- **We frame construction progress monitoring as a few-shot, process-aware perception problem.** This formulation enables robust recognition under limited annotations, strong scene variability, and continuously evolving site conditions.
- **We develop a fully autonomous UAV-based framework for process-aware construction monitoring.** The dock-enabled system integrates autonomous flight, image acquisition, few-shot process recognition, and reporting into a unified fly–sense–reason–report pipeline.
- **We establish a five-month real-world multi-view UAV benchmark for process-aware monitoring.** The longitudinal dataset captures diverse viewpoints, workflows, and environmental conditions on a 32-story high-rise project.
- **We validate the framework through long-term real-world deployment.** The system generalizes across unseen floors and evolving construction states, achieving 95.62% daily-level accuracy while recovering 21 working days and reducing delay-related losses by approximately 575,000 RMB.

RELATED WORK

Evolution of vision-based construction monitoring

Early vision-based construction monitoring relied primarily on handcrafted visual descriptors, such as HOG, LBP, and color features, combined with conventional classifiers including SVM,[28] KNN,[29] and ANN[35] for detecting workers, equipment, and materials. In activity recognition, dense trajectory features coupled with SVMs, particularly HOG/HOF/MBH descriptors, improved robustness to occlusion and viewpoint variation.[9,36] However, these methods remained sensitive to illumination changes and typically required substantial manual annotation.

The rise of deep learning substantially improved the accuracy and robustness of construction vision systems. Multi-stream CNNs, CNN–LSTM architectures, and lightweight real-time networks enhanced spatio-temporal representation learning,[36,37] while HRFPN-based segmentation strengthened semantic scene understanding.[38] In parallel, photogrammetric point clouds[12] and video-based statistical analysis[10] were introduced for productivity monitoring, and the integration of BIM with point cloud data enabled semi-supervised classification and knowledge-driven reconstruction.[10]

More recently, deep learning has been widely adopted for safety and resource monitoring, including helmet detection,[39] intrusion management,[40] voxel-based 3D reconstruction, and YOLOv8-based resource identification.[41,42] Despite these advances, construction progress monitoring still depends heavily on manual assessment.[21] More objective approaches, such as BIM-based comparison between planned and as-built models[13,43,44] and vision-based updating of 4D CAD models,[45] have shown promise, but persistent challenges remain, including class imbalance, high intra-class variation, and limited annotated data. These limitations highlight the need for more data-efficient and generalizable monitoring frameworks.

FSL in data-scarce and heterogeneous scenarios

Few-shot learning (FSL) is designed to achieve generalization under extremely limited annotation.[24,46] Representative approaches include Prototypical Networks (ProtoNet),[27] Relation Networks,[47] and Matching Networks,[48] which commonly rely on metric learning between support and query samples.[49,50] Among these, ProtoNet has been widely adopted because of its simplicity, interpretability, and strong performance under limited supervision.[27] Earlier work on semi-supervised and metric-based learning also provided important foundations for the development of modern FSL methods.[28]

Recent extensions have further strengthened FSL under heterogeneous and cross-domain conditions. In remote sensing and hyperspectral imaging, Stable ProtoNets,[30] multi-stage relation modeling,[51] dual-metric learning,[52] and contrastive attention mechanisms[29] have improved robustness to distribution shifts. Hierarchical architectures such as HiReNet[53] and DM-MRN[51] have further enhanced generalization ability.[26,54] Beyond these domains, FSL has shown strong potential across computer vision, agriculture, medicine, and industrial monitoring,[25] including SAR target detection,[55] contrastive representation learning,[56] crop disease recognition with vision–language models,[57] soil classification through prototype refinement,[58] medical image segmentation,[59,60] and industrial fault diagnosis.[61]

In construction, however, site heterogeneity[62,63] and complex multi-scale backgrounds[64] continue to hinder robust feature learning. Recent studies nevertheless suggest the promise of FSL in this domain. Seo and Kim[32] reported over 90% accuracy for irregular opening detection using minimal samples, Kim and Chi[31] achieved strong performance with only 20 images, and Wang and El-Gohary[33] integrated FSL with attribute recognition for hazard monitoring. Zero-shot learning has also been explored for equipment task recognition,[34] although it still depends heavily on pretrained vision models. Collectively, these studies highlight the potential of FSL to alleviate annotation scarcity and scene variability, while also indicating that its application to process-level construction recognition remains largely underexplored.

Process and progress monitoring in construction

Although UAV-based vision for construction monitoring has advanced rapidly, most existing studies remain focused on productivity assessment, earthwork measurement, and safety monitoring rather than process-level understanding and quality evaluation.[14,19] Typical approaches rely on detection–tracking pipelines to monitor equipment, components, and workers.[15,38] More fine-grained efforts have concentrated on geometric progress comparison,[11] earthwork cycle analysis,[20] or the recognition of specific finishing activities such as masonry and tiling.[65] Recent work has also explored the integration of digital twin concepts into vision-based progress monitoring,[66] highlighting the potential for more closed-loop construction control.

Related progress-monitoring frameworks have been developed for precast wall installation[67] and soil-foundation construction,[68] achieving strong segmentation performance through instance segmentation and BIM-based mapping. At the same time, hybrid and self-supervised approaches such as ConPro-NET,[69] together with intelligent schedule evaluation systems for prefabricated construction,[70] reflect a broader shift toward label-efficient and context-aware monitoring.

Despite these advances, most UAV-based systems still rely on detection–segmentation–tracking–statistics pipelines, which provide productivity indicators but offer limited process-level recognition and temporal reasoning.[22]

Many supervised approaches also require substantial task-specific annotations and remain sensitive to complex and heterogeneous site conditions.[23]

Moreover, autonomous data acquisition and automated analysis are typically treated as separate stages, while viewpoint, process continuity, and recurring floor structures remain underused, limiting adaptation across floors.[33]

The present study is designed to address these gaps by integrating dock-enabled UAV acquisition, few-shot process recognition, and daily multi-view fusion within a unified process-aware monitoring workflow.

MATERIALS AND METHODS

Overall framework

To address process diversity, limited annotations, and complex acquisition conditions, we propose a multi-stage recognition framework based on UAV imagery (Figure 1). Each UAV image is divided into patches and encoded by a Vision Transformer (ViT)[71] to generate latent feature tokens shared across three tasks. These features are first routed in parallel to camera-position classification and region-of-interest (ROI) detection. The position branch compares the query representation with position prototypes and outputs a discrete camera-position label. In parallel, the ROI branch predicts bounding-box coordinates and a confidence score, which define the process-relevant image region. The predicted position is embedded by a multilayer perceptron, while the cropped ROI is re-encoded by a ViT. These representations are then

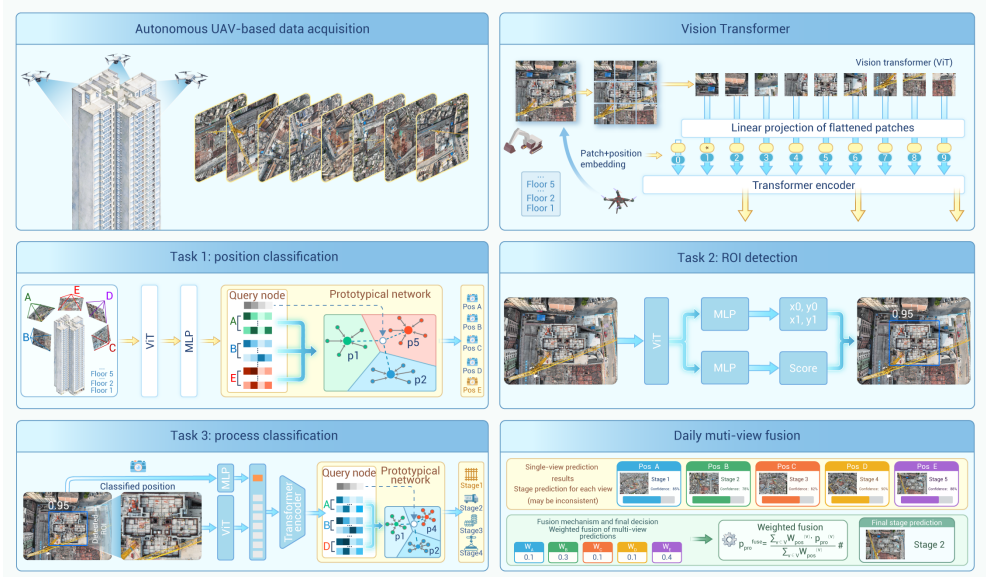

Figure 1. Overall framework for construction process recognition using FSL UAV images are first encoded by a ViT to obtain latent features, which are then utilized across three tasks: (1) ROI detection, (2) camera position classification, and (3) construction process classification. The detected ROI and position priors are integrated with the latent feature to enhance the robustness and accuracy of process recognition.

fused by a transformer encoder and passed to a Prototypical Network, which assigns the query image to a predefined construction process category. The resulting process predictions are subsequently aggregated across viewpoints through daily multi-view fusion to estimate the process status of each floor.

Camera position recognition

UAV images of the same construction process can exhibit substantially different visual feature distributions across viewpoints. We therefore treat camera position as a spatial prior for modeling viewpoint-dependent appearance, rather than as an indicator of construction process state. To address this issue, we employ a Vision Transformer (ViT)–based embedding extractor,[71] followed by a multilayer perceptron (MLP) projection layer and a Prototypical Network (ProtoNet)[27] classifier. Specifically, ViT-S/16 encodes each image into 384-dimensional latent representations through patch-based self-attention, effectively capturing global visual dependencies under varying viewpoints.

The extracted features are further projected into a compact 128-dimensional space using a two-layer MLP with ReLU activation. ProtoNet constructs class prototypes from $K_{pos} = 5$ support samples per position and classifies query images based on Euclidean distance. Five recurrent UAV positions, comprising three side views and two top views labeled A–E, are defined to represent different observation geometries rather than construction states. The discrete position label (A-E) is represented as a one-hot vector and mapped by a two-layer MLP (5*128*128 with ReLU). This design reduces cross-view ambiguity, provides reliable spatial priors, and improves the robustness of downstream process recognition under few-shot conditions.

ROI detection

UAV images of construction sites often contain substantial irrelevant background, including equipment, scaffolding, temporary facilities, and surrounding structures. These distractors can obscure process-relevant features and degrade recognition accuracy, particularly under multi-view conditions. To address this issue, we introduce a region-of-interest (ROI) detection module that localizes construction areas directly from UAV imagery. The module operates on shared ViT embeddings and predicts bounding boxes in the form of (x_0, y_0, x_1, y_1) along with a confidence score, ensuring that only semantically relevant regions are passed to the subsequent process recognition stage.

The ROI detection head is implemented as a lightweight two-layer MLP with a hidden dimension of 256 and ReLU activation, followed by a linear layer for joint prediction of bounding box coordinates and confidence scores. By leveraging adaptive feature learning rather than fixed cropping rules, the module adapts to variations in camera viewpoints, construction progress, and site layout. As a result, it produces stable and semantically focused regions across diverse acquisition conditions, providing reliable inputs for downstream process recognition.

Construction process recognition

Following position classification and ROI detection, the framework performs process recognition by integrating spatial context and focused visual features. Image patches within the detected ROIs are first encoded into 384-dimensional latent representations using ViT-S/16 embeddings. In parallel, the classified camera position is projected into a 128-dimensional embedding through a two-layer MLP with ReLU activation. These features are concatenated to form a 512-dimensional joint representation, which is further processed by a Transformer encoder with four layers and four attention heads to capture contextual dependencies between visual content and spatial priors.

The resulting representations are classified using a Prototypical Network (ProtoNet). Class prototypes are computed as the mean of $K_{pro} = 4$ support embeddings per class, and query samples are assigned based on Euclidean distance. The model recognizes four representative construction processes: slab rebar binding (Stage 1), concrete casting and curing (Stage 2), rebar column construction (Stage 3), and composite slab hoisting (Stage 4). By jointly leveraging position priors, ROI-focused features, and metric-based classification, the framework maintains stable decision boundaries under few-shot conditions, enabling accurate and robust process recognition in dynamic construction environments while significantly reducing annotation requirements.

Daily multi-view fusion

Since different areas of the same level may progress asynchronously, predictions from individual images can be inconsistent. To obtain a unified daily progress estimation, we fuse classification outputs from all captured views. For each image, the predicted class probability vector is $P_{Pro}^{(v)} \in \mathbb{R}^c$ (where $C = 4$ for Stage 1-4), with confidence implicitly encoded in its magnitude. Each viewpoint is further assigned a reliability weight $W_{pos}^{(v)}$. The fused probability distribution is computed as a weighted average:

$$p_{pro}^{fuse} = \frac{\sum_{v \in \mathcal{V}} W_{pos}^{(v)} \cdot p_{pro}^{(v)}}{\sum_{v \in \mathcal{V}} W_{pos}^{(v)}} \quad (1)$$

where $\mathcal{V} = \{A,B,C,D,E\}$ denotes the set of viewpoints. The final daily progress class is then determined as:

$$C_{pro}^{fuse} = \arg\max_{c \in \{1,\cdots,C\}} p_{pro,c}^{fuse}. \quad (2)$$

In deployment, the viewpoint weights were empirically assigned as $W_{pos} = \{1,3,1,1,4\}$ for Pos A–E, respectively, based on an initial assessment of observation stability, structural visibility, and susceptibility to occlusion. Pos E was assigned the highest weight because its top view provided broad and stable coverage of the working area. Pos B showed intermediate reliability, whereas Pos A, C, and D were more affected by occlusion or oblique perspective. The weights were fixed before evaluation and were not optimized using test-set labels. They therefore represent deployment-specific reliability coefficients rather than globally optimal parameters.



Figure 2. System context DJI Dock 2 deployment: (left) Dock 2 with M3TD UAV; (middle) charging station; (right) construction site overview.

Loss functions

The overall framework is trained end-to-end with a shared ViT encoder and three task-specific heads. Each training episode includes *support* and *query* sets for ProtoNet-based classification, along with per-image supervision for ROI detection. The total objective is a weighted sum of the three task losses:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{pos} + \beta\mathcal{L}_{roi} + \gamma\mathcal{L}_{proc} \tag{3}$$

where $\mathcal{L}_{pos}$, $\mathcal{L}_{roi}$, and $\mathcal{L}_{proc}$ present loss of camera position recognition, ROI detection, and process classification, respectively. And $(\alpha,\beta,\gamma) = (1,1,2)$ in our experiments to emphasize process recognition.

For **camera position recognition**, given position classes $\mathcal{C}_{pos}$ (A-E), prototypes are computed as the mean of $K_{pos} = 5$ support embeddings:

$$c_k = \frac{1}{|S_k|}\sum_{(x_i,y_i)\in S_k} f_{\theta(x_i)}, k \in C_{pos} \tag{4}$$

Query embeddings are classified by a temperature-scaled softmax over squared Euclidean distances with $\tau = 0.7$, and optimized via episodic negative log-likelihood:

$$\mathcal{L}_{pos} = -\frac{1}{|Q_{pos}|}\sum_{(x,y)\in Q_{pos}} logp_{\tau(y|x)} \tag{5}$$

For **ROI detection**, on top of the shared ViT embedding $\boldsymbol{z} = f_\theta(x)$, a two-layer MLP predicts one bounding box $\hat{\boldsymbol{b}} = (\hat{x}_0,\hat{y}_0,\hat{x}_1,\hat{y}_1)$ and confidence $\hat{s}$. With ground-truth bounding box $\boldsymbol{b}$ and label $s \in \{0,1\}$, the detection loss combines objectness (Binary Cross Entropy, BCE) and localization (GIoU loss):

$$\mathcal{L}_{roi} = \frac{1}{N}\sum_{n=1}^{N} BCE\left(\hat{s^{(n)}}, s^n\right) + \lambda_{box}\, GIoU\left(\hat{\boldsymbol{b}^{(n)}}, \boldsymbol{b}^n\right) \tag{6}$$

with $\lambda_{box} = 2.0$.

For **process classification**, ROI features are fused with position embeddings and passed through a Transformer encoder to produce process representations $\tilde{z}$. Class prototypes are computed from $K_{pro} = 4$ support samples:

$$\tilde{\boldsymbol{c}}_k = \frac{1}{|S_k|}\sum_{(x_i,y_i)\in S_k} \tilde{z}_i, k \in \mathcal{C}_{pro} \tag{7}$$

Query samples are classified via the same metric-based softmax as in Task 1 (with $\tau = 0.7$), and the loss is

$$\mathcal{L}_{proc} = -\frac{1}{|Q_{proc}|}\sum_{(x,y)\in Q_{proc}} logp_\tau(y|x) \tag{8}$$

This joint formulation integrates spatial priors (Task 1), region focusing (Task 2), and few-shot metric learning (Task 3) within a unified optimization framework, enabling accurate and robust process recognition under limited supervision.

EXPERIMENTS

Data and setup

Dataset. The objective of this study is to enable deployable construction process recognition directly from UAV imagery. Data were collected using a Matrice 3TD (M3TD) UAV equipped with a 5280×3956-pixel M3D camera and operated via the DJI Dock 2 system[72] on an active residential construction site in Guangzhou, China. The dock autonomously manages takeoff, landing, charging, and data transmission, enabling fully automated daily missions and near real-time progress reporting without on-site human intervention (Figure 2).

Images were acquired using a stop-and-shoot strategy with predefined viewpoints, with minor deviations of approximately 0.3 m in position and 8° in orientation. The monitored structure is a 32-story residential tower (~ 600 m^2 footprint, 3 m floor height) constructed using a hybrid prefabricated and cast-in-place system. Each floor cycle spans approximately one week and includes four representative processes (Stage 1–4), followed by a curing period, forming recurring progress patterns across floors.

From March to July 2025, UAV missions were conducted daily (excluding adverse weather) at approximately 11:00 AM, capturing images from five predefined viewpoints: three side views (A, C, D) and two top views (B at 150 m and E at 110 m). Each image is annotated with both position and process labels (Figure 3A). Construction progressed from the 3rd to the 18th floor during this period. Images from floors 3–7 (100 images) were used for training, whereas those from the subsequently constructed floors 8–18 (570 images) formed a floor-disjoint test set. This split simulates a realistic deployment scenario in which limited data from early lower floors are annotated before the model is applied to later higher floors. It also avoids the information leakage that could arise from randomly assigning images with similar floor layouts and near-duplicate visual conditions to both sets.

The dataset contains 670 images in total, distributed as 171 (25.5%) Stage 1, 83 (12.4%) Stage 2, 244 (36.4%) Stage 3, and 172 (25.7%) Stage 4. Training data are augmented via random cropping, color perturbation, and affine transformations to reduce overfitting. This longitudinal, multi-view dataset captures realistic variations in structure, lighting, and workflow, providing both a robust benchmark for few-shot process recognition and a practical testbed for long-term autonomous construction monitoring.

Training setup. All models are implemented within a unified ViT–ProtoNet framework. Unless otherwise specified, the backbone is ViT-S/16 (12 layers, 6 heads, 384-dimensional embeddings). Camera position recognition adopts ProtoNet with $K_{pos} = 5$ support samples per class. ROI detection is implemented using a lightweight two-layer MLP head operating on shared ViT features to predict bounding boxes and confidence scores, $(x_0, y_0, x_1, y_1,$ score). For process recognition, 384-dimensional ROI features are concatenated with a 128-dimensional position embedding (generated by a two-layer MLP, 5→128→128 with ReLU) to form a 512-dimensional joint representation. This representation is processed by a 4-layer Transformer encoder (hidden size 512, 4 attention heads), followed by ProtoNet with $K_{pro} = 4$ support samples per class. To balance efficiency and adaptability, only the last two layers of ViT-S/16 are fine-tuned, resulting in approximately 34.9M trainable parameters.

Training follows an episodic few-shot paradigm, where classification is performed using Euclidean distance with a softmax over negative distances.

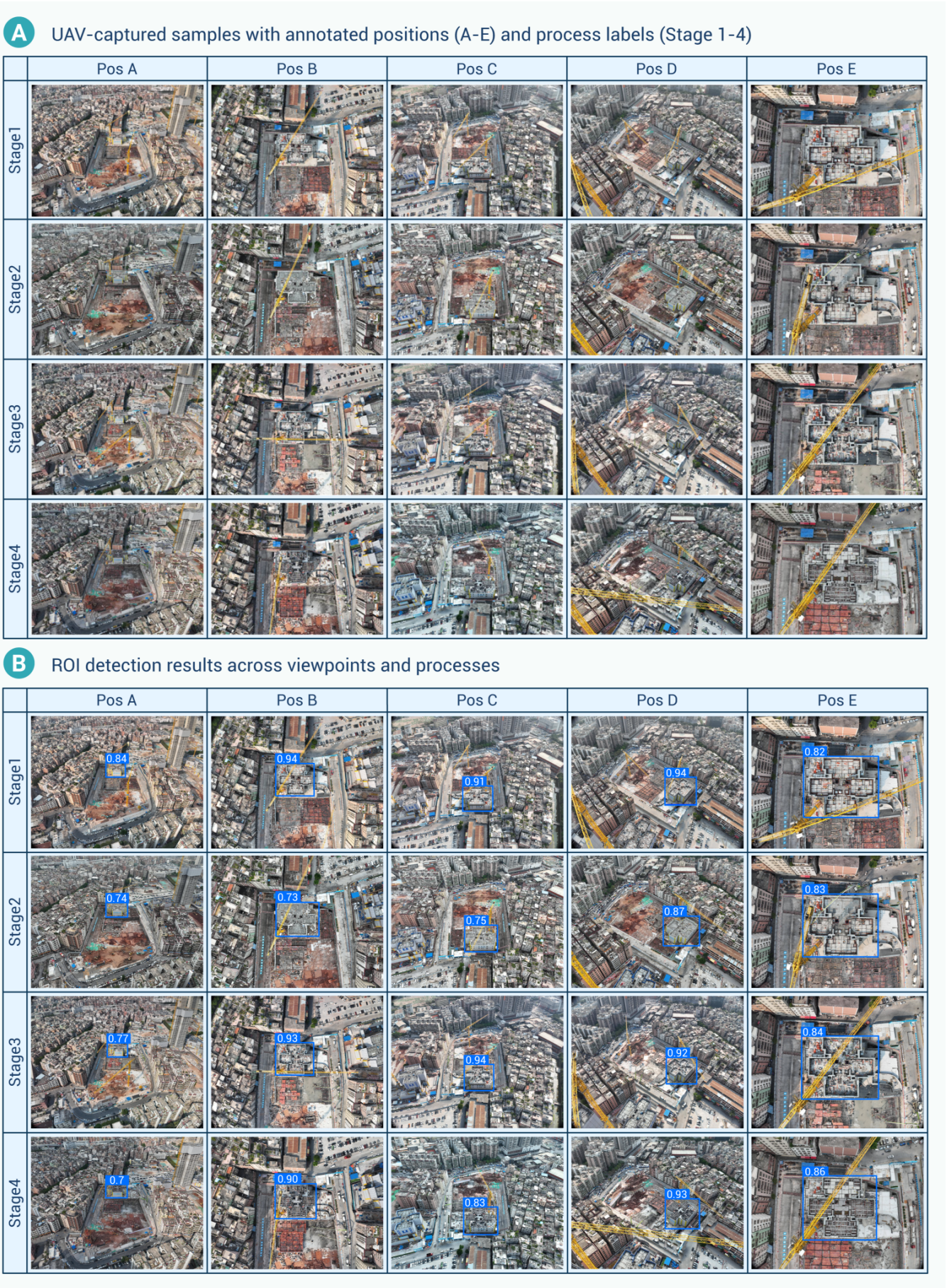

Figure 3. Data overview.

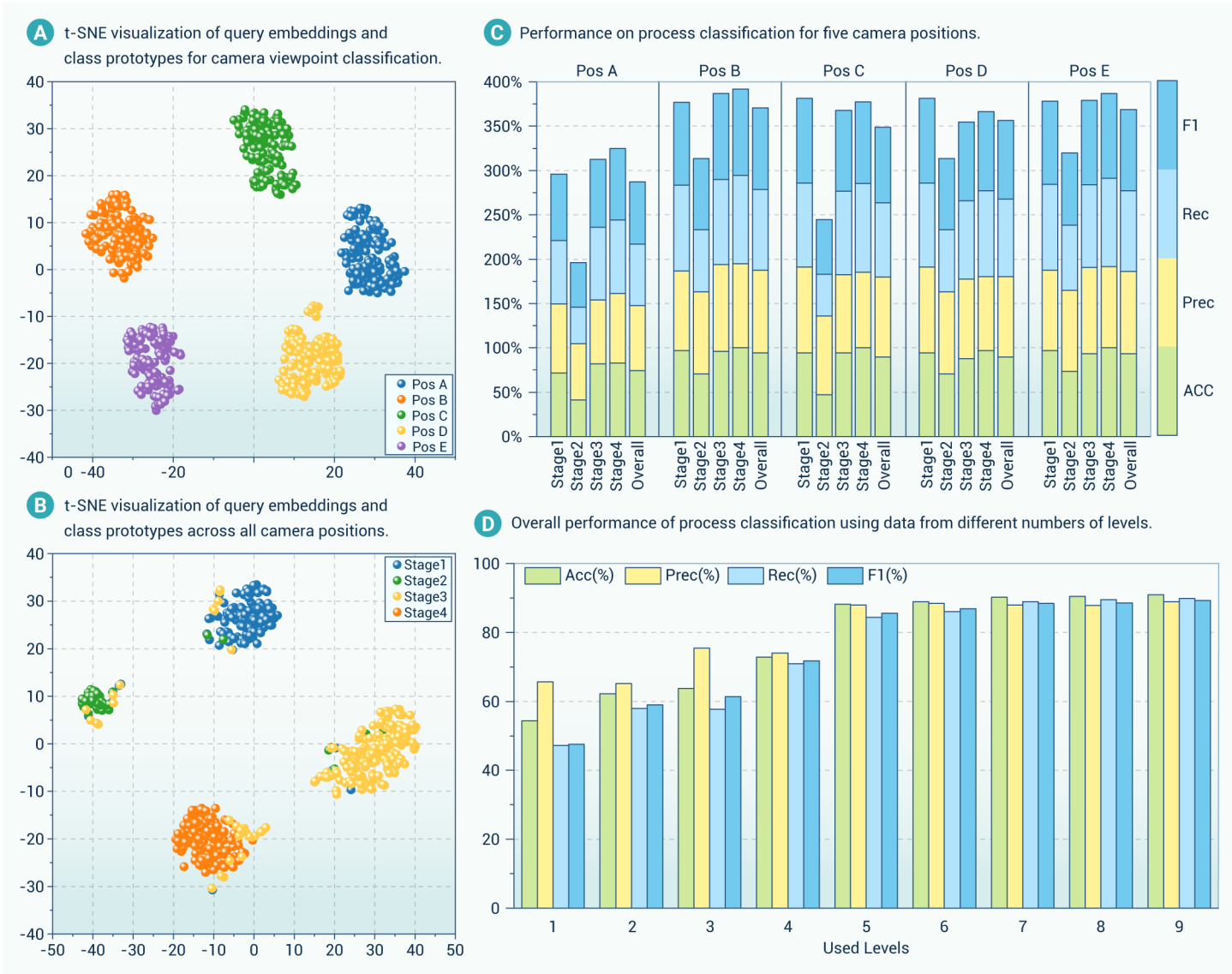

Figure 4. Embedding visualizations and recognition performance (A–B) t-SNE visualizations of few-shot embeddings; (C–D) performance analysis across viewpoints and training levels.

Optimization is conducted using AdamW with a learning rate of 1×10^{-4} and a batch size of 8, with mixed-precision (AMP) enabled for computational efficiency. All experiments are performed on a single NVIDIA RTX 4070 GPU with approximately 4.1 GB memory usage, using PyTorch 2.7.1, CUDA 12.9, and Python 3.13.5. This setup demonstrates that the proposed framework can be efficiently trained and deployed on a single mid-range GPU, supporting practical adoption in real-world construction environments.

Task-wise evaluation

Camera position performance. Camera position classification provides critical spatial priors for downstream process recognition. The proposed model achieves an overall accuracy of **93.73%** and an F1-score of **93.65%**, demonstrating consistent performance across viewpoints. Among the five positions, Pos D (100.00%), Pos E (99.18%), and Pos C (96.35%) achieve the highest performance, benefiting from stable viewpoints and clear structural visibility. In contrast, Pos A (91.24%) and Pos B (82.48%) show relatively lower performance due to oblique viewing angles and partial occlusions. Nevertheless, all positions maintain F1-scores above 90%, indicating that the ViT–ProtoNet framework effectively captures viewpoint-dependent features and provides reliable spatial priors.

The t-SNE visualization in Figure 4A further supports these findings.[73] As a nonlinear dimensionality reduction method, t-SNE projects high-dimensional embeddings into a two-dimensional space while preserving local structure. The resulting feature distributions form compact and well-separated clusters aligned with their respective prototypes. Notably, embeddings for Pos A are more dispersed, consistent with its lower accuracy, whereas Pos E exhibits the most compact cluster, reflecting its superior stability. From a deployment perspective, this reliable position inference enables consistent viewpoint identification across missions, which is essential for generating repeatable, floor-level progress records and aligning predictions with BIM or digital twin coordinates.

ROI detection performance. The ROI detection module is applied after camera position classification to suppress background clutter and focus on construction-relevant regions. Its robustness is evaluated on a long-span dataset covering multiple construction stages, viewpoints, and environmental conditions.

Performance is assessed using standard detection metrics, including mean precision (mp_{bbox}), mean recall (mr_{bbox}), and mean average precision ($mAP^{0.5}_{\text{bbox}}$) at different IoU thresholds. The module achieves $mp_{\text{bbox}} = 0.9905$, $mr_{\text{bbox}} = 0.9931$, and $mAP^{0.5}_{\text{bbox}} = 0.9948$, indicating near-perfect detection with minimal false positives and negatives. Under stricter localization criteria, $mAP^{0.5:\,0.95}_{\text{bbox}}$ reaches 0.8634, demonstrating accurate bounding box regression across varying overlap thresholds.

Representative examples in Figure 3B show that the predicted regions consistently capture active construction areas while excluding irrelevant elements such as scaffolding and equipment. The model maintains high

Table 1. Daily-level summary of per-class and overall performance across three fusion strategies.

Class	Pro_{tv}				Pro_{max}				Pro_{fuse}			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Stage1	97.14	89.47	97.14	93.15	100.0	87.50	100.0	93.33	100.0	92.11	100.0	95.89
Stage2	70.59	92.31	70.59	80.00	70.59	85.71	70.59	77.42	76.47	92.86	76.47	83.87
Stage3	96.00	97.96	96.00	96.97	92.00	95.83	92.00	93.88	96.00	97.96	96.00	96.97
Stage4	100.0	94.59	100.0	97.22	94.29	94.29	94.29	94.29	100.0	97.22	100.0	98.59
Overall	**94.16**	**93.58**	**90.93**	**91.84**	**91.97**	**90.83**	**89.22**	**89.73**	**95.62**	**95.04**	**93.12**	**93.83**

confidence under challenging lighting and oblique viewpoints, confirming its stability over long-term UAV deployment. In practice, this robustness reduces the need for manual image filtering and ensures that downstream analysis focuses on semantically meaningful regions, improving the efficiency and reliability of remote construction monitoring.

Process recognition performance. The construction process recognition module integrates position priors with ROI-focused features to classify four representative processes (Stage 1–4). As shown in Figure 4C, performance varies across viewpoints. The model achieves the highest accuracy at Pos B (94.16%) and Pos E (93.44%), benefiting from clearer structural visibility in top-down views. Intermediate performance is observed at Pos C (89.78%) and Pos D (89.78%), where partial occlusions and cross-view similarities introduce ambiguity. Pos A remains the most challenging viewpoint (74.45% accuracy, 70.45% F1), primarily due to oblique perspectives and increased background clutter. Across all viewpoints, Stage 2 is consistently more difficult to recognize, reflecting its visual similarity to adjacent transitional states.

The t-SNE visualization in Figure 4B further explains these patterns. Embeddings for Stage 1 and Stage 4 form compact, well-separated clusters, indicating strong intra-class consistency and discriminability. In contrast, Stage 2 and Stage 3 exhibit partial overlap, driven by similar visual textures such as concrete surfaces and reinforcement structures. Overall, the ViT–ProtoNet framework effectively distinguishes most construction processes, while residual ambiguities are concentrated in visually similar transitional stages. From a practical perspective, this suggests that the system can reliably identify dominant workflows at the floor level, with remaining uncertainties aligned with inherently ambiguous stages that benefit from additional temporal or multi-view context.

Fusion and monitoring

Daily fusion performance. UAV images captured on the same day often exhibit variability due to lighting changes, occlusions, and viewpoint differences, leading to inconsistent predictions across individual views. Daily multi-view fusion combines predictions from the available views of the same monitored floor on each observation day, producing one daily-level process label. The reported 95.62% daily-level accuracy evaluates these fused labels, whereas the 88.21% image-level accuracy in the ablation study evaluates individual images before fusion. The overall F1-score is macro-averaged across the four process classes. Three strategies are evaluated: (1) top-view only (Pro_{tv}), (2) maximum-confidence selection (Pro_{max}), and (3) weighted multi-view fusion (Pro_{fuse}) based on Equation (1).

As shown in Table 1, the proposed fusion strategy (Pro_{fuse}) consistently outperforms single-view baselines. Stage 1 and Stage 4 achieve the highest reliability, with F1-scores of 95.89% and 98.59%, respectively, while Stage 3 also demonstrates strong performance (Acc. = 96.00%, F1 = 96.97%). In contrast, Stage 2 remains more challenging, with a lower recall (76.47%) due to its visual similarity to adjacent stages. Overall, the fusion approach achieves a daily-level accuracy of 95.62% and an F1-score of 93.83%, indicating improved robustness compared with individual predictions.

A comparison of fusion strategies further highlights these gains. The Pro_{max} baseline yields the lowest accuracy (91.97%), reflecting its sensitivity to noisy predictions. The top-view strategy (Pro_{tv}) performs better (94.16%) due to its stable perspective, but remains limited by single-view information. The proposed weighted fusion (Pro_{fuse}) achieves the best overall performance (daily-level Acc. = 95.62%, F1 = 93.83%) by effectively balancing precision and recall across views. In practice, this mechanism acts as a multi-view stabilizer, suppressing occasional view-specific mispredictions and producing consistent daily-level process labels for integration into Gantt charts or BIM-based management systems.

Project monitoring. The proposed framework enables automated construction progress monitoring through daily multi-view fusion of UAV imagery. Process predictions are aggregated across all viewpoints to track the evolution of each floor and compared against the planned schedule, as illustrated in Figure 5. The resulting Gantt chart visualizes planned versus actual timelines for the four representative workflows (Stage 1–4), enabling clear identification of deviations across building levels.

In the early phase, Levels 3, 4, and 8 experienced delays of approximately two weeks due to coordination inefficiencies and equipment downtime. The UAV-based monitoring system detected these deviations at an early stage, allowing timely feedback and intervention. Following corrective actions, such as reallocation of labor, machinery, and materials, workflow efficiency improved, and from Level 9 onward, progress returned to the planned rate of approximately one level per week. Overall, the project management team estimated that monitoring-assisted corrective actions, together with other management measures, helped recover 21 working days of previously accumulated schedule delay (approximately 14% of the planned project duration). Based on an average daily cost of 25,702.94 RMB, the corresponding reduction in delay-related costs was estimated at 539,762 RMB in direct costs and approximately 575,000 RMB including indirect costs.

Beyond quantitative gains, the system enhances management transparency by providing objective, visual, and traceable records of daily site activities. The fused process labels can be directly integrated with BIM or digital twin platforms, enabling intuitive visualization of delays, bottlenecks, and out-of-sequence operations. This capability transforms UAV imagery from passive documentation into an active decision-support tool, supporting more timely, data-driven schedule control and risk management in complex construction projects.

Ablation and errors

In the ablation study, we adopt photo-based performance metrics rather than process-, viewpoint-, or day-based metrics, since photo-level evaluation directly reflects the model's recognition capability on the raw data without additional aggregation. This provides a more objective measure of the proposed pipeline's performance, while the fusion and monitoring experiments quantify how this capability translates into stable, day-level progress sensing for engineering practice.

Training levels ablation. To evaluate the impact of training data volume under few-shot conditions, we conduct an ablation study by varying the number of floors used for training while reserving the remaining floors for testing. Given the bottom-up construction process, the first $n$floors (e.g., Levels 3–5) are used for training, while higher floors (Levels 6–18) are used for evaluation. This setup reflects a realistic deployment scenario, where only limited early-stage data are annotated before applying the model to unseen construction progress.

As shown in Figure 4D, performance improves consistently with increasing training data before reaching saturation. Accuracy increases from 54.32% with one floor to 72.84% with four floors, and further to 88.21% with five floors. Beyond this point, improvements become marginal (less than 3% up to nine floors), indicating convergence. A similar trend is observed for the F1-

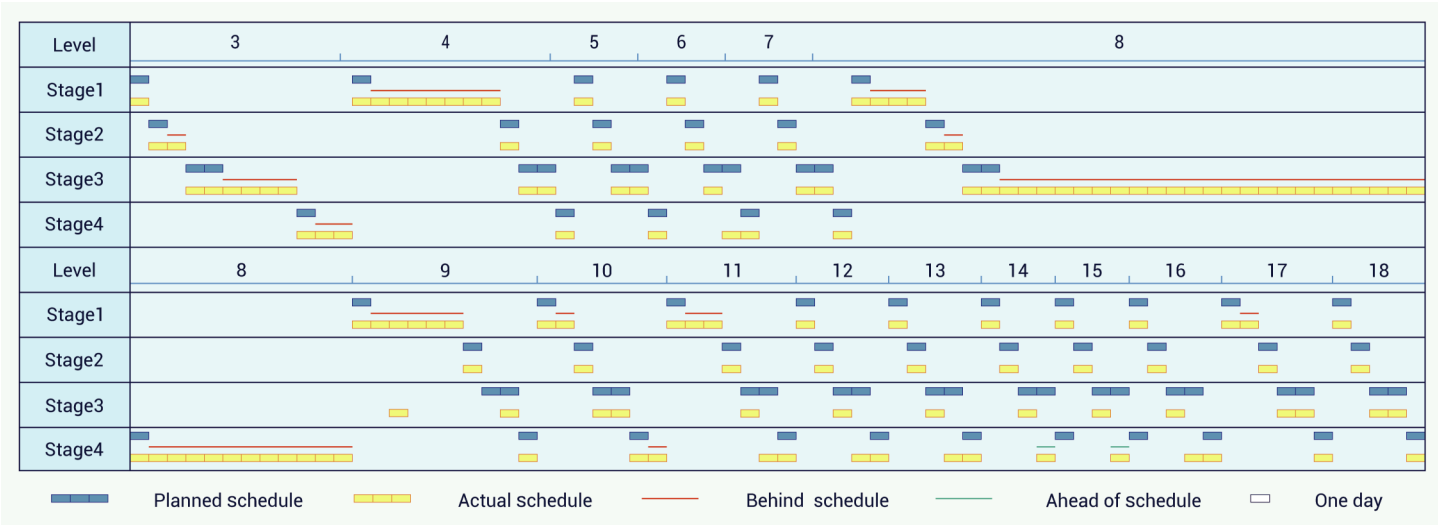

Figure 5. Comparison of planned and detected construction schedules across multiple levels and processes The Gantt chart illustrates planned versus actual progress, highlighting stages that are ahead of schedule, behind schedule, or under construction. Representative processes include the four defined workflows (Stage 1-4).

Table 2. Image-level ablation study on task combinations for construction process classification. Tasks: ① camera position classification, ② ROI detection, ③ process classification. An asterisk (*) indicates that the ProtoNet is replaced by an MLP classifier.

Task Combination	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
③*	27.75	29.59	30.57	27.15
③	55.50	55.12	57.44	53.91
①*+③*	29.01	29.49	30.52	28.18
①+③	58.20	56.77	58.43	56.29
②+③*	28.83	29.94	34.10	27.21
②+③	70.45	67.29	67.80	66.24
①*+②+③*	32.61	32.80	37.72	31.48
①+②+③	**88.21**	**87.91**	**84.35**	**85.59**

score, which rises from 47.54% to 85.59% at five floors and stabilizes near 88.95%. These results indicate that training on five floors provides an effective balance between annotation cost and model performance. From a practical perspective, this suggests that only a short initialization phase, spanning a few early floors, is sufficient to enable reliable deployment for automated monitoring on subsequent levels and similar projects.

Module ablation. To quantify the contribution of each component, we conduct an ablation study across different task combinations and classifier designs (Table 2). The three tasks include: (1) camera position classification, (2) ROI detection, and (3) process classification. An asterisk (*) denotes replacing the ProtoNet with a standard MLP classifier to evaluate the impact of metric-based learning.

The results highlight the complementary roles of the proposed modules. Using only process classification (③) yields moderate performance (55.50% accuracy, 53.91% F1), while replacing ProtoNet with an MLP (③*) leads to a substantial drop (27.75% accuracy), confirming the effectiveness of few-shot metric learning. Incorporating camera position priors (①+③) provides a modest improvement (58.20% accuracy), indicating that spatial context helps mitigate viewpoint ambiguity. In contrast, adding ROI detection (②+③) results in a significant performance gain (70.45% accuracy, 66.24% F1), demonstrating the importance of spatial focusing in suppressing background noise and enhancing feature discriminability.

The full model (①+②+③) achieves the best performance, with 88.21% image-level accuracy before fusion, 87.91% precision, 84.35% recall, and 85.59% F1. Compared with the process-only baseline, this represents improvements of over 29 percentage points in both accuracy and F1-score, underscoring the strong synergy between viewpoint priors, ROI focusing, and few-shot classification. From a deployment perspective, these results suggest that ROI detection and position priors are essential for maintaining robustness, while the ProtoNet-based classifier is critical for preserving performance under limited annotation conditions.

Error analysis. To assess the limitations of the proposed framework, we analyze representative misclassification cases (Figures 6 A–C). Most errors occur during transitional phases, where one process is concluding and another is emerging, resulting in incomplete or ambiguous visual cues. In Example A, early Stage 2 is misclassified as Stage 1 because exposed reinforcement remains visually dominant despite the onset of concrete casting. In Example B, post-Stage 2 imagery is incorrectly classified as Stage 3 due to prominent steel bars prepared for subsequent work, introducing misleading structural cues. In Example C, Stage 4 is confused with Stage 3 because slab installation is only partially completed, and distinctive slab features are not yet fully visible. These cases indicate that errors primarily arise when salient but ambiguous features dominate the embedding space, or when ROI regions contain overlapping construction activities.

From an operational perspective, such ambiguities align with phases where even human experts struggle to assign a definitive process label. Rather than representing system failure, these cases highlight inherent uncertainty in transitional states. Future work should therefore focus on explicitly modeling process transitions, incorporating temporal context, and expanding datasets to include intermediate states. Additional improvements may include finer-grained spatial modeling and the integration of multimodal data (e.g., point clouds, video, or construction logs), which could enhance temporal consistency and reduce ambiguity in real-world deployment.

CONCLUSION

This study formulates construction progress monitoring as a few-shot,



Figure 6. Examples of misclassification in process recognition (A) Stage 2 misidentified as Stage 1 due to exposed rebar; (B) Stage 2 misidentified as Stage 3 due to misleading steel bars on site; (C) Stage 4 misidentified as Stage 3 because only partial slab installation had been completed.

process-level recognition problem and develops a fully autonomous UAV-based framework for its implementation. The framework uses camera-position classification as a spatial prior, ROI detection to focus on process-relevant regions, and a ViT–ProtoNet architecture to recognize construction processes from limited labeled data. Daily multi-view fusion further aggregates image-level predictions into stable daily-level process labels. In a five-month deployment on a 32-story project, the framework achieved **95.62% daily-level accuracy** and a **93.83% F1-score** on previously unseen floors. During the same deployment, comparison with the baseline schedule indicated an observed recovery of **21 working days**, approximately 14% of the planned duration, and an estimated reduction of **575,000 RMB** in delay-related losses. These engineering outcomes support the practical value of integrating autonomous UAV sensing with data-efficient, process-aware recognition, although they should not be attributed solely to the recognition model.

Despite these advances, several challenges remain. Transitional workflow stages often exhibit ambiguous or overlapping visual cues, while long-term construction introduces appearance drift that is difficult to capture with limited annotations. Addressing these issues requires expanding datasets to include intermediate states, incorporating finer-grained spatial representations,

and explicitly modeling temporal dependencies in process evolution.

Future work will focus on three directions. First, multimodal integration, combining UAV imagery with video, point clouds, BIM models, and construction logs, will enable richer temporal reasoning and more reliable anomaly detection. Second, extending the system to indoor sensing and multi-robot coordination will support whole-building monitoring beyond outdoor observations. Third, enabling continuous knowledge accumulation will allow the system to adapt across projects and learn new workflows with minimal supervision. Together, these advances will move toward a scalable and adaptive framework for autonomous, process-aware construction monitoring, supporting the development of fully data-driven and intelligent jobsite management systems.

REFERENCES

1. Xu S., Wang J., Shou W., et al. (2021). Computer vision techniques in construction: A critical review. *Arch. Comput. Methods Eng.* **28**:3383–3397. DOI:10.1007/s11831-020-09504-3
2. Nakanishi Y., Kaneta T. and Nishino S. (2022). A review of monitoring construction equipment in support of construction project management. *Front. Built Environ.* **7**:632593. DOI:10.3389/fbuil.2021.632593
3. Kim P., Price L.C., Park J., et al. (2019). UAV-UGV cooperative 3D environmental mapping. In *Computing in Civil Engineering 2019* (American Society of Civil Engineers), pp:384–392. DOI:10.1061/9780784482438.049
4. Rao A.S., Radanovic M., Liu Y., et al. (2022). Real-time monitoring of construction sites: Sensors, methods, and applications. *Autom. Constr.* **136**:104099. DOI:10.1016/j.autcon.2021.104099
5. Zhang Z. and Zhu L. (2023). A review on unmanned aerial vehicle remote sensing: Platforms, sensors, data processing methods, and applications. *Drones* **7**:398. DOI:10.3390/drones7060398
6. Xiong B., Han S., Messerschmidt T.C., et al. (2024). Early detection of invasive Phragmites australis at the tidal marsh-forest ecotone with airborne LiDAR. *Ecol. Indic.* **167**:112651. DOI:10.1016/j.ecolind.2024.112651
7. Zhang J. and Jiang S. (2026). Review of artificial intelligence applications in construction management over the last five years. *Eng. Constr. Archit. Manag.* **33**:361–379. DOI:10.1108/ECAM-03-2024-0313
8. Kim J. and Chi S. (2020). Multi-camera vision-based productivity monitoring of earthmoving operations. *Autom. Constr.* **112**:103121. DOI:10.1016/j.autcon.2020.103121
9. Luo H., Xiong C., Fang W., et al. (2018). Convolutional neural networks: Computer vision-based workforce activity assessment in construction. *Autom. Constr.* **94**:282–289. DOI:10.1016/j.autcon.2018.06.007
10. Xiong B., Jin Y., Li F., et al. (2023). Knowledge-driven inference for automatic reconstruction of indoor detailed as-built BIMs from laser scanning data. *Autom. Constr.* **156**:105097. DOI:10.1016/j.autcon.2023.105097
11. Wei W., Lu Y., Zhang X., et al. (2024). Fine-grained progress tracking of prefabricated construction based on component segmentation. *Autom. Constr.* **160**:105329. DOI:10.1016/j.autcon.2024.105329
12. Rezazadeh Azar E. (2017). Semantic annotation of videos from equipment-intensive construction operations by shot recognition and probabilistic reasoning. *J. Comput. Civ. Eng.* **31**:04017042. DOI:10.1061/(ASCE)CP.1943-5487.0000693
13. Han K.K. and Golparvar-Fard M. (2017). Potential of big visual data and building information modeling for construction performance analytics: An exploratory study. *Autom. Constr.* **73**:184–198. DOI:10.1016/j.autcon.2016.11.004
14. Han J.-Y., Hsu C.-R. and Huang C.-J. (2024). Automated progress monitoring of land development projects using unmanned aerial vehicles and machine learning. *Autom. Constr.* **168**:105827. DOI:10.1016/j.autcon.2024.105827
15. Han M., Zhang J., Huang Y., et al. (2025). Enhancing worker monitoring and management on large-scale construction sites with UAVs and digital twin modeling. *Autom. Constr.* **174**:106108. DOI:10.1016/j.autcon.2025.106108
16. Paneru S. and Jeelani I. (2021). Computer vision applications in construction: Current state, opportunities & challenges. *Autom. Constr.* **132**:103940. DOI:10.1016/j.autcon.2021.103940
17. Arashpour M., Ngo T. and Li H. (2021). Scene understanding in construction and buildings using image processing methods: A comprehensive review and a case study. *J. Build. Eng.* **33**:101672. DOI:10.1016/j.jobe.2020.101672
18. Braun A. and Borrmann A. (2019). Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning. *Autom. Constr.* **106**:102879. DOI:10.1016/j.autcon.2019.102879
19. Asadi K., Suresh A.K., Ender A., et al. (2020). An integrated UGV-UAV system for construction site data collection. *Autom. Constr.* **112**:103068. DOI:10.1016/j.autcon.2019.103068
20. Ersoz A.B. and Pekcan O. (2025). UAV-based automated earthwork progress monitoring using deep learning with image inpainting. *Autom. Constr.* **175**:106211. DOI:10.1016/j.autcon.2025.106211
21. Yang J., Park M.-W., Vela P.A., et al. (2015). Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. *Adv. Eng. Inform.* **29**:211–224. DOI:10.1016/j.aei.2015.01.011
22. Reja V.K., Varghese K. and Ha Q.P. (2022). Computer vision-based construction progress monitoring. *Autom. Constr.* **138**:104245. DOI:10.1016/j.autcon.2022.104245
23. Lyu C., Lin S., Lynch A., et al. (2025). UAV-based deep learning applications for automated inspection of civil infrastructure. *Autom. Constr.* **177**:106285. DOI:10.1016/j.autcon.2025.106285
24. Wang Y., Yao Q., Kwok J.T., et al. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **53**:63:1-63:34. DOI:10.1145/3386252
25. Song Y., Wang T., Cai P., et al. (2023). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* **55**:1–40. DOI:10.1145/3582688
26. Antonelli S., Avola D., Cinque L., et al. (2022). Few-shot object detection: A survey. *ACM Comput. Surv.* **54**:1–37. DOI:10.1145/3519022
27. Snell J., Swersky K. and Zemel R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, Guyon I., Von Luxburg U., Bengio S., et al. (eds).(Curran Associates, Inc.), pp:4077–4087. https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html
28. Xiong B., Zhang X. and Jiang W. (2010). Semi-supervised classification based on gaussian mixture model for remote imagery. *Sci. China Technol. Sci.* **53**:85–90. DOI:10.1007/s11431-010-3211-5
29. Xiong B., Jiang W. and Li L. (2011). Gauss mixture model based semi-supervised classification for remote sensing image. *Geomat. Inf. Sci. Wuhan Univ.* **36**:108–112. https://ch.whu.edu.cn/cn/article/pdf/preview/431.pdf
30. Pal D., Bundele V., Banerjee B., et al. (2022). SPN: Stable prototypical network for few-shot learning-based hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **19**:1–5. DOI:10.1109/LGRS.2021.3085522
31. Kim J. and Chi S. (2021). A few-shot learning approach for database-free vision-based monitoring on construction sites. *Autom. Constr.* **124**:103566. DOI:10.1016/j.autcon.2021.103566
32. Seo M. and Kim H. (2025). Irregular openings identification at construction sites based on few-shot learning. *Buildings.* **15**:1834. DOI:10.3390/buildings15111834
33. Wang X. and El-Gohary N. (2024). Few-shot object detection and attribute recognition from construction site images for improved field compliance. *Autom. Constr.* **167**:105539. DOI:10.1016/j.autcon.2024.105539
34. Jeoung J., Jung S. and Hong T. (2025). Zero-shot framework for construction equipment task monitoring. *Comput.-Aided Civ. Infrastruct. Eng.* **40**:3218–3238. DOI:10.1111/mice.13506
35. Tais de Gouveia L., Costa L. da F., Senger L.J., et al. (2011). Entropy-based approach to analyze and classify mineral aggregates. *J. Comput. Civ. Eng.* **25**:75–84. DOI:10.1061/(ASCE)CP.1943-5487.0000071
36. Yang J., Shi Z. and Wu Z. (2016). Vision-based action recognition of construction workers using dense trajectories. *Adv. Eng. Inform.* **30**:327–336. DOI:10.1016/j.aei.2016.04.009
37. Ding L., Fang W., Luo H., et al. (2018). A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Autom. Constr.* **86**:118–124. DOI:10.1016/j.autcon.2017.11.002
38. Bang S., Kim H. and Kim H. (2017). UAV-based automatic generation of high-resolution panorama at a construction site with a focus on preprocessing for image stitching. *Autom. Constr.* **84**:70–80. DOI:10.1016/j.autcon.2017.08.031
39. Ren H., Fan A., Zhao J., et al. (2024). Lightweight safety helmet detection algorithm using improved YOLOv5. *J. Real-Time Image Process.* **21**:125. DOI:10.1007/s11554-024-01499-5
40. Zhang M., Liu X. and Li Y. (2025). Identity-based proactive human intrusion management in hazardous areas at construction sites: A deep learning–based method. *J. Constr. Eng. Manag.* **151**:04025045. DOI:10.1061/JCEMD4.COENG-15515
41. Zhang J., Wang D., An X., et al. (2025). A Voxel-Based 3D reconstruction and action recognition method for construction workers. *Adv. Eng. Inform.* **65**:103203. DOI:10.1016/j.aei.2025.103203
42. Zhang G., Wang J., Zhang J., et al. (2025). Construction resource identification under complex conditions of navigation–power junction project based on improved yolov8 and monocular vision. *J. Comput. Civ. Eng.* **39**:04025018. DOI:10.1061/JCCEE5.CPENG-5990
43. Fang Y., Cho Y.K. and Chen J. (2016). A framework for real-time pro-active safety assistance for mobile crane lifting operations. *Autom. Constr.* **72**:367–379. DOI:10.1016/j.autcon.2016.08.025
44. Li C.Z., Xue F., Li X., et al. (2018). An Internet of Things-enabled BIM platform for on-site assembly services in prefabricated construction. *Autom. Constr.* **89**:146–161. DOI:10.1016/j.autcon.2018.01.001
45. Puri N. and Turkan Y. (2020). Bridge construction progress monitoring using lidar and 4D design models. *Autom. Constr.* **109**:102961. DOI:10.1016/j.autcon.2019.102961
46. Fei-Fei L., Fergus R. and Perona P. (2006). One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**:594–611. DOI:10.1109/TPAMI.2006.79
47. Sung F., Yang Y., Zhang L., et al. (2018). Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE), pp:1199–1208. DOI:10.1109/CVPR.2018.00131
48. Vinyals O., Blundell C., Lillicrap T., et al. (2016). Matching networks for one shot

learning. In *Advances in Neural Information Processing Systems*, Lee D., Sugiyama M., Luxburg U., et al. (eds)., pp:3630–3638. https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html
49. Martinez-Lopez F., Santana L. and Rahouti M. (2024). Learning in multiple spaces: Few-shot network attack detection with metric-fused prototypical networks. DOI:10.48550/arXiv.2501.00050
50. Guo Z., Ao S. and Ao B. (2024). Few-shot learning based oral cancer diagnosis using a dual feature extractor prototypical network. *J. Biomed. Inform.* **150**:104584. DOI:10.1016/j.jbi.2024.104584
51. Zeng J., Xue Z., Zhang L., et al. (2023). Multistage relation network with dual-metric for few-shot hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **61**:1–17. DOI:10.1109/TGRS.2023.3271424
52. Liu C., Yang L., Li Z., et al. (2024). Multi-level relation learning for cross-domain few-shot hyperspectral image classification. *Appl. Intell.* **54**:4392–4410. DOI:10.1007/s10489-024-05384-3
53. Tian F., Lei S., Zhou Y., et al. (2024). HiReNet: Hierarchical-relation network for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **62**:1–10. DOI:10.1109/TGRS.2023.3348464
54. Matindife L., Sun Y. and Wang Z. (2022). Few-shot learning for image-based nonintrusive appliance signal recognition. *Comput. Intell. Neurosci.* **2022**:2142935. DOI:10.1155/2022/2142935
55. Zhou Z., Cui Z., Tian Y., et al. (2025). Dynamic semantics-guided meta-transfer learning for few-shot SAR target detection. *IEEE Trans. Geosci. Remote Sens.* **63**:1–17. DOI:10.1109/TGRS.2025.3561682
56. Xu Y., Bi H., Yu H., et al. (2024). Attention-based contrastive learning for few-shot remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **62**:1–17. DOI:10.1109/TGRS.2024.3385655
57. Zhou Y., Yan H., Ding K., et al. (2024). Few-shot image classification of crop diseases based on vision–language models. *Sensors* **24**:6109. DOI:10.3390/s24186109
58. Zeng S., Xia Y., Gu S., et al. (2025). Few-shot classification for soil images: Prototype correction and feature distance enhancement. *Comput. Electron. Agric.* **233**:110162. DOI:10.1016/j.compag.2025.110162
59. Shen Y., Fan W., Wang C., et al. (2024). Dual-guided prototype alignment network for few-shot medical image segmentation. *IEEE Trans. Instrum. Meas.* **73**:1–13. DOI:10.1109/TIM.2024.3411136
60. Madan S., Chaudhury S. and Gandhi T.K. (2023). Explainable few-shot learning with visual explanations on a low resource pneumonia dataset. *Pattern Recognit. Lett.* **176**:109–116. DOI:10.1016/j.patrec.2023.10.013
61. Zhao Z., Jiao Y., Xu Y., et al. (2025). A fault diagnosis framework using unlabeled data based on automatic clustering with meta-learning. *Eng. Appl. Artif. Intell.* **139**:109584. DOI:10.1016/j.engappai.2024.109584
62. Xu J. and Pan W. (2024). Deep learning-based object detection for dynamic construction site management. *Autom. Constr.* **165**:105494. DOI:10.1016/j.autcon.2024.105494
63. Shamsollahi D., Moselhi O. and Khorasani K. (2024). Automated detection and segmentation of mechanical, electrical, and plumbing components in indoor environments by using the YOLACT++ architecture. *J. Constr. Eng. Manag.* **150**:04024100. DOI:10.1061/JCEMD4.COENG-15115
64. Zhan Q., Wang B., Jiang A., et al. (2024). A two-stage spiking meta-learning method for few-shot classification. *Knowl.-Based Syst.* **284**:111220. DOI:10.1016/j.knosys.2023.111220
65. Oh J., Hong S., Choi B., et al. (2025). Integrating text parsing and object detection for automated monitoring of finishing works in construction projects. *Autom. Constr.* **174**:106139. DOI:10.1016/j.autcon.2025.106139
66. Pal A., Lin J.J., Hsieh S.-H., et al. (2023). Automated vision-based construction progress monitoring in built environment through digital twin. *Dev. Built Environ.* **16**:100247. DOI:10.1016/j.dibe.2023.100247
67. Wang Z., Zhang Q., Yang B., et al. (2021). Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase. *J. Comput. Civ. Eng.* **35**:04020056. DOI:10.1061/(ASCE)CP.1943-5487.0000933
68. Wei W., Lu Y., Lin Y., et al. (2023). Augmenting progress monitoring in soil-foundation construction utilizing SOLOv2-based instance segmentation and visual BIM representation. *Autom. Constr.* **155**:105048. DOI:10.1016/j.autcon.2023.105048
69. Reja V.K., Goyal S., Varghese K., et al. (2024). Hybrid self-supervised learning-based architecture for construction progress monitoring. *Autom. Constr.* **158**:105225. DOI:10.1016/j.autcon.2023.105225
70. Yan X., Zhang H. and Zhang W. (2023). Intelligent monitoring and evaluation for the prefabricated construction schedule. *Comput.-Aided Civ. Infrastruct. Eng.* **38**:391–407. DOI:10.1111/mice.12838
71. Dosovitskiy A., Beyer L., Kolesnikov A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint. DOI:10.48550/arXiv.2010.11929
72. DJI. (2025). DJI dock 2. https://enterprise.dji.com/cn/dock-2
73. Van der Maaten L. and Hinton G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**:2579–2605. https://jmlr.org/papers/v9/vandermaaten08a.html

FUNDING AND ACKNOWLEDGMENTS

This research was funded by the Major Research Project of Hubei Province ((JD)2023BAA007), and the National Natural Science Foundation of China (Grant No. 61901311). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Y.Q.Z. and T.X.L. conceived and designed the study. T.X.L. and I.B. developed the methodology. T.X.L. and M.W. were responsible for the software development and visualization. T.X.L. and I.B. performed the formal analysis. T.X.L. carried out the investigation and data curation with support from Z.D.X. and M.W. T.X.L. wrote the original draft of the manuscript. B.X., I.B., and Y.Q.Z. supervised the research and provided critical revisions. Y.Q.Z., Z.D.X., and B.X. were responsible for project administration and funding acquisition. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AND CODE AVAILABILITY

Data are available from the corresponding author upon reasonable request.