Construction Site Layout Monitoring Using Instance Segmentation

Vimal Bharathi and Jochen Teizer

Department of Civil and Mechanical Engineering, Technical University of Denmark
vvebh@dtu.dk, teizerj@dtu.dk

Abstract

Construction sites are complex and dynamic work environments with frequent changes in layout due to material handling and a diverse workforce presence, necessitating adaptable safety measures. Falls from height remain the primary cause of fatalities in construction, particularly impacting small firms with limited resources. Conventional manual construction site layout inspections are subjective, relative infrequent, and difficult to scale, often resulting in noncompliance with fall protection standards. This paper introduces an automated, geo-aware monitoring system that leverages high-resolution orthomosaic images from consumer drones and a novel Adaptive Multi-Scale DETR (AMS-DETR) model for panoptic segmentation. AMS-DETR enhances traditional DETR performance through dynamic query generation and multi-scale attention yielding lower inference time and segmentation Quality. This approach enables proactive, AI-driven site monitoring, improves regulatory adherence, and offers scalable safety management suitable for resource-constrained projects, advancing intelligent construction safety practices.

Keywords –
DETR, instance segmentation, orthomosaic images, site layout monitoring, unmanned aerial vehicle.

1 Introduction

Construction sites are among the most variable work environments, where complex operations lead to frequent shifts in personnel, materials, and equipment [1]. This constant change poses major challenges in keeping safety measures current and aligned with standards. Unlike stable factory settings, construction sites evolve daily through stages like excavation, framing, and finishing, requiring quick setup of temporary protections such as guardrails to address fall risks at heights over 1.8 m [2]. In 2023, falls caused 421 deaths (39.2% of all U.S. construction fatalities, totaling 1,075) [3], alongside more than 20,000 nonfatal fall-related cases annually [4]; and over 70% of fatal falls happened at companies with fewer than 11 employees, highlighting gaps in monitoring for smaller operations. On a global scale, the International Labor Organization projects over 60,000 construction deaths yearly, with falls responsible for 30–50%, varying by region due to differences in regulations and site conditions [5]. These trends reveal a key problem: even with clear rules (e.g., OSHA's 1.07 m guardrail requirements), unexpected removal of safety gear—due to task demands, worker fatigue, or weather issues quickly makes plans outdated, and over 70% of violations involve absent or incorrectly placed barriers.

As a result, regular compliance checks are essential, but current manual methods using checklists and measurements by safety staff are reactive, subject to human error, and limited by site size and access difficulties. Research shows that up to 65% of fall incidents arise from gaps missed during routine (e.g., weekly) checks [6], worsened by emerging threats like proximity issues: for instance, cranes within 2 m of walkways (violating ANSI Z359.2 buffers) or unguarded gaps between paths and pits, which can significantly increase the risk of secondary impacts [7].

Tools like digital twins, remote sensing, and AI can transform the above issues with real-time tracking that could spot issues early, link to evolving 4D-BIM models for predictions, and fit into standard workflows. However, past efforts are often isolated for example, 2D image detectors miss depth details, while basic Orthomosaic classifiers overlook the fine distinctions needed for tracking specific "things" (e.g., individual guardrail sections) versus backgrounds ("stuff" like debris floors) limiting geospatial tracking and risk calculations [8].

This paper addresses these limitations by proposing a location-aware monitoring pipeline. It integrates high-resolution orthomosaic maps generated from overhead imagery captured by DJI Mini 4 Pro drone and processed via photogrammetry with an advanced Adaptive Multi-Scale DETR (AMS-DETR) model for instance segmentation of safety features (e.g., guardrails, Safety poles, etc.) and site elements (e.g., cranes, excavations, steel plates, etc.). AMS-DETR overcomes key DETR limitations, including rigid queries, scale mismatches, and boundary inaccuracies, through adaptive query generation and multi-level attention mechanism, delivering higher segmentation quality while enabling real-time inference.

It enables the generation of predictive compliance scores and interactive hazard visualizations within a user-friendly SafeBIM platform [36]. The proposed framework provides a scalable, proactive approach to AI-supported safety oversight on highly dynamic construction sites, with strong potential to enhance fall prevention, regulatory compliance, and overall site management practices.

2 Background

Construction sites present unique challenges for safety monitoring due to their dynamic nature, with frequent changes in layout, equipment, and personnel that heighten risks such as falls from height. Traditional manual inspections are inadequate for real-time compliance, prompting the adoption of AI-driven technologies like Unmanned Aerial Vehicles (UAV), deep learning for image segmentation, and Building Information Modeling (BIM) integration. This section reviews existing DEtection TRansformer (DETR) models for segmentation a critical tool for analyzing site images and related work on AI-based site layout monitoring, highlighting gaps addressed by our proposed Adaptive Multi-Scale DETR (AMS-DETR) framework.

2.1 Existing DETR Models for Panoptic Segmentation

The DETR family has revolutionized object detection by formulating it as a set prediction task using transformer architectures, bypassing traditional components like non-maximum suppression [9]. Rooted in the original transformer model [10], DETR variants have been adapted for panoptic segmentation, which combines instance segmentation of "things" with semantic segmentation of "stuff" [11]. This is essential for processing drone-generated orthomosaic images in construction, where variability in scale, occlusion, and object density demands robust, real-time performance.

Despite advancements, DETR models exhibit persistent limitations in segmentation tasks, including slow convergence, inefficient attention, poor multi-scale handling, and query mismatches. The base transformer [10] lacks spatial hierarchies, leading to high computational costs ($O(n^2)$ complexity) and memory issues for high-resolution images (>1024×1024 pixels), impeding edge-device deployment. DETR [9] introduces fixed queries and bipartite matching for end-to-end detection but requires 300–500 epochs for convergence and struggles with small or occluded objects; panoptic extensions [12] add mask heads, yielding Panoptic Quality ≈43.4 on COCO due to heuristic merging errors.

Deformable DETR [13] sparsifies attention for faster training and better small-object detection but fails in boundary accuracy for crowded scenes and incurs latency from convolutional decoders. DAB-DETR [14] uses dynamic anchors for density adaptation but limits query flexibility, causing over/under-prediction in sparse/dense layouts. DN-DETR [15] employs denoising for quicker convergence but generalizes poorly to sparse scenes and multi-modal data. Lite-DETR [16] prioritizes lightweight design but sacrifices accuracy on textures and small objects. RT-DETR [17] achieves real-time speeds with intra-scale attention but trades PQ for efficiency.

Common flaws include inefficiency (>100 ms inference), mask inaccuracies, fixed queries and weak multi-scale adaptation. Our AMS-DETR mitigates these through adaptive query generation (dynamic counts for crowded and less crowded images) and hierarchical multi-scale attention, achieving higher segmentation quality and less inference time for construction orthomosaic images.

2.2 Site Layout Monitoring via Image Segmentation

Recent studies have advanced site layout monitoring by integrating UAV, AI segmentation, and BIM to enable real-time hazard detection and compliance in dynamic environments. Drone-based photogrammetry generates orthomosaic images for scalable analysis, surpassing manual methods prone to errors in large sites.

Tárrago Garay et al. [18] fused drone point clouds with 4D-BIM for progress/deviation detection (70% reduced verification), supporting unguarded edge identification but without AI segmentation. Han et al. [19] used SAM-based panoramic segmentation for indoor layout, highlighting adaptive segmentation value but limited to indoors. Akinsemoyin et al. [20] applied Faster R-CNN on UAV imagery for hardhat/activity detection, enabling machinery-proximity rerouting but instance-only. Alexandre et al. [21] combined AI-drones with LiDAR for anomaly/progress tracking, focusing on compliance without panoptic depth. Recent YOLOv8-seg improvements have advanced construction site monitoring Bhai et al. [37], offering good real-time performance on standard images. However, they face challenges with extreme scale variations and global context in large orthomosaics. Our method addresses these via adaptive queries and hierarchical multi-scale attention, enabling superior generalization for drone-based orthomosaic analysis.

In recent studies Hong and Teizer [22] introduced RAM module in DTCS, integrating SafeBIM rule-checking with RTK-GNSS trajectories for real-time struck-by hazard detection, proximity analysis, and dynamic walkway optimization, though sensor-based rather than vision-segmentation. Pfitzner et al. [23] developed KG-enhanced CV (YOLOv8x + GNN) for concrete pouring progress, linking BIM volumes to site images via spatial-temporal graphs for bottleneck identification, but activity-specific without broad safety. Pfitzner et al. [24] proposed grid-based spatial complexity analysis with ViTPoseActivity and BIM projection for rebar duration prediction, enabling localized productivity insights but focused on labor efficiency over fall hazards.

These approaches show progress in BIM-CV fusion and real-time tracking but rely on isolated detectors, lack unified instance segmentation for multi-scale features in orthomosaics, or require extra sensors, limiting geospatial risk mapping (e.g., precise guardrail/pathway buffers).Our AMS-DETR framework advances by delivering real-time instance segmentation of drone orthomosaics (guardrails, cranes, edges, pathways) with geo-spatial information for real time 4D-BIM comparison for the onsite safety Managers.

3 Methodology

To address persistent limitations in existing DETR-based instance segmentation models particularly fixed query cardinality and inefficient multi-scale feature fusion we propose an Adaptive Multi-Scale extension to the SOLQ network (AMS-DETR) for real-time analysis of high-resolution construction orthomosaic images. Building upon SOLQ's end-to-end query-based instance segmentation formulation [25], the proposed approach integrates adaptive query modulation and scale-aware hierarchical attention to improve robustness across highly variable construction scenes. Unlike earlier DETR variants that rely on a fixed number of object queries (e.g., 100 queries in [9, 17]) or convolution-heavy mask heads [26], AMS-DETR preserves SOLQ's fully transformer-based decoding while enabling dynamic allocation of representational capacity based on scene complexity. This design is particularly suited for drone-based site monitoring, where images range from sparse excavation zones to dense equipment clusters within the same orthomosaic images.

The novelty of AMS-DETR lies in its targeted architectural extensions to the SOLQ framework rather than introducing a new segmentation paradigm. Similarly, approaches such as [27] focus on global mask transformers but introduce redundant memory paths that scale poorly in occluded, high-resolution scenes. In contrast, the proposed method augments SOLQ with complexity-driven query adaptation and hierarchical multi-scale attention, allowing the decoder to adjust both the number and scale sensitivity of active queries without introducing heuristic post-processing or convolutional bottlenecks. These additions improve instance separation and boundary accuracy for small and overlapping objects, yielding measurable gains over baseline SOLQ and Mask DINO [26], while maintaining sub-50 ms inference on 1024×1024 inputs. AMS-SOLQ further reduces memory consumption by approximately 40%, making it suitable for edge deployment in construction site monitoring scenarios.

3.1 Adaptive Query Mechanism

DETR-based instance segmentation models employ a fixed number of object queries (e.g., 100 learnable embeddings in DETR [9] and DAB-DETR [14]), which leads to inefficiencies when processing scenes with large variations in object density. In construction orthomosaic imagery, this manifests as missed detections in densely

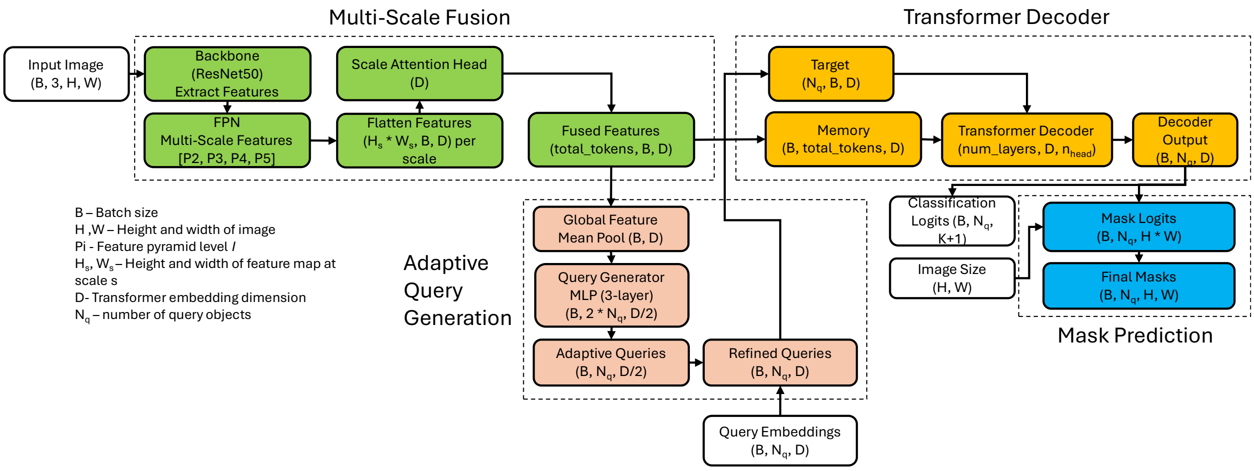

Figure 1: AMSDETR network architecture

populated regions (e.g., clustered equipment) and redundant predictions in sparse areas (e.g., open work zones), as also observed in Co-DETR [28] and RT-DETR [17].

To address this limitation within the SOLQ framework, AMS-DETR introduces an adaptive query mechanism that dynamically modulates the effective number of queries based on scene complexity. Encoder feature maps are globally aggregated and passed through a lightweight multi-layer perceptron (MLP) to estimate image complexity using normalized feature entropy:

$$H(I) = -\sum_{\mathrm{i}} pi \; log \; pi \quad (1)$$

Where p_i represents normalized activation responses. The number of active queries is then determined as:

$$N *= clip(N0 + \alpha H(I), Nmin, Nmax) \quad (2)$$

With N_0=150 base queries and $N_{\max}$=300. Queries beyond N^* are progressively suppressed during decoding using sigmoid-gated residual refinement to ensure stable optimization.

Unlike prior approaches that rely on fixed query sets [9], [14], the proposed mechanism allows the SOLQ decoder to allocate representational capacity in proportion to scene complexity, improving instance separation in crowded regions while preventing over-prediction in sparse layouts. This adaptive strategy enhances matching efficiency and reduces query redundancy without introducing heuristic filtering or post-processing. Empirically, the mechanism improves recall for small and overlapping objects compared to DAB-DETR [14] and Co-DETR [28], while maintaining real-time inference performance comparable to RT-DETR [17].

3.2 Multi-Scale Hierarchical Attention

DETR-based models exhibit limited robustness to scale variation due to rigid global attention mechanisms in DETR [9] or insufficient cross-scale fusion in Deformable DETR [13], leading to degraded performance for small objects and occluded instances in high-density construction scenes. This limitation exists mainly in large orthomosaic images, where objects span multiple orders of magnitude in spatial scale.

To address this, AMS-DETR augments the SOLQ feature encoding stage that combines deformable attention [13], [17] with global context-guided scale aggregation inspired by GM-DETR [29]. Feature maps extracted from a multi-scale backbone (e.g., ResNet-50) at strides $s \in \{8,16,32\}$ are processed in parallel and fused through scale-aware attention.

Each scale is first projected using a lightweight 1×1convolution to obtain scale-specific embeddings. A global context vector q, obtained via global feature aggregation, is then used to compute adaptive attention weights across scales. Specifically, for each scale s, a scale attention weight $\alpha_s$is computed as:

$$\alpha_s = Softmax\left(\frac{\mathrm{QK_s}}{\sqrt{d}}\right) \quad (3)$$

where $\mathrm{k}_s$denotes the pooled key embedding of scale s, and $d$is the embedding dimension. The final scale-aware representation is obtained via a weighted aggregation of scale-specific features:

$$F = \sum_s \alpha_s \mathrm{V}_s \quad (4)$$

with $\mathrm{V}_s$denoting the projected feature map at scale s. This formulation enables the encoder to dynamically emphasize high-resolution features for small objects (e.g., safety posts) and lower-resolution features for larger structures (e.g., work zones), thereby improving instance localization across a wide range of object scales.

Overall, the proposed hierarchical attention improves small-object recall and occlusion robustness compared to standard DETR [9] and Deformable DETR [13], while maintaining efficient inference through sparsified attention similar to RT-DETR [17]. Compared to convolution-heavy multi-scale fusion approaches such as Mask DINO [26], the proposed method achieves improved boundary precision without introducing high-resolution bottlenecks, making it suitable for real-time orthomosaic analysis.

4 Training Dataset

The data set used for this paper was collected weekly over a period of 1 year using DJI Mini 4 Pro. The data was collected in a time-series using manual path planning. The images collected per flight are processed into Orthomosaic images using Agisoft Metashape Pro [30] which can be seen in Figure 3. This resulted in approximately 45 Orthoimages with a resolution of approx. 14.000 x 14.000 pixels which was then split into 1024x1024 pixels and processed for training. The objects in the images were labelled using LabelMe software.

The labeled classes in Figure 2 were selected based on their direct relevance to fall-from-height hazards and site layout compliance monitoring, as guided by OSHA [2] and ANSI [7] standards. Key safety elements include guardrails, poles, and excavations (to detect unguarded edges), while site elements such as cranes, steel plates, and work areas support geospatial risk analysis (e.g., proximity to pathways) and integration with 4D-BIM models. These classes represent the most frequent and critical objects observed across the one-year dataset.

The resulting dataset contains approximately 1500 images which are filled with objects used for this work after filtering out the images that do not contain any objects. The dataset is split into 80% training and 20% validation data.

Figure 2: Labelled objects used for this work to monitor the construction site layout activity.

Figure 3: Cropped orthoimages from drone flights.

5 Results

Performance is evaluated using standard COCO-style Average Precision (AP) metrics. Bounding box AP and mask AP are reported at IoU thresholds from 0.5 to 0.95 (AP@0.5:0.95), following common practice in instance segmentation benchmarks. The performance of our model during training can be seen in the Table 1.

The overall Box AP and Mask AP are primarily affected by the relatively low performance on smaller object classes such as safety poles, cranes, and certain equipment categories. These objects occupy fewer pixels in high-resolution orthomosaic images and often appear with limited visual detail, making them more difficult for the model to detect and segment accurately.

Per-class results in Table 2 confirm that larger and well-defined objects (e.g., office containers, rollers, storage containers) achieve high AP scores, while smaller or thin-structured objects (e.g., safety poles) exhibit significantly lower performance. This imbalance explains the gap between overall metrics and the strong qualitative segmentation results observed for larger and safety-critical structures in Figure 5.

This shows that the instance segmentation over a large orthoimages of approx. 14000x14000 pixels has proved to be a fruitful way for monitoring site layout of construction site. The images also show the number of instances of each object at the bottom right. The training metrics can be seen in Figure 4 which shows the Box loss

Table 1: Performance metrics.

Model	Epochs	Box AP (0.50:0.95)	Mask AP (0.50:0.95)
AMS-DETR	140	0.54	0.42
SOLQ	170	0.42	0.29

Table 2: Per-Class AP

Class	Box AP (0.5:0.95)	Mask AP (0.5:0.95)
Articulated truck	0.54	0.47
Crane	0.37	0.29
Fence	0.61	0.41
Gate	0.45	0.23
Hydraulic excavator	0.57	0.39
Office container	0.89	0.79
Protective barrier	0.54	0.49
Roller	0.90	0.90
Safety pole	0.34	0.27
Stairs	0.68	0.58
Steel plates	0.38	0.36
Storage container	0.77	0.71
Work area	0.76	0.74

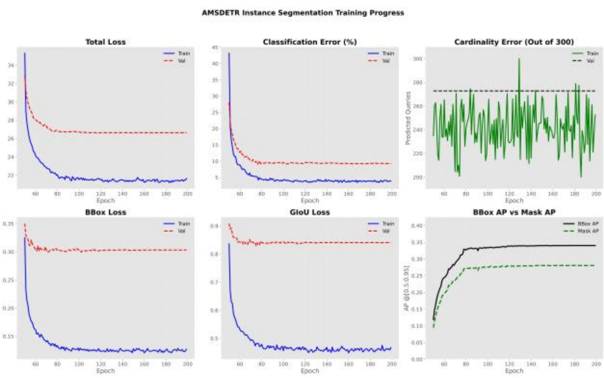

Figure 4: Training metrics.

and GIoU loss along with the BBox AP and Mask AP. The classes that most significantly affect the overall metrics are small-scale and thin-structured objects, particularly safety poles and similar site elements. Due to their limited spatial footprint in large orthomosaic images (~14,000 × 14,000 pixels), these objects are harder to localize and segment precisely, leading to lower AP values and a disproportionate impact on the aggregated performance metrics.

6 Limitations and Conclusion

The results demonstrate the efficacy of the proposed AMS-DETR model in addressing the challenges of real-time instance segmentation for construction site orthomosaic images, particularly in dynamic environments where fall hazards evolve rapidly. As shown in Table 1, the model's performance metrics highlight significant improvements in segmentation quality and inference time compared to baseline DETR variant. This mechanism dynamically adjusts query counts based on scene entropy (as quantified in the methodology), enabling the model to efficiently process sparse scenes (e.g., isolated excavations) without redundant predictions and dense clusters (e.g., equipment-crowded zones) without under-detection.

Figure 4 training metrics further illustrate the model's convergence advantages: the box loss and generalized IoU (GIoU) loss stabilize within 70–100 epochs, faster than traditional DETR [9] or Deformable DETR [13], Mask AP improvements are evident in Figure 5, where boundary precision is maintained even in crowded predictions, with a confidence threshold of 80% effectively filtering noise while preserving accurate instance separation. However, the overall metrics are impacted by challenges in detecting and segmenting small-scale objects, such as safety poles and certain equipment instances. These objects often occupy very few pixels in orthomosaic images and may suffer from low contrast, occlusion, or scale variation, making them difficult for transformer-based models to capture effectively. This limitation is consistent with known challenges in instance segmentation, where small objects contribute disproportionately to performance degradation despite accurate detection of larger structures which align with observations in related works [23, 24], where knowledge graph enhancements or grid-based analyses mitigate similar issues in activity-specific segmentation, suggesting potential synergies for future iterations. This observation highlights a known limitation of transformer-based detectors in high-resolution imagery, where small-object representation and feature resolution remain critical bottlenecks despite improvements in multi-scale attention.

Figure 5: Full ortho instance segmentation for construction site layout status on 2025-05-17.

Practically, the full orthomosaic instance segmentation in Figure 5 exemplifies the system's geospatial utility by geo-aligning segmented features with 4D-BIM models, deviations (e.g., unguarded edges <10 cm from pathways) can be detected with cm-level accuracy, facilitating Voronoi-based risk heatmaps and proximity analyses (e.g., cranes within 2 m of walkways) which is part of the future works. This proactive capability surpasses reactive manual inspections [6], offering scalable compliance monitoring for small firms. Integration with the SafeBIM platform [36] enables interactive visualizations, such as predictive compliance scores derived from temporal changes across the 1-year dataset, empowering site managers to prioritize interventions and reduce secondary risks.

Despite these strengths there are certain limitations. The dataset, while comprehensive (1500 labelled images from 45 orthomosaics), is site-specific and collected via manual drone paths, potentially introducing biases in viewpoint and weather conditions. Model inference, though real-time (<50 ms/frame), may degrade on ultra-high-resolution inputs without further optimization, and the reliance on LabelMe for annotations could amplify human error in ambiguous regions like "work areas." Compared to sensor-fusion approaches [22, 31] our vision-only pipeline is cost-effective but lacks depth cues, limiting 3D hazard assessment. Future work could incorporate multimodal data (e.g., LiDAR) or advanced refinements like diffusion-based boundary sharpening [32] to enhance amorphous class handling. Additionally, expanding the dataset with diverse global sites could improve generalization, aligning with ILO's call for region-specific safety adaptations [5].

Furthermore, aligning construction site layout monitoring using instance segmentation and using it in a Digital Twin for Construction Safety (DTCS) framework has potential for practitioners to detect and track relevant progress while comparing as-performed vs. as-planned status effortlessly [33-36]. Providing previously not available information in safety meetings (e.g., toolbox and safety planning or review meetings) adds value to select and apply the best control measures.

Acknowledgment

The research presented in this paper was funded by the European Union Horizon 2020 research and innovation program under grant agreement no. 101162318 (AM2PM) and Erhvervsfyrtårn Femern (FF6S). The Business Lighthouse Femern is co-financed by the European Union and the Danish Business Development Agency.

References

[1] J. Teizer, "Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 225–238, Apr. 2015. https://doi.org/10.1016/j.aei.2015.03.006

[2] OSHA, "Fall protection standards," 2023. https://www.osha.gov/fall-protection/standards

[3] "National census of fatal occupational injuries in 2023", U.S. Bureau of Labor Statistics (BLS), 2024. https://www.bls.gov/news.release/cfoi.nr0.htm

[4] "Nonfatal occupational injuries and illnesses," U.S. Bureau of Labor Statistics (BLS), 2023. https://www.bls.gov/iif/

[5] "Work-related fatalities in construction," International Labour Organization (ILO), 2023. https://www.ilo.org

[6] H. J. Lipscomb, A. L. Schoenfisch, W. Cameron, K. L. Kucera, D. Adams, and B. A. Silverstein, "How well are we controlling falls from height in construction? Experiences of union carpenters in Washington State, 1989–2008," *American J Industrial Med*, vol. 57, no. 1, pp. 69–77, Jan. 2014, https://doi.org/10.1002/ajim.22234

[7] "Z359.2 Fall protection code," American National Standards Institute (ANSI), 2023. https://webstore.ansi.org

[8] M. Golparvar-Fard, F. Peña-Mora, C. A. Arboleda, and S. Lee, "Visualization of Construction Progress Monitoring with 4D Simulation Model Overlaid on Time-Lapsed Photographs," *J. Comput. Civ. Eng.*, vol. 23, no. 6, pp. 391–404, Nov. 2009. https://doi.org/10.1061/(ASCE)0887-3801(2009)23:6(391)

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," May 28, 2020, *arXiv*. https://arxiv.org/abs/2005.12872

[10] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*. https://arxiv.org/abs/1706.03762

[11] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic Segmentation," Apr. 10, 2019, *arXiv*: arXiv. https://arxiv.org/abs/1801.00868

[12] Z. Li *et al.*, "Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers," Mar. 18, 2022, *arXiv*. https://arxiv.org/abs/2109.03814

[13] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," Mar. 18, 2021, *arXiv*: arXiv. https://arxiv.org/abs/2010.04159

[14] S. Liu *et al.*, "DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR," Mar. 30, 2022, *arXiv*: arXiv. https://arxiv.org/abs/2201.12329

[15] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L.

Zhang, “DN-DETR: Accelerate DETR Training by Introducing Query DeNoising,” Dec. 08, 2022, *arXiv*. https://arxiv.org/abs/2203.01305

[16] F. Li *et al.*, “Lite DETR : An Interleaved Multi-Scale Encoder for Efficient DETR,” Mar. 13, 2023, *arXiv*. https://arxiv.org/abs/2303.07335

[17] Y. Zhao *et al.*, “DETRs Beat YOLOs on Real-time Object Detection,” Apr. 03, 2024, *arXiv*. https://arxiv.org/abs/2304.08069

[18] N. Tárrago Garay, J. C. Jimenez Fernandez, R. San Mateos Carreton, M. A. Montes Grova, O. Kruth, and P. Elguezabal, “Integration of Drone-Based 3D Scanning and BIM for Automated Construction Progress Control,” *Buildings*, vol. 15, no. 19, p. 3487, Sep. 2025. https://doi.org/10.3390/buildings15193487

[19] J. Han, S. Yoon, M. Kang, and T. Kim, “Approach to Enhancing Panoramic Segmentation in Indoor Construction Sites Based on a Perspective Image Segmentation Foundation Model,” *Applied Sciences*, vol. 15, no. 9, p. 4875, Apr. 2025. https://doi.org/10.3390/app15094875

[20] A. Akinsemoyin, I. Awolusi, D. Chakraborty, A. J. Al-Bayati, and A. Akanmu, “Unmanned Aerial Systems and Deep Learning for Safety and Health Activity Monitoring on Construction Sites,” *Sensors*, vol. 23, no. 15, p. 6690, Jul. 2023. https://doi.org/10.3390/s23156690

[21] A. Oliveira, A. Dias, T. Santos, P. Rodrigues, A. Martins, and J. Almeida, “LiDAR-Based Unmanned Aerial Vehicle Offshore Wind Blade Inspection and Modeling,” *Drones*, vol. 8, no. 11, p. 617, 2024. https://doi.org/10.3390/drones8110617

[22] K. Hong and J. Teizer, “Digital construction site layout planning and real-time trajectory analysis for proactive safety monitoring and control of struck-by hazards,” *Automation in Construction*, vol. 177, p. 106353, Sep. 2025. https://doi.org/10.1016/j.autcon.2025.106353

[23] F. Pfitzner, S. Hu, A. Braun, A. Borrmann, and Y. Fang, “Monitoring concrete pouring progress using knowledge graph-enhanced computer vision,” *Automation in Construction*, 174, p. 106117, Jun. 2025. https://doi.org/10.1016/j.autcon.2025.106117

[24] F. Pfitzner, A. Braun, and A. Borrmann, “Object Detection-Based Knowledge Graph Creation: Enabling Insight into Construction Processes,” in *Computing in Civil Engineering 2023*, Corvallis, Oregon: American Society of Civil Engineers, Jan. 2024, pp. 186–193. https://doi.org/10.1061/9780784485224.023

[25] B. Dong, F. Zeng, T. Wang, X. Zhang, and Y. Wei, “SOLQ: Segmenting Objects by Learning Queries,” Sep. 30, 2021, *arXiv*. https://arxiv.org/abs/2106.02351

[26] F. Li *et al.*, “Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation,” Dec. 12, 2022, *arXiv*. https://arxiv.org/abs/2206.02777

[27] A. Kirillov *et al.*, “Segment Anything,” Apr. 05, 2023, *arXiv*. https://arxiv.org/abs/2304.02643

[28] Z. Zong, G. Song, and Y. Liu, “DETRs with Collaborative Hybrid Assignments Training,” Aug. 10, 2023, *arXiv*. https://arxiv.org/abs/2211.12860

[29] X. Liu, X. Yang, L. Shao, X. Wang, Q. Gao, and H. Shi, “GM-DETR: Research on a Defect Detection Method Based on Improved DETR,” *Sensors*, vol. 24, no. 11, p. 3610, Jun. 2024. https://doi.org/10.3390/s24113610

[30] Agisoft LLC, *Agisoft Metashape professional*. (2023). https://www.agisoft.com/pdf/metashape-pro_2_0_en.pdf

[31] W. Choi, S. Na, and S. Heo, “Integrating Drone Imagery and AI for Improved Construction Site Management through Building Information Modeling,” *Buildings*, vol. 14, no. 4, p. 1106, Apr. 2024. https://doi.org/10.3390/buildings14041106

[32] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Dec. 16, 2020, *arXiv*. https://arxiv.org/abs/2006.11239

[33] S. Siebert, J. Teizer, “Mobile 3D Mapping for Surveying Earthwork Projects using an Unmanned Aerial Vehicle (UAV) System,” Automation in Construction, 2014, vol. 41, pp. 1–14, http://dx.doi.org/10.1016/j.autcon.2014.01.004

[34] V. Bharathi, J. Teizer, “Construction site layout planning and progress monitoring using computer vision on drone images,” European Conference on Computing in Construction, CIB W78 Conference on IT in Construction, Porto, Portugal, 2025, https://doi.org/10.35490/EC3.2025.313

[35] K.W. Johansen, R. Pimentel de Figueiredo, O. Golovina, J. Teizer, “Autonomous Safety Barrier Inspection in Construction: An Approach Using Unmanned Aerial Vehicles and SafeBIM,” 38th International Symposium on Automation and Robotics in Construction, Dubai, UAE, 2021, https://doi.org/10.22260/ISARC2021/0085

[36] J. Teizer, K.W. Johansen, C.L. Schultz, K. Speiser, K. Hong, O. Golovina, “A Digital Twin Model for Advancing Construction Safety,” J. Fottner et al. (Eds.): CLEaR 2023, Lecture Notes in Civil Engineering (LNCE 390), 2024, ISBN: 978-3-031-44020-5, pp. 201–212, https://doi.org/10.1007/978-3-031-44021-2_22

[37] R. Bai, M. Wang, Z. Zhang, J. Lu, and F. Shen, "Automated Construction Site Monitoring Based on Improved YOLOv8-seg Instance Segmentation Algorithm," IEEE Access, vol. 11, pp. 139082-97, 2023, https://ieeexplore.ieee.org/document/10348569