A Critical Review of Supervised and Reduced-Dataset Vision-Based Construction Activity Recognition Methods

Ali Ghelmani[1], Amin Hammad[1]*

[1]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada
ali.ghelmanirashidabad@concordia.ca
amin.hammad@concordia.ca*

Abstract

Tracking and monitoring construction activities (e.g., equipment and workers) is crucial for improving site productivity and safety. Manual monitoring, however, is time-consuming, error-prone, and costly. To address these challenges, computer vision (CV) techniques have emerged for automated detection and classification of construction activities. Traditional supervised CV-based methods often rely on extensive annotated datasets, making it impractical to accommodate the wide range of equipment types and activities. Reduced-dataset methods, such as self-supervised, semi-supervised, and generative approaches, have recently surfaced as promising alternatives. Furthermore, construction digital twins are rapidly gaining traction as an integral component of Construction 4.0. However, their practical implementation requires efficient monitoring systems capable of handling vast data streams and dynamic site conditions using edge devices. Unmanned Aerial Vehicles (UAVs) are an example of such devices that have been increasingly utilized to provide monitoring of dynamic site conditions and consequently, necessitate the development of corresponding activity recognition methods. Previous reviews have primarily focused on supervised methods with stationary cameras, as such, this paper fills a critical gap by examining recent reduced-dataset approaches and also providing a forward-looking review of CV-based activity monitoring on edge devices.

1 Introduction

Monitoring the activities of workers and equipment on construction sites is critical for improving project productivity, safety, and overall efficiency. Traditionally, this monitoring has relied on manual observations by site superintendents, a time-consuming and error-prone approach that struggles to keep pace with the complexity of modern construction projects [1]. As the demand for timely, high-quality project delivery increases, so does the need for automated and accurate continuous monitoring systems [2]. Furthermore, the concept of Construction Digital Twin (CDT) has gained traction as a promising platform for real-time data integration, simulation, and optimization [3]. However, one of the main challenges in CDT implementation, is the availability of accurate, real-time monitoring data, further exacerbating the need for the development of corresponding monitoring methods [3].

Recent prevalence of surveillance cameras in most construction sites has triggered the development of various automated vision-based activity recognition methods for addressing the above challenges [4]. Early vision-based approaches relied heavily on hand-crafted features [4], but the advent of deep learning, particularly Convolutional Neural Networks (CNNs), has shifted the focus towards more robust and scalable solutions [5].

Initial works employed 2D CNNs combined with rule-based logic or recurrent architectures [6, 7] for sequential processing of spatial and temporal information. More recently, 3D CNNs that jointly model spatiotemporal features have emerged to obtain improved performance [8–10]. However, the supervised approach of these methods requires large, annotated datasets for each piece of equipment and activity that is infeasible to collect, hence limiting their scalability [11]. To address these data-related bottlenecks, reduced-dataset methods (e.g., self-supervised [12], semi-supervised [13], and generative techniques [14]) have begun to gain traction. Self-supervised learning involves training models on unlabeled data using labels extracted from the data itself [12] (e.g., predicting rotations or temporal consistency); while semi-supervised learning leverages a small amount of labeled data in conjunction with unlabeled data, typically through techniques such as pseudo-labeling, to enhance learning efficiency [13]. Generative approaches, however, focus on synthesizing additional artificial data that mimic real-world conditions to expand the dataset size and diversity, thereby improving model generalization [14]. While some studies have explored these approaches for equipment detection [11, 13], few have tackled the more complex problem of activity recognition [12].

Additionally, majority of the state-of-the-art activity recognition methods utilize videos collected from stationary cameras, however, CDT requirement for the availability of real-time data, given the dynamic nature of construction sites, necessitates the development of activity recognition methods employing edge devices such as Unmanned Aerial Vehicles (UAVs). As such, while previous review papers have focused on traditional supervised activity recognition methods (e.g., [4, 15]), there remains a need to thoroughly review these emerging technologies and discuss their potential in construction, particularly for activity recognition. This paper directly addresses these identified gaps by comprehensively reviewing recent advancements in reduced-dataset construction activity recognition methods and exploring the application and potential of Computer Vision (CV)-based monitoring solutions deployed on edge devices. The considered aspects aim to enhance the scalability of current supervised methods and address the challenge of providing dynamic construction site monitoring data essential for the development and implementation of CDTs.

2 Review Method

This study employed a systematic approach to identify and review the existing literature on vision-based construction activity recognition methods. Three main steps were followed: (1) paper collection, (2) initial screening, and (3) in-depth review and synthesis. First, an exhaustive search was conducted using multiple academic databases (e.g., Web of Science, Scopus, Google Scholar) to retrieve publications spanning approximately the last decade. The search keywords were designed to cover three broad categories: (1) terms related to construction (e.g., "construction," "vision-based methods," "automated monitoring," "activity recognition," "productivity analysis,"), (2) terms related to reduced-dataset deep learning methods (e.g., "self-supervised learning," "semi-supervised learning," "few-shot learning", "generative adversarial networks,", "synthetic data"), and (3) terms related to edge devices (e.g., "moving camera activity recognition," "UAV-based monitoring," "aerial construction monitoring").

Following the initial collection, titles and abstracts were screened to remove works focusing solely on non-construction domains or unrelated technologies. However, where gaps existed, particularly for activity recognition using moving edge devices, the search scope was broadened to include highly relevant and applicable methods from the general CV domain. Thus, papers discussing supervised or reduced-dataset approaches for construction activity recognition using stationary cameras, as well as relevant CV methods employing moving edge devices, were selected. Subsequently, a brief content review was then performed to ensure that selected papers substantially contributed to the current review scope. Finally, an in-depth qualitative analysis of the remaining papers was conducted to identify key trends and challenges in the literature.

3 Construction Activity Recognition Using Stationary Cameras

This section reviews supervised and reduced-dataset construction activity recognition methods using stationary cameras. Moreover, Figure 1 presents a categorization of these methods, along with representative publications for each approach.

3.1 Supervised Approaches

In the past decade, many vision-based methods have been proposed for the task of construction activity recognition. With the advent of deep learning, 2D CNNs became central to extracting spatial features from images, while temporal dependencies were increasingly modeled using additional modules such as Hidden Markov Models [16], or Long Short-Term Memory (LSTMs) networks [17]. Such combinations improved performance but often introduced computational overhead by separating spatial and temporal processing into multiple stages [8].

The 3D CNN-based methods alleviated this issue by incorporating the spatiotemporal data extraction into a single architecture. For instance, Chen et al. [1] proposed a three-stage excavator activity recognition method comprising of detection, tracking, and a final activity recognition stage using a 3D CNN network. In another study [5], they used a similar approach to detect excavator and dump truck activities, and to identify reasons for excavator idling by analyzing their interactions. A similar multi-stage approach for worker activity recognition was also proposed by Lou et al. [18]. However, the main limitations of these methods include not being fully optimized and also the propagation of errors from earlier stages to the later ones; thus, degrading the overall performance [19].

To address the abovementioned limitations, Jung et al. [19] proposed a 3D CNN-based single-stage method for simultaneous detection of multiple construction equipment and recognizing their activities. Others have also proposed single-stage methods based on the You Only Watch Once (YOWO) [20] method for joint detection and classification of construction workers' [8] and excavators' [21] activities. Despite improved performance, the main drawback of these methods is the

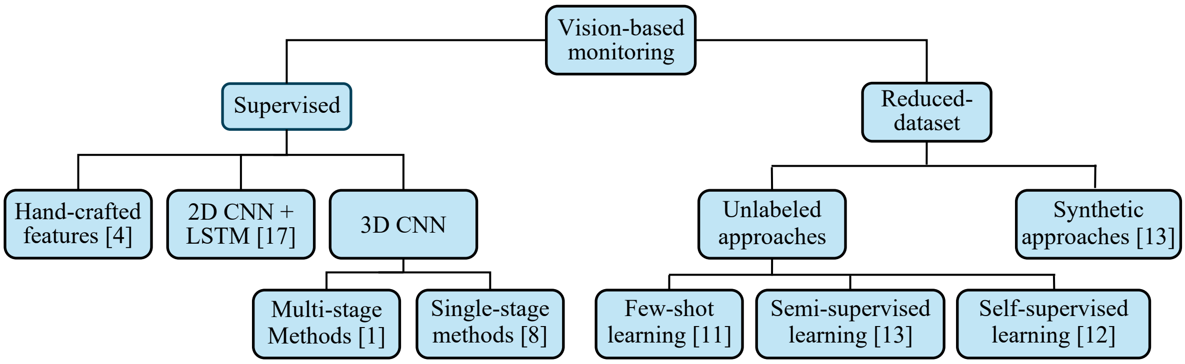

Figure 1. Categories of supervised and reduced-dataset construction activity recognition methods

necessity of post-processing due to their low per-frame activity recognition and localization performance.

Recently, a combination of multiple data modalities for activity recognition has also been suggested by some works. For example, Luo et al. [22] developed a CNN network by integrating RGB, optical flow, and gray-scale images to monitor workers' activities. Yang et al. [23] proposed a two-stream architecture for recognizing workers' unsafe behavior using a transformer model and a 3D CNN network to process RGB and Temporal Gradient (TG) data, respectively. Ghelmani and Hammad [10], also developed a multi-stream method, which leveraged TG data through knowledge distillation to achieve accurate, real-time per-frame activity recognition and localization performance for excavator activities.

Even though the state-of-the-art construction activity recognition methods can achieve high performance, their supervised approach necessitates extensive, annotated datasets for every piece of equipment and each specific activity. Since the types of construction equipment frequently change from one project phase to another [11], generating such comprehensive labeled dataset becomes a significant obstacle for the development of a general activity recognition method.

3.2 Reduced-Dataset Approaches

Researchers in the construction domain have proposed several methods for reducing the time and effort required for dataset collection and annotation. Liu and Golparvar-Fard [24] proposed crowdsourcing the annotation task using Amazon Mechanical Turk. However, such approaches still require human effort for data annotation and for removing the mistakes in the prepared dataset due to crowdsourcing [13]. Generally, reduced-dataset methods can be broadly categorized into methods that utilize a synthetic approach for data generation and methods that utilize unlabeled data for reducing the data annotation cost.

3.2.1 Synthetic Approaches

Calderon et al. [25] employed a virtual and kinematically articulated 3D model of excavators to generate synthetic 2D pose data for excavator activity recognition. Assadzadeh et al. [26] also generated a synthetic dataset for excavator pose estimation using domain randomization. Soltani et al. [27] developed a dataset of synthetically generated images of deformable excavator parts, which were used to estimate the excavator's 3D pose and consequently its kinematic posture. Data augmentation schemes using Generative Adversarial Networks (GANs) were also proposed by some researchers to enhance the variability and size of a collected dataset for construction equipment detection [28], activity recognition [14], and site monitoring [29]. Although synthetic approaches improve model's generalization through generation of challenging scenarios such as viewpoint variation, or partial occlusion, however a major limitation of these methods is the inherent visual differences between synthetic and real data [13], hence limiting their real-world applicability.

3.2.2 Unlabelled-Data Approaches

Few-shot learning. Kim and Chi [11] developed a few-shot learning approach by first training a model on a base dataset with sufficient training images of multiple construction equipment. Afterwards, the trained model was used in a transfer learning scheme to learn to detect new equipment with only a few available labeled images. In another study, Kim et al. [30] proposed a three-stage active learning method in which uncertainty evaluation was used to select an informative subset from the unlabeled data. Then, the selected small subset was labeled manually by the user and used for model training. Recently, Chen et al. [31] developed a multi-stage approach for recognizing excavator activities in a zero-shot learning setting in which the detected non-idling frames were input into a pre-trained CLIP (Contrastive Language-Image Pre-training) model [32] for activity recognition. A similar CLIP-based method for construction safety inspection was also proposed by Tsai et al. [33]. While few-shot and zero-shot learning techniques aim to minimize reliance on large, labeled datasets, they remain restricted to learning only from the limited number of provided annotated images. Consequently, their final performance can fluctuate greatly depending on the quality and characteristics of the provided data [11].

Semi-supervised learning. Xiao et al. [13] introduced a teacher-student architecture in which a teacher network was first trained on a small, labeled dataset using weak data augmentation (e.g., horizontal translation). The trained teacher then generated pseudo-labels for unlabeled data that underwent stronger augmentations (e.g., cropping) to train the student network (the final output of this method). Such strong augmentations can simulate challenging scenarios, such as occlusions and significant viewpoint variations, potentially enhancing the student network's robustness to these real-world challenges. However, the main drawback of this method is the reliability of the generated pseudo-labels. In other words, Since the teacher network itself is trained in a supervised manner, it requires a sufficiently large, labeled dataset to produce accurate pseudo-labels, which greatly limits the method’s applicability for reducing the annotation requirements. Furthermore, pseudo-labeling inherently risks introducing biases into the training data, as incorrect predictions made by the teacher model might be confidently assigned as accurate labels. Such biases could subsequently skew the model toward frequently occurring or easily identifiable examples, potentially diminishing performance on less common, complex, or ambiguous scenarios.

Self-supervised learning. The previous methods leveraged unlabeled data solely for construction equipment detection. However, the collection of an annotated dataset for activity recognition is even more challenging, as they often require frame-level labeling. Thereby increasing the necessity of developing methods that exploit unlabeled data for activity recognition. Nevertheless, only one study has addressed this need: Ghelmani and Hammad [12] proposed a self-supervised, contrastive learning approach for pretraining a 3D CNN network using unlabeled data, followed by fine-tuning using a small amount labeled data for excavator activity recognition. Although this method reduced annotation demands, it only applied self-supervised learning to activity recognition and not localization. No research to date has employed unlabeled data to facilitate the localization of construction equipment, even though this step is among the most annotation-intensive processes. Advancing research in this area could thus pave the way for more general and scalable construction activity recognition methods.

4 CV-based Activity Recognition on Edge Devices

As mentioned in Section 1, an indispensable component for CDT implementation is real-time monitoring using edge devices which include stationary and moving cameras (e.g., on UAVs). Methods developed for stationary cameras were reviewed in Section 3 while this section reviews methods developed for activity recognition using moving edge devices. Moreover, given the lack of activity recognition methods in the construction domain using edge devices, this section also presents a comprehensive review of the corresponding methods developed in the CV domain.

Activity recognition methods using moving edge devices can generally be divided into ground-based and aerial approaches.

4.1 Ground-based Methods

Ground-based methods in the CV domain are mainly concerned with applications related to self-driving cars. For instance, Yu et al. [34] proposed a method for real-time activity recognition in unconstrained video streams by utilizing frame-level foreground segmentation to obtain spatiotemporal proposals for activity recognition. Zhuang et al. [35] proposed a multi-stage road active agent detection and activity classification by initially passing the optical flow and RGB frames through corresponding backbones for object detection, followed by tracking and activity recognition stages. However, the main limitations of ground-based methods include: (1) the need for automated or manual navigations systems, which can be particularly challenging in a construction site; and (2) the limited field of view and ground-level camera angle, which can greatly increase the chance for occlusion and hence, reduce the model performance.

4.2 Aerial Methods

Compared with ground-based methods, a greater volume of research has been conducted on activity recognition using aerial imagery [36, 37]. Unlike traditional stationary cameras, UAVs capture videos from higher altitudes, allowing for a broader field of view. Additionally, UAVs can cover an large surveillance area due to their significant range of movement [36]. However, these factors introduce complex challenges related to detection, tracking, and activity recognition [38]. These challenges include the small resolution of objects of interest due to the high camera altitude. As a result, most of the pixels in a video frame are occupied by the background, with only a small fraction representing the targeted actions. Furthermore, the size and scale of the video can vary dramatically depending on the UAV's flying altitude, making the extraction of spatiotemporally representative features quite challenging with the possible necessity of a preprocessing stage for scene stabilization [38]. Current activity recognition techniques are computationally demanding and memory-intensive, making it challenging to respect the time constraints in real-world applications. This is particularly true for edge devices with limited computational resources, such as UAVs [39]. To mitigate these constraints, existing methods typically adopt lightweight neural network architectures or hybrid inference schemes, where initial lightweight processing occurs onboard, and computationally intensive processing is delegated to remote servers or cloud-based resources (e.g., [38]). Nonetheless, effectively balancing computational efficiency with high model performance remains an active area of open research.

Dataset collection. The first step in addressing the above challenges using deep learning-based methods is the availability of extensive datasets. To this end, Barekatain et al. [40] proposed one of the first aerial video datasets for activity recognition comprising of 13 different types of human activities. Perera et al. [41] also developed a drone video dataset for activity recognition and pose estimation including 240 video clips recorded at low altitudes and low drone speed to provide high resolution human pose details.

Supervised approaches. Mliki et al. [42] proposed one of the first UAV activity recognition methods comprising of an offline training stage and an inference stage. Their approach utilized a scene stabilization technique for preprocessing followed by a human and non-human detector to filter out moving non-human entities. Finally, the detecting human agents were used for activity recognition. To address the problem of small-scale human targets in high altitude drone videos, Wang et al. [38] proposed the auto zoom and temporal reasoning method, which initially scales and crops human detections. Subsequently, a hybrid approach combining a lightweight 3D CNN network on edge devices and a more powerful 3D CNN network on local devices were employed for activity recognition.

Recently, many approaches have suggested a combination of feature-based and deep learning methods for aerial video activity recognition. For instance, Gundu and Syed [43] utilized Histogram of Oriented Gradients features for pattern extraction followed by a combination of CNN and LSTM networks for human activity recognition. Peng and Razi [44] used speeded up robust features (SURF) and Lucas-Kanade method for motion stabilization of the UAV captured videos, followed by detection and activity recognition stages. Kothandaraman et al. [45] employed Fourier analysis to separate out the human agent from the background in the frequency domain. Afterwards, a novel Fourier space-time attention module was utilized for activity recognition. Despite their improved performance, these methods still depend on manually engineered feature extractors, which introduce biases and constrain the representational capacity of the subsequent deep learning networks. The prevalence of such hybrid approaches can be attributed, in part, to the scarcity of large, high-quality aerial video datasets as reviewed earlier. As such, further development in this domain warrants the availability of large, annotated datasets.

Reduced-dataset approaches. As stated in Section 3.2, another possible approach for addressing the abovementioned limitations can be to utilize reduced-dataset approaches. To this end, Sultani and Shah [46] introduced a multi-task learning approach for activity recognition using aerial images, which leveraged a GAN network to synthetically transform ground activity video features into their aerial counterparts. To further improve performance, a small number of real aerial activity videos were also incorporated alongside the synthetic data. Choi et al. [47] proposed unsupervised and semi-supervised domain adaptation approaches to leverage the existing action recognition datasets in combination with unannotated or only few annotated drone videos for activity recognition. Although these approaches lessen the need for annotated aerial data, they mostly focus on domain adaptation from ground videos to aerial counterparts, a process that inevitably introduces errors and restricts the final model's accuracy. As such, there is a need in this area for new methods that can directly exploit unlabeled aerial videos to facilitate both activity recognition and localization, without relying solely on cross-domain transformations.

Aerial approaches in construction. Several studies have addressed the task of detecting construction entities from aerial images. For example, D. Kim et al. [48] proposed a UAV-assisted monitoring method as a proactive safety measure against struck-by hazards by utilizing a YOLOv3 [49] network for object localization and an image rectification method to measure actual distance from collected images. Guo et al. [50] proposed a method for detecting construction equipment from UAV-captured images. The proposed method combined orientation-aware feature fusion with a 2D CNN network for detecting equipment with arbitrary orientations. Comparison between different object detection methods were also carried out by Mutis et al. [51] for construction resource detection and by Pi et al. [52] to detect various construction site entities using aerial top-view images. A few studies have also developed UAV-based methods to improve construction site safety by detecting small and medium objects to avoid struck-by accidents [53–55] and for monitoring worker safety compliance [56].

To address dataset requirements for training deep learning methods Ersoz et al. [57] developed an aerial image dataset for construction machinery detection.

Others have developed methods to reduce annotated data requirements for aerial construction entity detection. For example, Bang et al. [58] proposed a variety of data augmentation techniques including moving-and-inpainting, cut-and-paste, and image-variation to improve the performance of a Faster R-CNN [59] network for detecting construction resources from UAV-captured images. Wang et al. [60] employed a 3D reconstruction technique using a UAV and photogrammetry to synthetically generate an image dataset for the analysis of crane productivity. Specifically, they reconstructed a crane bucket and used this synthetic data to conduct their study. However, as discussed in Section 4, all current aerial approaches in construction have focused solely on equipment or resource detection. To enable accurate, real-time monitoring that supports CDT implementation, methods must also address activity recognition, an area that remains unexplored for aerial-based methods.

5 Summary and Contributions

This paper provided a critical review of the state-of-the-art in vision-based construction activity recognition, examining both traditional supervised methods and the emerging reduced-dataset methods that leverage self-supervised, semi-supervised, and generative approaches. The data annotation barriers in current supervised methods were identified, followed by highlighting promising reduced-dataset strategies that mitigate these challenges. Moreover, given the growing demand of emerging CDT systems for the availability of accurate, real-time monitoring data, this review addressed a previously underexplored area: activity recognition using moving cameras (e.g., on UAVs) on edge devices, which is essential for capturing dynamic site conditions. Finally, the existing research gaps were identified, especially in reduced-dataset activity localization and aerial-based applications, hence proposing future directions for development.

References

[1] C. Chen, Z. Zhu, and A. Hammad, “Automated excavators activity recognition and productivity analysis from construction site surveillance videos,” *Automation in Construction*, vol. 110, Feb. 2020.

[2] J. Kim, S. Chi, and J. Seo, “Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks,” *Automation in Construction*, vol. 87, pp. 297–308, Mar. 2018.

[3] D.-G. J. Opoku, S. Perera, R. Osei-Kyei, and M. Rashidi, “Digital twin application in the construction industry: A literature review,” *Journal of Building Engineering*, vol. 40, Aug. 2021.

[4] C. Chen, Z. Zhu, and A. Hammad, “Critical Review and Road Map of Automated Methods for Earthmoving Equipment Productivity Monitoring,” *Journal of Computing in Civil Engineering*, vol. 36, no. 3, May 2022.

[5] C. Chen, Z. Zhu, A. Hammad, and M. Akbarzadeh, “Automatic Identification of Idling Reasons in Excavation Operations Based on Excavator–Truck Relationships,” *Journal of Computing in Civil Engineering*, vol. 35, no. 5, Sep. 2021.

[6] J. Zhang, L. Zi, Y. Hou, M. Wang, W. Jiang, and D. Deng, “A Deep Learning-Based Approach to Enable Action Recognition for Construction Equipment,” *Advances in Civil Engineering*, vol. 2020, Nov. 2020.

[7] M. Younesi Heravi, Y. Jang, I. Jeong, and S. Sarkar, “Deep learning-based activity-aware 3D human motion trajectory prediction in construction,” *Expert Systems with Applications*, vol. 239, Apr. 2024.

[8] G. Torabi, A. Hammad, and N. Bouguila, “Two-Dimensional and Three-Dimensional CNN-Based Simultaneous Detection and Activity Classification of Construction Workers,” *Journal of Computing in Civil Engineering*, vol. 36, no. 4, Jul. 2022.

[9] S. Jung, J. Jeoung, D. Lee, H. Jang, and T. Hong, “Visual–auditory learning network for construction equipment action detection,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 14, pp. 1916–1934, Feb. 2023.

[10] A. Ghelmani and A. Hammad, “Improving single-stage activity recognition of excavators using knowledge distillation of temporal gradient data,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 39, no. 13, Jul. 2024.

[11] J. Kim and S. Chi, “A few-shot learning approach for database-free vision-based monitoring on construction sites,” *Automation in Construction*, vol. 124, Apr. 2021.

[12] A. Ghelmani and A. Hammad, “Self-supervised contrastive video representation learning for construction equipment activity recognition on limited dataset,” *Automation in Construction*, vol. 154, Oct. 2023.

[13] B. Xiao, Y. Zhang, Y. Chen, and X. Yin, “A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation,” *Advanced Engineering Informatics*, vol. 50, Oct. 2021.

[14] Y. Shen, J. Wang, S. Mo, and X. Gu, “Data augmentation aided excavator activity recognition using deep convolutional conditional generative

adversarial networks," *Advanced Engineering Informatics*, vol. 62, Oct. 2024.

[15] B. Sherafat, C. R. Ahn, R. Akhavian, A. H. Behzadan, M. Golparvar-Fard, H. Kim, *et al.*, "Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review," *Journal of Construction Engineering and Management*, vol. 146, no. 6, Jun. 2020.

[16] X. Luo, H. Li, D. Cao, F. Dai, J. Seo, and S. Lee, "Recognizing Diverse Construction Activities in Site Images via Relevance Networks of Construction-Related Objects Detected by Convolutional Neural Networks," *Journal of Computing in Civil Engineering*, vol. 32, May 2018.

[17] J. Kim and S. Chi, "Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles," *Automation in Construction*, vol. 104, Aug. 2019.

[18] X. Luo, H. Li, Y. Yu, C. Zhou, and D. Cao, "Combining deep features and activity context to improve recognition of activities of workers in groups," *Computer-Aided Civil and Infrastructure Engineering*, vol. 35, no. 9, pp. 965–978, 2020.

[19] S. Jung, J. Jeoung, H. Kang, and T. Hong, "3D convolutional neural network-based one-stage model for real-time action detection in video of construction equipment," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 1, pp. 126–142, Jan. 2022.

[20] O. Köpüklü, X. Wei, and G. Rigoll, "You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization." arXiv, Oct. 18, 2021.

[21] M.-Y. Cheng, M.-T. Cao, and C. K. Nuralim, "Computer vision-based deep learning for supervising excavator operations and measuring real-time earthwork productivity," *The Journal of Supercomputing*, vol. 79, no. 4, Mar. 2023.

[22] H. Luo, C. Xiong, W. Fang, P. E. D. Love, B. Zhang, and X. Ouyang, "Convolutional neural networks: Computer vision-based workforce activity assessment in construction," *Automation in Construction*, vol. 94, pp. 282–289, Oct. 2018.

[23] M. Yang, C. Wu, Y. Guo, R. Jiang, F. Zhou, J. Zhang, *et al.*, "Transformer-based deep learning model and video dataset for unsafe action identification in construction projects," *Automation in Construction*, vol. 146, Feb. 2023.

[24] K. Liu and M. Golparvar-Fard, "Crowdsourcing Construction Activity Analysis from Jobsite Video Streams," *Journal of Construction Engineering and Management*, vol. 141, no. 11, Nov. 2015.

[25] W. T. Calderon, D. Roberts, and M. Golparvar-Fard, "Synthesizing Pose Sequences from 3D Assets for Vision-Based Activity Analysis," *Journal of Computing in Civil Engineering*, vol. 35, no. 1, 2021.

[26] A. Assadzadeh, M. Arashpour, I. Brilakis, T. Ngo, and E. Konstantinou, "Vision-based excavator pose estimation using synthetically generated datasets with domain randomization," *Automation in Construction*, vol. 134, 2022.

[27] M. M. Soltani, Z. Zhu, and A. Hammad, "Skeleton estimation of excavator by detecting its parts," *Automation in Construction*, vol. 82, Oct. 2017.

[28] F. Baek, S. Park, and H. Kim, "Data Augmentation Using Adversarial Training for Construction-Equipment Classification." Nov. 26, 2019.

[29] H.-S. Kim, J. Seong, and H.-J. Jung, "Optimal domain adaptive object detection with self-training and adversarial-based approach for construction site monitoring," *Automation in Construction*, vol. 158, Feb. 2024.

[30] J. Kim, J. Hwang, S. Chi, and J. Seo, "Towards database-free vision-based monitoring on construction sites: A deep active learning approach," *Automation in Construction*, vol. 120, Dec. 2020.

[31] C. Chen, B. Xiao, Y. Zhang, and Z. Zhu, "Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition," *Automation in Construction*, vol. 146, Feb. 2023,.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," In *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021, vol. 139, pp. 8748–8763.

[33] W.-L. Tsai, P.-L. Le, W.-F. Ho, N.-W. Chi, J. J. Lin, S. Tang, *et al.*, "Construction safety inspection with contrastive language-image pre-training (CLIP) image captioning and attention," *Automation in Construction*, vol. 169, p. 105863, Jan. 2025.

[34] L. Yu, Y. Qian, W. Liu, and A. G. Hauptmann, "Argus++: Robust Real-time Activity Detection for Unconstrained Video Streams with Overlapping Cube Proposals," In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Jan. 2022, pp. 112–121.

[35] A. Zhuang, E. Zhou, Q. Li, R. Dempster, A. Budhwani, M. Al-Sharman, *et al.*, "RADACS: Towards Higher-Order Reasoning using Action Recognition in Autonomous Vehicles." arXiv, Sep. 28, 2022.

[36] K. Nguyen, C. Fookes, S. Sridharan, Y. Tian, F. Liu, X. Liu, *et al.*, "The State of Aerial Surveillance: A Survey." arXiv, Jan. 12, 2022.

[37] G. U. Sai Theja, M. S. Murari, M. F. Singha, R. Patgiri, and A. Choudhury, "A Survey on

Surveillance Using Drones," In *Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing*, New York, USA, 2022.

[38] X. Wang, R. Xian, T. Guan, C. M. de Melo, S. M. Nogar, A. Bera, *et al.*, "AZTR: Aerial Video Action Recognition with Auto Zoom and Temporal Reasoning." arXiv, Mar. 02, 2023.

[39] M. Ding, N. Li, Z. Song, R. Zhang, X. Zhang, and H. Zhou, "A Lightweight Action Recognition Method for Unmanned-Aerial-Vehicle Video," In *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*, Dec. 2020, pp. 181–185.

[40] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, *et al.*, "Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection," In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 2153–2160.

[41] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition," *Drones*, vol. 3, 2019.

[42] H. Mliki, F. Bouhlel, and M. Hammami, "Human activity recognition from UAV-captured video sequences," *Pattern Recognition*, vol. 100, p. 107140, Apr. 2020.

[43] S. Gundu and H. Syed, "Vision-Based HAR in UAV Videos Using Histograms and Deep Learning Techniques," *Sensors*, vol. 23, no. 5, 2023.

[44] H. Peng and A. Razi, "Fully Autonomous UAV-Based Action Recognition System Using Aerial Imagery," In *Advances in Visual Computing*, Cham, 2020, pp. 276–290.

[45] D. Kothandaraman, T. Guan, X. Wang, S. Hu, M. Lin, and D. Manocha, "FAR: Fourier Aerial Video Recognition," In *Computer Vision – ECCV 2022*, Cham, 2022, pp. 657–676.

[46] W. Sultani and M. Shah, "Human action recognition in drone videos using a few aerial training examples," *Computer Vision and Image Understanding*, vol. 206, p. 103186, May 2021.

[47] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and Semi-Supervised Domain Adaptation for Action Recognition from Drones," In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 1706–1715.

[48] D. Kim, M. Liu, S. Lee, and V. R. Kamat, "Remote proximity monitoring between mobile construction resources using camera-mounted UAVs," *Automation in Construction*, vol. 99, pp. 168–182, Mar. 2019.

[49] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[50] Y. Guo, Y. Xu, and S. Li, "Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network," *Automation in Construction*, vol. 112, Apr. 2020.

[51] I. Mutis, V. A. Joshi, and A. Singh, "Object Detectors for Construction Resources Using Unmanned Aerial Vehicles," *Practice Periodical on Structural Design and Construction*, vol. 26, no. 4, Nov. 2021.

[52] Y. Pi, N. D. Nath, and A. H. Behzadan, "Convolutional neural networks for object detection in aerial imagery for disaster response and recovery," *Advanced Engineering Informatics*, vol. 43, p. 101009, Jan. 2020.

[53] Y. Lu, W. Qin, C. Zhou, and Z. Liu, "Automated detection of dangerous work zone for crawler crane guided by UAV images via Swin Transformer," *Automation in Construction*, vol. 147, Mar. 2023.

[54] H. Kim, J. Seong, and H.-J. Jung, "Real-Time Struck-By Hazards Detection System for Small- and Medium-Sized Construction Sites Based on Computer Vision Using Far-Field Surveillance Videos," *Journal of Computing in Civil Engineering*, vol. 37, no. 6, Nov. 2023.

[55] S. Kim, S. H. Hong, H. Kim, M. Lee, and S. Hwang, "Small object detection (SOD) system for comprehensive construction site safety monitoring," *Automation in Construction*, vol. 156, p. 105103, Dec. 2023.

[56] 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, 2023.

[57] A. B. Ersoz, O. Pekcan, and E. Akbas, "AIDCON: An Aerial Image Dataset and Benchmark for Construction Machinery," *Remote Sensing*, vol. 16, no. 17, 2024.

[58] S. Bang, F. Baek, S. Park, W. Kim, and H. Kim, "Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques," *Automation in Construction*, vol. 115, p. 103198, Jul. 2020.

[59] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[60] D. Wang, X. Wang, B. Ren, J. Wang, Z. Zeng, D. Kang, *et al.*, "Vision-Based Productivity Analysis of Cable Crane Transportation Using Augmented Reality-Based Synthetic Image," *Journal of Computing in Civil Engineering*, vol. 36, no. 1, Jan. 2022.