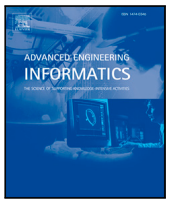
Advanced Engineering Informatics 68 (2025) 103691

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Temporal sequence-based object detection and action recognition for mobile machinery on construction sites

Bobo Helian [a,b,*], Gen Huang [a], Marcus Geimer [a]

[a] Institute of Mobile Machines, Karlsruhe Institute of Technology, 76131, Karlsruhe, Germany
[b] State Key Laboratory of Fluid Power and Mechatronic System, Zhejiang University, 310027, Hangzhou, China

ARTICLE INFO

Keywords:
Object detection
Action recognition
SlowFast
Transformer
Construction site

ABSTRACT

Automation of mobile machinery is critical in the construction industry to improve efficiency and ensure safety. Perception technologies, particularly for detecting and monitoring the actions of construction machinery, are essential for optimizing workflows and mitigating accident risks. However, the complex nature of construction environments, the variety of machines, and the dynamic interactions at construction sites pose significant challenges for reliable object detection and action recognition. This study introduces a deep learning approach using temporal vision information for object detection and action recognition of mobile machinery in construction environments. In particular, a novel strategy called Integrated YL-SF is proposed, which integrates the YOLOv8 framework with the SlowFast model enhanced by Transformers to achieve robust action recognition and motion analysis of construction machinery. The proposed method is evaluated on a custom dataset with a variety of machine types and real-world operating environments, and it is benchmarked against the standard YOLOv8 model. The results show that the Integrated YL-SF framework outperforms existing methods and effectively addresses challenges such as dynamic scenarios, object occlusion, and multi-machine interactions in complex environments.

1. Introduction

1.1. Motivation

Ensuring safety while implementing advanced automation in the construction industry is a critical challenge, particularly with the widespread use of mobile machines [1]. Although the construction sector employs only about 7% of the global workforce, it accounts for a staggering 30%–40% of all workplace fatalities [2]. In complex and dynamic construction sites, real-time detecting and monitoring of machinery movements, such as excavators, are essential to prevent accidents, optimize workflows, and protect workers [3].

Traditional monitoring approaches, which often rely on human supervision, are labor-intensive, time-consuming, and costly [4]. In addition, the inherent complexities of construction sites — such as dynamic backgrounds, varying lighting conditions, and frequent object occlusions — make traditional methods inadequate for accurately tracking and detecting machine movements in real time. With the increasing use of cameras on construction sites, image and video data has become a reliable and cost-effective source of information. Computer vision and machine learning techniques have demonstrated their potential to address object detection and action recognition challenges in these environments [5]. Compared to alternative sensing technologies such as RFID, GPS, and UWB, which require sensors to be installed on each monitored entity and provide limited information (e.g., location) [6–9], computer vision offers a more comprehensive and scalable solution. It enables efficient and detailed understanding of construction tasks and interactions in complex environments [10].

Current research in construction machine monitoring predominantly focuses on object detection from static single-frame images. While such methods are computationally efficient and suitable for certain scenarios [11], they lack the temporal context needed to interpret complex, sequential activities. This limitation becomes critical in dynamic and cluttered environments like construction sites, where machines frequently interact, move intermittently, or remain partially occluded. Static-frame methods often fail to differentiate between visually similar but temporally distinct actions, such as swinging versus hauling, especially under occlusion or varying lighting conditions [12, 13].

To overcome these challenges, this study presents a vision-based method that integrates YOLOv8 for spatial object detection with the SlowFast framework for temporal action recognition [14–16]. The

* Corresponding author at: Institute of Mobile Machines, Karlsruhe Institute of Technology, 76131, Karlsruhe, Germany.
E-mail address: bobo.helian@kit.edu (B. Helian).

https://doi.org/10.1016/j.aei.2025.103691
Received 6 January 2025; Received in revised form 7 July 2025; Accepted 20 July 2025
Available online 5 August 2025
1474-0346/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

YOLOv8 model provides high-precision detection of construction machines, while the SlowFast architecture captures both slow-changing contextual cues and fast motion patterns. This dual-stream design enables the system to understand time-dependent operations and machine behavior at a fine-grained level. Furthermore, leveraging temporal features enhances the model's generalization ability in out-of-distribution (OOD) settings, such as varying viewpoints, lighting, or machine types, supporting safer and more robust automation in real-world construction scenarios.

This approach is tailored to the complex conditions of construction sites, where multiple machines operate concurrently in dynamic and often obstructed environments. By validating the method on both training and out-of-distribution test data, we demonstrate its robustness and applicability in practical scenarios. Rather than attempting full automation, this work focuses on establishing a reliable foundation for machine-level activity understanding, which is essential for improving safety monitoring and enabling future automation in construction operations.

1.2. Related works

1.2.1. Typical object detection methods

Object detection has undergone significant advancements in computer vision and machine learning, particularly through deep learning approaches, such as YOLO and Region Convolutional Neural Networks (R-CNN). R-CNN, introduced by Girshick et al. in 2014, marked a major breakthrough in static image-based object detection [17]. R-CNN demonstrated the potential of deep learning in object detection and ignited widespread interest in the field.

Subsequently, Faster R-CNN, introduced in 2016, integrated key processes such as feature extraction, region proposal generation, bounding box regression, and classification into a unified network structure. This innovation resulted in substantial improvements in detection speed and overall performance, making Faster R-CNN a benchmark for object detection [18].

1.2.2. Two-stream approaches using optical flow

Two-stream networks have shown as a potential solution for action recognition tasks, using both static and temporal information from videos. These networks utilize two streams — one for RGB frames and another for optical flow — to enhance the understanding of dynamic content in videos [12]. For example, the study in [19] proposed a temporal- and appearance-guided object detection method for construction sites that integrates RGB images with optical flow, leading to improved recognition accuracy and enhanced generalization to out-of-distribution (OOD) data. However, two-stream approaches have notable limitations. The need to process two independent data streams significantly increases computational requirements, especially for high-resolution video data [20]. Additionally, while the combination of static and temporal information improves accuracy, there remains room for enhancement in both efficiency and precision.

1.2.3. Temporal action recognition methods

Construction sites present a wide range of dynamic actions, from fast and abrupt motions, such as a truck reversing or an excavator swinging, to slow and sustained activities, such as an excavator digging soil. The dynamic operations of construction machinery exhibit complex temporal characteristics, and effective extraction and analysis of these features are crucial for enhancing the reliability and generalizability of neural networks.

For instance, in [21], a three-dimensional convolutional neural network (3D CNN) is employed to accurately recognize and classify excavator activities into detailed categories, preserving both spatial and temporal information from video data. However, the study notes that its accuracy is constrained by challenges such as interference from nearby excavators and variations in lighting conditions, which impact the robustness of activity recognition. In addition, the study [22] proposed a hybrid deep learning algorithm, combining CNN and Long Short Term Memory to learn the sequential patterns of excavator working actions. However, the deep learning model built in this study required excessive computational training time and a large amount of training data. Furthermore, Temporal Segment Networks (TSN) represent an improvement over the typical two-stream approach, addressing its limitations in modeling long-duration videos. TSN divides a video into equal-length segments and randomly samples one frame (or a smaller segment) from each. This strategy enables the network to capture global temporal context while significantly reducing computational overhead [23,24]. While TSN achieves better results on datasets like HMDB51 [25] and UCF101 [26], which lack strong temporal dependencies, its ability to model temporal motion remains limited. TSN relies on pre-extracted optical flow information, making it less effective as a fully end-to-end video modeling approach. Moreover, its performance diminishes on datasets with complex temporal dynamics [27,28].

1.2.4. Multi-sensory approaches

Multi-sensor fusion represents a promising direction for enhancing perception tasks in mobile machinery by integrating complementary modalities such as LiDAR, radar, and visual data. For example, the study in [29] proposes an end-to-end multimodal fusion framework based on Transformers, incorporating deformable attention and residual structures within the fusion encoding module. This design allows for simultaneous sampling from 2D image features and 3D voxel features, offering greater flexibility and adaptability. Similarly, the work in [30] presents a multi-sensory guidance system that generates navigation maps for global obstacle avoidance by combining ORB-SLAM and YOLO-based object detection. Furthermore, the study in [31] introduces a customized LSTM-based classifier for real-time excavation workload classification. This system utilizes multi-sensor signals, including position, velocity, and hydraulic pressures from both chambers of the bucket cylinder actuator, demonstrating the practical effectiveness of sensor fusion in recognizing external operating conditions. In contrast, the present study focuses exclusively on vision-based methods, considering the widespread availability and desired capability of camera sensors for object detection and action recognition in mobile construction equipment.

As discussed above, deep-learning-based action recognition methods struggle to model both high-speed motion patterns and long-duration temporal dependencies within a unified framework. The SlowFast framework, proposed by Feichtenhofer et al. [14], addresses this limitation through its dual-path architecture: a fast pathway operating at high temporal resolution captures fine-grained motion details, while a slow pathway processes semantically rich features over a longer timescale. This dual design is particularly well-suited to the construction domain, where both types of motion often occur simultaneously or in close sequence.

Furthermore, the SlowFast-based model, compared to conventional action recognition models that rely on computationally expensive 3D convolutions or optical flow (e.g., two-stream networks), significantly reduces resource requirements while maintaining strong recognition performance. For instance, methods that compute optical flow explicitly or use 3D CNNs over all frames often involve high memory and time costs, which make them less practical for real-time deployment in dynamic construction environments.

1.3. Technical problem formulation

Although existing methods for object detection and action recognition have made significant progress in many fields, designing an efficient, robust algorithm that effectively combines static and temporal information remains a major challenge in complex construction site environments.

1.3.1. Limitations of current methods

Object Detection Methods: R-CNN and Faster R-CNN methods rely on extracting candidate regions and features for detection. However, they struggle to meet real-time requirements due to computational inefficiency, making them unsuitable for dynamic construction site scenarios with multiple objects. YOLO series (e.g., YOLOv8) have optimized the balance between detection speed and accuracy, making them better suited for real-time object detection. However, YOLO-based methods are limited to single-frame static information, lacking the ability to incorporate temporal dynamics for action recognition.

Action Recognition Methods: Two-Stream networks (e.g., RGB-optical flow) use parallel streams to extract RGB images (static information) and optical flow (temporal information) for action recognition. However, computing optical flow incurs high computational costs and is susceptible to noise and lighting changes, resulting in reduced robustness. Temporal Segment Networks (TSN) attempt to capture global temporal dynamics by segmenting videos and sampling keyframes. While effective for coarse-grained analysis, TSN struggles to model fine-grained dynamic actions, limiting its performance in complex scenarios.

1.3.2. Challenges in detection on construction sites

Detection and action recognition of construction machinery in dynamic construction site environments present significant challenges due to the following limitations in existing methods: (1) Insufficient static-temporal data fusion: Current approaches, relying on either single-frame static analysis or purely temporal methods, fail to comprehensively capture the intricate interplay of static and temporal features in machinery motion, resulting in limited accuracy for complex and dynamic actions. (2) Trade-off between real-time performance and accuracy: While methods like R-CNN or optical flow-based methods achieve high precision, they come at the cost of substantial computational overhead, making them unsuitable for real-time monitoring and decision-making in construction scenarios. (3) Limited robustness in challenging environments: Variations in lighting, dynamic backgrounds, frequent occlusions, and interaction of different types of machines significantly degrade the performance of existing models, hindering their ability to reliably detect and recognize actions across diverse and unpredictable construction site conditions.

1.4. Contributions

This study proposes a novel framework, Integrated YL-SF, for recognizing multiple machines, actions, and objects on construction sites. By integrating YOLOv8 for object detection and SlowFast for action recognition, the framework effectively processes both dynamic and static information, addressing challenges such as complex environments, occlusions, and multi-machine interactions. The key contributions of this study are as follows:

(1) This study introduces an innovative temporal sequence-based approach for object detection and action recognition of construction machinery in complex environments. The proposed Integrated YL-SF, integrates the YOLOv8 framework for high-performance object detection with the SlowFast model, enhanced by Transformers, to analyze temporal information for dynamic action recognition. This design effectively captures both dynamic and static features, enabling reliable perception in challenging multi-object, dynamic scenarios.

(2) To enhance the generalization and adaptability of the proposed framework, a comprehensive custom dataset was developed. This dataset encompasses a wide variety of construction machinery, diverse operating environments, and machine actions, providing a solid foundation for training and validating the framework in real-world conditions.

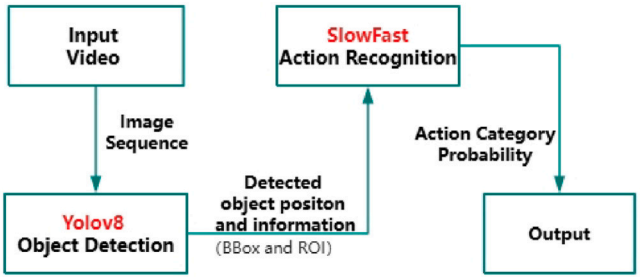


Fig. 1. Workflow of integrated YL-SF network.

(3) Rigorous multi-scenario evaluations were conducted to assess the performance of the Integrated YL-SF framework. The results demonstrate superior accuracy and generalization capabilities in challenging conditions, such as occlusions, complex backgrounds, and multi-machine interactions, showcasing its effectiveness in addressing the perception challenges of construction sites.

2. Design of the integrated YL-SF network

This section mainly introduces the framework of Integrated YL-SF, as well as how to use YOLOv8 for object detection and SlowFast for action recognition. The overall process is shown in Fig. 1. The workflow illustrates the proposed pipeline, where input video sequences are processed through the YOLOv8 framework for object detection to generate bounding boxes and regions of interest (ROIs), which are subsequently fed into the SlowFast model for action recognition, ultimately producing class probability outputs.

2.1. Tasks distribution

Industrial automation on construction sites requires addressing two key tasks:

- **Object Detection:** Detecting and identifying various equipment and vehicles (e.g., excavators, loaders, trucks) on the construction site is essential. This requires accurately determining their locations within the environment and real-time image processing to get the classification results (i.e., type of the machine).
- **Action Recognition:** After detecting these objects, the task is to recognize their actions or working status, such as whether an excavator is “digging” or “swinging” and whether a truck is “transporting” or “unloading”.

The SlowFast model is not designed for object detection; its primary purpose is action recognition in videos. It processes a complete sequence of video frames to extract features, capture temporal information in dynamic actions, and identify actions occurring throughout the video. However, this approach does not include explicit object detection or localization, making it unsuitable for determining the specific positions of machines on construction sites.

In this section, to address these requirements, a strategy is proposed that utilizes YOLOv8 for object recognition and a SlowFast model to perform robust action recognition.

2.2. Object detection with YOLOv8

The one-stage YOLO model proposed by Joseph Redmon et al. changes the traditional object detection framework by predicting the bounding box and category probabilities directly from the whole image through a single neural network. This type of approach uses the idea of regression, using the whole image as the input to the network, and directly regresses the object bounding box at this location, and the

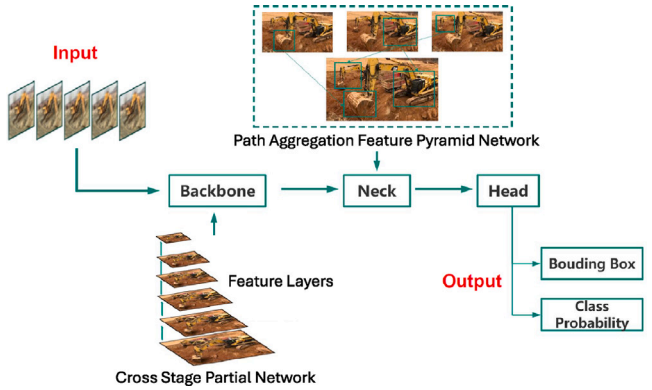


Fig. 2. Object detection with YOLOv8.

category to which the object belongs, at multiple locations in the image, enabling YOLO to provide high accuracy while achieving true real-time object detection [32]. In this paper, we utilize YOLOv8 to perform object detection. YOLOv8 is the newest object detection model in the YOLO family, which is known for its real-time performance and high accuracy [33].

The main task of YOLOv8 is to perform object detection on each frame of the video, pinpointing and labeling objects of interest (e.g., excavators, trucks, etc.). Object detection provides not only the class of the object, but also information about the position of each object in the image, i.e., the bounding box. This provides precise inputs for subsequent action recognition, allowing the system to focus on the detected object detection area, thereby reducing unnecessary computation and improving overall efficiency and accuracy. As shown in Fig. 2. The process begins by preprocessing the excavator video to extract RGB images from each frame. Once the images are prepared, they are annotated. After annotation, the RGB images are input into a backbone network, Darknet53, which is responsible for feature extraction. This backbone produces five different feature layers, referred to as P1 through P5. The P1 layer, with the highest resolution, captures fine details like edges and textures, while the P5 layer, with the lowest resolution, captures more abstract and complex structural information. These multi-scale feature maps are then merged using a mechanism called the Path Aggregation Feature Pyramid Network (PAFPN). Finally, the head of the network carries out two key tasks: first, it identifies the current operational patterns of the excavator, and second, it determines the location of the excavator within the image by generating bounding boxes.

2.3. Action recognition with SlowFast framework

2.3.1. SlowFast network architecture design

After object detection with YOLOv8 and passing through the multi-Object tracking algorithm, ROI (Region of Interest) extraction is performed: each tracked object will have its corresponding bounding box. These frame sequences (images in consecutive frames for each Object) will be used as input to SlowFast.

The entire processing flow is shown in Fig. 3 below:

(1) Two-path Processing

- Slow Path

  Input: Low frame rate segments of a video sequence. In this work, we use 5 frames from a 3-s video sequence (from the beginning, the end, and the middle three frames), which are fed into the Slow Path for processing.

  Feature Output (C, T): After 3D convolutional processing, the output feature contains the time dimension (T) and the number of channels (C), representing long-term action features in the low frame rate video.

- Fast Path

  Input: A high frame rate segment of the video sequence. In this work, we take a 3-s video with 90 frames per second. These frames are input into the Fast Path, which is designed to capture short, rapid changes and is suitable for recognizing fast movements or transient actions in videos, such as the instantaneous rotation of an excavator or the acceleration of a truck.

  Feature Output (βC, αT): The Fast Path outputs a feature map with a reduced number of channels (βC) and a downsampled temporal frame rate (αT). This ensures that although it processes more frames, it outputs fewer channels of features to keep the model computationally efficient.

(2) 3D Convolution

3D convolution is the foundational module in this framework, focusing on extracting local spatiotemporal features. By performing convolution operations simultaneously across static and temporal dimensions, it captures static features within individual frames (such as shapes and textures) and temporal dynamics between frames (such as the continuity of actions). In the **Slow Path**, 3D convolution specializes in extracting long-term action features from low frame-rate videos, such as slow movements or sustained states, helping the model comprehend overall action trends. In the **Fast Path**, 3D convolution emphasizes capturing short-term rapid dynamics in high frame-rate videos, such as instantaneous fast movements or abrupt changes. By generating feature maps that retain the temporal dimension, 3D convolution provides a rich, locally sensitive spatiotemporal representation that lays the groundwork for subsequent global modeling.

(3) Transformer

The Transformer module, on the other hand, focuses on global feature modeling and capturing complex temporal dependencies. Using the self-attention mechanism, the Transformer dynamically evaluates the relationships between all temporal frames in a video sequence, enabling the model to globally focus on key frames, such as the beginning, end, or peak of an action. The multi-head attention mechanism processes multiple feature patterns in parallel, ensuring a comprehensive understanding of different phases of an action's evolution. Positional encoding is incorporated to help the Transformer accurately perceive the temporal order of frames, preventing the loss of sequence information. In the Slow Path, the Transformer further refines the long-term features extracted by 3D convolution, allowing the model to capture temporal dependencies over a global scope. In the Fast Path, it focuses on modeling the details of rapid short-term changes, enabling the model to identify more complex and fast-paced action patterns. Thus, the Transformer and 3D convolution form a complementary relationship: 3D convolution emphasizes the extraction of local dynamics, while the Transformer focuses on the integration and modeling of global temporal information.

(4) Lateral Connection

Lateral connection is represented by the dotted arrow from the Fast Path to the Slow Path. This connection transfers the short-term dynamics captured by the Fast Path to the Slow Path, enabling the Slow Path to incorporate these short-term dynamics when processing long-term information. This unidirectional transfer of information helps the Slow Path better understand changes in the object's behavior.

(5) Feature Fusion

After the Slow Path and Fast Path feature maps are processed, they enter the feature fusion stage. Here, features from different time scales (long and short) are combined to produce a feature representation that contains both long-term information and short-term details. This fusion helps the model capture both global and local rapid changes in the action, improving the understanding of complex actions.

(6) Global Average Pooling (GAP)

After feature fusion, the model performs dimensionality reduction on the feature map through Global Average Pooling (GAP). GAP globally aggregates the information in all feature maps to form a more

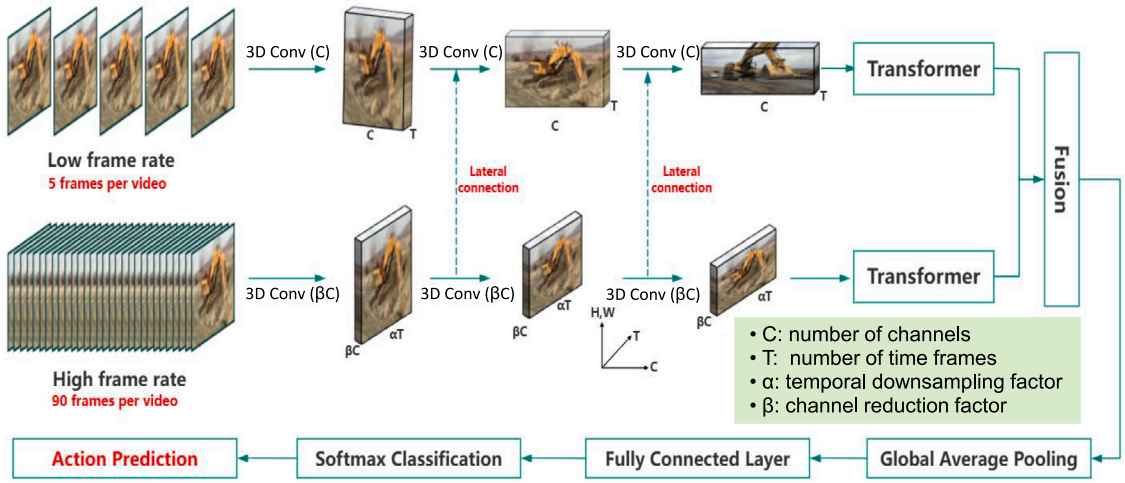


Fig. 3. SlowFast network architecture design for action recognition.

compact feature vector, reducing data dimensionality while retaining global semantic information. This step prepares the features for subsequent classification operations.

(7) Fully Connected Layer (FCL)

The pooled feature vectors are passed through a fully connected layer (FCL), which further compresses the features and generates more discriminative features for the final classification task.

(8) Softmax Classification

The final step in action recognition is Softmax classification. In this step, the model classifies the Object actions in each input video clip, generating probability distributions for each category. For example, the action of an excavator may be categorized as "digging" or "swinging", while a truck may be categorized as "transporting" or "unloading". The detailed dataset distribution across three machine types and eight action types is given in Table 1 and later sections.

(9) Action Prediction

After classification, the system outputs the action category of the Object, which is the result of action prediction. Combined with the previous YOLOv8 object detection, the system can output action recognition results for each individually detected object.

3. Comparative experiments and analysis

3.1. Dataset generation on the construction site

3.1.1. Data source

The training dataset we use as an action recognition model comes from MegaMachinesChannel [34] on YouTube, a channel that features videos of various construction machinery. As a foundational study, we focused on three representative types of construction machinery: excavators, dump trucks, and wheel loaders. These machines were selected because they are among the most commonly used on construction sites and are frequently involved in coordinated operations, which makes their action recognition tasks both practically important and technically challenging.

We selected some of the most typical tasks in construction. For example, the actions in the dataset include digging, dumping, hauling, swinging (performed by an excavator), transporting, unloading (performed by a dump truck), and loading and unloading (performed by a loader), as shown in Fig. 4. The action recognition of these actions can effectively help us provide improved construction efficiency and safety. In addition, these machines often perform overlapping or sequential actions, as shown in Fig. 13, such as digging and unloading, in shared workspaces, leading to complex visual and temporal interactions. In total, eight distinct action categories were defined and annotated across these machines.

Fig. 4. Action category examples.

3.1.2. Keyframe extraction and data volume

To capture both fine-grained motion and longer-term temporal context, we adopt a dual-pathway frame sampling strategy following the principles of the SlowFast architecture. For the Slow Path, we extracted 5 keyframes from each 3-s video clip and labeled them using YOLO's labeling format. These keyframes were evenly sampled across the start to end of the clip, specifically, at 0.0 s, 0.6 s, 1.2 s, 1.8 s, and 2.4 s—ensuring that critical stages of the action were captured. This sampling strategy enables the Slow Path to capture high-level semantic features and longer-term temporal evolution, which is especially important for recognizing sustained or gradual actions in construction tasks. In parallel, the Fast Path processes all 90 frames from the same 3-s video, sampled at 30 frames per second. This dense sampling allows the model to detect fast, transient movements and subtle motion changes that occur between keyframes.

The choice of a 3-s window is based on empirical observation and domain knowledge: most atomic actions of construction machinery (such as an excavator swinging or a truck unloading) typically unfold within a 2–4 s range. Thus, 3 s provides a reasonable balance between capturing the full action context and maintaining computational efficiency. Additionally, the model processes video using a sliding window updated every second (e.g., 1–3 s, 2–4 s, 3–5 s), enabling timely and continuous action recognition for real-time perception, as shown in Fig.

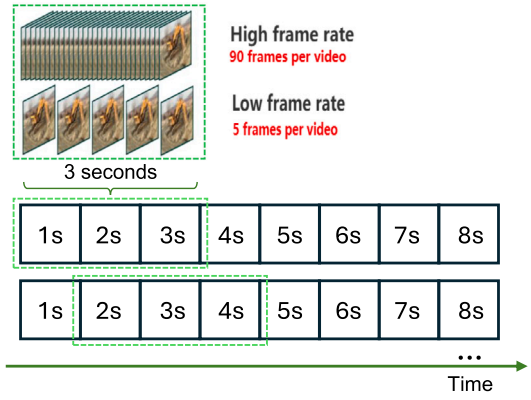


Fig. 5. Sliding window strategy for parallel keyframe extraction.

Table 1
Single-machine single-action dataset.

Machinery	Action	Videos (3s)	Keyframes	Train	Val
Excavator	Digging	75	375	335	40
Excavator	Dumping	75	375	335	40
Excavator	Hauling	75	375	335	40
Excavator	Swinging	75	375	335	40
Dump truck	Transport	75	375	335	40
Dump truck	Unloading	75	375	335	40
Loader	Loading	75	375	335	40
Loader	Unloading	75	375	335	40
Total		**600**	**3000**	**2680**	**320**

Table 2
Multi-machine multi-action dataset.

Machinery	Action	Videos (3s)	Keyframes	Train	Val
Excavator, Dump truck	Digging, Dumping, Hauling, Swinging, Transport, DT_Unloading	100	500	420	80
Dump truck, Loader	Transport, DT_Unloading, LD_Loading, LD_Unloading	100	500	420	80
Excavator, Dump truck, Loader	Digging, Dumping, Hauling, Swinging, Transport, DT_Unloading, LD_Loading, LD_Unloading	100	500	420	80
Total		**300**	**1500**	**1260**	**240**

5. The processing time for a single 3-s video clip (comprising 90 frames for the Fast Path and 5 frames for the Slow Path, as depicted in Fig. 5) is approximately one second (i.e., 1 s per video or $1/95 \approx 0.01$ s per frame).

The total number of videos is 600 (75 videos per action), which translates into 3000 keyframes for the Slow input. The dataset is divided into 2680 training images and 320 validation images, distributed across the eight action classes as shown in Table 1.

Additionally, to enhance the model's generalization in mixed and time-varying scenarios involving different machine types and overlapping actions, we constructed a Multi-Machine Multi-Action Dataset. As shown in Table 2, this dataset comprises 1500 keyframes extracted from 300 video clips, capturing various cooperative behaviors among multiple construction machines.

3.1.3. Data augmentation

Construction sites are visually complex environments where machinery often occludes one another, lighting varies dramatically, and camera viewpoints are inconsistent. These factors can severely impact recognition accuracy, as illustrated in Fig. 6.

Fig. 6. Examples of recognition challenges due to occlusion, lighting, and clutter.

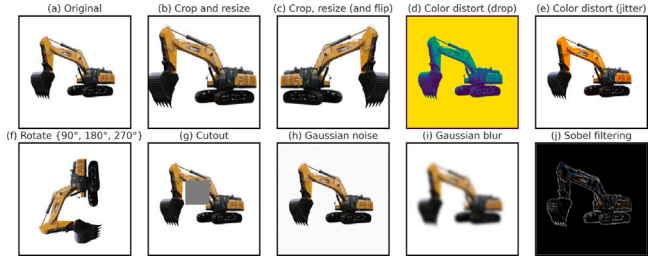


Fig. 7. Data augmentation techniques applied to construction machinery.

To improve model robustness, we apply a series of data augmentation techniques commonly used in deep learning, including flipping, cropping, rotation, brightness adjustment, and noise addition as shown in Fig. 7. These augmentations address the following challenges:

- **Occlusion Handling:** Techniques like random cropping and rotation allow the model to recognize partially visible machinery.
- **Lighting Variability:** Adjustments in brightness and color help the model generalize across different lighting conditions.
- **Viewpoint Diversity:** Random rotations and flips simulate different camera angles, enhancing recognition from multiple perspectives.
- **Scale Variation:** Scaling helps the model detect machinery at different distances and frame sizes.
- **Generalization:** Additional transformations (e.g., Gaussian blur and noise) simulate real-world variation without requiring extra data collection.

Together, these augmentations enhance the generalization ability of the integrated YL-SF model, enabling reliable action recognition across dynamic and challenging construction scenarios.

3.2. Training configuration for transformer

The Transformer module is trained end-to-end using Cross-Entropy Loss, which is standard for multi-class classification tasks such as action recognition. We fine-tuned the entire model on our construction-specific dataset. The model was initialized using a pre-trained SlowFast backbone (SLOWFAST_32 × 2_R101_50_50.pkl) from the AVA dataset. This pre-trained model provides a strong foundation by leveraging previously learned features from a broad dataset, significantly accelerating the training process and enhancing the model's ability to generalize from the outset. The model was then fine-tuned on our custom dataset with the following training configuration (see Table 3):

This configuration balances training stability and learning efficiency, particularly with the warm-up strategy that avoids gradient instability during early epochs. Fine-tuning the model with a learning rate schedule and decay helps adapt the pretrained network to construction-specific visual and motion patterns without overfitting.

Table 3
Training configuration for the transformer module.

Parameter	Value	Description
base_lr	0.01	Initial learning rate
lr_policy	steps_with_relative_lrs	Learning rate schedule
steps	[0, 20, 40, 60]	Epochs for lr change
lrs	[1, 0.1, 0.01, 0.001]	Corresponding lr multipliers
max_epoch	300	Max training epochs
momentum	0.9	Momentum for SGD
weight_decay	1×10^{-4}	Regularization factor
warmup_epochs	10.0	Warm-up period duration
warmup_start_lr	0.00001	Initial lr for warm-up
optimizer	SGD	Optimization algorithm
pretrained	SLOWFAST_32 × 2_R101_50_50.pkl	Pre-trained model used
loss	Cross-entropy	Loss function for classification

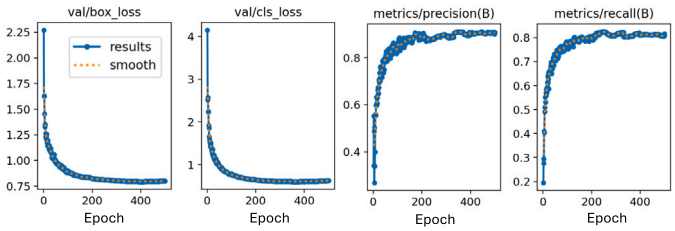


Fig. 8. Training results of integrated YL-SF.

In this study, the YL-SF model for action recognition was trained on the self-generated dataset using an RTX 3090 GPU (24 GB), 15 vCPUs (AMD EPYC 7642 48-Core Processor), and 80 GB memory, with a total runtime of under three hours.

3.3. Training results in YOLOv8 and integrated YL-SF networks

This section compares the performance of object detection and action recognition using two approaches: (1) YOLOv8 applied independently for both tasks of object detection and action recognition, and (2) the proposed hybrid approach where YOLOv8 is used for object detection and the integrated YL-SF network is used for action recognition. All models are trained and evaluated on the same dataset, with identical training, validation, and test splits to ensure a fair comparison. The training curves of the integrated YL-SF network, shown in Fig. 8, demonstrate stable loss convergence as well as high precision and recall, indicating effective training.

3.3.1. Confusion matrix

The following analysis compares the performance of Integrated YL-SF and YOLOv8 for action recognition in a construction site scenario using confusion matrices, as shown in Figs. 9 and 10. We discuss the differences in action recognition accuracy, how they handle time-based information, and their applicability to complex tasks.

(1) Accuracy Comparison: The Integrated YL-SF framework demonstrates exceptional accuracy across all action categories, particularly excelling in distinguishing similar actions. Minimal confusion is observed, even for complex scenarios. In contrast, YOLOv8 performs well on simpler actions but shows significantly lower accuracy for actions that require temporal analysis.

(2) Confusion Analysis: Integrated YL-SF exhibits almost no confusion between action classes, excelling in identifying subtle differences between similar actions, such as *Swinging* and *Dumping*. However, YOLOv8 struggles with considerable confusion between action classes, especially for actions with similar static features, such as *Digging* and *Dumping* or *Hauling* and *Swinging*.

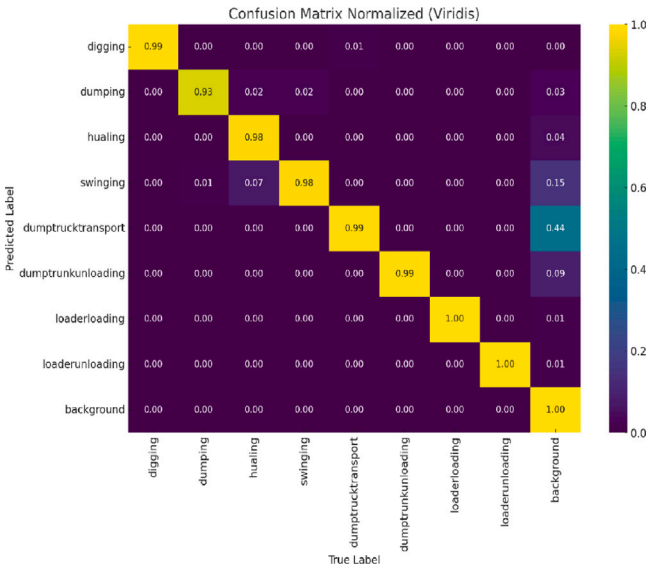


Fig. 9. Integrated YL-SF confusion matrix.

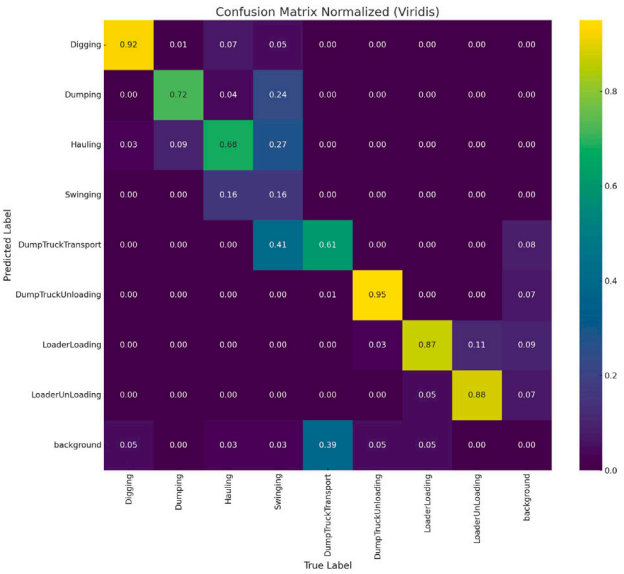


Fig. 10. YOLOv8 (baseline) confusion matrix.

(3) Temporal Information Processing: Integrated YL-SF effectively captures both short-term and long-term action dynamics through its dual-pathway design, resulting in highly accurate recognition of complex and continuous actions. By contrast, YOLOv8 lacks the ability to process time-series data and focuses primarily on frame-by-frame detection, making it ineffective for actions requiring temporal progression.

3.3.2. Precision-Confidence curve analysis

Figs. 11 and 12 represent the precision confidence curve for Integrated YL-SF and YOLOv8 in various action classes in a construction site scenario. Precision measures the proportion of true positive predictions out of all positive predictions, while confidence reflects the model's certainty about its predictions.

(1) Precision: Integrated YL-SF consistently maintains higher precision across all action classes, with minimal variance as confidence increases. This stability is attributed to its dual-path temporal processing, which enables it to handle time-sensitive action sequences

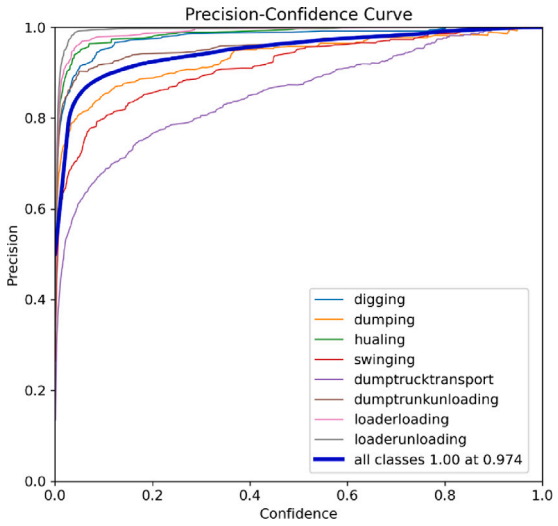


Fig. 11. Precision-Confidence curve of the integrated YL-SF.

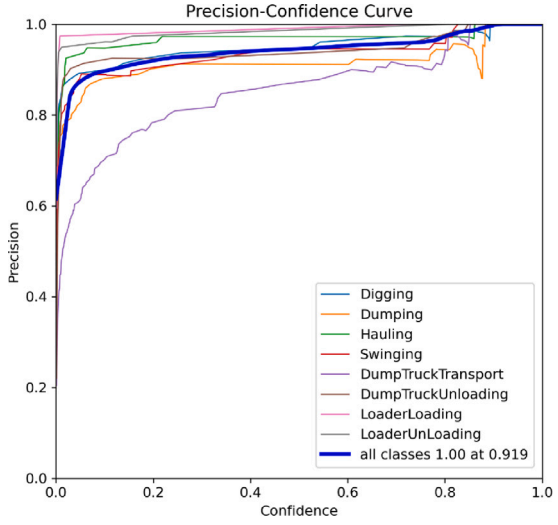


Fig. 12. Precision-Confidence curve of baseline YOLOv8.

effectively. In contrast, YOLOv8 exhibits greater fluctuation in precision, particularly for actions involving subtle or dynamic motion changes, highlighting its limitations in complex action recognition tasks.

(2) Performance Across Confidence Levels: For Integrated YL-SF, precision grows rapidly and stabilizes at a high level even for lower confidence values, demonstrating robust performance even when predictions are less confident. By comparison, YOLOv8 requires higher confidence levels to achieve acceptable precision, especially for actions requiring temporal context, such as *Swinging* and *LoaderUnloading*.

3.4. Test results in YOLOv8 and integrated YL-SF networks

In this study, we carefully designed the test set to include out-of-distribution (OOD) samples that differ from the training set in terms of construction environments, machinery types, and visual conditions. This allows us to assess the model's robustness in practical deployment scenarios. Specifically, the OOD test data includes:

- Unseen Construction Sites: Videos recorded at different locations not included in the training set, featuring distinct backgrounds, terrain, and construction layouts.

Fig. 13. Comparison single action recognition.

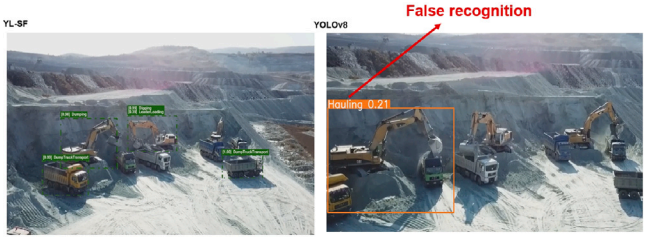


Fig. 14. Comparison of multiple action recognition.

- Different Machine Brands and Appearances: Although the action classes remain consistent, the test data involves machines of the same type but from different manufacturers or with different visual appearances (e.g., color, wear, and decals), which were not seen during training.
- Environmental and Lighting Variations: The test set includes clips taken under diverse lighting conditions (e.g., dusk, backlight, cloudy weather) and environmental factors such as fog or dust, which affect the visual clarity and introduce domain shift.
- Input Quality Shift: Some test videos were collected using different cameras or at varied distances, resulting in changes in resolution, motion blur, or occlusion.

To investigate the performance differences between YOLOv8 and Integrated YL-SF in object recognition tasks, we conducted evaluations using out-of-distribution (OOD) data. The dataset involves challenging scenarios, including: (1) single-object recognition in complex environments, (2) multi-objective recognition under multi-angle and long-distance conditions, (3) potential contextual confusion due to visually similar actions across objects.

As shown in Fig. 13, Integrated YL-SF consistently achieves higher confidence scores and recognition accuracy compared to YOLOv8. For instance, in single-object recognition tasks, Integrated YL-SF achieves a confidence score of 1.00 for "Dumping", outperforming YOLOv8, which records a lower score of 0.87.

Similarly, As shown in Fig. 14, multi-objective recognition scenarios, YOLOv8 struggles with false recognition (e.g., labeling "Hauling" with a confidence of 0.21), whereas Integrated YL-SF effectively distinguishes multiple objects with accurate classifications and high confidence across all detections.

Further quantitative evaluations are depicted in Fig. 15 and Fig. 16. The Integrated YL-SF model demonstrates superior generalization and recognition capabilities under complex environmental conditions.

These results clearly demonstrate the advantages of incorporating the SlowFast temporal modeling into YOLOv8. The Integrated YL-SF model significantly enhances detection robustness, particularly in multi-objective, multi-angle, and long-distance recognition tasks, showcasing its superior performance in real-world industrial scenarios.

Regarding scalability, the YL-SF framework is designed with modularity and computational efficiency. Its dual-branch architecture — comprising Slow and Fast pathways — enables selective frame processing, avoiding the need for exhaustive 3D convolutions

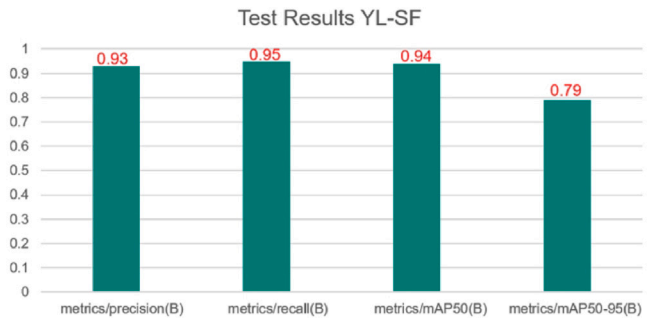


Fig. 15. Performance indicators of integrated YL-SF.

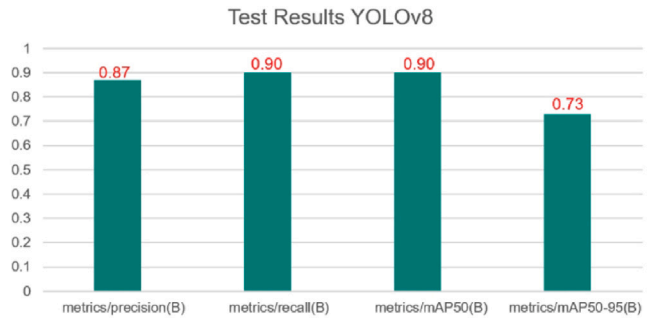


Fig. 16. Performance indicators of YOLOv8 (baseline).

across entire video sequences. This design significantly reduces computational overhead and allows flexibility in adapting the framework to various hardware environments. Key parameters such as frame sampling rate or input resolution, can be adjusted to meet the resource constraints of edge devices. The proposed model was tested with an NVIDIA RTX 3090 GPU in this study. For industrial applications with high requirements for real-time performance, more powerful industrial GPUs can be applied. Moreover, its modular structure facilitates targeted refinement of individual components, such as the object detector or temporal model for action recognition.

4. Limitations and future work

While the Integrated YL-SF framework demonstrates strong performance in recognizing complex machine actions, several limitations remain. The testing and evaluation were conducted primarily on a single custom dataset collected from publicly available YouTube videos. Although this dataset includes diverse machinery types and operating conditions, it does not comprehensively represent the full range of activities, environmental variations, and site-specific challenges encountered across different construction projects worldwide. Additionally, the framework depends on fully labeled data, which is a common issue that might restrict scalability and practical deployment in large-scale, real-world scenarios.

To address these limitations, future work will focus on expanding the dataset to cover a wider variety of construction machinery, action categories, and environmental contexts, including multi-machine and cooperative scenarios. Moreover, to facilitate real-world applicability and reduce dependence on manual labeling, we intend to explore advanced learning paradigms such as weakly supervised and self-supervised methods. For example, methods such as pseudo-label generation based on pretrained models or consistency learning can be used to automatically label large volumes of unlabeled video data, enhancing the scalability and adaptability of the framework. These efforts will collectively advance the practical deployment of the Integrated YL-SF framework as a reliable perception module for autonomous and semi-autonomous mobile machinery in construction automation. We also plan to explore multi-sensory fusion by integrating data from LiDAR, radar, or onboard sensors to improve perception robustness in complex scenes.

5. Conclusion

This study designed an integrated framework, named Integrated YL-SL, of YOLOv8 and SlowFast to achieve multi-machine, multi-action, and multi-object recognition on construction sites. In complex environments such as construction sites, YOLOv8 was employed for precise static object detection, while SlowFast processed temporal information, significantly enhancing the model's ability to recognize dynamic and multi-object actions. Additionally, a self-created dataset was developed, covering various construction machine actions such as digging, hauling, and unloading, and capturing various cooperative behaviors among multiple construction machines. This dataset provided diverse scene data for model training and facilitated action recognition in real-world construction scenarios. Comparative experiments validated the high accuracy and generalization ability of the Integrated YL-SF framework, demonstrating its superior performance in distinguishing complex and similar actions, such as "digging" and "unloading", with significantly higher precision than YOLOv8 alone. Further experimental analysis demonstrated the benefits of temporal information processing by comparing YOLOv8 and the Integrated YL-SF across various action categories. The proposed Integrated YL-SF consistently achieved higher precision and reliability, reaching a precision of 0.93 compared to YOLOv8's 0.87. This comparative validation provides valuable insights and practical guidance for future related research. The proposed Integrated YL-SF strategy contributes to the advancement of automated construction systems and the enhancement of overall construction site safety and operational efficiency.

CRediT authorship contribution statement

Bobo Helian: Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Gen Huang:** Writing – original draft, Visualization, Validation, Software, Data curation, Conceptualization. **Marcus Geimer:** Supervision, Resources, Project administration, Funding acquisition.

Research elements

The dataset customized in this study can be accessed at: https://drive.google.com/file/d/1tpdyE4Q-luAfED73M8Cq8rwTkZ6LOhZq/view?usp=sharing.

Declaration of competing interest

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.

Acknowledgments

This study is funded by the Open Foundation of the State Key Laboratory of Fluid Power and Mechatronic Systems, China under Grant GZKF-202317.

Data availability

Data will be made available on request.

References

[1] M. Geimer, Mobile Working Machines, SAE International, Warrendale, Pennsylvania (USA), 2020, http://dx.doi.org/10.4271/9780768094329.

[2] W. Fang, L. Ding, P.E. Love, H. Luo, H. Li, F. Peña-Mora, B. Zhong, C. Zhou, Computer vision a in construction safety assurance, Autom. Constr. 110 (2020) 103013.

[3] Y. Wang, B. Xiao, A. Bouferguene, M. Al-Hussein, H. Li, Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning, Adv. Eng. Inform. 53 (2022) 101699.

[4] M. Golparvar-Fard, A. Heydarian, J.C. Niebles, Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers, Adv. Eng. Inform. 27 (4) (2013) 652–663.

[5] J. Yang, Z. Shi, Z. Wu, Vision-based action recognition of construction workers using dense trajectories, Adv. Eng. Inform. 30 (3) (2016) 327–336.

[6] K.M. Rashid, J. Louis, Times-series data augmentation and deep learning for construction equipment activity recognition, Adv. Eng. Inform. 42 (2019) 100944.

[7] I. Brilakis, M.-W. Park, G. Jog, Automated vision tracking of project related entities, Adv. Eng. Inform. 25 (4) (2011) 713–724.

[8] N. Pradhananga, J. Teizer, Automatic spatio-temporal analysis of construction site equipment operations using gps data, Autom. Constr. 29 (2013) 107–122.

[9] S.M. Shahandashti, S.N. Razavi, L. Soibelman, M. Berges, C.H. Caldas, I. Brilakis, J. Teizer, P.A. Vela, C. Haas, J. Garrett, et al., Data-fusion approaches and applications for construction engineering, J. Constr. Eng. Manag. 137 (10) (2011) 863–869.

[10] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inform. 29 (2) (2015) 239–251.

[11] L. Li, W. Huang, I.Y.-H. Gu, Q. Tian, Statistical modeling of complex backgrounds for foreground object detection, IEEE Trans. Image Process. 13 (11) (2004) 1459–1472.

[12] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 27 (2014).

[13] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[14] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6202–6211.

[15] J. Terven, D. Cordova-Esparza, A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas, 2023, arXiv preprint arXiv:2304.00501.

[16] D. Wan, R. Lu, B. Hu, J. Yin, S. Shen, T. xu, X. Lang, Yolo-mif: Improved yolov8 with multi-information fusion for object detection in gray-scale images, Adv. Eng. Inform. 62 (2024) 102709, http://dx.doi.org/10.1016/j.aei.2024.102709, URL https://www.sciencedirect.com/science/article/pii/S1474034624003574.

[17] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[18] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).

[19] K. Wang, B. Helian, V. Fischer, M. Geimer, Temporal- and appearance-guided object detection in construction machines considering out-of-distribution data, J. Comput. Civ. Eng. 39 (2) (2025) 04024057, http://dx.doi.org/10.1061/JCCEE5.CPENG-5590.

[20] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.

[21] C. Chen, Z. Zhu, A. Hammad, Automated excavators activity recognition and productivity analysis from construction site surveillance videos, Autom. Constr. 110 (2020) 103045.

[22] J. Kim, S. Chi, Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles, Autom. Constr. 104 (2019) 255–264.

[23] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 20–36.

[24] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks for action recognition in videos, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2018) 2740–2755.

[25] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: a large video database for human motion recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2556–2563.

[26] K. Soomro, A.R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.

[27] H. Wang, D. You, S. Zhang, Exploiting relation of video segments for temporal action detection, Adv. Eng. Inform. 62 (2024) 102585, http://dx.doi.org/10.1016/j.aei.2024.102585, URL https://www.sciencedirect.com/science/article/pii/S1474034624002337.

[28] Y. Shen, J. Wang, S. Mo, X. Gu, Data augmentation aided excavator activity recognition using deep convolutional conditional generative adversarial networks, Adv. Eng. Inform. 62 (2024) 102785, http://dx.doi.org/10.1016/j.aei.2024.102785, URL https://www.sciencedirect.com/science/article/pii/S1474034624004336.

[29] C. Hu, H. Zheng, K. Li, J. Xu, W. Mao, M. Luo, L. Wang, M. Chen, Q. Peng, K. Liu, Y. Zhao, P. Hao, M. Liu, K. Yu, FusionFormer: A multi-sensory fusion in bird's-eye-view and temporal consistent transformer for 3D object detection, 2023, arXiv preprint arXiv:2309.05257. arXiv:2309.05257.

[30] Z. Xie, Z. Li, Y. Zhang, J. Zhang, F. Liu, W. Chen, A multi-sensory guidance system for the visually impaired using YOLO and ORB-SLAM, Information 13 (7) (2022) 343, http://dx.doi.org/10.3390/info13070343.

[31] B. Helian, X. An, Y. Zhou, Z. Chen, M. Geimer, LSTM-based workload recognition for hydraulic actuators: A case study on excavator digging process, in: BATH/ASME 2024 Symposium on Fluid Power and Motion Control, Fluid Power Systems Technology, 2024, V001T01A013, http://dx.doi.org/10.1115/FPMC2024-140092.

[32] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[33] Ultralytics, Ultralytics github repository, https://github.com/ultralytics/ultralytics. (Accessed: Day-Month-Year).

[34] M.M. Channel, Mega machines channel, 2024, https://www.youtube.com/@MegaMachinesChannel. (Accessed 03 October 2024).