Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.081546

ARTICLE

Monitoring of Drill-and-Blast Workflows at the Tunnel Face Using Computer Vision and Context Reasoning

Chuanjiang Chen[1], Junyong Zhou[1,*], Binbin Du[1], Miaosi Dong[2,*], Liwen Zhang[1] and Bitang Zhu[3]

[1]School of Civil Engineering and Transportation, Guangzhou University, Guangzhou, China
[2]Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
[3]School of Civil Engineering and Architecture, East China Jiaotong University, Nanchang, China

*Corresponding Authors: Junyong Zhou. Email: jyzhou@gzhu.edu.cn; Miaosi Dong. Email: miaosid@andrew.cmu.edu

Received: 04 March 2026; Accepted: 02 May 2026; Published: 27 May 2026

ABSTRACT: Computer vision has been widely adopted in intelligent construction monitoring; however, existing studies primarily focus on identifying individual construction elements or isolated activities, with limited capability for integrated monitoring of complete construction workflows. Such workflow-level automation is a prerequisite for intelligent construction and unmanned job sites. To address the challenge of reliable visual recognition in drill-and-blast tunnel environments characterized by uneven illumination, localized glare, and dust interference, this study proposes a methodological framework for construction workflow recognition at the tunnel face using computer vision and context reasoning. The framework consists of three components: (1) a construction workflow model with a sequence library database, (2) a robust construction element recognition model combining an enhanced YOLOv11 with Segment Anything Model 2 (SAM2), and (3) a hierarchical workflow reasoning mechanism driven by domain knowledge. A hierarchical workflow model embedding procedural logic is established through field investigation and normative analysis. SAM2 is employed for automated dataset annotation, while YOLOv11 is structurally enhanced with Convolutional Block Attention Module (CBAM), Adaptive Feature Enhancement (AFE), and Swin Transformer modules to improve feature representation and adaptability to degraded visual conditions. Workflow identification is finally achieved by integrating visual perception outputs with hierarchical context reasoning. Validation in an active drill-and-blast tunnel shows that the proposed method attains an average detection precision of 91.1% across 11 construction element categories, exceeding 95% for large equipment, and an average workflow recognition accuracy of 94%. The results demonstrate the effectiveness of the proposed framework for monitoring the tunnel construction workflow and supporting construction management.

KEYWORDS: Drill-and-blast tunnel; construction workflow; intelligent construction monitoring; computer vision; workflow recognition; context reasoning

1 Introduction

1.1 Research Background

Advances in artificial intelligence and construction machinery have accelerated the transformation of the construction industry toward digitalization and intelligence. Under the Industry 4.0 paradigm, technologies such as data intelligence, robotics, and building information modeling are driving construction sites toward highly interconnected smart environments, with significant potential to improve productivity, safety, and sustainability [1]. As a critical component of transportation infrastructure, tunnel engineering plays an essential role in regional and national development, and among available construction methods,

the drill-and-blast approach remains widely adopted under complex geological conditions [2]. However, drill-and-blast tunnel construction involves sequential and highly interdependent processes conducted in confined spaces and high-risk operating environments, making accurate and timely construction progress monitoring crucial for safety assurance, cost control, and schedule management [3]. In practice, tunnel construction sites are affected by severe dust interference, uneven and unstable illumination, strong glare from equipment-mounted lighting, and frequent process transitions, under which traditional approaches based on manual inspection and experiential judgment are labor-intensive, subjective, and lack real-time responsiveness, limiting their effectiveness in dynamic construction scenarios [4]. These challenges have driven growing interest in intelligent construction monitoring techniques capable of continuously perceiving and interpreting construction activities [5]. However, despite these advances, substantial challenges remain in bridging robust visual perception with high-level workflow understanding in complex tunnel environments. Accordingly, this study addresses the following research questions: (1) how to enhance the robustness of visual identification of critical entities (e.g., machinery and workers) under harsh tunnel conditions; and (2) how to effectively integrate low-level perception results of entities and activities with construction logic to enable automated monitoring of high-level construction workflows.

1.2 Literature Review

Construction workflow recognition integrates multiple techniques to identify construction entities and machinery states, and map observable features (i.e., entities and machinery states) to construction stages. It can be viewed as a hierarchical process across semantic levels. At the lowest level, element-level perception identifies entities such as machinery and workers. Activity-level perception then interprets their states as meaningful construction activities. At a higher level, workflow-level inference captures phase-level semantics by linking activities to predefined sequences. For example, a drill-and-blast cycle can be inferred from the progression of drilling, blasting, ventilation, mucking, and initial support. However, existing studies mainly focus on element detection and activity recognition, typically using video or sensor data for object detection, behavior analysis, and classification. While effective for capturing localized states, these approaches are limited in representing overall workflow progression. In structured and cyclic scenarios such as drill-and-blast operations, monitoring workflows is more critical than identifying isolated activities, as it directly supports process coordination, resource allocation, and productivity assessment. Nevertheless, workflow recognition remains underexplored.

(1) Construction entities and activities recognition

The use of computer vision (CV) technologies for the recognition of construction entities and activities has grown rapidly over the past decade [6]. Representative object detection algorithms, such as the you only look once (YOLO) family, faster region-based convolutional neural network (R-CNN), and detectron2, have demonstrated strong detection performance in general visual scenarios and have been progressively introduced into construction environments for tasks including construction element detection and localization [7,8], object tracking and pose estimation [9–11], construction activity recognition [12–14], and productivity analysis [15–17]. Early studies in this field [18–20] primarily relied on handcrafted features, such as histograms of oriented gradients, local binary patterns, and color histograms, in combination with traditional machine learning models, including support vector machines and random forests, to achieve recognition and tracking of construction resources, demonstrating the feasibility of visual information for construction monitoring.

With the advancement of deep learning, CNN-based end-to-end vision models have gradually become the dominant research paradigm in this domain. For example, Fang et al. [21] proposed an improved Faster R-CNN method for automatic detection of workers and heavy construction equipment, such as excavators,

on construction sites. By constructing a large-scale construction image dataset, their approach achieved high-accuracy object detection under near-real-time conditions, with significantly improved performance compared to traditional CV methods. Chen et al. [22] further developed a CV-based automated construction monitoring framework that integrates object detection, object tracking, construction activity recognition, and productivity analysis. This framework employs Faster R-CNN to detect the location and type of excavators and applies the deepsort tracking algorithm to establish temporal associations of the same equipment across video sequences. By analyzing dynamic variations in bounding box features, the framework enables recognition of excavator operational states and assessment of production efficiency. Lin et al. [23] proposed a four-stage image analysis framework that analyzes sequential images using Faster R-CNN integrated with a Feature Pyramid Network (FPN). By jointly modeling deviations in action sequences and anomalies in operation cycle duration, and incorporating statistical methods such as Gaussian distributions and box plots, their approach enables automatic identification of abnormal construction events in earthmoving operations.

Subsequent studies introduced improved YOLO models, temporal analysis techniques, and zero-shot learning strategies to enhance detection performance and model generalization under complex working conditions. Zeng et al. [24] proposed a detection and localization method for large-scale construction equipment in long-distance surveillance videos by combining an improved YOLOv3 model with an enhanced Extreme Learning Machine named Grey Wolf Optimization. Chen et al. [25] presented a visual analysis approach based on YOLOv5 with zero-shot learning, establishing a unified framework that integrates object detection, object tracking, and zero-shot activity recognition to automatically evaluate excavator earthmoving productivity, while identifying idle states through pixel-level positional variations of the excavator. For nighttime construction scenarios, Hua et al. [26] proposed an excavator recognition method based on an improved YOLOv7 model, which incorporates an unsupervised Night Enhancement algorithm together with a layer decomposition network and a light effect suppression network to enhance image brightness and contrast, thereby significantly improving detection accuracy under low-light conditions.

Recent studies have explored Transformer-based methods for entity and activity recognition. In weak and self-supervised settings, Sun et al. [27] proposed a Deep Convolutional Transformer Contrastive Self-Supervised (DCTCSS) framework that reduces reliance on labeled data, achieving an F1-score of 99.00% using only 80% of labels on the UCI-HAR dataset. For video-based temporal modeling, Núñez-Marcos and Arganda-Carreras [28] adopted Uniformer, which integrates convolution and self-attention, to process short video clips via a sliding window, achieving an F1-score of 93.39% on the UP-Fall dataset. To improve robustness, Nabi et al. [29] developed the BiTransAct model by combining EfficientNet-B0 with a Transformer encoder, achieving 97% validation accuracy on the SPHAR dataset. Beyond general activity recognition, Transformers have been extended to engineering and multi-modal contexts. In construction, Baek et al. [30] proposed a two-stream vision Transformer for modular construction productivity analysis, decomposing hoisting operations into six states and achieving an average F1-score of 0.9769. For multi-modal fusion, Aidarova et al. [31] integrated sensor and audio data using self-attention to capture long-range dependencies, achieving Macro-F1 scores of 0.914 and 0.909 on the Extrasensory and UCI-HAR datasets, respectively.

Overall, most existing studies focus on identifying isolated construction elements or short-term activities, facilitating intelligent recognition of construction risks and improved site management. However, the application of CV techniques in complex tunnel construction environments for intelligent monitoring requires further investigation. Moreover, extending visual perception toward higher-level semantic understanding of construction workflows remains an open research challenge.

(2) Construction process and workflow recognition

A construction workflow is a logically structured sequence of interdependent activities that delivers a project component or phase under defined technical, temporal, and managerial constraints [32]. In existing studies, the term "process recognition" is widely used, but "process" is often ambiguously defined. It may denote either specific techniques or operational steps (e.g., drilling or shotcreting), or broader cyclic activities at the project level, such as excavation–support cycles in tunneling. This dual usage obscures the distinction between low-level operations and higher-level construction organization. Accordingly, this study differentiates the two concepts: a "process" refers to a localized sequence of task-specific operations, whereas a "workflow" denotes a higher-level structure that organizes multiple processes into a temporally and logically sequenced alignment with overall construction objectives.

Process and workflow recognition and modeling in construction can be broadly categorized into four approaches: rule-based or knowledge-driven modeling, probabilistic temporal modeling, deep learning-based temporal methods, and hybrid frameworks. Rule-based models provide strong interpretability through predefined expert logic [18,33,34] but lack flexibility to handle dynamic site variations. Probabilistic temporal models such as Hidden Markov Models (HMM) [35,36] capture sequential dependencies but rely on simplifying short-term memory assumptions. In contrast, vision-based activity recognition and deep learning temporal models (e.g., LSTM networks) extract spatiotemporal patterns in a data-driven manner but operate as "black boxes", requiring large-scale annotated sequential datasets and lacking explicit construction logic representation.

Extending visual perception to this higher semantic level is particularly challenging in drill-and-blast tunnel construction. These operations exhibit tightly coupled, hierarchically organized workflows governed by strict temporal and spatial constraints. Unlike general settings, tunnel states are not defined by stable visual targets but by dynamic configurations of equipment, interactions, and evolving environments over a long duration. As a result, vision-based methods are vulnerable to occlusion, poor lighting, and fail to capture long-term logical transitions. Likewise, temporal models such as HMMs and LSTMs struggle: HMMs cannot represent long-range logic, while LSTMs require extensive frame-level annotations that are not robust in harsh underground conditions [37].

To address these limitations and reduce reliance on large annotated datasets, hybrid approaches integrating vision-based perception with knowledge-driven reasoning [24,38–40] have emerged as a promising direction. By aligning low-level observations with structured workflow semantics, they improve interpretability and robustness in cyclic tunnel operations. Notably, despite harsh visual conditions, drill-and-blast operations follow well-defined sequences (e.g., drilling, blasting, mucking, support installation) with strong coupling between activities and equipment states. Leveraging this structure, this study adopts a data–knowledge hybrid framework that integrates a predefined workflow library, a vision-based perception module, and a contextual knowledge base. This design enables robust element extraction and logically constrained inference under degraded visibility, ensuring interpretable and reliable workflow recognition in complex tunnel environments.

1.3 Contribution of This Study

To address these challenges, this study proposes an automatic construction workflow identification framework that integrates vision-based perception with hierarchical, workflow-driven knowledge reasoning using monocular video data captured by a fixed camera. The contributions of this study are twofold:

(1) Unlike prior studies focused on element detection or single-activity recognition, this study develops a hierarchical workflow-level recognition framework for drill-and-blast tunnel face construction.

A three-level representation (element–activity–workflow) is introduced to enable semantic abstraction from low-level entities perception to high-level workflow understanding. A construction sequence library is further established to formalize standard operational procedures, representing workflows as logically constrained processes rather than isolated events. In addition, a context-aware reasoning mechanism is developed by integrating temporal dependencies, object-state relationships, and rule-based constraints, enabling structured inference, state evolution analysis, and robust workflow identification.

(2) An enhanced YOLOv11 model (YOLOv11-CSA) is developed to improve the perception of multi-scale construction elements under challenging tunnel conditions, including low illumination, heavy dust, and frequent occlusions. Serving as the foundational perception module, it provides reliable low-level observations for subsequent reasoning. By strengthening feature representation and environmental robustness, the model ensures accurate extraction of construction entities and their operational states, enabling higher-level semantic inference.

This enables automatic identification and state evolution inference of construction workflows. Validation in an active drill-and-blast tunnel project demonstrates the effectiveness and engineering applicability of the proposed framework for robust construction workflow monitoring. By shifting the analytical focus from isolated equipment actions to structured workflow interpretation, this study extends current construction monitoring research toward process-aware and management-oriented intelligence, supporting construction progress tracking, productivity evaluation, and schedule compliance assessment in intelligent construction management systems.

2 Methodology

The overall framework of the proposed method is illustrated in Fig. 1 and consists of four core modules: (1) construction workflow sample database for the drill-and-blast tunnel face construction, (2) an automatic annotation module based on Segment Anything Model 2 (SAM2), (3) a construction elements (including construction equipments, workers, and equipment components) recognition module based on an enhanced YOLOv11 model (YOLOv11-CSA), and (4) a hierarchical construction workflow context reasoning module. These components jointly form a pipeline driven by CV for workflow recognition for drill-and-blast tunnel construction, enabling construction workflow perception, monitoring, and management in intelligent construction environments.

The construction workflow sample database is established through hierarchical workflow modeling. Discrete construction workflows are systematically classified to clarify workflow boundaries and transition logic, forming the foundation for workflow recognition scheme design. Drill-and-blast tunnel construction is a multi-stage, multi-task coupled workflow that typically includes construction preparation, tunnel face operations, and quality inspection. Among these stages, tunnel face operations exhibit strong cyclicality, continuity, and dynamic evolution, involve coordinated operation of multiple equipment types, and impose strict quality control requirements. Therefore, tunnel face construction is selected as the primary research focus. Based on construction specifications, industry standards, expert knowledge, and field investigations, detailed operational workflows and stage transition relationships are formalized and further represented as a knowledge graph, explicitly modeling temporal dependencies and workflow constraints to support subsequent automated workflow recognition.

In the visual perception module, the automatic annotation capability of SAM2 is integrated with the YOLOv11-CSA model. SAM2 provides high-quality instance segmentation and contour extraction under complex visual conditions, significantly reducing manual annotation effort. However, due to its high computational cost, SAM2 is unsuitable for real-time deployment in tunnel construction scenarios.

In contrast, YOLOv11 offers favorable real-time performance and engineering deployability, but its detection accuracy is highly dependent on data quality and is challenged by tunnel environments characterized by low illumination, dust interference, and glare. To address these issues, targeted structural enhancements are introduced to the YOLOv11 backbone and neck to improve feature extraction and multi-scale fusion. Combined with high-quality training data generated by SAM2, coordinated optimization is achieved at both data and model levels, enhancing detection accuracy and robustness for construction elements.

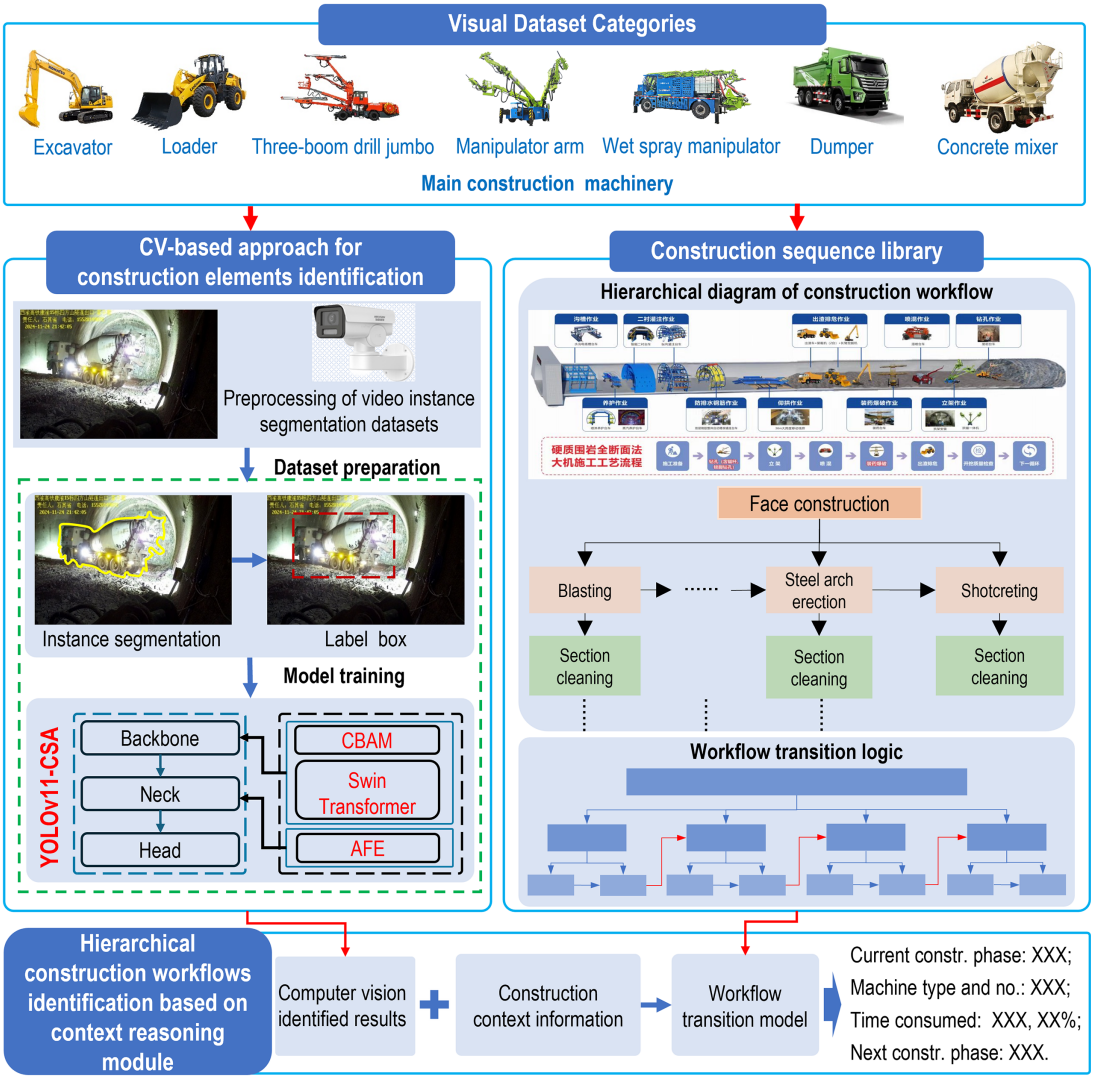

Figure 1: Methodology framework of construction workflow recognition in drill-and-blast tunnels.

At the workflow recognition decision reasoning module, a hierarchical model incorporating domain knowledge is developed to capture contextual relationships among construction workflows. Detection results from the visual perception module are mapped onto the hierarchical workflow model to enable automatic identification of the current construction workflow, prediction of subsequent workflow, and statistical analysis of workflow durations. By integrating CV with hierarchical construction workflows based on a context reasoning module that fuses workflow features, construction context information, and workflow

transition model, the proposed method provides an effective technical solution for continuous workflow monitoring and intelligent management of tunnel face construction.

3 Establishment of Construction Sequence Library

The systematic decomposition of tunnel face construction workflows in drill-and-blast tunnel forms the basis for constructing the construction workflow sample database. This decomposition is conducted using the Work Breakdown Structure (WBS) method, which enables complex construction workflows to be hierarchically organized into clearly defined and logically related components. Oriented toward automatic construction workflow recognition, the tunnel face construction workflow is organized into a three-level hierarchy consisting of primary, secondary, and tertiary workflows.

The primary workflow corresponds to tunnel face construction, representing the macro operational unit of a complete drill-and-blast construction cycle. Secondary workflows subdivide the primary workflow into key technological stages that are executed sequentially in accordance with standardized construction procedures, collectively forming a closed operational cycle. These secondary workflows are relatively independent in terms of construction objectives, equipment configuration, and operational organization. Tertiary workflows further refine secondary workflows into specific, executable construction actions that directly correspond to observable equipment operations, worker activities, or construction events. As such, tertiary workflows constitute the fundamental analysis units for vision-based element detection, state recognition, and workflow reasoning. Based on this hierarchical decomposition, tunnel face construction is defined as the primary workflow. At the secondary level, six workflow categories are identified: blasting, mucking, setting out, steel arch erection, drilling, and shotcreting. Each secondary workflow is further decomposed into tertiary workflows according to specific construction actions, forming a complete construction sequence library, as shown in Table 1.

Table 1: The sample library of construction workflow in drill-and-blast tunnels.

Primary Workflow	Secondary Workflow	Tertiary Workflow	Construction Elements	CV Recognition Scheme
Face construction	Blasting	Face cleaning	Excavator	Detect “excavator”
		Machinery entry	Loader + arch installation trolley	Detect “loader + steel arch installation jumbo”
		Charging & blasting		Detect “boom raising action of loader + arch installation trolley”
	Mucking	Ventilation & hazard removal	Excavator	Detect “excavator”
		Mucking operation	Loader + dumper + excavator	Detect “two loaders + dumper” or “excavator + dumper”
	Setting out	Surveying & layout work	Total station + worker	Detect “worker + total station”
	Steel arch erection	Arch transportation	Loader	Detect “loader”
		Arch jumbo entry	Arch installation trolley + total station + worker + handle rock drill	Detect “arch installation trolley”
		Surveying		Detect “arch installation trolley + total station + worker”
		Arch assembly		Detect “boom raising action of arch installation trolley”
		Arch welding & linking		Detect “worker + boom raising action of arch installation trolley”
		Foot locking		Detect “worker + handle rock drill”

(Continued)

Table 1 (continued)

Primary Workflow	Secondary Workflow	Tertiary Workflow	Construction Elements	CV Recognition Scheme
	Drilling	Face scaling Site leveling	Excavator	Detect "excavator"
		Drilling jumbo entry	Double-boom drilling jumbo	Detect "double-boom jumbo boom raising"
		Drilling		Detect "boom raising action of double-boom drilling jumbo"
	Shotcreting	Concrete transportation	Concrete mixer truck	Detect "concrete mixer truck"
		Spraying	Wet shotcrete manipulator	Detect "wet spray manipulator boom raising"

For example, the secondary workflow of blasting consists of three tertiary workflows: face cleaning, machinery entry, and charging & blasting. Face cleaning involves removing shotcrete residues that block blast holes using an excavator to ensure proper explosive loading. Machinery entry refers to the entry of blasting equipment into the tunnel face area, followed by the coordinated execution of explosive charging and detonation. The Mucking workflow includes ventilation and hazard removal, and the mucking operation. After blasting, harmful gases and dust are removed through ventilation, loose rocks are cleared, and the stability of the surrounding rock is inspected to ensure safe working conditions. Subsequently, the blasted rock debris is transported out of the tunnel through coordinated operations involving two loaders, a mucking dumper, and an excavator. Initially, two loaders operate in coordination with the dumper to remove large muck fragments; subsequently, an excavator cooperates with the dumper to clear the remaining finer muck.

Setting out corresponds to a single tertiary workflow in which construction workers use a total station to determine and mark blast hole positions on the tunnel face. Steel arch erection is decomposed into six tertiary workflows: arch transportation, arch jumbo entry, surveying, arch assembly, arch welding & linking, and foot locking. These workflows collectively describe the delivery, positioning, assembly, welding, and fixation of arches to form a stable initial support structure. The drilling workflow consists of face scaling, site leveling, drilling jumbo entry, and drilling. These steps ensure safe and stable drilling conditions and enable the formation of blast holes using the drilling jumbo. Shotcreting includes concrete transportation and spraying, involving the entry of a concrete mixer truck, docking with the wet shotcrete manipulator, and spraying concrete onto steel arches and surrounding rock surfaces to complete initial support. Table 1 presents the tunnel face construction workflow sample database and its corresponding CV recognition scheme based on tertiary workflow features. By mapping construction actions to observable visual elements, the proposed construction sequence library provides structured prior knowledge for subsequent automatic workflow recognition and hierarchical decision reasoning.

4 Hierarchical Construction Workflows Identification Based on Context Reasoning

The construction workflow context reasoning module is developed based on a hierarchical workflow modeling paradigm. By integrating a hierarchical construction event sample library (Table 1), construction context information, and a dynamically adaptive workflow sequence library, the proposed method establishes a unified decision reasoning module tailored to tunnel face construction scenarios. This module supports four core functions: current construction workflow identification, subsequent workflow prediction, dynamic construction workflow monitoring, and construction element quantity statistics with consistency verification.

4.1 Construction Workflow Context Information

Construction workflow context information provides structured background knowledge and dynamic state constraints required for construction workflow identification and decision reasoning. Unlike isolated object detection or instantaneous activity recognition, workflow context introduces semantic, temporal, and procedural constraints that enable a more advanced interpretation of construction workflows and their evolution over time.

In this study, construction workflow context information is modeled along three complementary dimensions: (1) Temporal and workflow context, which encodes prior knowledge of construction workflow within a construction cycle, logical dependencies among workflows, and typical duration distributions. (2) Object and state context, which describes the category, functional role, and operational states of construction elements, including equipment and workers, within specific time windows. (3) Rule and logic context, which formalizes construction procedures, safety regulations, and technological constraints that restrict feasible workflow transitions. By jointly modeling these contextual dimensions, object detection results and equipment state observations can be mapped to admissible workflow states with explicit logical consistency. Temporal context supports next workflow prediction within a construction cycle; object and state context enables workflow affiliation and suspension detection based on resource availability; and rule and logic context constrains workflow evolution to comply with procedural and safety requirements.

Based on the rule and logic context, an initial workflow transition sequence is extracted and formalized, as illustrated in Fig. 2. In drill-and-blast tunnel construction, workflow execution follows a highly standardized order to ensure operational safety and construction quality. For example, blasting is obligatorily followed by ventilation and hazard removal, and subsequently by mucking, to restore safe working conditions. Surveying must be conducted after ventilation but before drilling, as it defines blast hole positions. Shotcreting is performed after steel arch erection to ensure effective interaction between sprayed concrete, steel arches, and surrounding rock.

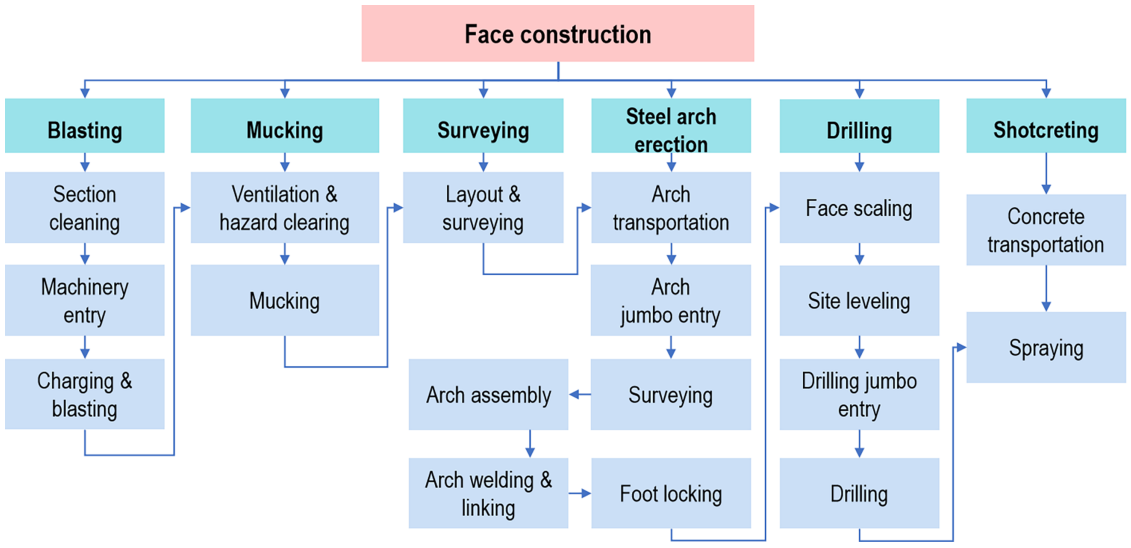

Figure 2: Initialized workflow transition sequence of tunnel face operations in drill-and-blast tunnel.

In practice, local deviations from the standard workflow sequence may occur due to geological variability, construction organization, or equipment scheduling. However, such deviations remain strictly governed by construction logic and safety constraints. Leveraging this certainty, a workflow sequence library

is constructed, consisting of a default initialized sequence and a limited set of constrained alternative sequences. When deviations are authorized by construction management, the workflow recognition system dynamically updates its temporary sequence library and prioritizes the corresponding transition rules during decision reasoning.

Through the integration of standardized workflow logic and adaptive sequence updating, the proposed context decision reasoning module maintains consistency with construction procedures while supporting robust workflow identification under complex and dynamic tunnel construction conditions.

4.2 Construction Workflow Decision Reasoning

The construction workflow decision reasoning module is developed by integrating the hierarchical construction event sample library, construction workflow context information, and a dynamically adaptive workflow sequence repository. The module enables automatic identification, prediction, and state evaluation of tunnel face construction workflows.

The recognition of the current construction activity forms the basis of the decision reasoning workflow and is primarily achieved through tertiary workflow identification. For each tertiary workflow, a set of discriminative workflow features is defined, and a mapping is established between feature combinations and workflow categories (Table 1). Through joint feature evaluation, the active tertiary workflow is determined and subsequently mapped upward to the corresponding secondary and primary workflows according to the predefined hierarchical structure. Workflow prediction is performed by jointly considering workflow transition rules and the adaptive workflow sequence repository. Under the constraints of construction logic, safety regulations, and contextual consistency, the algorithm determines feasible candidate workflows for the subsequent stage. When multiple candidates are admissible, they are ranked based on historical execution patterns, current workflow state, and contextual consistency, and the workflow with the highest likelihood is selected.

Construction workflow monitoring is conducted based on recognized workflow states using a multi-granularity strategy. When the detection model achieves sufficient recognition performance, the tertiary level workflows can be directly identified. If tertiary workflows are difficult to detect using CV techniques, monitoring is instead performed at the secondary workflow level to ensure stable and robust process tracking. If no valid construction targets are detected, or all construction elements remain stationary within a continuous time window, the workflow state is classified as idle. This study introduces a temporal smoothing mechanism. Under this constraint, short-term target disappearance caused by sudden illumination changes or occlusion does not immediately trigger workflow state transitions. Instead, the system waits until recognition results stabilize above a preset threshold and then performs logical inference using contextual information. A process is considered complete only when primary construction elements show a clear movement trend away from the operational area (e.g., the bounding box center continuously shifts from the tunnel face toward the gantry until disappearance). Similarly, process initiation is confirmed only when key machinery enters the gantry and moves toward the tunnel face. This fault-tolerant strategy, combining spatial positioning and temporal logic, effectively filters visual noise and enhances robust workflow reasoning. In addition, construction element quantity monitoring is implemented to verify consistency between observed equipment configurations and planned resource allocation. It continuously counts construction element instances and analyses their spatial distribution. Persistent discrepancies between observed and planned configurations trigger anomaly alerts, enabling timely detection of resource scheduling deviations or organizational inconsistencies.

Overall, the decision reasoning module integrates CV outputs with a hierarchical workflow identification model and contextual constraints, forming a closed-loop process for workflow identification,

prediction, progress monitoring, and construction resource allocation consistency verification in tunnel face construction. The pseudo-code of the decision-making algorithm is provided in Algorithm 1.

Algorithm 1: Construction workflow decision reasoning using YOLOv11-CSA and context information

Input: Continuous video frame sequence V at the tunnel face; Trained YOLOv11-CSA model Mcsa; Hierarchical construction event sample library L_{event}; Context information base C_{base} (contains rule/logic context C_{rule}, object/state context C_{state}, adaptive sequence library C_{seq}, historical execution patterns H_{pattern}, and planned resource configurations R_{plan}); Temporal smoothing threshold T_{smooth}; Confidence threshold T_{conf}.

Output: Current hierarchical workflow W_{curr}; Predicted subsequent workflow W_{next}; Previous Workflow W_{prev}; Hierarchical construction workflow $W_{\text{tertiary}} W_{\text{secondary}} W_{\text{primary}}$Anomaly alerts A_{alert}; Element quantity statistics S_{qty}.

Initialize: W_{curr} = Null, W_{prev} = Null, W_{next}= Null; T_{missing} = 0 // Timer for missing or stationary targets.

FOR each frame $f_t \in V$ **DO**
// Step 1: Visual Perception & Element Quantity Statistics
$D_t \leftarrow M_{\text{csa}}(f_t)$ (Extract bounding boxes, categories, and operational states)
$S_{\text{qty}} \leftarrow \text{CountInstances}(D_t)$
// Step 2: Resource Consistency Verification
IF DetectDeviation(S_{qty}, R_{plan}) == True **THEN**
Trigger anomaly alert $A_{\text{alert}} \leftarrow$ "$Resourceschedulingdeviation$"
END IF
// Step 3: Dynamic Workflow Monitoring (Temporal Smoothing & Idle Detection)
IF $D_t == \varnothing$ **or** All targets in D_t are stationary **THEN**
$T_{missing} \leftarrow T_{missing} + 1$
IF $T_{\text{missing}} > \tau_{\text{smooth}}$ **THEN**
$W_{curr} \leftarrow$ "$Idle$"
ELSE
$W_{\text{curr}} \leftarrow W_{\text{prev}}$ (Temporal smoothing: ignore transient noise/occlusion)
END IF
ELSE
$T_{\text{missing}} \leftarrow 0$ (Reset smoothing timer)
$F_{\text{context}} \leftarrow \text{ExtractFeatures}(\text{D}_\text{t}, \text{C}_{\text{base}})$
// Step 4: Current Workflow Identification (Multi-granularity Strategy)
$W_{tertiary} \leftarrow \text{JointFeatureEvaluation}(F_{context}, L_{event})$
IF $\text{Confidence}(W_{\text{tertiary}}) > T_{\text{conf}}$ **THEN**
// Bottom-up hierarchical mapping
$W_{\text{curr}} \leftarrow \text{MapUpward}(W_{\text{tertiary}}, L_{\text{event}})$
ELSE
// Fallback to secondary level for robustness
$W_{\text{secondary}} \leftarrow \text{EvaluateAtSecondaryLevel}(C_{\text{base}}, L_{\text{event}})$
$W_{\text{curr}} \leftarrow \text{MapUpward}(W_{\text{secondary}}, L_{\text{event}})$
END IF
END IF
// Step 5: Workflow Transition & Subsequent Prediction

(Continued)

Algorithm 1 (continued)

IF $W_{curr} \neq W_{prev}$ **AND** $W_{curr} \neq$ "Idle" **THEN**
  // Dynamic Sequence Update
  IF Transition $(W_{prev} \rightarrow W_{curr})$ deviates from the default but is authorized **THEN**
    UpdateAdaptiveSequence$(C_{seq}, W_{prev}, W_{curr})$
  END IF
  // Subsequent Workflow Prediction
  $List_{candidates} \leftarrow$ GetFeasibleTransitions(W_{curr}, C_{rule})
  $W_{next} \leftarrow$ RankByLikelihood$(List_{candidates}, C_{seq}, H_{pattern})$
END IF
$W_{prev} \leftarrow W_{curr}$
Output $W_{curr}, W_{next}, A_{alert}, S_{qty}$
END FOR

5 Tunnel Construction Elements Recognition Based on YOLOv11 and SAM2

Accurate recognition of construction elements at the tunnel face is a prerequisite for subsequent construction workflow reasoning and progress monitoring. To this end, this study develops a construction element recognition framework that combines the YOLOv11-CSA model with an efficient automatic annotation strategy based on the SAM2.

5.1 Construction of Labeled Datasets Using SAM2

(1) SAM2-Based Automatic Annotation

SAM2 is a second-generation general-purpose instance segmentation model proposed by Meta AI, featuring strong zero-shot generalization capability [41]. Without category-specific supervised training, SAM2 can achieve high precision instance segmentation for arbitrary targets using only lightweight user prompts, such as points, bounding boxes, text, or historical masks. Its overall architecture is illustrated in Fig. 3. Structurally, SAM2 adopts a temporal segmentation framework that integrates visual perception with an explicit memory mechanism. The image encoder extracts multi-scale visual features, while the prompt encoder processes prompts provided by the user. During inference, a memory encoder constructs a cross-frame memory bank, and a memory attention module propagates and fuses temporal information. Based on the current visual features, prompt embeddings, and historical memory, the mask decoder outputs high-quality instance segmentation results, enabling end-to-end inference.

In video instance segmentation scenarios, a single prompt is required only in the initial frame. Subsequent frames are processed sequentially, with the memory attention module maintaining temporal consistency of object masks across frames. Segmentation results are generated in real time and stored in the memory bank to support subsequent inference, allowing SAM2 to achieve stable and robust performance in long-term, dynamic scenes such as construction videos. Leveraging the precise segmentation and temporal consistency of SAM2, this study further proposes an automated annotation method for training YOLO-based object detection models. The method converts pixel-level instance segmentation masks into bounding box annotations required by object detectors, significantly reducing manual labeling effort while preserving annotation quality. Specifically, selected tunnel construction videos are input into the annotation system, and target objects are prompted through the SAM2 interface to generate instance masks. The pixel-level contours are extracted to compute extrema along horizontal and vertical directions, from which minimum enclosing

bounding boxes are derived. Each bounding box is then assigned a category label according to a predefined construction semantic taxonomy. During output, bounding box coordinates are normalized following the YOLO annotation format, generating label files containing class indices, object center coordinates, and relative width and height, paired with the corresponding images. The overall framework of the automated annotation process is shown in Fig. 3.

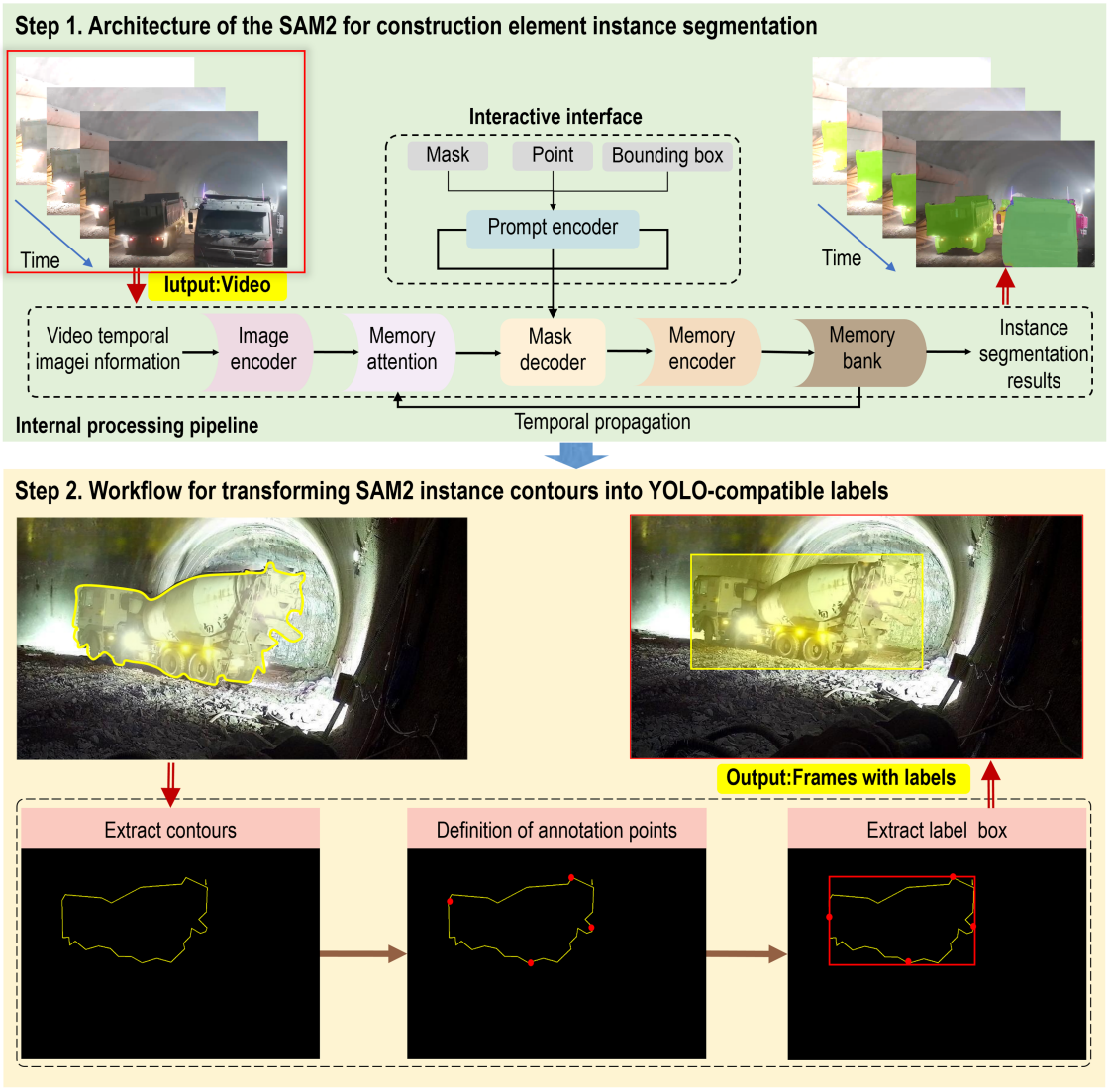

Figure 3: Architecture of the SAM2 and the instance-level annotation workflow.

Furthermore, the proposed approach naturally extends to batch annotation of video sequences. Benefiting from SAM2's temporal modeling and object tracking capability, the system can automatically extract consistent instance contours of the same construction target across consecutive frames, ensuring temporal coherence of annotations. This significantly improves labeling efficiency for large-scale construction video datasets and provides high-quality structured training data for subsequent YOLO-based construction element detection models.

(2) Annotation Quality Evaluation

The evaluation was conducted by randomly sampling 500 images from the full dataset of 22,268 images, covering 11 categories (Excavator, Loader, Dumper, Worker, Arch installation trolley, Drilling jumbo, Jumbo arm, Wet spray manipulator, Manipulator arm, Concrete mixer, and Guardrail). For this subset, tightly fitted bounding boxes were manually annotated to establish ground truth. The bounding boxes generated by SAM2 were then compared against these ground truth annotations. Annotation quality was quantitatively evaluated using instance-averaged Intersection over Union (IoU) and mean four-corner pixel distance for each image. The specific formulas are defined as follows:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{|B_{sam}^{i} \cap B_{gt}^{i}|}{|B_{sam}^{i} \cup B_{gt}^{i}|} \tag{1}$$

where $mIoU$ denotes the mean Intersection over Union (IoU) of all instances in a single image. N is the total number of instances in the image. $i \in \{1, 2, \ldots, N\}$ indexes each instance. B_{sam}^{i} and B_{gt}^{i} represent the bounding boxes generated by SAM2 and the manually annotated ground-truth, respectively. $|B_{sam}^{i} \cap B_{gt}^{i}|$ and $|B_{sam}^{i} \cup B_{gt}^{i}|$ denote their intersection and union areas.

$$D = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{4}\sum_{j=1}^{4}\sqrt{(x_{ij}^{sam} - x_{ij}^{gt})^2 + (y_{ij}^{sam} - y_{ij}^{gt})^2}\right) \tag{2}$$

where D denotes the mean pixel distance between corresponding bounding-box corners across all instances in a single image. $j \in \{1, 2, 3, 4\}$ indexes the four corners of each bounding box (i.e., top-left, bottom-left, top-right, and bottom-right). $(x_{ij}^{sam}, y_{ij}^{sam})$ and $(x_{ij}^{gt}, y_{ij}^{gt})$ are the pixel coordinates of the j-th corner of the SAM2-predicted and ground-truth bounding boxes, respectively.

Fig. 4 illustrates the miss-rate distribution of the SAM2-based auto-annotation pipeline across construction element categories. Among 1733 sampled instances, the overall miss rate is 10.96%, indicating a notable class-wise imbalance. SAM2 achieves high annotation quality for large, rigid machinery (e.g., Loader, Dumper, and Excavator), with miss rates below 3%. In contrast, performance degrades significantly for slender structures with weak visual salience (e.g., Guardrail, 28.37%) and articulated, occlusion-prone components (e.g., Jumbo Arm and Wet Spray Manipulator, 19.05% and 16.67%, respectively). These results indicate that under harsh tunnel conditions (low illumination, heavy dust, and unclear boundaries), zero-shot foundation models without domain-specific fine-tuning are prone to fine-grained semantic loss. While SAM2 can substantially accelerate dataset construction, its performance is unstable under severe occlusion and complex scenes. Therefore, a rigorous manual verification stage is performed in practical dataset development, with missed objects re-annotated manually, to ensure labeling accuracy for subsequent CV model training.

Fig. 5a,b presents frame-wise evaluations of mIoU and corner pixel deviations, respectively. Across the 500-frame sample at 1920 × 1080 resolution, the average IoU reaches 0.805, and the mean corner distance is D = 21.19 px, confirming the high annotation accuracy of SAM2. However, analysis of high-error frames shows that part of the deviation arises not from model failure but from a systematic discrepancy between manual annotation rules and algorithmic inference. For small targets (e.g., workers and guardrails), frequent occlusions by soil piles or machinery lead human annotators to adopt a strict "visible-only" policy (e.g., labeling only visible helmet regions). In contrast, SAM2 exploits spatio-temporal consistency to infer full object extents across occlusions, which reduces IoU under this inconsistent definition of object boundaries. Similarly, in multi-equipment interactions, machinery is often visually fragmented (e.g., excavator body and bucket separated by foreground occlusion). Human annotation typically covers only

the main visible part, whereas SAM2 predicts complete object boundaries, resulting in large box-size discrepancies and corresponding spikes in pixel distance (Fig. 5b). Overall, although large-model-based auto-annotation significantly accelerates dataset construction, it remains sensitive under severe occlusion and dusty conditions. Therefore, a hybrid pipeline combining machine-generated coarse annotations with rigorous manual verification is required to ensure reliable training data for tunnel construction scenarios.

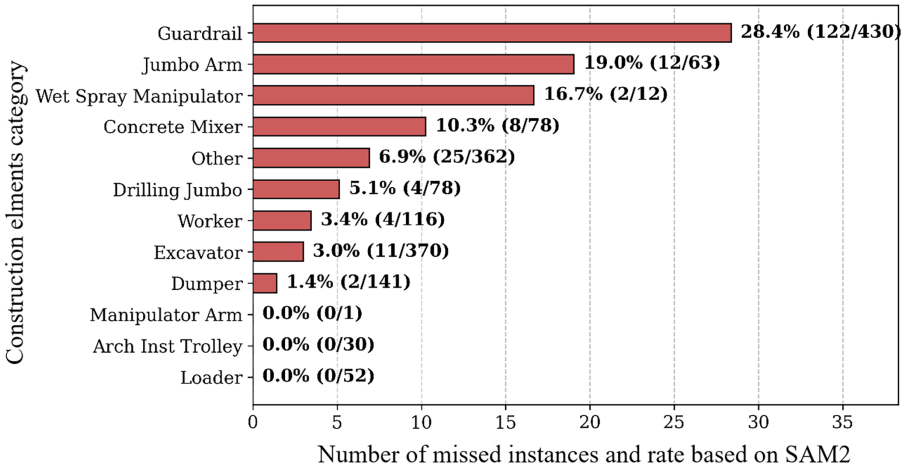

Figure 4: Quantitative evaluation of missed detections by the SAM2 auto-annotation pipeline.

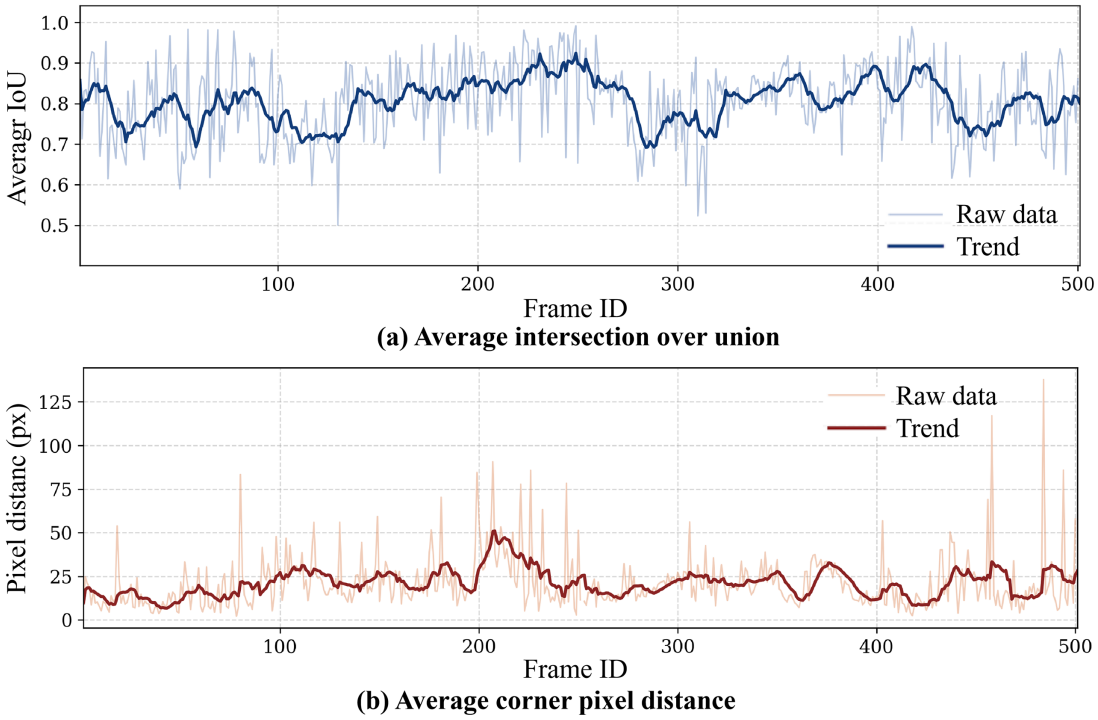

Figure 5: Consistency evaluation of mIoU and *D* between ground-truth and SAM2-generated annotations.

5.2 YOLOv11-CSA: An Enhanced YOLOv11-Based Object Detection Model

YOLOv11 is a recent generation real-time object detection model proposed by the Ultralytics team. It achieves a favorable balance between detection accuracy and inference efficiency and has been widely applied in scenarios such as autonomous driving, industrial inspection, and robotic vision [42]. The YOLOv11 baseline incorporates: (1) C3k2 blocks leveraging depthwise separable convolutions and lightweight feature recalibration mechanisms for local feature extraction; (2) Spatial Pyramid Pooling Fast for multi scale context aggregation via parallel max pooling; (3) C2PSA modules combining Cross Stage Partial connections and Pyramid Squeeze Attention; and (4) detection heads enhanced with depthwise convolutions for improved parameter efficiency.

Although YOLOv11 performs well under conventional conditions, its detection accuracy and stability degrade noticeably in the tunnel construction environment. Data augmentation alone is insufficient to fundamentally address this issue, and targeted optimization at the model architecture level is therefore required. To address the challenges of tunnel construction scenarios, this study enhances YOLOv11 by focusing on three key capabilities, including local saliency enhancement, cross-scale dynamic fusion, and long-range global modeling. Based on these principles, a cooperative suppression attention detection model referred to as YOLOv11 CSA is constructed, as illustrated in Fig. 6.

At the end of the backbone, the original SPPF to C2PSA module is replaced with an SPPF to Convolutional Block Attention Module [43] in the Swin Transformer structure [44]. The Convolutional Block Attention Module reduces interference caused by dust and overexposure while enhancing responses to critical structural features through combined channel and spatial attention mechanisms. The Swin Transformer captures global semantic relationships along the tunnel depth direction using a shifted window self-attention approach. In the deep feature fusion paths of the neck, the Convolutional Block Attention Module and Swin Transformer are incorporated to improve the consistency and stability of multi-scale semantic representations under complex illumination conditions. Additionally, an Adaptive Feature Enhancement module [45] is inserted into the upsampling path, where a learnable dynamic weighting mechanism strengthens high-resolution feature representations and mitigates semantic suppression of distant small targets as well as feature distortion in near field regions affected by overexposure.

Through the coordinated deployment of attention mechanisms and global modeling modules in both the backbone and the neck, YOLOv11 CSA improves detection accuracy and robustness under severely degraded visual conditions in the tunnel construction environment while preserving the real-time inference capability of YOLOv11. The final model produces three detection heads at different scales, corresponding to high recall for small targets under low illumination, high discriminative capability for medium-scale equipment, and precise localization for large construction machinery.

6 Engineering Validation

A high-definition explosion-proof fixed camera was deployed in the trestle area of the drill-and-blast tunnel, where the lower space accommodates invert construction, and the upper deck serves as a passage for construction equipment. The camera was installed approximately 50 m away from the tunnel face at a height of about 12 m, providing a stable, panoramic viewpoint that continuously covers the tunnel face and the primary operational zones throughout the construction cycle. Engineering validation demonstrates that the proposed method can operate continuously and robustly across different tunnel face workflows, including blasting, mucking, setting out, steel arch erection, drilling, and shotcreting. Even under complex site conditions, such as significant illumination fluctuations, dust generated by blasting and mucking, and the concurrent operation of multiple types of equipment. The method maintains reliable recognition of construction elements and accurate perception of tunnel construction workflows.

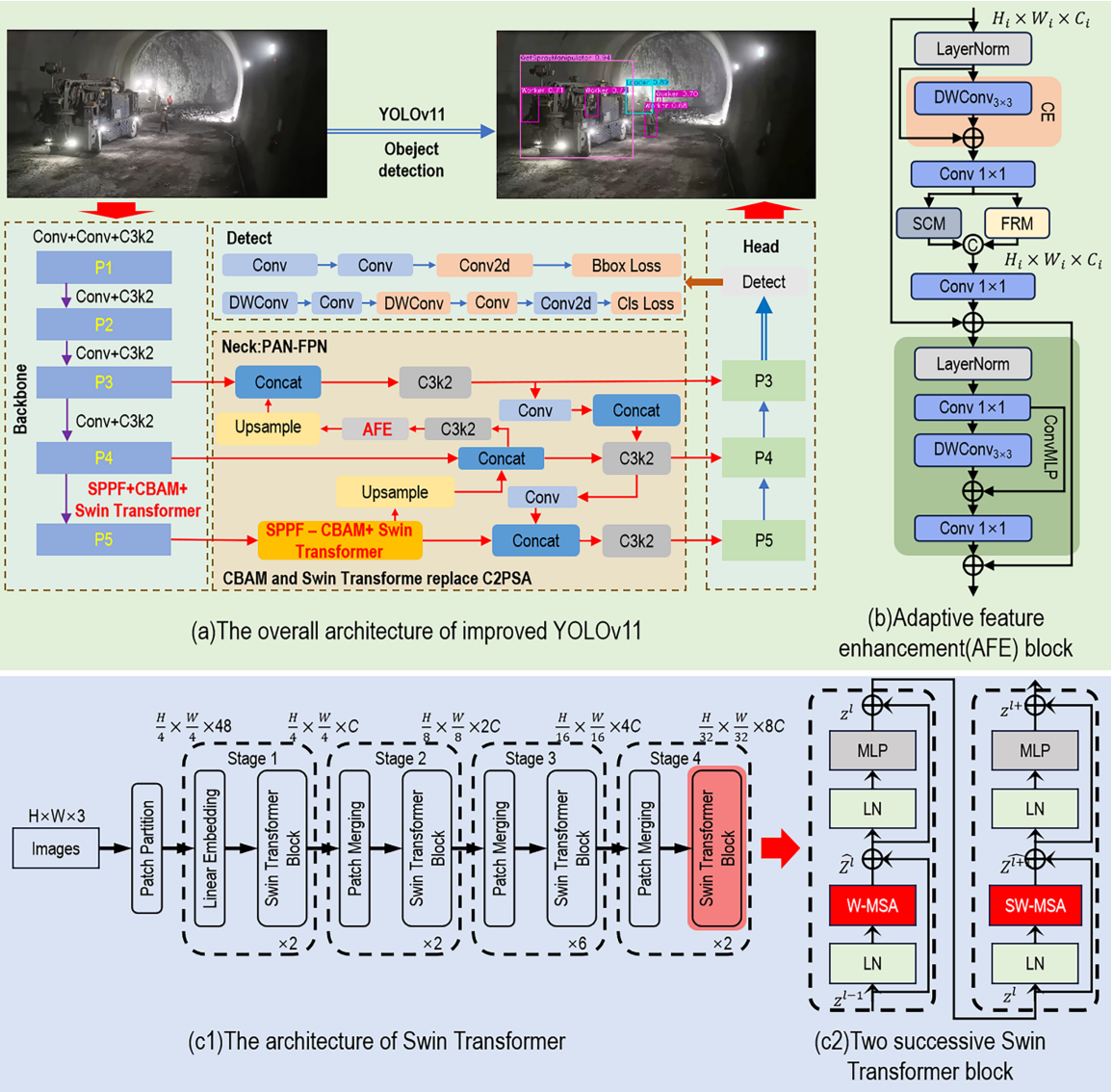

Figure 6: The architecture of the YOLOv11-CSA.

6.1 Evaluation of Construction Element Recognition Performance

To evaluate the recognition performance of YOLOv11-CSA, a dedicated image dataset for drill-and-blast tunnel construction was curated from one year of onsite video recordings, covering approximately 100 construction cycles. The dataset spans diverse operating conditions and construction workflows to enhance robustness and generalization capability. Using the SAM2, fine-grained instance-level annotations were generated for 11 representative object categories involved in tunnel face construction, including major construction equipment, key mechanical components, workers, and fixed structural elements. In total, 22,268 images were annotated and randomly divided into training and validation sets at a ratio of 8:2. Model training was conducted on an NVIDIA GeForce RTX 5090D GPU for 500 epochs. The dataset composition and training configuration are summarized in Table 2.

To evaluate the effectiveness and synergistic mechanisms of the proposed modules (CBAM, Swin Transformer, and AFE), comprehensive ablation experiments were conducted on the tunnel construction

dataset. As shown in Table 3, the performance of YOLOv8, the YOLOv11 baseline, individual module variants, and the final integrated model (YOLOv11-CSA) was compared in terms of Precision, Recall, mAP50, and mAP50–95. The YOLOv11 model slightly outperforms YOLOv8 across all metrics, owing to its improved network architecture. Compared with the YOLOv11 baseline, the CSA model achieves superior performance across all metrics, with the most notable gains observed in Recall and mAP50–95, increasing from 0.977 to 0.981 (+0.4%) and from 0.913 to 0.924 (+1.1%), respectively. These results indicate enhanced detection completeness and improved localization accuracy.

Table 2: Construction workflow and element distribution in the model training dataset.

Workflow	Elements	Training Instances	Validation Instances
Blasting	Excavator	12,813	3192
	Loader	2127	542
Mucking	Dumper	4877	1218
Setting out	Worker	13,255	3325
Steel arch erection	Arch installation trolley	1020	265
Drilling	Drilling jumbo	2694	672
	Jumbo arm	3915	1000
Shotcreting	Wet spray manipulator	316	72
	Manipulator arm	2153	529
	Concrete mixer	2623	645
N/A	Guardrail	11,712	2960
All		57,505	14,420

Table 3: Ablation study on the contribution of each module to the performance of the proposed YOLOv11-CSA.

Elements	V8	V11	+CBAM	+ST	+AFE	+CBAM +ST	V11-CSA	V8	V11	+CBAM	+ST	+AFE	+CBAM +ST	V11-CSA
	Precision							Recall						
Excavator	0.993	0.992	0.992	0.993	0.994	0.993	0.990	0.991	0.992	0.992	0.993	0.992	0.992	0.991
Loader	0.979	0.982	0.976	0.974	0.973	0.964	0.968	0.990	0.990	0.983	0.989	0.989	0.989	0.989
Arch installation trolley	0.986	0.990	0.993	0.992	0.992	0.992	0.992	0.985	0.989	0.985	0.992	0.987	0.985	0.989
Dumper	0.997	0.997	0.997	0.996	0.997	0.997	0.996	0.998	0.998	0.998	0.998	0.998	0.997	0.998
Drilling jumbo	0.981	0.983	0.966	0.969	0.972	0.978	0.972	0.973	0.975	0.973	0.976	0.963	0.971	0.973
Wet spray manipulator	0.974	0.975	0.999	0.989	0.986	0.981	0.986	0.957	0.958	0.972	0.958	0.967	0.972	0.969
Concrete mixer	0.992	0.991	0.992	0.992	0.992	0.992	0.987	0.992	0.995	0.989	0.991	0.994	0.992	0.994
Manipulator Arm	0.964	0.973	0.965	0.972	0.962	0.974	0.971	0.985	0.991	0.970	0.983	0.987	0.987	0.991
Jumbo arm	0.971	0.974	0.956	0.955	0.963	0.969	0.968	0.978	0.980	0.976	0.977	0.980	0.981	0.980
Worker	0.942	0.940	0.924	0.923	0.946	0.946	0.947	0.888	0.885	0.822	0.843	0.906	0.881	0.913
Guardrail	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.999	0.999
All	0.980	0.981	0.978	0.978	0.980	0.980	0.980	0.976	0.977	0.969	0.973	0.978	0.977	0.981

(Continued)

Table 3 (continued)

Elements	V8	V11	+CBAM	+ST	+AFE	CBAM +ST	V11-CSA	V8	V11	+CBAM	+ST	+AFE	CBAM +ST	V11-CSA
	mAP50							mAP50-95						
Excavator	0.994	0.994	0.994	0.994	0.995	0.994	0.994	0.946	0.947	0.933	0.933	0.937	0.946	0.942
Loader	0.994	0.995	0.994	0.994	0.994	0.994	0.994	0.954	0.959	0.941	0.946	0.949	0.956	0.965
Arch installation trolley	0.994	0.995	0.990	0.995	0.992	0.994	0.995	0.943	0.944	0.916	0.920	0.930	0.950	0.951
Dumper	0.994	0.994	0.995	0.995	0.995	0.995	0.995	0.988	0.990	0.988	0.988	0.989	0.990	0.990
Drilling jumbo	0.992	0.992	0.991	0.991	0.992	0.992	0.992	0.954	0.954	0.924	0.931	0.946	0.954	0.961
Wet spray manipulator	0.989	0.990	0.985	0.979	0.978	0.984	0.991	0.865	0.866	0.810	0.807	0.822	0.875	0.904
Concrete mixer	0.993	0.994	0.994	0.994	0.994	0.994	0.994	0.969	0.971	0.959	0.960	0.965	0.970	0.971
Manipulator Arm	0.991	0.992	0.991	0.991	0.993	0.993	0.992	0.857	0.858	0.766	0.779	0.829	0.853	0.872
Jumbo arm	0.990	0.991	0.988	0.988	0.991	0.990	0.990	0.916	0.917	0.851	0.859	0.893	0.907	0.923
Worker	0.947	0.948	0.921	0.928	0.961	0.954	0.960	0.643	0.644	0.554	0.565	0.636	0.650	0.686
Guardrail	0.995	0.995	0.995	0.995	0.995	0.995	0.995	0.993	0.994	0.993	0.992	0.992	0.994	0.994
All	0.988	0.989	0.985	0.986	0.989	0.989	0.990	0.912	0.913	0.876	0.880	0.899	0.913	0.924

From a general perspective, the introduction of individual modules does not consistently improve performance. For mAP50–95, CBAM (0.876), Swin Transformer (0.880), and AFE (0.899) all underperform the baseline (0.913). This suggests that under complex tunnel conditions—characterized by dust, strong illumination, and frequent occlusion—single mechanisms are susceptible to noise interference and insufficient feature representation. For instance, CBAM's local attention mechanism may be misled by high-brightness dust regions; Swin Transformer, despite its capability for global modeling, may weaken local boundary details without task-specific enhancement; and AFE's multi-scale fusion remains limited without sufficient contextual guidance. Therefore, individual modules alone cannot provide stable performance improvements under such challenging conditions.

In contrast, the CSA integrated architecture consistently improves performance across all metrics, with the most significant gain in Recall. In construction monitoring, Recall directly reflects the miss rate; therefore, its improvement indicates more complete detection of construction elements, which is critical for safety monitoring and workflow recognition. Moreover, CSA achieves the highest mAP50–95 value (0.924), demonstrating improved bounding-box regression accuracy under complex backgrounds. These improvements arise from the synergistic interaction among modules: the Swin Transformer provides global contextual information to mitigate occlusion effects, CBAM enhances attention to critical regions under global guidance, and AFE strengthens fine-grained feature representation through multi-scale aggregation, thereby enabling more robust feature extraction in complex environments.

At the category level, the CSA model demonstrates clearer advantages for challenging targets. For example, *Worker* detection remains the most challenging task due to small target size, low contrast against the background, and frequent occlusions. Nevertheless, YOLOv11-CSA demonstrates a substantial performance improvement for this category, with mAP50–95 increasing from 0.644 to 0.686 and Recall improving from 0.885 to 0.913. This improvement is particularly important for construction safety monitoring and real-time worker awareness in tunnel environments. For large-scale and structurally complex construction equipment, YOLOv11-CSA consistently achieves higher or more stable detection performance. The mAP50–95 for the *Arch installation trolley* increases from 0.944 to 0.951, while that for the *Drilling jumbo* improves

from 0.954 to 0.961. A particularly notable improvement is observed for the *Wet spray manipulator*, whose mAP50–95 increases by 3.8 percentage points. These results confirm the effectiveness of the enhanced multi-scale feature fusion and attention mechanisms in capturing local structural details, pose variations, and partial occlusions in cluttered tunnel environments. For highly visible targets with stable geometric and appearance characteristics, including *Dumper*, *Concrete mixer*, and *Guardrail*, both YOLOv11 and YOLOv11-CSA achieve near-saturated detection performance. The *Dumper* maintains an mAP50–95 of 0.990 under both models, while the *Concrete mixer* and *Guardrail* remain stable at 0.971 and 0.994, respectively. This indicates that, for such targets, detection performance is primarily constrained by data characteristics rather than network architecture.

In terms of computational efficiency, YOLOv11-CSA achieves an average inference time of approximately 6 ms per image, including 0.1 ms for preprocessing, 5.4 ms for inference, and 0.9 ms for postprocessing. Although this inference time is slightly higher than that of the original YOLOv11 model, it fully satisfies the real-time monitoring requirements of tunnel construction sites, demonstrating a favorable balance between detection accuracy and computational cost.

Overall, the proposed YOLOv11-CSA model significantly improves detection accuracy and robustness in visually degraded tunnel environments while maintaining real-time performance. In particular, it demonstrates superior performance for complex equipment, slender mechanical components, and construction workers, thereby providing a reliable visual perception foundation for subsequent construction workflow recognition and reasoning. Representative detection results under dim tunnel lighting conditions are illustrated in Fig. 7, where many missed and false detections produced by the original YOLOv11 model are effectively mitigated by the proposed YOLOv11-CSA model.

Figure 7: Comparison of construction element recognition performance between the two models.

6.2 Construction Workflow Identification Results

To evaluate the construction workflow identification performance of the proposed method under real engineering conditions, a new drill-and-blast tunnel construction cycle video was selected for testing. This cycle follows a typical operating pattern in which four key secondary workflows, namely steel arch erection, drilling, shotcreting, and mucking, account for approximately 80 percent of the total cycle duration, which

is about 20 to 21 h, and therefore play a decisive role in overall construction progress. These four workflows were accordingly selected as the primary targets for identification and evaluation.

Fig. 8 presents the construction workflow identification results at the drill-and-blast tunnel face for construction workflow monitoring. A clipping video showing the effectiveness of the proposed approach in identifying drill-and-blast workflows at the tunnel face is provided in Supplementary Materials. Given the long duration and strong temporal continuity of the raw construction video, a segmented sampling strategy was adopted. Key operational segments corresponding to each workflow, including the transition phases between adjacent workflows, were extracted and concatenated into an integrated test video for validation, with playback acceleration applied. The editing duration represents the manually extracted clip length for evaluation and does not exactly correspond to the effective workflow duration. The edited durations of the workflow segments were 150.4 s for steel arch erection, 211.9 s for drilling, 231.7 s for shotcreting, and 282.9 s for mucking. Notably, ground truth is defined as the moment when construction elements enter the trestle and initiate the corresponding workflow, as determined through manual monitoring. In contrast, the identified results refer to the moment when the recognition system outputs a detection and infers the corresponding workflow based on the contextual relationship model. With the introduction of a temporal smoothing mechanism, when a key construction target temporarily disappears for several consecutive frames (unless there is a clear trend indicating that the target is leaving the construction area), the system does not immediately trigger a workflow state transition. Instead, it maintains the workflow decision from the previous time step. As a result, the transitional gap intervals (2 s) between tertiary workflows within each construction stage are successfully detected, while the overall workflow progress monitoring remains continuous, without temporal discontinuities. Due to inherent observational bias associated with manual observation and camera viewpoints, the identified results consistently lag behind the ground truth. The duration of this delayed recognition depends on the time required for the construction elements to move from the manual observation reference point (defined as the cross-section where the construction elements leave the trestle bridge and enter the construction zone, roughly corresponding to its first entry into the camera's field of view) until it is fully within the camera's field of view. Accuracy is defined as the ratio between the identified workflow duration and the corresponding ground truth duration.

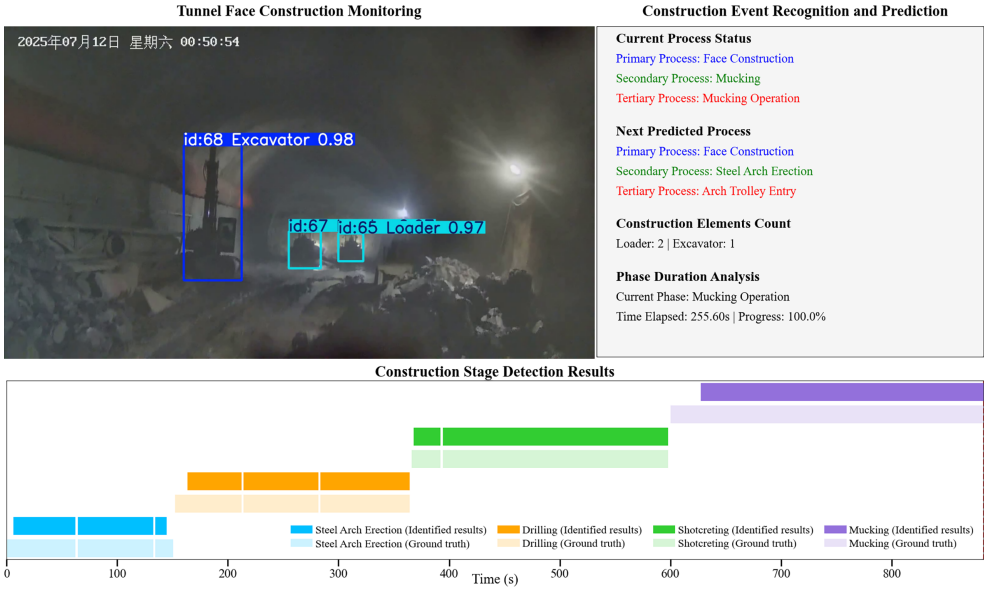

Figure 8: Workflows identification results of drill-and-blast tunnel face construction.

Table 4 compares the automatically identified workflow durations with the corresponding ground truth records. Overall, the identified durations of the four workflows show a high degree of agreement with the actual values. Among them, shotcreting exhibits the smallest identification error of 1.63 s, corresponding to an accuracy of 0.99. The identification errors for drilling and steel arch erection are both controlled within 12 s, with accuracies of 0.95 and 0.92, respectively. The mucking workflow shows relatively larger deviation due to frequent equipment entry and exit, as well as complex phase transitions. Nevertheless, its identification accuracy remains above 0.90. The automatically identified duration for the four workflows is 134.60 s for steel arch erection, 196.77 s for drilling, 228.07 s for shotcreting, and 255.67 s for mucking. When the ratio of identified duration to actual duration is used as the evaluation metric, the recognition accuracies for the four workflows are 92 percent, 95 percent, 99 percent, and 90 percent, respectively. These results demonstrate that the proposed method provides stable and reliable workflow discrimination capability under complex construction conditions. A further comparison between the automatically generated progress curve and the ground truth confirms a high level of consistency between the proposed method and manual judgment in both workflow identification and time estimation.

Table 4: Comparison between identified and ground truth workflow duration.

Workflow	Editing Duration (s)	Workflow Time (s)			Accuracy
		Ground Truth Time (s)	Identified Time (s)	Error (s)	
Steel arch erection	150.40	146.37	134.60	11.77	0.92
Drilling	211.90	207.93	196.77	11.16	0.95
Shotcreting	231.70	229.70	228.07	1.63	0.99
Mucking	282.90	282.93	255.67	27.26	0.90

In practical recognition results, the automatically identified workflow durations are generally slightly shorter than the ground truth duration. This phenomenon can mainly be attributed to one factor. During construction elements entry and exit stages, targets may not be fully within the camera field of view or may exhibit incomplete visual features, which prevent timely detection and workflow confirmation. Taking the steel arch erection workflow as an example, represented by the blue region in Fig. 6, stable identification is triggered only after the arch installation trolley has largely entered the field of view. This results in delayed recognition at the beginning of the workflow. Similarly, at the end of the workflow, as the trolley exits the scene, its visual features gradually weaken and no longer support reliable detection. Consequently, the system terminates the workflow state earlier than the actual completion time, leading to a shortened accumulated duration. For the mucking workflow, shown by the purple region in Fig. 6, the identification deviation is more pronounced. This is primarily because the system requires the simultaneous detection of two loaders and one excavator to confirm the workflow state. In actual operations, the excavator often blocks the gantry exit at the beginning of the workflow, preventing loaders from entering the camera view. Only after the excavator moves away can the loaders enter the working area, and this delay causes the system to satisfy the workflow confirmation conditions significantly later than the actual workflow start.

It should be noted, however, that in real tunnel construction, the effective operation duration of each workflow is typically much longer than the transition periods between workflows, and therefore, the overall workflow identification accuracy is expected to be higher in continuous long-term deployment. In this study, due to the use of manually edited video clips for experimental evaluation, the relative proportion of workflow transition segments is artificially enlarged. As a result, misalignments at the beginning and end of workflows

have a greater impact on the calculated accuracy, which explains why the recognition accuracy of certain workflows is limited to approximately 90%. This issue is primarily attributable to the data preparation strategy rather than intrinsic limitations of the proposed method.

7 Conclusion

This study addresses the problem of intelligent identification of construction workflows at the tunnel face in a drill-and-blast tunnel. A unified workflow recognition framework that integrates visual perception with hierarchical decision reasoning is systematically developed and validated through real engineering applications. The main conclusions are summarized as follows.

(1) A workflow intelligent recognition framework that integrates visual perception with domain knowledge is proposed, enabling structured understanding of the construction workflow. To address the hierarchical complexity and strong temporal characteristics of drill-and-blast tunnel construction workflows, the framework adopts hierarchical modeling to dynamically combine the construction event sample library, workflow transition logic, and visual recognition results. Experimental validation demonstrates that the proposed framework can achieve automatic workflow identification and state evolution reasoning under complex construction conditions, providing a reliable theoretical and structural basis for intelligent tunnel construction monitoring.

(2) A YOLOv11-CSA object detection method tailored to harsh tunnel visual environments is proposed, significantly improving model robustness and detection accuracy. In response to challenges such as low illumination, strong glare, and large variations in object scale within tunnels, this study employs SAM2 to generate high-quality automated annotations and integrates CBAM, AFE, and Swin Transformer modules into YOLOv11 to construct a multi-level feature enhancement mechanism. Experimental results show that the improved model achieves an overall mAP50-95 of 92.4% on the validation set, representing a 1.1 percentage point improvement over the baseline model, while the average detection accuracy across all construction elements exceeds 98%. These results confirm the strong performance of the proposed method under visually degraded conditions.

(3) The engineering applicability and recognition stability of the proposed system are validated in a real tunnel project, meeting the requirements of real-time site monitoring. System validation based on monocular video data collected from an active drill-and-blast tunnel demonstrates that the automatic recognition accuracy for four key workflows—steel arch erection, drilling, shotcreting, and mucking exceeds 90%, with shotcreting achieving an accuracy of 99% and an absolute error of only 1.63 s. In addition, the average single-frame inference time is approximately 6 ms, satisfying real-time deployment requirements and providing a practical and effective technical solution for intelligent monitoring of tunnel construction progress.

A limitation of this study is that validation is based on data from a single tunnel project, without cross-site generalization. This constraint arises from strict safety requirements and the high cost of deploying explosion-proof monitoring systems in drill-and-blast tunnels, where hazardous conditions severely limit data availability and sharing. In addition, extreme occlusion, low illumination, and dust introduce inherent perceptual limits, making it difficult for vision-only approaches to capture fine-grained construction processes. Future work will focus on improving generalization and data diversity by collecting multi-site datasets under varying geological and construction conditions. Moreover, multimodal sensing (e.g., equipment positioning and state sensors) will be integrated to overcome visual blind spots. Efforts will also be directed toward developing standardized benchmark datasets for construction workflow recognition, enabling more rigorous evaluation and comparison of alternative models.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Key Research and Development Program Project [2023YFB2603900].

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Junyong Zhou; methodology, Junyong Zhou and Chuanjiang Chen; validation, Chanjiang Chen, Miaosi Dong and Binbin Du; formal analysis, Chanjiang Chen and Binbin Du; investigation, Chanjiang Chen, Junyong Zhou and Binbin Du; resources, Binbin Du; data curation, Binbin Du and Chuanjiang Chen; writing—original draft preparation, Chuanjiang Chen and Binbin Du; writing—review and editing, Junyong Zhou, Miaosi Dong and Liwen Zhang; supervision, Bitang Zhu; project administration, Junyong Zhou; funding acquisition, Junyong Zhou and Bitang Zhu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, [Junyong Zhou], upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/cmes.2026.081546/s1.

References

1. Turner CJ, Oyekan J, Stergioulas L, Griffin D. Utilizing industry 4.0 on the construction site: challenges and opportunities. IEEE Trans Ind Inform. 2021;17(2):746–56. doi:10.1109/TII.2020.3002197.
2. He B, Armaghani DJ, Lai SH, He X, Asteris PG, Sheng D. A deep dive into tunnel blasting studies between 2000 and 2023—a systematic review. Tunn Undergr Space Technol. 2024;147:105727. doi:10.1016/j.tust.2024.105727.
3. Seo J, Han S, Lee S, Kim H. Computer vision techniques for construction safety and health monitoring. Adv Eng Inform. 2015;29(2):239–51. doi:10.1016/j.aei.2015.02.001.
4. Xu Z, Wang Z, Ren P, Zhang X, Li T. Framework and construction methodology of underground engineering domain knowledge large language model: undergrGPT. Smart Constr. 2024;1(2):1–15. doi:10.55092/sc20240012.
5. Jradi M. Transforming construction: digital twin technology for site monitoring and optimization in Denmark. Smart Constr. 2024;1(3):0015. doi:10.55092/sc20240015.
6. Xu S, Wang J, Shou W, Ngo T, Sadick AM, Wang X. Computer vision techniques in construction: a critical review. Arch Comput Meth Eng. 2021;28(5):3383–97. doi:10.1007/s11831-020-09504-3.
7. Armaghani DJ, Liu Z, Khabbaz H, Fattahi H, Li D, Afrazi M. Tree-based solution frameworks for predicting tunnel boring machine performance using rock mass and material properties. Comput Model Eng Sci. 2024;141(3):2421–51. doi:10.32604/cmes.2024.052210.
8. Cheng J, Wang D, Zheng W, Wang H, Shen Y, Wu M. Position measurement technology of boom-type roadheader based on binocular vision. Meas Sci Technol. 2024;35(2):026301. doi:10.1088/1361-6501/ad0958.
9. Soltani MM, Zhu Z, Hammad A. Framework for location data fusion and pose estimation of excavators using stereo vision. J Comput Civ Eng. 2018;32(6):04018045. doi:10.1061/(asce)cp.1943-5487.0000783.
10. Tang J, Wang M, Luo H, Wong PK, Zhang X, Chen W, et al. Full-body pose estimation for excavators based on data fusion of multiple onboard sensors. Autom Constr. 2023;147(1):104694. doi:10.1016/j.autcon.2022.104694.
11. Vahdatikhaki F, Hammad A, Siddiqui H. Optimization-based excavator pose estimation using real-time location systems. Autom Constr. 2015;56(1):76–92. doi:10.1016/j.autcon.2015.03.006.
12. Cheng CF, Rashidi A, Davenport MA, Anderson DV. Activity analysis of construction equipment using audio signals and support vector machines. Autom Constr. 2017;81(11):240–53. doi:10.1016/j.autcon.2017.06.005.

13. Ahn CR, Lee S, Peña-Mora F. Application of low-cost accelerometers for measuring the operational efficiency of a construction equipment fleet. J Comput Civ Eng. 2015;29(2):04014042. doi:10.1061/(asce)cp.1943-5487.0000337.
14. Akhavian R, Behzadan AH. Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. Adv Eng Inform. 2015;29(4):867–77. doi:10.1016/j.aei.2015.03.001.
15. Vahdatikhaki F, Hammad A. Framework for near real-time simulation of earthmoving projects using location tracking technologies. Autom Constr. 2014;42(4):50–67. doi:10.1016/j.autcon.2014.02.018.
16. Kim J, Chi S, Ahn CR. Hybrid kinematic-visual sensing approach for activity recognition of construction equipment. J Build Eng. 2021;44:102709. doi:10.1016/j.jobe.2021.102709.
17. Sherafat B, Rashidi A, Asgari S. Sound-based multiple-equipment activity recognition using convolutional neural networks. Autom Constr. 2022;135(2):104104. doi:10.1016/j.autcon.2021.104104.
18. Kim J, Chi S, Seo J. Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks. Autom Constr. 2018;87(6):297–308. doi:10.1016/j.autcon.2017.12.016.
19. Rezazadeh Azar E, McCabe B. Part based model and spatial-temporal reasoning to recognize hydraulic excavators in construction images and videos. Autom Constr. 2012;24(4):194–202. doi:10.1016/j.autcon.2012.03.003.
20. Yang J, Shi Z, Wu Z. Vision-based action recognition of construction workers using dense trajectories. Adv Eng Inform. 2016;30(3):327–36. doi:10.1016/j.aei.2016.04.009.
21. Fang Q, Li H, Luo X, Ding L, Rose TM, An W, et al. A deep learning-based method for detecting non-certified work on construction sites. Adv Eng Inform. 2018;35:56–68. doi:10.1016/j.aei.2018.01.001.
22. Chen C, Zhu Z, Hammad A. Automated excavators activity recognition and productivity analysis from construction site surveillance videos. Autom Constr. 2020;110(4):103045. doi:10.1016/j.autcon.2019.103045.
23. Lin ZH, Chen AY, Hsieh SH. Temporal image analytics for abnormal construction activity identification. Autom Constr. 2021;124:103572. doi:10.1016/j.autcon.2021.103572.
24. Zeng T, Wang J, Cui B, Wang X, Wang D, Zhang Y. The equipment detection and localization of large-scale construction jobsite by far-field construction surveillance video based on improving YOLOv3 and grey wolf optimizer improving extreme learning machine. Constr Build Mater. 2021;291(2):123268. doi:10.1016/j.conbuildmat.2021.123268.
25. Chen C, Xiao B, Zhang Y, Zhu Z. Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition. Autom Constr. 2023;146:104702. doi:10.1016/j.autcon.2022.104702.
26. Hua T, Lang B, Sun J, Lin G, Wang J, Li D, et al. Excavator identification method based on computer vision technology in nighttime scenes. Eng Constr Archit Manag. 2025;32(12):1–22. doi:10.1108/ecam-02-2025-0176.
27. Sun Y, Xu X, Tian X, Zhou L, Li Y. Efficient human activity recognition: a deep convolutional transformer-based contrastive self-supervised approach using wearable sensors. Eng Appl Artif Intell. 2024;135(7):108705. doi:10.1016/j.engappai.2024.108705.
28. Núñez-Marcos A, Arganda-Carreras I. Transformer-based fall detection in videos. Eng Appl Artif Intell. 2024;132(5):107937. doi:10.1016/j.engappai.2024.107937.
29. Nabi AU, Shi J, Kamlesh, Jumani AK, Ahmed Bhutto J. Hybrid transformer-EfficientNet model for robust human activity recognition: the BiTransAct approach. IEEE Access. 2024;12:184517–28. doi:10.1109/ACCESS.2024.3506598.
30. Baek J, Ban J, Kim H, Kim D, Choi B. Dual-stream transformer-based activity classification for off-site construction productivity analysis. Autom Constr. 2025;180(24):106505. doi:10.1016/j.autcon.2025.106505.
31. Aidarova S, Nurmakhan T, Myrzakhan R, Fazli S, Yazici A. Advancing activity recognition with multimodal fusion and transformer techniques. IEEE Sens J. 2025;25(11):19632–49. doi:10.1109/JSEN.2025.3555663.
32. Project Management Institute. A guide to the project management body of knowledge (PMBOK® guide). 7th ed. Newtown Square, PA, USA: Project Management Institute; 2021.
33. Chi S, Caldas CH. Image-based safety assessment: automated spatial safety risk identification of earthmoving and surface mining activities. J Constr Eng Manag. 2012;138(3):341–51. doi:10.1061/(asce)co.1943-7862.0000438.
34. Kim J, Chi S. Multi-camera vision-based productivity monitoring of earthmoving operations. Autom Constr. 2020;112:103121. doi:10.1016/j.autcon.2020.103121.

35. Zhang T, Lee YC, Scarpiniti M, Uncini A. A supervised machine learning-based sound identification for construction activity monitoring and performance evaluation. In: Proceedings of the Construction Research Congress 2018; 2018 Apr 2–4; New Orleans, LA, USA. p. 358–66.
36. Roberts D, Golparvar-Fard M. End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level. Autom Constr. 2019;105(12):102811. doi:10.1016/j.autcon.2019.04.006.
37. Lee JG, Hwang J, Chi S, Seo J. Synthetic image dataset development for vision-based construction equipment detection. J Comput Civ Eng. 2022;36(5):04022020. doi:10.1061/(asce)cp.1943-5487.0001035.
38. Sherafat B, Rashidi A, Lee YC, Ahn CR. Automated activity recognition of construction equipment using a data fusion approach. In: Proceedings of the Computing in Civil Engineering 2019; 2019 Jun 17–19; Atlanta, Georgia. p. 1–8.
39. Jeong G, Jung M, Park S, Park M, Ahn CR. Contextual multimodal approach for recognizing concurrent activities of equipment in tunnel construction projects. Autom Constr. 2024;158(2):105195. doi:10.1016/j.autcon.2023.105195.
40. Jung S, Jeoung J, Lee DE, Jang H, Hong T. Visual-auditory learning network for construction equipment action detection. Comput Aided Civ Infrastruct Eng. 2023;38(14):1916–34. doi:10.1111/mice.12983.
41. Ravi N, Gabeur V, Hu YT, Hu R, Ryali C, Ma T, et al. SAM 2: segment anything in images and videos. arXiv:2408.00714. 2024.
42. Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements. arXiv:2410.17725. 2024.
43. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Proceedings of the Computer Vision—ECCV 2018; 2018 Sep 8–14; Munich, Germany. p. 3–19.
44. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 9992–10002.
45. Ali M, Javaid M, Noman M, Fiaz M, Khan S. FANet: feature amplification network for semantic segmentation in cluttered background. arXiv:2407.09379. 2024.