An Ontology-Guided Attribute-Based Model for Automated Construction Hazard Description

Wen-der Yu[1*], Wen-ta Hsiao[2] and Tao-ming Cheng[3]

[1]Department of Civil and Construction Engineering, Chaoyang University of Technology, Taiwan.
[2] Department of Civil and Construction Engineering, Chaoyang University of Technology, Taiwan.
[3] Department of Civil and Construction Engineering, Chaoyang University of Technology, Taiwan.
*wenderyu@cyut.edu.tw, wdshiau@cyut.edu.tw; tmcheng@cyut.edu.tw

Abstract–

Automated understanding of complex construction hazard scenarios is essential for intelligent site safety management. However, existing computer vision and image captioning approaches often lack interpretability and semantic consistency required for engineering decision support. This study proposes a Hazard Ontology-Guided Attribute-Based Model (HOGAM) for automated construction hazard description. The framework decouples visual perception from language generation by first predicting six hazard attributes—worker, location, activity, medium, condition, and hazard type—using Inception v3–based transfer learning models. A construction safety ontology and a template-driven generation mechanism are then employed to produce structured and interpretable hazard descriptions. Experiments conducted on 2,660 construction hazard images demonstrate an average attribute prediction accuracy of 92.29%, while the generated descriptions achieve a BLEU score of 0.93. The results confirm that HOGAM provides an interpretable and engineering-usable pathway from visual hazard recognition to automated safety assessment and decision support.

Keyword –

Construction safety; Hazard ontology; Hazard description; Computer vision; Image captioning

1 Introduction

The construction industry remains one of the most hazardous sectors due to its open, dynamic, and resource-intensive work environments, where workers and equipment interact continuously under changing site conditions. Despite decades of safety management efforts, construction still contributes a disproportionately high share of occupational injuries and fatalities worldwide. Reducing these incidents requires not only identifying hazardous factors, but also understanding how multiple factors jointly form risky scenarios and what such scenarios imply for safety decisions [1–3].

In occupational health and safety management, a hazard refers to a potential source of harm that may lead to injury or adverse health outcomes [4]. Accordingly, effective accident prevention depends on timely hazard recognition and the ability to translate recognized hazards into actionable mitigation measures. However, prior studies have repeatedly shown that workers may fail to notice hazards, misinterpret their severity, or underestimate the consequences under time pressure, cognitive workload, or limited experience. This gap between "what exists on site" and "what is understood as risky" is a critical reason why hazard recognition and risk perception remain persistent challenges in practice [5–7].

With the rapid development of computer vision (CV), vision-based site monitoring has become a major direction for automated safety management. Existing studies have demonstrated promising results in personal protective equipment (PPE) compliance checking, worker pose or behavior monitoring, equipment activity tracking, and environment inspection [8–10]. These approaches typically output class labels, bounding boxes, or tracked trajectories, which are useful for detection and surveillance. Nevertheless, construction hazards in real environments are rarely defined by a single object or an isolated action. Instead, hazards often emerge from the interaction among multiple entities (e.g., trade type, location, activity, equipment, and temporary conditions), which must be interpreted in context to support safety management decisions. As a result, even accurate object or action detection may still be insufficient to answer the decision-oriented question: What safety risk does this situation represent and how should it be handled [11]?

Recently, image captioning and vision–language models have been introduced to generate natural-language descriptions of construction scenes. Such approaches can provide intuitive textual summaries and support semantic indexing and retrieval of site imagery [12–14]. However, most captioning methods rely on free-form generation and are optimized primarily for

linguistic similarity metrics. In construction safety, a useful hazard description must satisfy stricter requirements: it should be interpretable, semantically consistent with domain knowledge, and aligned with safety logic and regulations. When multiple hazard attributes and complex interactions are involved, free-form captions can become incomplete, ambiguous, or logically inconsistent, which limits their reliability for automated safety assessment and decision support [11,14].

Meanwhile, construction safety ontologies have been developed to formalize domain concepts and relationships, enabling structured representation and reasoning for risk identification and mitigation planning [15]. Although ontology-based approaches improve knowledge consistency, many existing studies treat ontologies as static repositories and do not sufficiently integrate them with real-time visual recognition outputs. Consequently, a practical gap remains between (i) vision-based recognition results and (ii) engineering-usable hazard descriptions that can directly support safety decision-making in dynamic site environments [11,15].

To bridge this gap, this study proposes a Hazard Ontology-Guided Attribute-Based Model (HOGAM) for automated construction hazard description. Instead of generating free-form captions end-to-end, HOGAM explicitly decouples visual perception from language generation and introduces an attribute-based intermediate representation. First, Inception v3–based transfer learning models are trained to predict six predefined hazard attributes—worker (W), location (L), activity (A), medium (M), condition (C), and hazard type (H)—from site images. Then, a construction safety ontology is used to constrain attribute combinations and enforce safety-consistent semantics. Finally, a sentence-template mechanism instantiates the predicted attributes into structured hazard descriptions, producing interpretable outputs that can be directly linked to subsequent safety assessment and preventive planning. Compared with existing studies, the novelty of this work lies in three aspects: (1) unlike end-to-end image captioning approaches that directly map visual inputs to free-form text, this study introduces an attribute-based intermediate representation that explicitly decouples visual perception from language generation; (2) while prior ontology-based studies primarily serve as static knowledge repositories, the proposed framework operationalizes ontology as both a semantic schema and a constraint mechanism to guide structured hazard description; (3) by integrating attribute-level prediction, ontology-guided reasoning, and template-based generation into a unified pipeline, HOGAM provides a controllable and interpretable approach for automated hazard understanding that is directly aligned with engineering decision-making requirements. Through this design, HOGAM provides a clear and automation-ready pathway from raw images to semantically consistent, engineering-usable hazard descriptions for construction safety management.

2 Related Work

This section reviews existing studies relevant to automated construction hazard understanding and description. The discussion is organized into three research streams: (1) computer vision for complex construction scene understanding, (2) image captioning for construction hazard description, and (3) construction hazard knowledge ontology. The limitations identified in these streams motivate the proposed HOGAM framework.

2.1 Computer Vision for Complex Construction Scene Understanding

Computer vision (CV) techniques have been extensively applied to construction safety monitoring, aiming to automatically identify hazardous elements and unsafe behaviors from site imagery. Existing studies primarily focus on tasks such as personal protective equipment (PPE) detection, worker pose estimation, equipment monitoring, and activity recognition [6–9]. These approaches have demonstrated strong performance in detecting isolated objects or predefined actions and have significantly reduced the manual workload associated with safety inspections.

However, construction hazards are rarely defined by single visual cues. Instead, they often arise from the interaction of multiple entities, including worker roles, spatial locations, work activities, equipment conditions, and temporary site configurations. To address such complexity, recent research has explored complex scene understanding (CSU), which integrates object detection, action recognition, and contextual reasoning to capture richer semantic information from images or videos [10,11].

Despite progress in general vision research, most construction-oriented CV methods still output low-level results such as class labels, bounding boxes, or trajectories. These outputs lack explicit semantic structure and are difficult to directly interpret in terms of safety risk or engineering implications. As a result, even accurate visual recognition does not necessarily translate into actionable hazard understanding, highlighting a critical gap between perception and decision support in construction safety applications [11].

2.2 Image Captioning for Construction Hazard Description

Image captioning combines visual perception with natural language generation to produce textual

descriptions of image content. In construction engineering and management, image captioning has been applied to describe site activities, construction resources, and general work conditions, supporting image retrieval, documentation, and progress monitoring [12–14].

Several studies have attempted to extend image captioning to construction safety contexts, generating natural-language descriptions of hazardous scenes using CNN-based visual encoders and recurrent or attention-based language models [13,14]. While these approaches offer intuitive textual outputs, their evaluation is typically driven by linguistic similarity metrics such as BLEU or CIDEr.

However, image captioning methods are fundamentally based on free-form language generation. In safety-critical domains, such descriptions may lack semantic consistency, omit key hazard attributes, or violate safety logic, particularly when multiple interacting risk factors are present. Consequently, free-form captions may be difficult to interpret, verify, or directly use for safety assessment and preventive planning. These limitations restrict the practical applicability of end-to-end captioning models in automated construction safety management systems [11,14].

2.3 Construction Hazard Knowledge Ontology

Knowledge ontologies provide formal representations of domain concepts, relationships, and rules, and have been widely adopted to support construction safety analysis and risk management. Existing ontology-based studies often integrate safety knowledge with Building Information Modeling (BIM) to enable structured risk identification, reasoning, and visualization [15].

Prior research has demonstrated the potential of safety ontologies for organizing hazard-related knowledge, linking construction activities to risk factors, and supporting rule-based safety checks. However, many ontology-driven approaches remain focused on design-stage risk assessment or static knowledge repositories. Their integration with real-time visual perception is limited, and they are rarely used to directly guide automated hazard description from site imagery.

As a result, current ontology-based methods are insufficient to bridge the gap between vision-based recognition outputs and engineering-usable hazard descriptions in dynamic construction environments. This limitation underscores the need for a framework that tightly couples visual perception, structured hazard representation, and ontology-guided semantic constraints.

2.4 Research Gap

In summary, existing CV-based approaches excel at detecting objects and actions but lack semantic depth for hazard interpretation; image captioning methods provide natural-language descriptions but suffer from ambiguity and limited safety consistency; and ontology-based studies offer structured knowledge representations but are weakly integrated with real-time visual recognition. These gaps motivate the proposed HOGAM framework, which integrates attribute-based visual inference with ontology-guided semantic constraints to generate interpretable and engineering-usable construction hazard descriptions. This study addresses these limitations by introducing an attribute-based, ontology-guided framework that bridges the gap between perception and decision-oriented hazard representation.

3 Hazard Ontology-Guided Attribute-Based Model (HOGAM)

This section presents the proposed Hazard Ontology-Guided Attribute-Based Model (HOGAM) for automated construction hazard understanding and description. The overall design rationale is grounded in the hazard medium classification and ontology structure (Figure 1–3), while the end-to-end information flow of HOGAM is summarized in Figure 4. Unlike end-to-end image captioning models, HOGAM explicitly decouples visual perception from language generation by introducing an attribute-based intermediate representation guided by construction hazard ontology.

3.1 Construction Hazard Knowledge Ontology

The construction hazard knowledge ontology provides a formalized representation of hazard-related concepts, entities, and their semantic relationships in construction safety management. In this study, the ontology is designed to support hazard description rather than risk quantification, with an emphasis on describing hazardous situations encountered during construction operations.

Following the hazard medium classification and safety reporting logic, the ontology organizes hazard knowledge around five core semantic elements: worker (W), location (L), activity (A), hazardous medium (M), and condition (C), which jointly determine the hazard type (H). The overall ontology structure is illustrated in Figure 1, enabling explicit representation of “who is performing what activity, where, under which conditions, and why it constitutes a hazard.” The ontology further serves as a semantic constraint layer to ensure that generated hazard descriptions remain logically consistent with construction safety knowledge. Specifically, the

ontology in this study plays a dual role: (1) it defines the semantic schema of hazard representation by structuring the relationships among the six attributes (W–L–A–M–C–H); (2) it acts as a constraint mechanism applied after attribute prediction, where incompatible or logically inconsistent attribute combinations are filtered or corrected before description generation. For example, certain hazard types are only valid under specific combinations of activity and condition; if an inconsistent combination is predicted, the ontology-guided rules prevent it from being directly instantiated into the final description. This mechanism ensures that the generated hazard descriptions remain semantically valid and aligned with construction safety knowledge, rather than relying solely on unconstrained predictions.

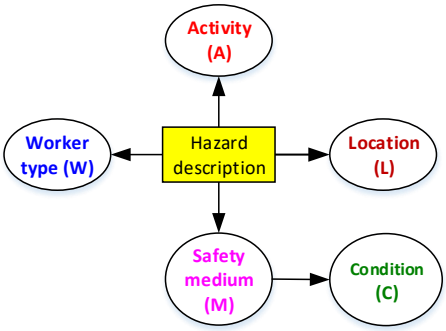

Figure 1. Construction hazard ontology defined in this study, illustrating the semantic relationships among worker (W), location (L), activity (A), medium (M), condition (C), and hazard type (H).

3.2 Attribute-Based Semantic Representation

To bridge the gap between visual recognition outputs and semantic hazard description, HOGAM adopts an attribute-based intermediate representation. Instead of directly generating free-form sentences from images, the proposed framework first converts visual information into a structured attribute set {W, L, A, M, C, H}, consistent with the ontology defined in Figure 2.

This representation corresponds to the annotation logic of the Construction Hazard Description System (CHDS) developed in this study and provides explicit semantic roles for each recognized element. By enforcing a fixed semantic structure, the attribute-based representation enhances interpretability and prevents ambiguity commonly observed in end-to-end captioning approaches.

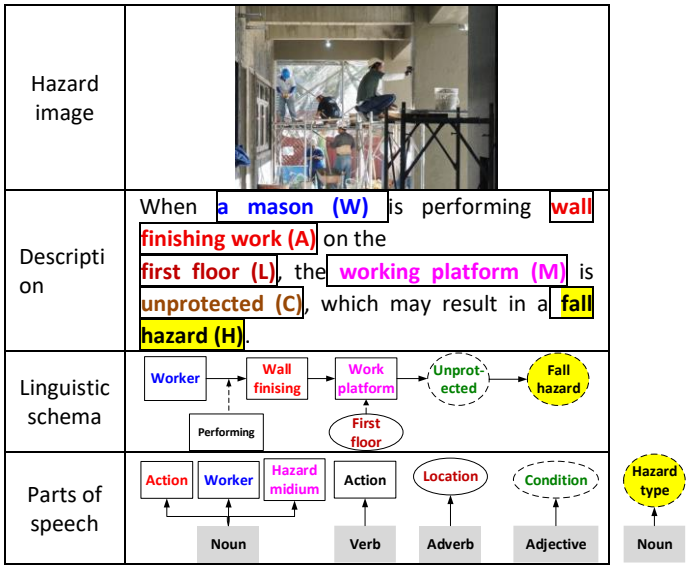

Figure 2. Example of a semantic description schema for construction hazard images, illustrating how visual information is structured according to the hazard ontology.

3.3 Template-Driven Hazard Description Generation

Based on the attribute-based representation, hazard descriptions are generated using predefined sentence templates (Figure 3). Each template encodes a standardized grammatical structure aligned with construction safety reporting practices, ensuring that the six key attributes are explicitly expressed in the final description. Attribute instantiation is then performed by filling the predicted attribute values into the corresponding template slots.

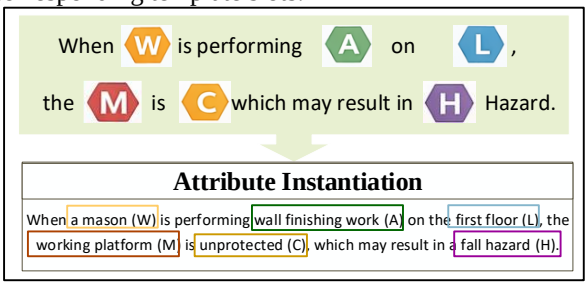

Figure 3. Sentence template for hazard description generation, defining a structured W–L–A–M–C–H instantiation mechanism.

This template-driven mechanism ensures that generated descriptions are complete, interpretable, and semantically aligned with construction safety logic. Unlike free-form language generation, the proposed approach reduces missing critical hazard information and avoids linguistically plausible but safety-inconsistent descriptions.

3.4 Construction Hazard Image Description System and Dataset

To support model training and evaluation, a Construction Hazard Description System (CHDS) was developed to construct a structured hazard image dataset. Using CHDS, hazard images were annotated according to the predefined attribute schema and sentence templates.

The resulting dataset contains 2,660 construction hazard images collected from 18 building construction sites, each paired with a structured hazard description. In terms of hazard-type distribution, fall-related hazards (H01) constitute the largest proportion of the dataset (47.44%), followed by falling-object hazards (H04, 26.62%), slip/trip hazards (H02, 15.60%), and collapse-related hazards (H05, 9.21%). A smaller portion corresponds to electrical hazards (H13, 1.13%). This distribution reflects the prevalence of major construction safety risks while maintaining coverage of less frequent but critical hazard types. Representative examples of the dataset are presented in Table 1.

Table 1. Examples of the construction hazard image description dataset

Hazard image	Hazard description
	When a **formwork worker** is performing **formwork assembly** on the **upper floor under construction**, the use of an **unqualified ladder or stepladder** may result in a **fall or rolling hazard**.
	When a **rebar worker** is performing **rebar assembly work** in the **basement level**, the use of an **unqualified ladder or stepladder** may result in a **fall or rolling hazard**.

3.5 Deep Learning–Based Visual Attribute Recognition

Visual attribute recognition is performed using transfer learning based on the Inception v3 architecture. To avoid semantic interference caused by multi-task learning, HOGAM adopts a single-task training strategy. Six independent Inception v3 models are trained in parallel to predict the six hazard attributes: W, L, A, M, C, and H.

3.6 Integrated HOGAM Framework

The complete HOGAM framework integrates visual perception, attribute-based representation, ontology-guided semantic constraints, and template-driven language generation into a unified pipeline, as shown in Figure 4.

After attribute prediction, an ontology-guided constraint step is applied to validate and refine attribute combinations before template-based description generation.

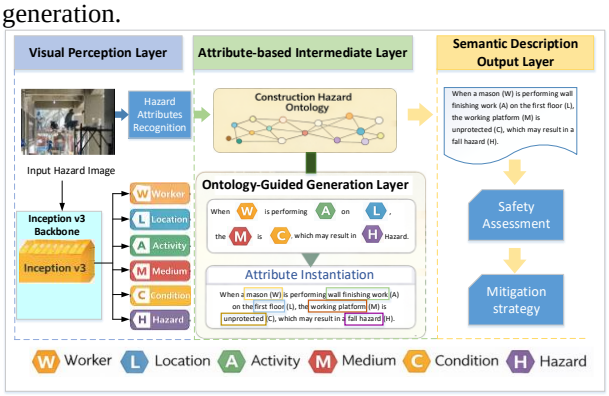

Figure 4. Conceptual framework of the proposed Hazard Ontology-Guided Attribute-Based Model (HOGAM), illustrating the transformation from visual perception to ontology-guided, engineering-usable hazard descriptions.

By decomposing hazard understanding into manageable sub-tasks, HOGAM ensures interpretability, semantic consistency, and engineering usability. Compared with end-to-end image captioning models, the proposed framework provides a clearer information flow from image perception to safety-oriented decision support, effectively bridging the gap between visual recognition and construction safety management.

4 Results and Discussion

This section evaluates the performance of the proposed HOGAM framework from two complementary perspectives: (1) hazard attribute recognition performance, and (2) quality of ontology-guided hazard description generation. The results are discussed with emphasis on interpretability, semantic consistency, and engineering usability rather than isolated metric optimization. To ensure clarity, it should be noted that validation performance is reported only for monitoring training convergence, while all quantitative results presented in Tables 2 and 3 are based on the held-out test set.

4.1 Performance of Hazard Attribute Recognition

To evaluate the visual perception component of HOGAM, six independent Inception v3 models were fine-tuned to predict the six predefined hazard attributes: worker (W), location (L), activity (A), medium (M), condition (C), and hazard type (H). A total of 2,162

images were used for training, while 498 images were reserved for testing.

The experimental results are shown in Table 2, which indicate that the proposed single-task training strategy yields stable and consistent performance across all attributes. The average classification accuracy across six attributes reached 0.919, with an average F1-score of 0.916. During training, the validation accuracy monitored in MATLAB reached 92.29%, indicating good convergence behavior. The final model performance is evaluated on the independent test set, as reported in Table 2.

Table 2. Visual Recognition Performance of Inception v3

Hazard attribute	Metrics			
	Accuracy	Precision	Recall	F1
Worker (W)	0.9016	0.8957	0.933	0.9098
Location (L)	0.9675	0.9018	0.986	0.9355
Activity (A)	0.9494	0.9062	0.9831	0.9398
Medium (M)	0.9063	0.9208	0.9405	0.9231
Condition (C)	0.8896	0.8794	0.8925	0.8831
Hazard type (H)	0. 9237	0. 9124	0.9012	0.9067
Average	0.9229	0.9008	0.9394	0.9163

Among the six attributes, location (L) and activity (A) achieved the highest recognition performance, with F1-scores of 0.936 and 0.940, respectively. This suggests that spatial context and activity patterns provide salient visual cues that are relatively robust to variations in viewpoint and background clutter. In contrast, condition (C) exhibited comparatively lower performance (F1-score = 0.883), which can be attributed to subtle and fine-grained visual differences and frequent occlusion in real construction environments. Specifically, condition-related attributes often depend on fine-grained visual cues that are less salient than object or activity features. For example, distinguishing whether a working platform is “protected” or “unprotected” may rely on small structural details such as guardrails, which can be partially occluded by workers or materials. Similarly, identifying an “improper ladder placement” condition may be challenging when the ladder is only partially visible or viewed from a suboptimal angle. These characteristics increase the ambiguity of condition recognition and contribute to the relatively lower performance compared to other attributes. Nevertheless, even for this more challenging attribute, the model maintained an accuracy close to 0.89, which is considered acceptable for practical deployment.

These results demonstrate that decomposing hazard understanding into attribute-specific recognition tasks effectively avoids performance degradation commonly observed in multi-label or end-to-end models, and provides a reliable semantic foundation for subsequent hazard description generation.

4.2 Quality of Hazard Description Generation

Following attribute recognition, the predicted hazard attributes were instantiated into structured natural-language descriptions using the predefined sentence templates and ontology constraints. The quality of generated descriptions was evaluated on the same 498 test images by comparing system-generated sentences with human-annotated references.

Standard image captioning evaluation metrics were adopted, including BLEU, ROUGE-L, METEOR, CIDEr, and SPICE. The testing results are shown in Table 3, which demonstrates that the proposed approach achieved a BLEU score of 0.931, ROUGE-L of 0.956, METEOR of 0.956, CIDEr of 0.895, and SPICE of 0.937.

Table 3. Performance of Hazard Description Generation

Metric	BLEU	ROUGE_L	METEOR	CIDEr	SPICE
Score	0.9306	0.9557	0.9558	0.8948	0.9369

The high ROUGE-L and METEOR scores indicate that the generated descriptions consistently preserved key semantic components, while the SPICE score confirms that core relational semantics among hazard attributes were correctly captured. The template-based mechanism ensures that all critical hazard elements are explicitly included.

It should be noted that the objective of this study is not to outperform state-of-the-art captioning models in purely linguistic metrics, but rather to produce semantically consistent, interpretable, and engineering-usable hazard descriptions. From this perspective, the achieved results demonstrate that the proposed ontology-guided generation mechanism effectively bridges visual recognition and safety-oriented textual representation.

5 Conclusion

This study presented HOGAM, an ontology-guided attribute-based framework for automated construction hazard description. By decoupling visual perception from language generation, the proposed approach enables interpretable and semantically consistent hazard understanding.

Experimental results demonstrate that the framework achieves reliable attribute-level recognition and generates structured hazard descriptions aligned with construction safety logic. Compared with end-to-end captioning approaches, HOGAM improves

interpretability and provides a more controllable pathway from visual perception to engineering-usable hazard representation.

The proposed framework offers a scalable solution for integrating automated hazard understanding into construction safety management workflows.

Acknowledgements

This research project was funded by the National Science and Technology Council, Taiwan, under project No. MOST 111-2221-E-324 -011 -MY3. Sincere appreciations are given to the sponsor by the authors.

References

[1] X. Xia, P. Xiang, S. Khanmohammadi, T. Gao, and M. Arashpour. Predicting safety accident costs in construction projects using ensemble data-driven models. *Journal of Construction Engineering and Management*, 150(7):04024054, 2024.

[2] International Labour Organization. *Safety and Health in the Construction Sector—Overcoming the Challenges*. ILO, Geneva, Switzerland, 2014.

[3] Eurostat. Accidents at work statistics. On-line: http://ec.europa.eu/eurostat/statistics-explained/index.php/Accidents_at_work_statistics, Accessed: 08/06/2024.

[4] International Organization for Standardization. *Occupational Health and Safety Management Systems (ISO 45001:2018)*. ISO, Geneva, Switzerland, 2018.

[5] P. Mitropoulos, T. S. Abdelhamid, and G. A. Howell. Systems model of construction accident causation. *Journal of Construction Engineering and Management*, 131(7):816–825, 2005.

[6] A. Perlman, R. Sacks, and R. Barak. Hazard recognition and risk perception in construction. *Safety Science*, 64:13–21, 2014.

[7] W.-T. Hsiao, W.-D. Yu, and C.-C. Shih. Proactive construction hazard prevention model using machine learning. *Journal of Construction Engineering and Management*, 2025.

[8] S. Kim, S. H. Hong, H. Kim, M. Lee, and S. Hwang. Small object detection system for comprehensive construction site safety monitoring. *Automation in Construction*, 156:105103, 2023.

[9] J. Zhou, W. Zhou, and Y. Wang. CWPR: An optimized transformer-based model for construction worker pose estimation on construction robots. *Advanced Engineering Informatics*, 62:102894, 2024.

[10] B. Zhong, L. Shen, X. Pan, and L. Lei. Visual attention framework for identifying semantic information from construction monitoring video. *Safety Science*, 163:106122, 2023.

[11] S. Chen and K. Demachi. Towards on-site hazard identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Automation in Construction*, 125:103619, 2021.

[12] H. Liu, G. Wang, T. Huang, P. He, M. Skitmore, and X. Luo. Manifesting construction activity scenes via image captioning. *Automation in Construction*, 119:103334, 2020.

[13] B. Xiao, Y. Wang, and S. C. Kang. Deep learning image captioning in construction management: A feasibility study. *Journal of Construction Engineering and Management*, 148(7):04022049, 2022.

[14] W.-T. Hsiao, W.-D. Yu, T.-M. Cheng, and A. Bulgakov. Preliminary study on image captioning for construction hazards. In *Proceedings of the 5th International Conference on Architecture, Construction, Environment, and Hydraulics*, pages 1–4, Taichung, Taiwan, 2023.

[15] R. Neches, R. E. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator , and W. R. Swartout. Enabling technology for knowledge sharing. *AI Magazine*, 12(3):36–45, 1991.