Automation in Construction 142 (2022) 104535

Contents lists available at ScienceDirect

Automation in Construction

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

Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge

Lite Zhang [a], Junjie Wang [a,*], Yanbo Wang [b], Hai Sun [a], Xuebing Zhao [a]

[a] College of Civil Engineering, Ocean University of China, Qingdao 266000, China
[b] College of Civil Engineering, Tongji University, Shanghai 200092, China

ARTICLE INFO

Keywords:
Scene graph
BERT
Hazards
Natural language processing
Safety

ABSTRACT

Hazards often arise from interactions of different parties and can cost great loss. Proactively identification can prevent it from happening. Computer vision currently can identify the entities and attributes inside the construction scenes but fail to generate interaction-level scene descriptions and integrate them with domain knowledge for hazards inference. This paper proposed an automatic hazard inference method using construction scene graphs and C-BERT network. First, computer vision was utilized to build the construction scene graphs with interaction-level scene descriptions including entities, attributes, and their interactions. Second, C-BERT network was designed to infer hazards by integrating scene graphs with domain knowledge like construction regulations. 5 different working scenes were used to demonstrate the validity of the proposed approach and reached 97.82% of hazard identification accuracy. It provided an efficient method for combining visual information and domain knowledge for automated safety monitoring, and path for the industry's massive multimodal information fusion.

1. Introduction

According to the U.S. Occupational Safety and Health Administration, more than 6000 fatal injuries are reported worldwide each year as related to construction projects and account for nearly 20% of fatalities in the United States, which make construction one of the most dangerous industries. As for China, there were 773 housing and municipal safety accidents causing 904 fatalities in 2019 [1]. 65% of these accidents are from falling from height and struck-by which often originate from non-compliant interactions between workers and other construction components. In addition to the great loss of life and the personnel impact of injuries, the financial burden is also significant. To pervert such losses, there is a need for systematic monitoring of construction scenes to eliminate hazards by the analyzation of construction scenes.

In practical applications, the safety monitoring of construction scenes has gone through three stages: side station monitoring, sensor-based monitoring and computer vision-based monitoring [2]. Side station monitoring is a common approach for hazards prevention. During monitoring, experienced experts are needed which are expensive, time-consuming and error-prone. Sensor-based methods utilize Radio Frequency Identification (RFID) [3], Global Positioning System GPS [4] and Ultra-Wide Band (UWB) [5] to collect data from construction scenes for hazard identification. Despite more accurate results than computer vision methods, they need to attach additional sensors to workers and other construction components thus being different for large-scale deployment. Computer vision-based methods are more practical for safety monitoring because they analyze data from existing image streams and require no other attached sensors or human labor.

Computer vision deployed on construction sites monitors safety through the detection and tracking of different construction components or the action analyzation such as PPE detection, worker and heavy machinery tracking [2]. Despite the wide application of computer vision methods to assist safety management by the analyzation of construction scenes, there are still some limitations. The detection of entities cannot represent the semantic information embedded in the construction scenes thoroughly. According to the trace intersecting theory, the occurrence of accidents arises from both the human unsafe behavior and the machinery's unsafe state in the same time and space, thus the hazards inference needs not just the recognition of construction components but also reasoning about construction scenes based on the semantic relationships between different components. The detection of semantic relationships helps to format unstructured image information into structured text

* Corresponding author.
E-mail address: wjj@ouc.edu.cn (J. Wang).

https://doi.org/10.1016/j.autcon.2022.104535
Received 16 January 2022; Received in revised form 11 August 2022; Accepted 18 August 2022
Available online 26 August 2022

containing high-level semantic understanding which allows the integration with external knowledge for hazard identification.

Although the high-level semantic understanding of construction scenes has been extracted, the computer vision system cannot understand the hazard meaning contained in the complex scene information. The complex scene information that contains multiple objects and their semantic relationships should be evaluated by the domain knowledge such as construction regulations, regulations, and standards. For hazard identification, the semantic gap between the knowledge-intensive construction regulations and the high-level scene understanding needs to be addressed. The automated hazard identification needs to distinguish the hazard information inherent in the intersection of segmentation scene information with construction regulations. Previous studies focus on adopting ontology in the construction domain and combing with additional inference software for the link between scene information and regulations. The construction regulations are first formatted into structural representation for knowledge extraction. The knowledge extracted from regulations should be further transferred into queries according to the inference software. The structural formatting, knowledge extraction and query generation progress of extensive construction regulations all consume a lot of experts' labor. Also when facing changes of the regulations, the domain ontology and queries need to be re-established.

Bert, a semantic tool, can be used to provide safety domain knowledge to support hazard identification. The Bert network can learn the dependencies and extract the association from the scene information generated by the computer vision algorithms and safety-related construction regulations. Bert technique provides a method for compressing the domain knowledge into the computer vision system and filling the semantic gap between scene information and regulations by dependency learning. Compared with the combination of ontology tools and additional inference software, Bert-based hazard inference saves the experts' labor to establish the domain ontology from extensive construction regulations. Also, the Bert-based hazard inference can be expanded to adapt to the changes in line with changing of regulations. In addition, the Bert-based method can be embedded into the computer vision system which saves the effort and resources for adaptation of additional inference software.

To realize safety monitoring through the analyzation of construction scenes, this paper aims at proposing a safety monitoring method for automated hazard identification for construction scenes. This requires the method can understand construction scenes based on the inherent interactions and inference hazards based on external knowledge such as specifications. In this paper, a construction scene graph module is adopted for construction scene understanding. The construction scene graph module consists of a Transformer based interaction detection network on top of the entities detection to convert unstructured scene information into graph-structured information which contains the entities and their interactions. After that, a Construction BERT-based hazards inference module named C-BERT identifies hazards in the construction scene through text classification of the concatenation of the detected construction scene graph and domain knowledge. Our main contributions are: (1) This paper explored the hazards inference process based on the complex semantic interaction relationship between different construction components since the hazards often arise from the interactions on construction sites. It can provide deeper insights by transferring the construction scenes into structural representations which contain valuable interaction information. Other than the spatial relationship which can be calculated by the coordinate information, the semantic interactions are generated by the novel Transformer-based interaction detection network based on the high-level scene features. (2) The hazards inference is often challenging as it requires significant domain knowledge encompassed in the system to analyze the construction scene information generated from the computer vision algorithms. Upon this time, the hazards inference on construction most relies on additional inference software like Neo 4j database or Protégé. This requires the ontology tools to format the construction regulations into structural representation and extract the restricted objects and restrictions. Then the restrictions are further encoded into queries or SWRL rules. With the complexity of interactions, the queries and rules will become complex and incomprehensible. The reliance on additional inference software, the alignment of different information formats between the computer vision algorithms and inference software, and the transformation of construction regulations by the domain ontology for query generation are all time-consuming, knowledge-intensive, and inefficient. This paper explored the utilization of BERT to compress domain knowledge into the construction scene information from computer vision algorithms for hazards inference. The C-BERT network performs hazards inference by mining hazards information from the concatenation of scene information and construction regulation which saves the effort for ontology development, data format alignment, and additional inference software. Also, the C-BERT network can be conveniently embedded into the computer vision algorithms and enhance the generality and automation of the construction safety management process.

2. Related work

Recent years, computer vision-based methods have been deployed widely on construction sites for hazard identification and they can be typically categorized into 3 types: entity detection, location limitation and action identification as shown in Table 1.

Entity detection methods assess hazards or non-compliance instances by utilizing computer vision algorithms to identify demanded construction components. This involves simple computer vision tasks (e.g., object detection, semantic segmentation), and the commonly used computer vision algorithms are YOLO, Faster RCNN, and Mask RCNN [6,7]. YOLO, as the representation of the one-stage method, provides end-to-end detection with only one CNN operation. YOLO can achieve real-time object detection with low computational power requirements thus being convenient for edge-side computing device deployment. The different variants of the YOLO have been widely deployed on construction sites. For example, Son et al. [8] adopted the fourth edition of YOLO with the Siamese network to identify workers and track them. Kim et al. [9] utilized the third edition of YOLO to detect workers, excavators, and wheel loaders from videos captured from UAVs. The detection results combined with the image rectification method can measure the actual distance of the objects and alert the struck-by accidents. Also, because of the low computation requirement, the modified YOLO algorithms can be embedded in the UAVs for safety monitoring. Semantic segmentation is splitting and clustering the pixels of different objects in a single construction scene. Compared with object detection with bounding boxes, the segmentation can generate more accurate classification results and enable computer vision-based safety management with better ability to analyze the spatial relationship such as within, overlap and away when the objects are in different shapes. For example, Fang et al. [10] adopted Mask R-CNN network to analyze the construction scenes to avoid falls from height. Due to the complex shapes of the structural support, Fang et al. generate both bounding boxes and semantic segmentation masks for the support and worker and then check the spatial relationship for hazard identification. Semantic segmentation can deal with complex shapes but requires huge computational power compared with bounding box generation.

Location limitation methods rely on the location detection of construction components to ensure they are in the right place and maintain a safe distance from potential hazards. This involves the tracking of workers [11], equipment [8], and other onsite components on construction sites [12] to ensure sufficient safety measures are in place and workers are maintaining the required safe distance from the potential hazards. On top of object detection, tracking is required to introduce an additional tracking method that infers new positions for the detected objects in the next frame. The tracking is divided into three types by Park et al. [13] by the way the methods present target objects for tracking:

Table 1
Samples from current vision-based method.

Methods	Sample Image	Schematic of the Process		Objectives
Entity detection		Visual information	Object	Detection of PPE, guardrails and other construction components
		Text information	Label of object's name	
Location limitation		Visual information	Object	Tracking the location of workers and heavy machinery
		Text information	Label of object's name & its location	
Activity recognition		Visual information	Activity	Detection of entities' posture and motions
		Text information	Expression of construction activity	

point-based, counter-based, and kernel-based methods. In addition, the object detection and tracking algorithms are based on 2D pixels locations, which are inefficient for depth information. There has emerged 3D vision for worker tracking. Lee et al. [14] utilized stereo cameras and fused visions to build 3D environment. By the combination of object detection and tracking method, Lee et al. realized 3D tracking for multiple onsite workers.

Action recognition has been used to identify the action or tasks that workers or machinery are engaged in, by the recognition of the motions, body movements, and position, thus detecting unsafe actions such as improper ladder-climbing [15]. The action of workers or heavy machinery exists in the sequence scenes. The action recognition should analyze the sequential information and operation cycles. For example, Kim et al. [16] encode the sequential patterns of visual features and operation cycles in the action recognition system. By the combination of convolutional neural network and double layer long short-term memory network, the proposed method can perform excavator's action recognition based on sequential visual information and explain the sequential working patterns.

The aforementioned methods showed their capability of assisting safety monitoring on construction sites, however, the information extracted by these methods concentrated on entity-related information like the location or category. When it comes to construction scene understanding which involves multiple construction components interacting with each other, these methods failed to provide sufficient information for that because they cannot show the relationships among components. To address this problem, there have emerged some attempts to explore the semantic relationships within construction scenes. For instance, zhang et al. [17] developed an automatic hazard identification method combining object detection and ontology to analyze semantic interpretations of construction scenes. Wu et al. [18] presented a conceptual framework that combines computer vision and ontology techniques to facilitate safety management by semantically reasoning hazards. The combination of computer vision and ontology showed potential for assistance in safety management, but the semantic analyzation is limited with spatial relationships such as under, within and beneath. Interactions that enable high-level scene understanding like "what is the worker is doing and if he is in danger" include detailed relationships such as actions between different construction components which need the integration of computer vision with natural language processing. For example, the text descriptions and scene graphs are used to describe the semantic information in the construction scenes on the base of object recognition as shown in Fig. 1. Liu et al. [19] developed an encoder-decoder framework with deep neural networks to generate image captioning for construction scenes. Tang et al. [20] explored worker-tool interactions for vision-based safety compliance checking. However, image captioning involves too much noisy information that cannot be processed by computers easily and it is a general description which cannot detect hazards when exists multiple entities, and the worker-tool interactions cannot fully present the event in the construction scenes like the interactions between heavy machinery. Thus, Xiong et al. [21] developed a hazard identification system to evaluate the operation descriptions generated from site videos against the safety guidelines extracted from specifications with the help of an ontology model. The formatted computer-friendly context extracted from construction scenes and specifications shows great potential at automatic hazard identification.

While the rich semantic interactions in construction scene graphs generated from computer vision systems have helped high-level construction scene understanding, there is still a missing link between the visual information with domain knowledge (e.g. regulations and specifications) [2]. Onsite interactions must comply with the requirements inherited in the construction regulations. As a general knowledge present method, ontology can generate formatted knowledge expression from domain knowledge [22] and it has been widely adopted in the construction domain to assist management [18,23–25]. Related to ontology, the development and reasoning process often utilize Protégé and ontology reasoners like Pallet that require additional effort to coordinate the data flows between different systems and design inference templates.

With the development of natural language processing techniques and pre-trained models on large-scale corpus, NLP-based methods have been widely adopted to perform compliance checks between given contexts [26]. Li et al. [27] used word frequency analysis and cluster analysis to identify safety risks from accident reports. Tixier et al. [28] proposed a natural language processing system to extract precursors and the outcomes from unstructured injury reports. However, there are limited researches about utilizing natural language processing techniques with computer vision for compliance check between image structured construction scene information with text structured specification information. Bert, which has been proved to be the most accurate language representation model, is pretrained using the masked language model

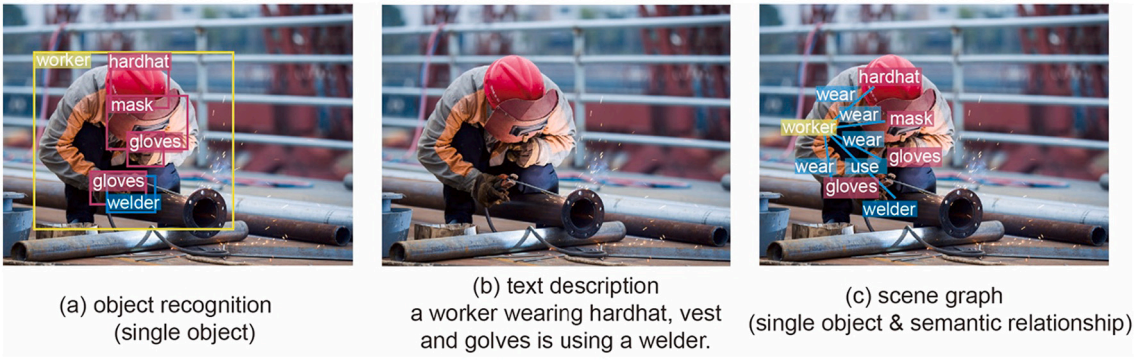


Fig. 1. Example of a construction site image, object recognition and scene graph.

(MLM) and next sentence prediction (NSP) [29]. The BERT model enables contextualized word embedding by the deep bi-directional representations from the MLM pre-training. Also, the NSP pre-training determines whether the two sentences are associated [30]. Bert model has been widely used for knowledge mining in the construction domain. Fang et al. [31] utilized the Bert model to perform text classification of near misses from safety reports. For construction hazards inference, the construction scene graph and construction regulations are encoded together to determine the association between them, which is well-suited for integrating domain knowledge with construction visual facts to perform hazards inference.

On top of the former work, we commerce our paper to develop a unified method for automatic hazard identification from construction scenes. For object detection, the objects tend to be in complex shapes that can result in errors in relationship analyzation when calculating the coordinate information for spatial relationships. In that condition, the detection requires semantic segmentation. For semantic relationships such as operate and wear, the relationship analyzation is conducted by analyzing the features of the detected objects with algorithms such as CNN and Transformer network. The features from bounding box-based detection can consider more environmental information and the bounding box detection is much more computational friendly compared with semantic segmentation. Based on that, this paper adopted Mask R-CNN network with bounding box detection for high precision and high features generation ability. Inspired by Xiong et al.'s [21] work, we format construction scenes into scene graph structure for hazard identification. Considering the interactions in construction scenes are complex and the environment is chaotic, we adopted Transformer network with attention mechanism for scene graph generation for the better contextual information extraction ability compared with Xiong's conditional random field network. The Transformer network can detect the rich semantic relationships in the construction scenes and thus enable the hazards inference based on interaction level of construction scene understanding. To integrate domain knowledge with construction scene graph for hazards inference, this paper designed a BERT-based hazards inference model. The model learns the dependencies between the construction scene graph with domain knowledge like regulations and outputs the hazards type based on the text classification result of the construction scene graph and the regulations.

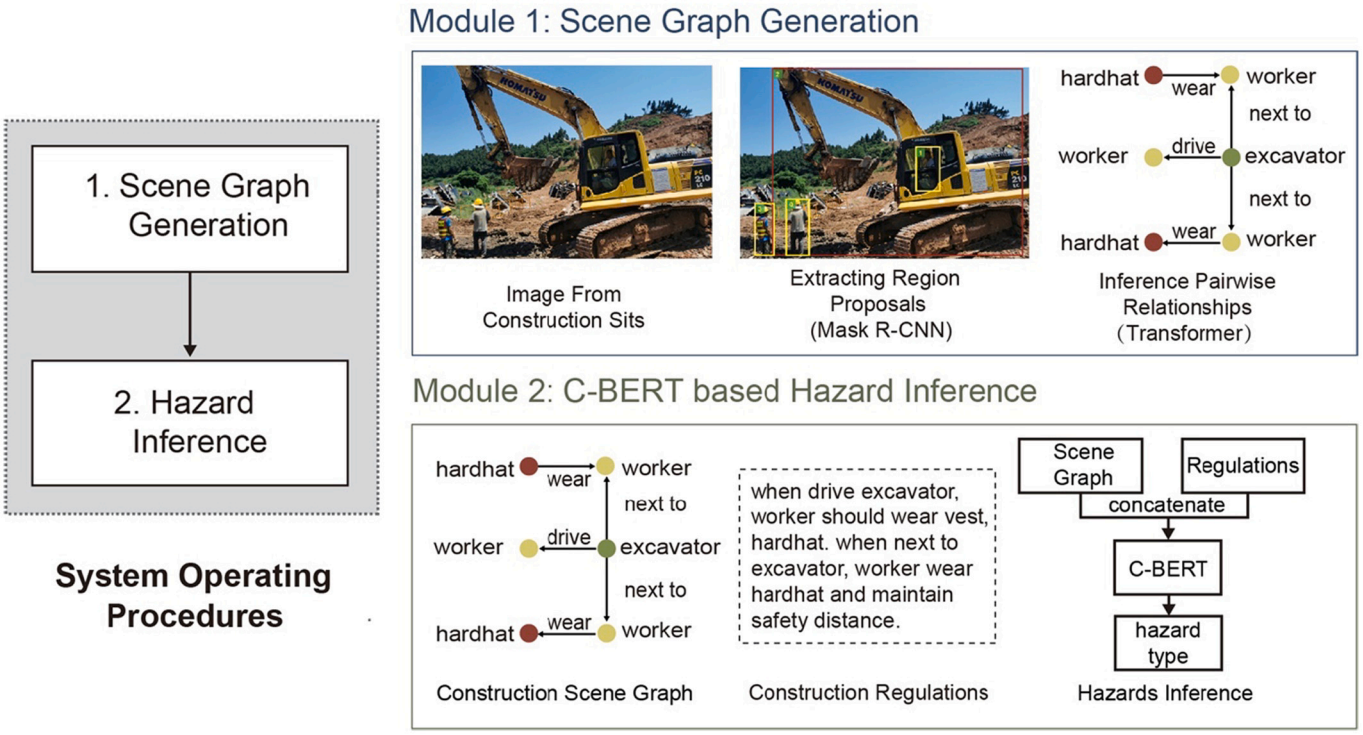


Fig. 2. Proposed framework for automatic hazard identification.

3. Methodology

To perform construction scene hazards inference based on the integration of construction scene graph with domain knowledge, this paper proposed an automatic hazard identification method for construction scenes. As shown in Fig. 2, the proposed method consists of two parts: construction scene graph generation and C-BERT based hazards inference. The construction scene graph generation model uses a Transformer based network to capture the entities and interactions inherited in the construction scenes and format them into construction scene graphs. For hazards inference, the C-BERT model can learn the relationship between the construction scene graph and construction regulations, and then perform hazards inference based on the integration of construction scenes with domain knowledge. For application, the proposed method is capable of identifying hazards from construction scenes with clear interactions between objects. Structural representation information can be extracted from the construction scenes and then adopted for hazards inference. Moreover, currently the method is suitable for single working condition with clear visual scenes and restricted by the multitasking and masking scenarios. In this paper, We focus on relations such as geometry (e.g., under, next to), actions (e.g., drive, wear) in the four different working conditions: welding, scaffolding, excavating and bricking to test the proposed method.

3.1. Construction scene graph generation

Construction scenes contain rich semantic information. The first step of hazards inference is to extract semantic descriptions from construction scenes. Fig. 3 provides three levels of construction scene understanding based on the scene analysis [17]. Fig. 3 (b) illustrates the construction components and their interactions in the construction scenes; Fig. 3 (c) contains three different levels of construction scene understanding. The entity-level detects the components in the scenes; the attribute level defines the details of the entities and the interaction level reveals their semantic relationships. Hazard inference based on cognition of construction scenes needs not just recognition but also reasoning the construction scenes which requires the union of three levels of scene understanding [32]. Semantic interactions, standing for the relationships of entities, play an important role in scene understandings [21]. The construction scene graph in this paper contains the entities, pixel locations, categories, spatial relationships and semantic relationships between entities. The generation process of construction scene graphs is challenging because interaction detection needs to extract contextual information from construction scenes. So, we adopted a Transformer network by Tang et al. [33] to extract contextual information to identify the interactions for the attention mechanism to capture global information from construction scenes.

The construction scene graph generation (CSGG) method involves three modules: 1) entity detection, 2) object representation and 3) relation representation. Given an image I, the probability of generating construction scene graph G is decomposed into three factors:

$$Pr(G|I) = Pr(B|I)Pr(O|B,I)Pr(R|B,O,I) \quad (1)$$

where I stands for the construction site images, B stands for a set of bounding boxes of detected entities, O stands for the object labels for each entity and R stands for the interaction between entities. $\Pr(B|I)$ indicates the process for entity detection from site images I. The images I, after being resized to 1024 * 1024 pixels, are being processed by the entity detection module to extract features from regions of interest B. $\Pr(O|B, I)$ indicates the process of contextualized object representation based on the features from entity detection B. $\Pr(R|B, O, I)$ indicates the process of contextualized relation presentation based on the object representation O and features from entity detection B. Finally, the entities and their interaction relationships are calculated with two different linear projection functions from the contextualized object and relation representation and then formatted into graph structure for compliance check to inference hazards.

Entity detection. Mask RCNN network with RPN is adopted to detect entities and extract features from construction scenes. The image is processed in Mask RCNN with two steps: 1) the construction site image is input for conventional layers and the residual and feature pyramid networks to extract feature maps for the entire image. Region proposal network is adopted to generate a series of regions of interest. 2) RoIAlign network takes the features of regions of interest as input to calculate the pixel locations for each candidate box and then performs bounding box

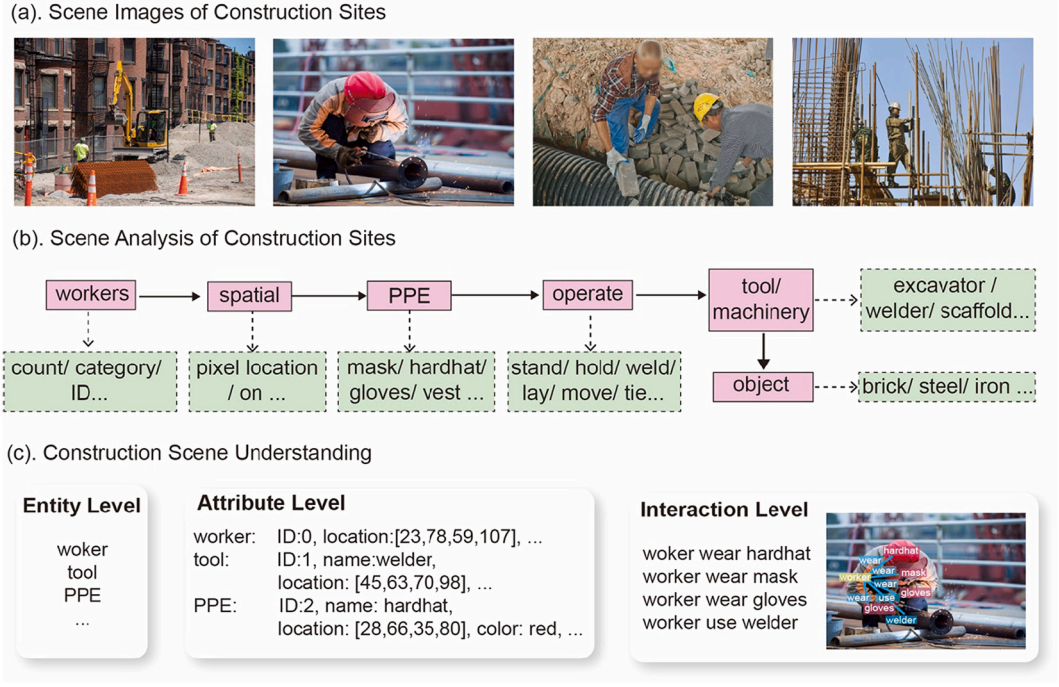


Fig. 3. Three levels of construction scene understanding.

regression. For each proposal b_i, the entity detection module also outputs a feature vector f_i and a non-contextualized object label probabilities l_i.

$$B, F, L = Mask\ R - CNN(I) \tag{2}$$

Object representation the next step is to generate contextual object representation for each entity in the construction scene. Scene graphs for construction scenes represent the interaction relationships among the whole scenes, it is crucial to consider features from all detected entities. Thus, this paper adopted $Transformer_{object}$ network to encode contextualized object representations as shown in Fig. 4.

$Transformer_{object}$ network uses attention mechanism to collect information from all the detected entities. As shown in fig, the inputs for transformer object network consist of three parts: label embedding, bounding box embedding, features of regions of interest and output the contextualized visual information $X = \{x_1, x_2, \ldots, x_n\}$ and labels L.

$$X, L = Transformer_{object}\left([b_i, f_i, l_i]_{i=1,\ldots n}\right) \tag{3}$$

The label embedding can be formulated with softmax function and a pretrained word matrix $w_{embedded}$ initialized on glove6B text.

$$obj_{embed} = softmax\ (l_i, dim = 1) \times w_{embedded} \tag{4}$$

e bounding box information b_i are transformed into embedding vector of 128 dimensions with two different sets of trainable matrixes A and bias B.

$$bbox_{embed} = \left(b_i \times A_1^T + B_1\right) \times A_2^T + B_2 \tag{5}$$

The final object representation obj_{rep} comes from the concatenation of features f_i from entity detection, obj_{embed} and $bbox_{embed}$ with multi-head self-attention mechanism. The feature vector f_i, together with obj_{embed} and $bbox_{embed}$ are concatenated together as a matrix $feature_{obj}$ and processed by 3 different fully connected layers ($Linear_{Q,\ K,\ V}$) to get matrixes Q, K, V, as shown in Eq. 6.

$$feature_{obj} = concat(f_i, obj_{embed}, bbox_{embed}) \tag{6}$$

$$Q, K, V = Linear_{Q,K,V}(concat(f_i, obj_{embed}, bbox_{embed})) \tag{7}$$

The attention mechanism as shown in Fig. 5 calculates the similarity between matrixes Q, K, V to get to feature vector z (Eq. 7). The multi-head attention utilizes different weight matrixes W to collect feature vectors z_i in different scales (Eq. 8). The output of multi-head attention mechanism Z is then sent into the feed-forward network to get the final result of contextualized object representation obj_{rep} (Eq. 9). Through a linear projection layer and a softmax function, the object label $label_{obj}$ is generated from object features (Eq. 10).

$$\mathrm{z} = \mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

$$z_i = \mathrm{head_i} = Attention\left(QW_i^Q, \mathrm{K}W_i^K, VW_i^V\right) \tag{9}$$

$$obj_{rep} = \mathrm{MultiHead}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = Concat(head_1, \cdots, \mathrm{head_h})W^O \tag{10}$$

$$label_{obj} = Softmax\left(\mathrm{feature_{obj}} \times A_3^T + B_3\right) \tag{11}$$

Relation representation the $Transformer_{relation}$ network is adopted to collect contextualized relation representation for relationship detection. For the relationship between i^{th} and j^{th} object, the relation representation adopts visual features e_{ij}^{vis} consisting of the union of two object boxes with union feature extractor from entity detection. The union area of b_i and b_j are calculated with the union function to get the pixel locations $b_{i,\ j}$ (as shown in Fig. 4).

The union feature extractor extract features of $b_{i,\ j}$ from different scales $b_{i,\ j(1\ldots n)}$. Now we need to get the visual feature for the relation prediction, therefore a pooler function is adopted to pool the features of different scales into the same scale and joins them together to get better feature representation b_{ij}.

$$b_{ij} = \sum f_{pool}\ \left(b_{i,j(m)}\right)\ m = 1\ldots n \tag{12}$$

Afterward, the spatial features b_{ij} are added with the concatenated label features e_{ij}^{sem} of both classes and visual features e_{ij}^{vis}. Subsequently, the relation detector uses a linear project layer (f_{elp}) to obtain the initial edge features $e_{i,\ j}^{in}$.

$$e_{ij}^{in} = f_{elp}\left(e_{ij}^{vis} + b_{ij} + e_{ij}^{sem}\right) = \left(e_{ij}^{vis} + b_{ij} + e_{ij}^{sem}\right) \times W_e + b_e \tag{13}$$

The initial edge features are adopted into the transformer_realtion network to get contextual edge features e_{ij}^{final} (Eq. 14). The multi-attention mechanism will help an edge, enriched with global information, to learn from edges with similar relational embeddings. Finally, with a softmax function, we get the distribution of the edge labels (Eq. 15).

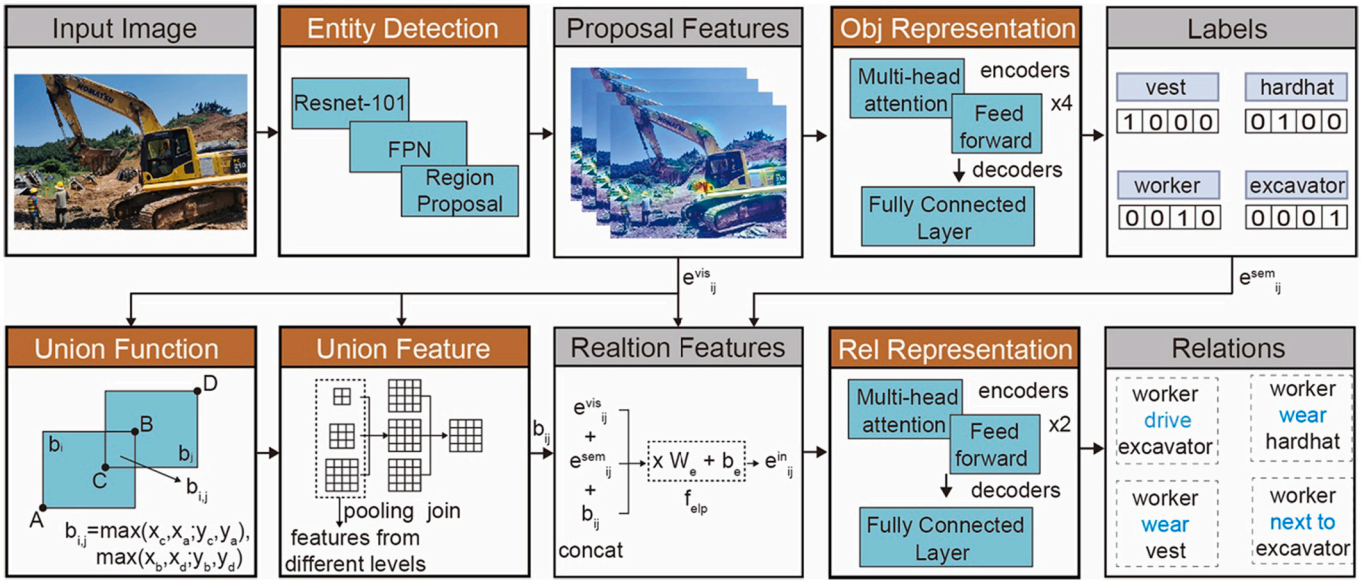


Fig. 4. Structure of Transformer network.

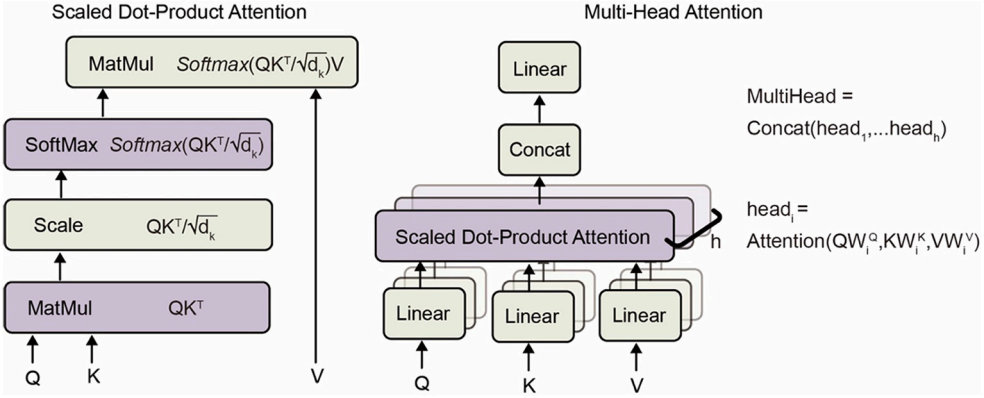


Fig. 5. Calculation in attention mechanism.

$$e_{ij}^{final} = Transformer_{relation}\left(e_{ij}^{in}\right) \tag{14}$$

$$Pr\left(x_{i\to j}|\mathrm{B},\mathrm{O}\right) = softmax\left(W_r e_{i,j}^{final} + w_{oi,oj}\right) \tag{15}$$

3.2. Hazards inference from C-Bert based text classification

The hazards in the construction scenes can be identified by comparing the visual facts with construction regulations. In addition, the visual facts generated from the construction scene graph model, the external domain knowledge like construction regulations should be integrated for hazards inference [21]. In this section, BERT, a widely used natural language processing technique, is utilized for automated hazards inference.

As shown in Fig. 6, the first step for hazards inference is to retrieve the regulations for current working condition. This step involves to identify the working condition in the construction scene graph based on the entities and interactions, and then utilize keyword mapping to get

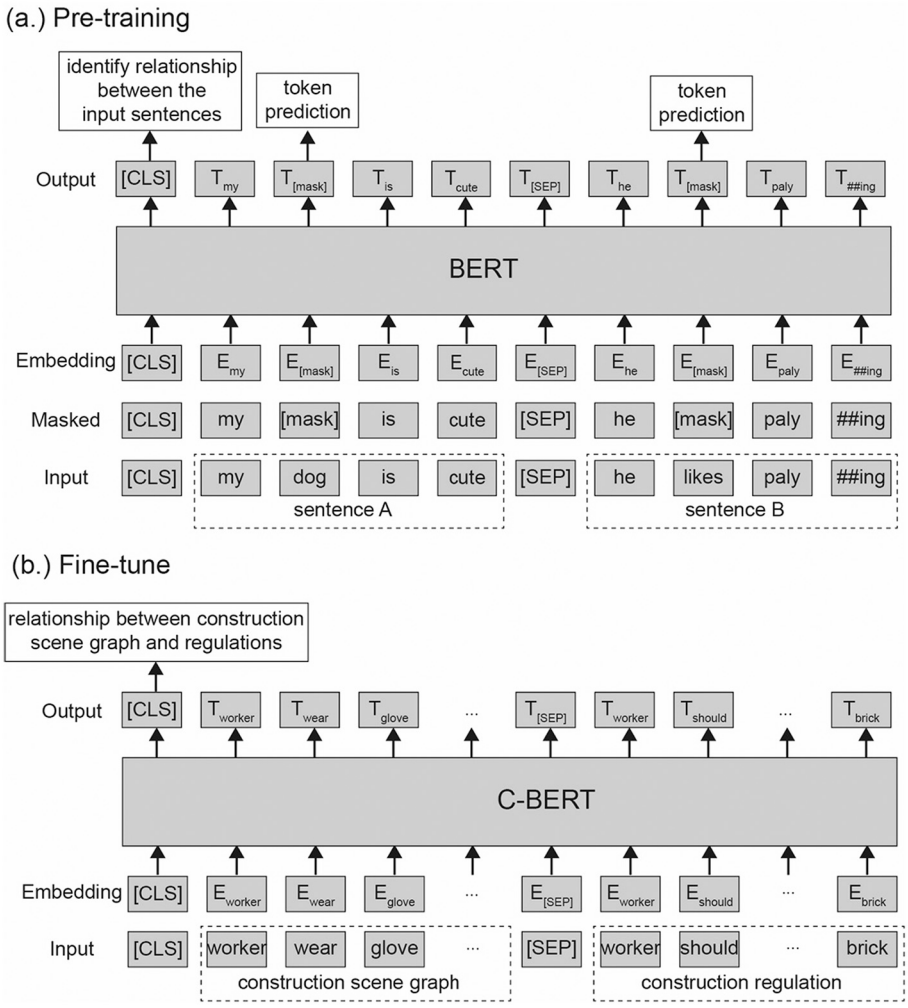


Fig. 6. Pre-training and fine tune of BERT model.

the regulations in the regulation base. The regulations in the base are encoded with working condition label. The second step is to integrate the construction scene graph with domain knowledge (the retrieved regulations) for hazards inference.

In the C-BERT model shown in Fig. 7, the construction scene graph and regulations are connected into a unified vector representation which begins with [*CLS*] and separates with [*SEP*] and the encoded by the C-BERT model. The C-BERT model utilizes Transformer to learn the relationships between the construction scene graph and the regulations and output the encoded output. The final step is to perform hazards classification. The encoded output is process with a fully connected layer and the a pooler function is used to generate the sentence level on the token [*CLS*] representation for classification. With a softmax function, the hazards type is identified.

Fig. 8 illustrates the model structure of C-BERT and it consists of the following parts: 1) keyword mapping to get the required regulations, 2) tokenizing the construction scene graph and regulations as input of the C-BERT model, 3) encoding the input with C-BERT as output and 4) classification of the encoded output. The C-BERT model [34] adopted in this paper is pre-trained on Wikipedia English text and we fine-tune it for our construction scene hazards inference.

Key-word mapping with the generated construction scene graph, according to the entities and interactions in the construction scene graph, the working condition w can be decided. With the working condition w, the required regulations S are retrieved for the construction scene graph G.

Tokenize As shown in Fig. 6, the tokens of construction scene graph $token_g$ with the corresponding specifications $token_s$ are concatenated together and tokenized into vector $E(x_1, \ldots, x_n, y_a, \ldots, y_m)$.

$$E = tokenizer\left(concat\left(token_g, [SEP], token_s\right)\right) \tag{16}$$

Encode The C-BERT model utilizes Transformer encoders with attention mechanism to encode the input. The encoders can obtain the dependencies between words and sentences which benefits from the stacking of Multi-Head Attention layer and Feed Forward Network layer of each sublayer. For the input vector E, the output of each sublayer is shown in Eq. 17.

$$output_{sublayer} = LayerNorm\left(E + (SubLayer(E))\right) \tag{17}$$

In each sublayer, first, the input E is transform into Q, K, V with matrix transformation (Eq. 18) and then processed by self-attention function (same as Eq. 8). After that, Q, K, V are projected into h dimensions with different linear transformation matrixes W_1, …, W_h. The final output $T(b_1, \ldots, b_n, b_a, \ldots, b_m)$ (Eq. 20) is the concatenation of the result from h dimensions attention mechanism. The output T then processed by the feed forward network to calculate the features of non-linear (Eq. 21).

$$Q = EW_Q, K = EW_K, V = EW_V \tag{18}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{19}$$

$$T(Q, K, V) = Concat(head_1, \ldots, head_h)W^0 \tag{20}$$

$$FFN(T) = max(0, TW_1 + b_1)W_2 + b_2 \tag{21}$$

Classification Through the C-BERT model, the vector E is transformed into encoded vector Z. With a pooler function and a softmax function, the encode vector Z is transformed into the [*CLS*] representation S. S is also the result of text classification and can project to the hazards type.

$$T = BERT\left(E_1, \ldots, E_N, E_{SEP}, E_1^{'}, \ldots, E_N^{'}\right) \tag{22}$$

$$S = softmax(pooler(T)) \tag{23}$$

4. Experiment

This section describes the implementation of the proposed method and we select welding, scaffolding, excavating and bricking conditions to demonstrate and test the feasibility of the proposed method.

4.1. Environment

The proposed method is programmed in Python 3.6, CUDA 10.0 and CUDNN 7.3.4. The proposed method is tested in an Ubuntu 16.04 64bit system environment equipped with Intel Core i5 9400F CPU, GeForce 1660 6G GPU and 64G memory.

4.2. Scene graph dataset of construction sites

Existing large-scale datasets like Visual Relationships dataset [35] and Visual Genome dataset [32] contains visual relationships mainly cover lifestyle scenes with no construction scenes. To address this problem, the self-built Scene Graph Dataset of Construction Sites is established. The dataset contains a total of 5060 pieces of construction site images with 9 types of construction components and 7 types of interactions. A total of 16,218 entities and 8300 valid interactions are labeled and the details of the dataset are shown in Table 2. To simply the annotation process, The VIA annotation software and a self-coded Python-based format conversion program are used to automate the annotation process.

The annotation procedure is composed of two parts: entity annotation and interaction annotation. As shown in Fig. 9, we set four attributes for every bounding box including subject_id, object_id, relationship and category. In entity annotation, the bounding box

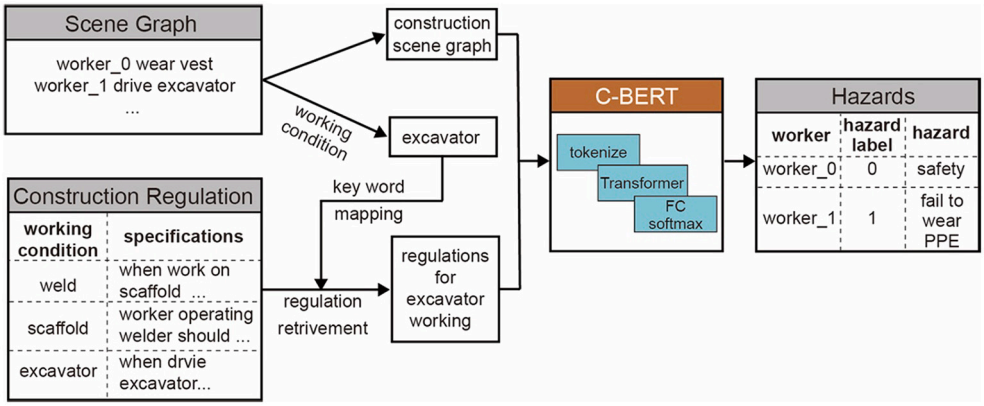


Fig. 7. Workflow of C-BERT based hazards inference.

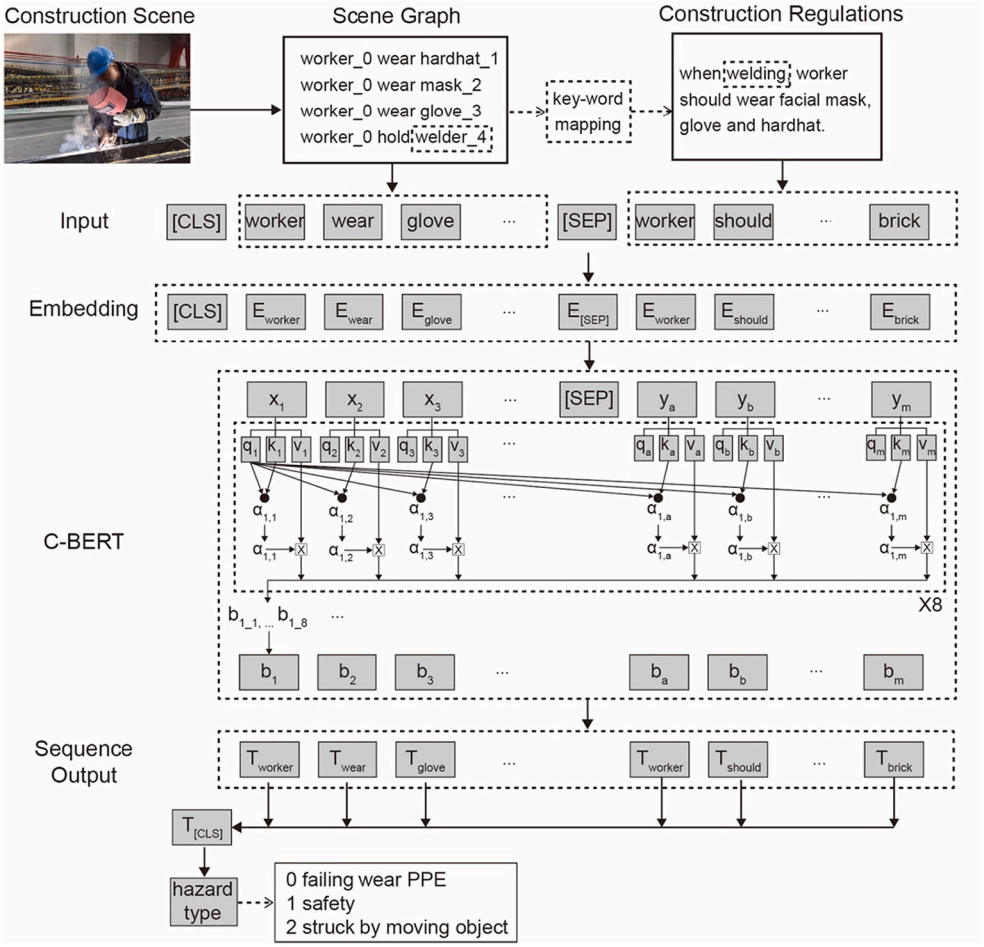


Fig. 8. Structure of C-BERT network

Table 2
Distribution of dataset.

Category	Total	Worker	Hardhat	Brick	Welder	Facial mask	Glove	Scaffold	Vest	Excavator
	16,218	7524	4475	877	156	156	934	1187	744	165

Category	Total	Drive	Under	Lay	Stand on	Operate	Wear	Next to
	8300	118	655	696	1246	275	4454	856

attribute and category attribute are adopted to describe the pixel location and the category of the labeled entity. In interaction annotation, the subject_id, object_id and relation attributes are adopted to describe the category and the two parties of current interaction. The results of entity and interaction annotation are processed to get object.json and relationships.json shown in Fig. 10.

4.3. Performance of scene graph generation

After the annotation, the dataset is split into two parts, 70% for training and 30% for testing. The performance of training was evaluated on the test dataset. For the task of object detection, average precision mean average precision and top-k recall score are employed to evaluate the performance of the object detector. The top-k recall (Eq. 24) is abbreviated as R@K, where TP means True Positive and FP means False Positive. The Mean Average Precision (Eq. 26) is abbreviated as mAp, where P indicates Precision, R indicates Recall and n is the number of categories. The object detector was trained for 19,000 times to get the best performance as shown in Table 3.

$$Recall = \frac{TP}{TP + FP} \tag{24}$$

$$AP = \int_0^1 PR\, dr \tag{25}$$

$$mAp = \frac{\sum_1^n AP}{n} \tag{26}$$

For relation detection, no graph constraint mean recall (ng-mR@K) is also adopted to evaluate the performance. In traditional recall computing, a pair of objects can only have one relationship involved in the ranking process. No graph constraint recall allows all the detected relationships getting involved in the ranking process. Also, to evaluate the performance of the adopted Transformer network, this paper compared the evaluation results with the classic scene graph generation method based on Bi-LSTM network by Zeller [36] as shown in Table 4.

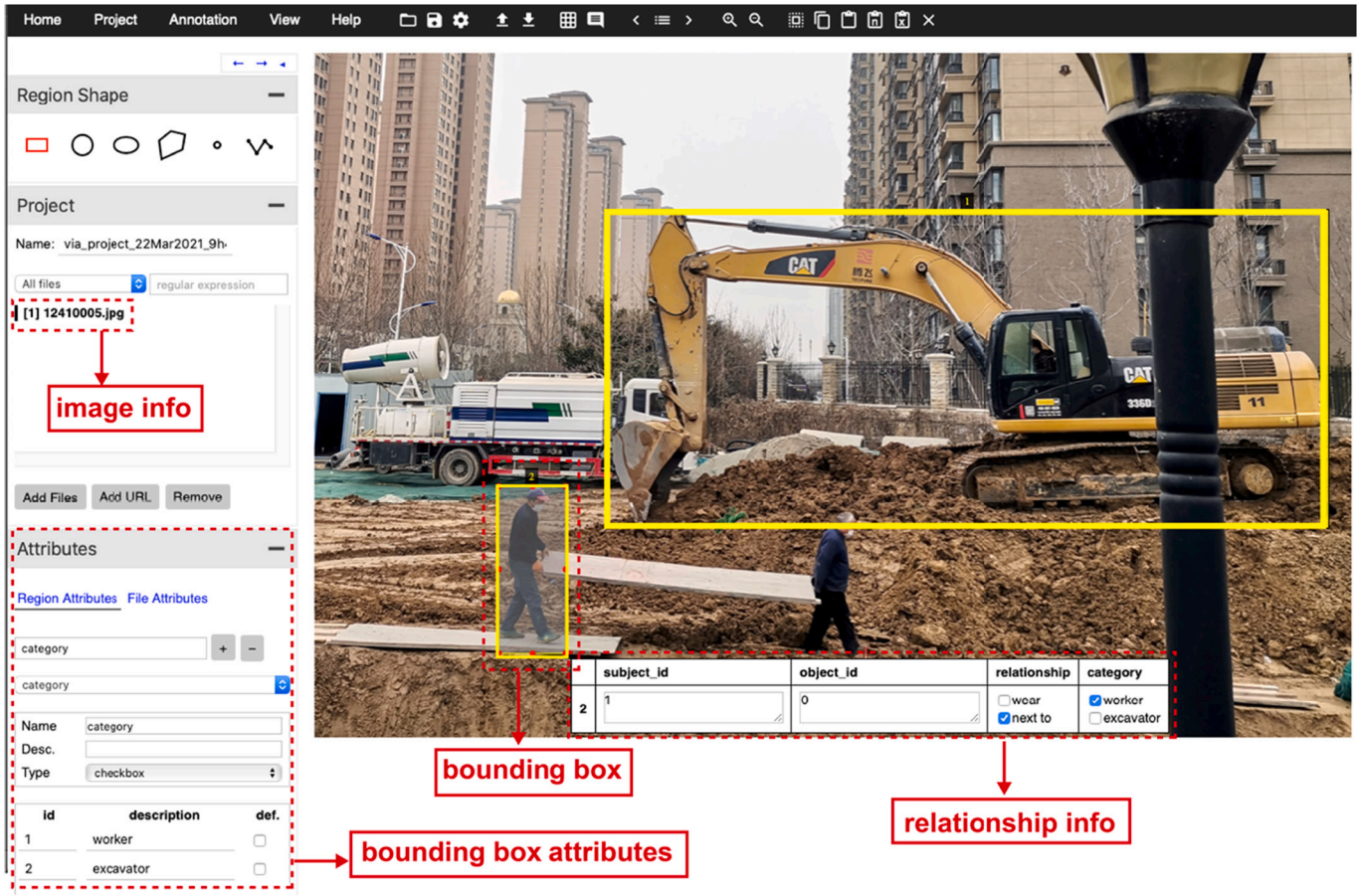


Fig. 9. Customed VIA annotation interface.

The construction scene graph module distills the pixel locations, categories and interactions from the detected construction components thus formatting the unstructured scene information into structured format for computer to process. The module generates triplets with attributes to describe the construction scene. Also, the spatial relationships between worker and heavy machinery are defined as next to and calculated with IoU function in Fang's work [23]. If the IoU result is larger than 0, we add an IoU tag to the worker. According to the entities and their interactions, this paper designs a projection function to generate the corresponding working condition with keyword mapping. With the working condition, the working requirements are retrieved from specifications. The construction scene graph and the working requirements are concatenated together and sent into the C-BERT model for text classification and hazards inference.

4.4. Hazards inference with C-BERT

To demonstrate and assess the validity of the C-BERT network for hazards inference, this paper selected 4 kinds of working scenes containing 5 kinds of unsafe behaviors and safety condition based on the self-built construction scene graph dataset and safety handbook for construction sites as shown in Table 5. For the promotion of construction scenes that contain much more complex interactions between a vast number of different components, and much more complex causations of different hazards, further researches are needed.

For hazards inference, the metric of accuracy is utilized for performance evaluation as shown in Eq. 27. The C-BERT network achieved ***97.82%*** of accuracy for hazard identification of 5 kinds as shown in Table 6.

$$Accuracy = \frac{Number\ of\ correctly\ classified\ classes}{Total\ number\ of\ classes} \times 100\% \tag{27}$$

We manually simulated 5 different working scenes to test the C-BERT network. The fusion matrix in Fig. 11a shows the detection result of the C-BERT model. Since the C-BERT model performs hazard identification based on the feature of [*CLS*] token, the attention visualization of the C-BERT model in Fig. 11b shows that the feature of [*CLS*] comes from the integration of the construction scene graph with construction regulations and also indicates the C-BERT can inference hazards from construction scene graphs integrating external domain knowledge. As shown in Fig. 11a, the C-BERT network can accurately identify the hazards in the simulation working conditions. 5 kinds of hazards and safety condition are identified with very rare misidentification. With total of 105 testing scenes, only 4 falling and combating hazards are misidentified as struck by. The high performance for hazards inference validated the C-BERT network for hazards inference.

4.5. Demonstrations of hazard identification

The hazard identification is divided into two steps: 1) match the construction scene graph with the required specifications and 2) perform hazard identification with C-BERT model. In the first step, the construction scene graphs from the construction scene graph generation are joined together as formatted representation for construction scenes. Also, to retrieve the required specifications, we set a key-word matching function according to the scene elements from the construction scene graph. With the formatted construction scene representation and the retrieved the specifications, the information are concatenated together as the input of the second step. The C-BERT model performs text

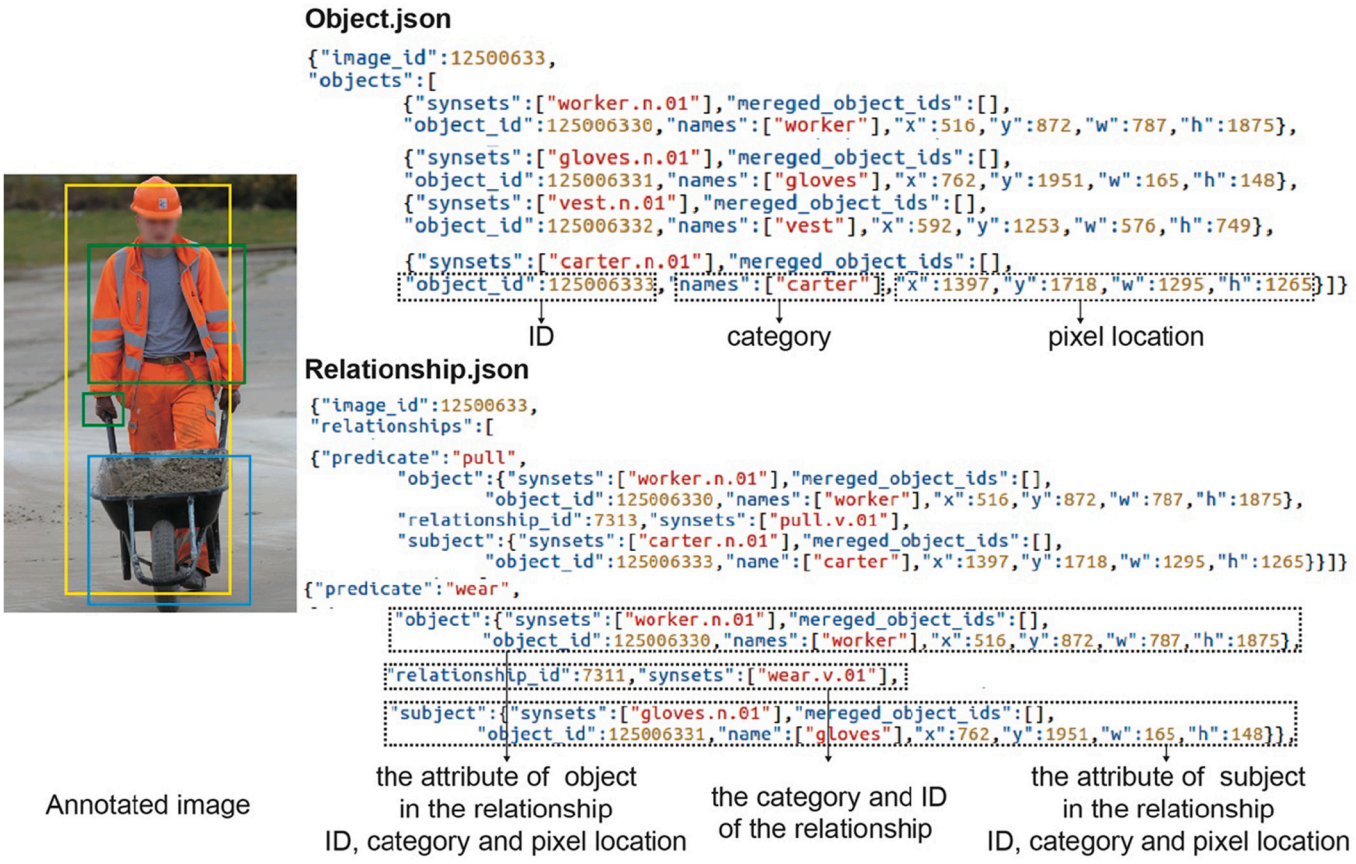


Fig. 10. Illustration of object.json and relationship.json.

Table 3
Results of entity detection.

Category	Mean	Worker	Hardhat	Brick	Welder	Facial mask	Glove	Scaffold	Vest	Excavator
AP	0.801	0.876	0.865	0.793	0.766	0.801	0.824	0.800	0.832	0.653
AR	0.810	0.895	0.887	0.821	0.745	0.812	0.842	0.811	0.783	0.692

Table 4
Results for relation detection.

Ng-mR@K	@20	@50	@100
Transformer	0.8235	0.9651	0.9892
Bi-LSTM	0.7830	0.9414	0.9822

Category	Drive	Under	Lay	Stand on	Operate	Wear	Next to
Recall	0.9545	0.9753	0.9963	0.9875	0.9956	0.9855	0.9873

classification of the concatenated input and outputs the label as the hazards inference result. To demonstrate the validity of the method, 4 different working conditions are selected and the testing results are shown in Fig. 12.

5. Discussion and limitation

To improve the efficiency of safety management and mitigate the accident originated from unsafe interactions between different construction components, this paper proposed a hazard identification method that integrates computer vision with natural language processing to discover the potential accidents from construction scenes. This approach provides safety management on construction sites with the analyzation of unsafe interactions thus can proactively mitigate the loss caused by accidents. Also, the realization of being monitoring will make the workers strong tendency to abide the requirements on construction sites.

In comparison with previous studies that had used computer vision for hazard identification, our study has the following advantages:

1) in our study, the state-of-art Transformer network is adopted to detect the interactions of different construction components in the construction scenes. Thanks to the self-attention mechanism, the Transformer network can generate contextualized object and

Table 5
Details of unsafe behaviors.

Working condition	Unsafe behavior	Regulations	Causation
Welding	ARC light exposure and burns	Worker should wear hardhat, gloves, facial mask when operates welder	Lack of PPE
Scaffolding	Falling, combating	Worker should wear hardhat and reflective vest on scaffold	Lack of PPE
Excavating	Struck by moving object	Worker should wear hardhat, reflective vest when next to excavator and obtain safety distance	Lack of PPE Distance approach
	Safe driving	Worker should wear hardhat, reflective vest when drive excavator	Lack of PPE
Bricking	Cross work on the same working surface	Worker should wear hardhat, gloves and reflective vest when bricking, also avoid cross work on the same working surface	Lack of PPE
	Falling, combating	Worker should wear hardhat, gloves and reflective vest when bricking	Lack of PPE

relation representations and detect the interactions of the entities in construction scenes. The test shows the Transformer network outperforms the classic Bi-LSTM models for relation detection. The detected entities with their interactions are formatted into graph structure for hazard identification.

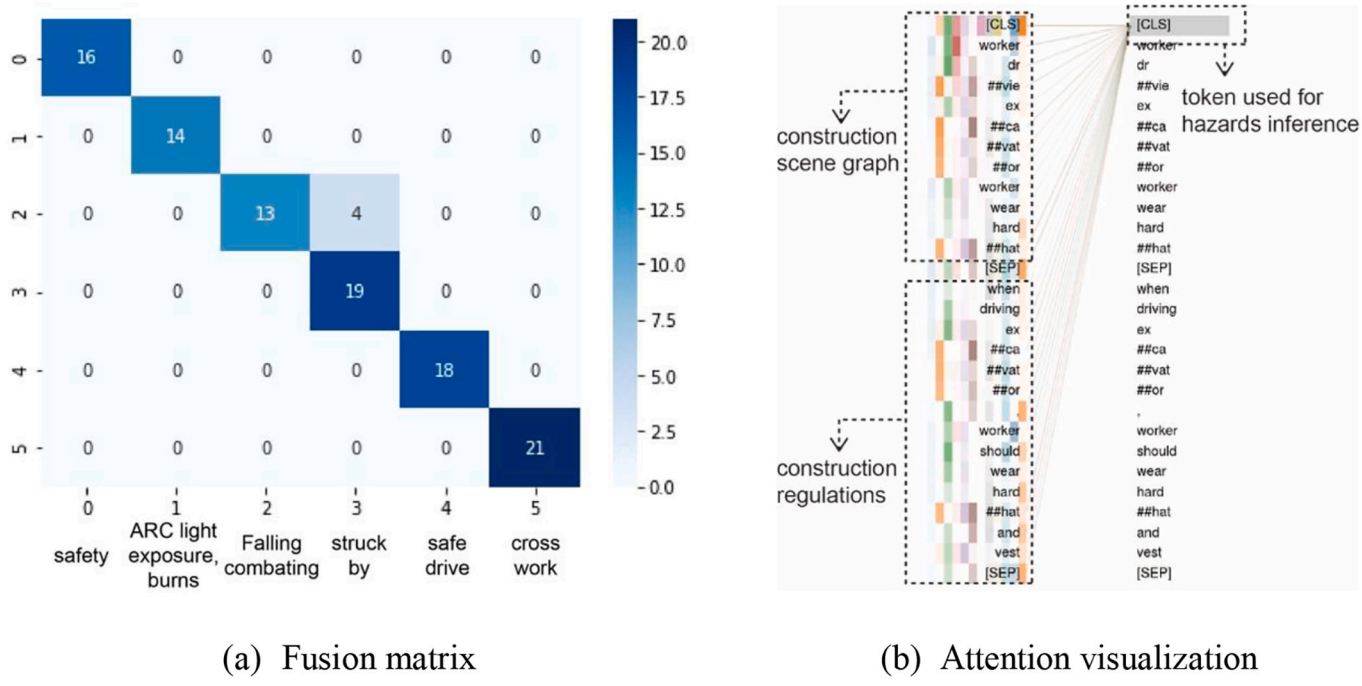
2) We designed a C-BERT model for hazards inference through text classification of the detected construction scene graphs with specifications. The ontology-based hazards inference needs to design ontology templates and huge human efforts to extract regulatory documents into structured computer-accessible format with additional inference software. The C-BERT model can realize automatic hazards inference of the construction scene graph with the required specifications which are retrieved from key-word mapping.

Despite the novelty of our proposed method, we need to admit that it has several limitations. Firstly, since entity detection is the foundation of hazard identification, the performance of entity detection will directly determine the accuracy of hazard identification. To improve the accuracy and robustness of entity detection, a much larger dataset with multiple construction scenes of different working, weather, and illumination conditions are needed to train the computer vision algorithms. Also, the Transformer block can be added to the Mask RCNN network in the future for robustness improvement. In addition, relation detection also relies heavily on the labeled dataset. Currently, the interactions in the labeled dataset are limited and tend to be simple because of the much higher cost of dataset annotation with relations compared with the annotation of entities. For scenes with multiple workers and machines, the interrelationships are complex. The dataset should be expanded with more semantic relationships to present the complex interactions with multiple workers and machines. The relation detection can be optimized by introducing a prior filtering criterion to deal with the huge computational power neediness with multiple workers and machines.

Secondly, the C-BERT model for hazards inference only covers the demonstration scenes. The safety monitoring for all construction scenes needs a much larger knowledge base to be embedded in the C-BERT model. There are some attempts for automatic information extraction in the construction domain to build the knowledge base. Moon et al. [37,38] made specification review with semantic text pairing and analyzed different properties of multiple construction specifications and utilized named entity recognition for construction specifications review

Table 6
Accuracy matrix of hazard inference.

Hazard type	Safety	ARC exposure burns	Falling combating	Struck by	Safe drive	Cross work	Average
Accuracy	100%	100%	100%	82.6%	100%	100%	**97.82%**



(a) Fusion matrix (b) Attention visualization

Fig. 11. Fusion matrix and attention visualization of C-BERT testing result.

Fig. 12. Construction scene graph and hazards inference

tasks. The utilization of advanced natural language processing (NLP) techniques shows great potential for building large scale knowledge based in the construction domain.

6. Conclusions

This paper proposed a novel hazard identification method for construction scenes based on the analyzation of interactions among different construction components. In the method, the construction scene is encoded into graph-structured representations by the construction scene graph generation module, and the required safety specifications are retrieved through keyword mapping in the construction scene graphs. For hazards inference, the construction scenes graphs and the retrieved safety specifications are concatenated together to perform hazards classification which enables hazards inference based on the integration of visual facts and domain knowledge. The classification result indicates the safety issues for the construction scenes. In particular, the scene graph generation module adopted a Transformer based network to detect the interactions in the construction scenes. The self-attention mechanism in the Transformer network can generate contextualized object and relation representations thus outputting the semantic interactions of the detected entities. The scene graph generation module then formats the unstructured construction scene into graph-structured triplets format. Also, this paper designed a C-BERT module for hazards inference. According to the generated construction scenes graphs, the required safety specifications are retrieved by keyword mapping and then concatenated together as input for the C-BERT module. The C-BERT model can perform hazards classification of the concatenated inputs and output the result of hazard identification.

The proposed hazards inference method was trained and tested on the self-built construction scene dataset. The testing results indicate two types of hazards are identified. Using no graph constraint mean recall (ng-mR@K) evaluation matrices, our model achieved 0.8235, 0.9651 and 0.9892, respectively on relation detection measured by ng-mR@20, ng-mR@50 and ng-mR@100, which outperform the classic Bi-LSTM model. For hazard inference, the proposed method achieved 97.82% of accuracy with 5 kinds of hazards. The high performance results show the proposed method is promising in detecting semantic interactions between construction components and hazard identification. As compared to text description methods, the scene graph performs semantic scene understanding with graph structure which is more friendly for computer to process.

The major contribution of this study is in two aspects. First, the feasibility of hazard inference based on the interactions between the components inside the construction scenes was demonstrated. Second, a creative way to infer hazards in a computer-automated way integrating two types of cross-modal information: domain knowledge and construction scene information was proposed thus reducing the human intervention. Overall, the method provides an efficient way to combine visual information and domain knowledge on construction sites for automated safety monitoring, and a path for the fusion of massive amounts of multi-modal information in the industry, which can be used as an automatic hazard identification tool to assist site managers in intervening ahead of site hazards. However, there still exit some

shortcomings. First, the proposed method identifies hazards based on the components and their interactions inside the construction scenes, and the huge size difference of different components, the occlusion of the components and the complex interactions between the components are restricting the performance of distilling structured scene information from the construction scenes. The performance can be improved by the advancement of computer vision algorithms and the establishment of lager scale domain related datasets. Second, the proposed method integrates domain knowledge into the computer vision system to identify hazard, and the domain knowledge is in natural state that lacks labeling and structuring. With the complexity of construction scenes, the need of knowledge base that contains the related labeled domain knowledge extracted from massive amount of regulations is arising, and the massive regulations need to be classified into different categories and structured to build the knowledge base that supports large scale knowledge retrieval and matching. Our future research will focus on improving the robustness of the algorithms and datasets to adapt to the more complex construction scenes, as well as structuring and normalizing the large volumes of construction regulations to enable the large-scale knowledge base.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Zhang lite reports financial support was provided by Ocean University of China.

Data availability

Data will be made available on request.

References

[1] Ministry of Housing and Urban Rural Development of China. https://www.mohurd.gov.cn/gongkai/fdzdgknr/tzgg/202006/20200624_246031.html, June 2020 (last access August 11, 2022).

[2] S. Paneru, I. Jeelani, Computer vision applications in construction: current state, opportunities & challenges, Autom. Constr. 132 (2021), 103940, https://doi.org/10.1016/j.autcon.2021.103940.

[3] W. Qin, R.Y. Zhong, H.Y. Dai, Z.L. Zhuang, An assessment model for RFID impacts on prevention and visibility of inventory inaccuracy presence, Adv. Eng. Inform. 34 (OCT.) (2017) 70–79, https://doi.org/10.1016/j.aei.2017.09.006.

[4] N. Pradhananga, J. Teizer, Automatic spatio-temporal analysis of construction site equipment operations using GPS data, Autom. Constr. 29 (JAN.) (2013) 107–122, https://doi.org/10.1016/j.autcon.2012.09.004.

[5] A. Shahi, A. Aryan, J.S. West, C.T. Haas, R.C.G. Haas, Deterioration of UWB positioning during construction, Autom. Constr. 24 (7) (2012) 72–80, https://doi.org/10.1016/j.autcon.2012.02.009.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, SSD: Single Shot MultiBox Detector, European Conference on Computer Vision, 2016, https://doi.org/10.1007/978-3-319-46448-0_2.

[7] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T.M. Rose, W. An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (2018) 1–9, https://doi.org/10.1016/j.autcon.2017.09.018.

[8] Z. Zhu, X. Ren, Z. Chen, Integrated detection and tracking of workforce and equipment from construction jobsite videos, Autom. Constr. 81 (sep.) (2017) 161–171, https://doi.org/10.1016/j.autcon.2017.05.005.

[9] D. Kim, M. Liu, S. Lee, V.R. Kamat, Remote proximity monitoring between mobile construction resources using camera-mounted UAVs, Autom. Constr. 99 (2019) 168–182, https://doi.org/10.1016/j.autcon.2018.12.014.

[10] W. Fang, B. Zhong, N. Zhao, P.E.D. Love, H. Luo, J. Xue, S. Xu, A deep learning-based approach for mitigating falls from height with computer vision: convolutional neural network, Adv. Eng. Inform. 39 (2019) 170–177, https://doi.org/10.1016/j.aei.2018.12.005.

[11] M. Neuhausen, J. Teizer, M. König, Construction worker detection and tracking in bird's-eye view camera images, ISARC, in: Proceedings of the International Symposium on Automation and Robotics in Construction 35, IAARC Publications, 2018, pp. 1–8, https://doi.org/10.22260/ISARC2018/0161.

[12] S. Chi, C.H. Caldas, D.Y. Kim, A methodology for object identification and tracking in construction based on spatial modeling and image matching techniques, Comp. Aided Civil Infrastruct. Eng. 24 (3) (2010) 199–211, https://doi.org/10.1111/j.1467-8667.2008.00580.x.

[13] Park Man-Woo, Makhmalbaf Atefe, Brilakis Ioannis, Comparative study of vision tracking methods for tracking of construction site resources, Autom. Constr. 20 (7) (2011) 905–915, https://doi.org/j.autcon.2011.03.007.

[14] Y.J. Lee, M.W. Park, 3D tracking of multiple onsite workers based on stereo vision, Autom. Constr. 98 (FEB.) (2019) 146–159, https://doi.org/j.autcon.2018.11.017.

[15] S.U. Han, S.H. Lee, A vision-based motion capture and recognition framework for behavior-based safety management, Autom. Constr. 35 (2013) 131–141, https://doi.org/10.1016/j.autcon.2013.05.001.

[16] J. Kim, S. Chi, Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles, Autom. Constr. (2019) 255–264, https://doi.org/10.1016/j.autcon.2019.03.025, 104 (AUG.).

[17] M. Zhang, M. Zhu, X. Zhao, Recognition of high-risk scenarios in building construction based on image semantics, J. Comput. Civ. Eng. 34 (4) (2020), 04020019. https://doi.org/(ASCE)CP.1943-5487.0000900.

[18] H. Wu, B. Zhong, H. Li, P. Love, X. Pan, N. Zhao, Combining computer vision with semantic reasoning for on-site safety management in construction, J. Build. Eng. 42 (2021), https://doi.org/10.1016/j.jobe.2021.103036.

[19] H. Liu, G. Wang, T. Huang, P. He, M. Skitmore, X. Luo, Manifesting construction activity scenes via image captioning, Autom. Constr. 119 (2020), https://doi.org/10.1016/j.autcon.2020.103334.

[20] S. Tang, D. Roberts, M. Golparvar-Fard, Human-object interaction recognition for automatic construction site safety inspection, Autom. Constr. 120 (2020), https://doi.org/10.1016/j.autcon.2020.103356.

[21] R. Xiong, Y. Song, H. Li, Y. Wang, Onsite video mining for construction hazards identification with visual relationships, Adv. Eng. Inform. 42 (2019), https://doi.org/10.1016/j.aei.2019.100966.

[22] S. Zhang, F. Boukamp, J. Teizer, Ontology-based semantic modeling of construction safety knowledge: towards automated safety planning for job hazard analysis (JHA), Autom. Constr. 52 (apr.) (2015) 29–41, https://doi.org/10.1016/j.autcon.2015.02.005.

[23] W. Fang, L. Ma, P.E.D. Love, H. Luo, L. Ding, A. Zhou, Knowledge graph for identifying hazards on construction sites: integrating computer vision with ontology, Autom. Constr. 119 (2020), https://doi.org/10.1016/j.autcon.2020.103310.

[24] L.Y. Ding, B.T. Zhong, S. Wu, H.B. Luo, Construction risk knowledge management in BIM using ontology and semantic web technology, Saf. Sci. 87 (2016) 202–213, https://doi.org/10.1016/j.ssci.2016.04.008.

[25] X. Xing, B. Zhong, H. Luo, H. Li, H. Wu, Ontology for safety risk identification in metro construction, Comput. Ind. 109 (2019) 14–30, https://doi.org/10.1016/j.compind.2019.04.001.

[26] S. Baek, W. Jung, S.H. Han, A critical review of text-based research in construction: data source, analysis method, and implications, Autom. Constr. 132 (2021), 103915, https://doi.org/10.1016/j.autcon.2021.103915.

[27] J. Li, J. Wang, N. Xu, Y. Hu, C. Cui, Importance degree research of safety risk management processes of urban rail transit based on text mining method, Information 9 (2) (2018) 26, https://doi.org/10.3390/info9020026.

[28] A.J.-P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports, Autom. Constr. 62 (2016) 45–56, https://doi.org/10.1016/j.autcon.2015.11.001.

[29] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv Preprint (2018), https://doi.org/10.48550/arXiv.1810.04805 arXiv:1810.04805.

[30] Y. Kim, S. Bang, J. Sohn, H. Kim, Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers, Autom. Constr. 134 (2022), 104061, https://doi.org/10.1016/j.autcon.2021.104061.

[31] W. Fang, H. Luo, S. Xu, P. Love, C. Ye, Automated text classification of near-misses from safety reports: an improved deep learning approach, Adv. Eng. Inform. 44 (6) (2020), 101060, https://doi.org/10.1016/j.aei.2020.101060.

[32] R. Krishna, Y. Zhu, O. Groth, J. Johnson, F.F. Li, Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017), https://doi.org/10.1007/s11263-016-0981-7.

[33] K. Tang, Y. Niu, J. Huang, J. Shi, H. Zhang, Unbiased scene graph generation from biased training, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, https://doi.org/10.48550/arXiv.2002.11949.

[34] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, H. Wu, Ernie: enhanced representation through knowledge integration, arXiv Preprint (2019), https://doi.org/10.48550/arXiv.1904.09223 arXiv:1904.09223.

[35] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual Relationship Detection with Language Priors, 2016, https://doi.org/10.1007/978-3-319-46448-0_51.

[36] R. Zellers, M. Yatskar, S. Thomson, Y. Choi, Neural motifs: scene graph parsing with global context, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2018) 5831–5840, https://doi.org/10.48550/arXiv.1711.06640.

[37] S. Moon, G. Lee, S. Chi, H. Oh, Automated construction specification review with named entity recognition using natural language processing, J. Constr. Eng. Manag. 147 (1) (2021), https://doi.org/10.1061/(asce)co.1943-7862.0001953.

[38] S. Moon, G. Lee, S. Chi, Semantic text-pairing for relevant provision identification in construction specification reviews, Autom. Constr. 128 (2021), 103780, https://doi.org/10.1016/j.autcon.2021.103780.