buildings

MDPI

Article

Automatic Construction Hazard Identification Integrating On-Site Scene Graphs with Information Extraction in Outfield Test

Xuan Liu [1,2,*], Xiaochuan Jing [2], Quan Zhu [1,2], Wanru Du [1,2] and Xiaoyin Wang [2]

[1] China Aerospace Academy of Systems Science and Engineering, Beijing 100048, China
[2] Aerospace Hongka Intelligent Technology (Beijing) Co., Ltd., Beijing 100048, China
* Correspondence: liux@cau.edu.cn

Abstract: Construction hazards occur at any time in outfield test sites and frequently result from improper interactions between objects. The majority of casualties might be avoided by following on-site regulations. However, workers may be unable to comply with the safety regulations fully because of stress, fatigue, or negligence. The development of deep-learning-based computer vision and on-site video surveillance facilitates safety inspections, but automatic hazard identification is often limited due to the semantic gap. This paper proposes an automatic hazard identification method that integrates on-site scene graph generation and domain-specific knowledge extraction. A BERT-based information extraction model is presented to automatically extract the key regulatory information from outfield work safety requirements. Subsequently, an on-site scene parsing model is introduced for detecting interaction between objects in images. An automatic safety checking approach is also established to perform PPE compliance checks by integrating detected textual and visual relational information. Experimental results show that our proposed method achieves strong performance in various metrics on self-built and widely used public datasets. The proposed method can precisely extract relational information from visual and text modalities to facilitate on-site hazard identification.

Keywords: construction hazard; information extraction; scene graph; safety inspection

Citation: Liu, X.; Jing, X.; Zhu, Q.; Du, W.; Wang, X. Automatic Construction Hazard Identification Integrating On-Site Scene Graphs with Information Extraction in Outfield Test. *Buildings* **2023**, *13*, 377. https://doi.org/10.3390/buildings13020377

Academic Editor: Svetlana J. Olbina

Received: 18 December 2022
Revised: 13 January 2023
Accepted: 17 January 2023
Published: 29 January 2023

1. Introduction

In China, the development of a national defense information infrastructure is accelerating, supporting the rapid growth of the military electronics industry. Outfield tests are crucial in the research and development process of military electronic devices since the laboratory environment differs from practical application conditions [1,2]. The outfield test in the military electronics industry is defined as electronic devices leaving the original research and production site for a specific site for practical application testing [3]. In order to conduct testing, it is necessary to transfer electronic equipment and products to the field. There are many construction safety hazards present in the process, particularly during the packaging and strengthening of the devices before transfer, the installation and erection phase, the product testing stage, and the post-test dismantling.

Construction is a hazardous occupation, with a high rate of occupational injuries. In 2020, construction was responsible for roughly 22% of fatal occupational injuries in the United States [4], and in China, 689 safety incidents involving housing and municipal engineering led to 794 fatalities [5]. The top two accident classifications with the highest proportion are falling from heights and object contact, which are commonly caused by failure to comply with the required specifications and improper interactions with the environment. The same risks are present for the outfield tests. The most common workplace accidents in outfield sites include falls, fires, lifting, electrocution, scalding, object contact, and mechanical injury. Most of these hazards result from inadequate precautions and

incorrect use of tools, which result in irreparable loss and significantly limit the development of the national science and technology industry. A thorough safety inspection is necessary for outfield tests at least once each week. To tackle hazards effectively and record the findings of the safety checks, the safety manager should perform daily on-site safety inspections. Before executing different operations, employees must also go through rigorous training and instruction. However, they might not be able to fully comply with the safety requirements because of stress, fatigue, or negligence. Regulations may also alter based on the workplace. Therefore, it is necessary to conduct an automatic text analysis study on redundant safety regulations and laws.

The Internet of Things (IoT) and advanced vision devices have been used in construction safety management to help monitor construction sites [6–8]. According to the outfield work safety requirements [3], key areas, such as the entrance, exit, test area, and equipment area, must be entirely covered by monitoring and guarded by personnel throughout the day. However, such a method cannot quickly detect and eliminate hazards. Additionally, manually checking surveillance for compliance with safety regulations is labor-intensive and error-prone. Thus, an automated safety inspection is needed to facilitate the safety monitoring of outfield construction workers and hazard identification. Visual data from key areas and safety regulations text can meet these objectives. Large amounts of image and text data make the automatic extraction of key information challenging, but the development of artificial intelligence has mostly solved this issue. Deep-learning-based models may automatically extract complex features and key information from a large amount of data, which contributes to more efficient safety management. Therefore, some researchers utilized deep learning convolutional neural networks [9–12] to process on-site images directly. These methods can be more effectively used on the construction site.

Deep-learning-based computer vision methods have been widely used for construction safety management in the past five years, with the majority of studies concentrating on object detection [10,13], proximity measurement [7,14], and action recognition [11,15]. However, these studies rely on visual feature extraction and lack semantic understanding to parse visual scenes accurately. Subsequently, deep learning image understanding is gradually developing from low-level feature extraction to high-level semantic learning (e.g., scene understanding [16], visual question and answer [17], image caption [18]). Recently, some researchers have proposed approaches to parse on-site construction scenes by detecting the semantic relations between objects [19,20]. The majority of the methods that are now available rely on vision models to identify construction hazards. However, they do not automatically extract textual information, such as safety standards and regulations. It will be a significant challenge to the practicability of current vision-based methods to adapt to the diversity of safety regulations. Thus, additional research on automatic safety inspection that integrates visual and textual information is required.

In this work, we aim to design a framework that integrates visual and linguistic information to enable outfield on-site safety inspection while addressing the challenges mentioned above. First, a BERT-based safety regulations processing method is presented for automatically extracting text information. The key semantic information related to hazard identification is represented in a structured form. The method presented here extracts and represents textual features based on natural language processing (NLP). Subsequently, a vision-based scene parsing approach is developed to process on-site images. The visual features of the on-site workers and their interacting objects (e.g., PPEs) are extracted based on Mask R-CNN [21]. The interaction relations between the object instances are further predicted by on-site scene graph generation. Additionally, an automatic safety checking process is established based on relational triples analysis by integrating extracted visual and textual relational information.

The key contributions are summarized as follows:

- We propose a BERT-based text processing approach to extract key textual information from Chinese safety regulations automatically.

- We develop a deep-learning-based scene parsing method for detecting visual interactions between objects in on-site images. The textual and visual information is then integrated to implement safety inspection and hazard detection.
- Experiments on self-built and public datasets show that the proposed approaches effectively extract relational information from visual and textual modalities. Demonstration of PPE compliance checking on two scenes show the feasibility of our proposed method.

2. Related Work

Recently, many studies have employed new technologies on construction sites to identify occupational hazards. Guo et al. utilized an RGB-D camera (Kinect v2 Sensor) to collect the worker behavior information by capturing the depth images [9]. Kelm et al. applied radio-frequency identification (RFID) to assess the PPE compliance of workers [22]. Yan et al. developed a real-time motion warning PPE based on wearable Inertial Measurement Units (IMUs) [23]. Gheisari et al. conducted a survey to indicate the safety activities that can be improved using unmanned aerial vehicles (UASs) by monitoring construction sites [24]. However, the RGB-D sensors are incompatible with complex and unstable outfield situations due to their sensitivity to solar radiation. Wearable sensors typically prohibit employees from performing their tasks effectively and utilizing UAVs to obtain complete surveillance coverage would be too costly. The restrictions have been greatly improved with the quick progress of deep-learning-based computer vision. Many studies developed the vision-based model to process on-site images to identify hazards. The majority of these methods focused on detecting on-site resources based on classic neural network approaches, such as Region Proposals (Faster R-CNN [25], Mask R-CNN [21]), Single Shot MultiBox Detector (SSD), [26], and You Only Look Once (YOLO) [27–29]. Fang et al. utilized the Faster R-CNN network to detect construction workers' non-hardhat-use in different site conditions [30]. Kim et al. proposed a UAV-assisted monitoring video method based on YOLO-V3 that enabled the detection of struck-by hazards [31]. Fang et al. utilized Mask R-CNN to identify workers' unsafe behavior to avoid falls from heights [32]. Wu et al. proposed a one-stage convolutional neural network based on SSD for hardhats wearing and corresponding color detection [33]. Wang et al. adopted the MobileNet as the backbone for detecting workers wearing hard hats on construction sites [34]. The aforementioned methods showed their capacity to support on-site safety checks. However, object detection methods based on neural networks only detected the categories and locations of objects. These methods also could not perform high-level visual semantic understanding of the on-site scene because they do not explore the rich semantic information and interactions between objects.

Many methods have been developed to explore the semantic relations between on-site objects to address these issues. Xiong et al. integrated a visual relationship detection network with construction safety otology to identify hazards in the workplace [19]. Zhang et al. [35] proposed an automatic hazard identification method combining object detection and ontology in the foundation pit excavation scene. Wu et al. [36] developed a method for hazard identification that integrated Mask R-CNN object detection and ontology to detect three types of spatial relations (on, overlap, and away). Unlike spatial-based interaction detection, Tang et al. [20] proposed a human–object interaction (HOI) recognition model for checking personal protective equipment compliance, which explored three action-based interactions (standing, using, and wearing). To extract richer semantic information, Wang et al. [37] proposed a semantic information extraction method integrating object detection and image captioning to facilitate on-site safety management. However, it did not perform automatic checking on regulatory rules.

Even though the rich semantic information for the vision-based model makes it easier to understand on-site scenes, there is still a lack of connection between visual information and domain-specific knowledge [38]. Ontology can generate formatted knowledge representations from domain-specific knowledge and has been widely adopted in the

architecture, engineering, and construction (AEC) industry. Both methods [19,35,36,39] processed regulatory rules based on ontology for on-site hazard identification. Ontology development and inference processes often utilize Protégé (software), the most popular and widely used ontology editor. However, it cannot be easily integrated into a unified model with deep neural networks. With the achievements of natural language processing (NLP) techniques and large text corpus pre-trained language models, NLP-based methods have been widely adopted to perform occupational accident analysis. Improving relation extraction and integrating multi-source information are major concerns and problems in on-site construction. Chen et al. [40] proposed a graph-based framework to process regulatory rule text and images for on-site occupational hazards identification, which integrated NLP-based syntactic structure analysis and a deep learning visual model. Zhang et al. proposed a cross-modal automatic hazard inference method integrating scene graph generation and BERT-based text classification. Domain knowledge combined with on-site images can make it easier to identify hazards by integrating on-site visual features and regulations textual features. In order to facilitate safety inspection and hazard inference, it is essential to narrow the semantic gap between natural language and vision. Based on previous studies, we propose a method that reduces the semantic gap by integrating visual and textual relational information. Additionally, we design an automatic control process for performing automatic security checks.

3. Methodology

3.1. Research Philosophy and Design

The purpose of this study is to develop an on-site hazard identification framework that automatically performs on-site scene safety inspections against outfield safety regulations. Based on the aforementioned literature review, deep-learning-based computer vision algorithms have recently gained popularity for identifying hazards on construction sites. Scene graph generation methods also showed promising results for construction scene understanding, but there is still a gap between visual information and domain-specific knowledge. Many research studies employed ontology to generate new forms of expression of domain knowledge. These ontology-related studies need to be conducted across different platforms, making it more challenging to combine with computer vision models. The development of NLP and large pre-trained language models has substantially improved knowledge mining in the construction industry. BERT is a language model with high accuracy and ease of fine-tuning that can implement automatic information extraction on safety regulation texts. However, few studies about employing NLP and vision-based methods for construction safety inspections integrate textual relational information with visual relational information.

Consequently, we integrate on-site scene graphs and BERT-based information extraction to develop a framework for identifying construction hazards. Our proposed deep-learning-based methodology consists of three main processes: data processing, model development, and model validation. (1) Data processing: A schema-based Chinese work safety regulation dataset was established by manual annotation. The on-site image dataset was established by open resources crawler and manual annotation. We also employed widely used public benchmark datasets to assist with model validation. (2) Model development: This paper builds a BERT-based model for extracting textual information and a scene graph generation model for parsing on-site visual scenes. Additionally, a simple automatic control method is applied to match the relational triples output from the linguistic and visual model for automatic safety checking and hazard identification. (3) Model validation: This paper uses qualitative and quantitative analytical methods to validate the performance of the models on the self-built and public datasets.

The overview of our proposed framework and the detailed procedures for data processing, model development, and validation are then described.

3.2. Overview of Our Proposed Framework

This paper proposes a framework for safety inspection based on the NLP-based regulations information extraction and the on-site scene graphs. The following functions are available through our framework: (1) The capacity to automatically parse regulations text: A BERT-based model is used to extract textual relational information from safety regulations to parse outfield work safety requirements. (2) The capacity to generate on-site scene graphs: A visual network and semantic modeling are used to detect the objects and relations in an image. The on-site scene graphs are generated as a result. (3) Safety checking and hazard inference capability: Visual and textual relational information are integrated to perform PPE compliance safety checking and on-site hazard identification. Figure 1 shows the overall pipeline of our proposed framework, which consists of a BERT-based information extraction module, an on-site scene parsing module, and an automatic safety checking process. We will detail each part of our framework.

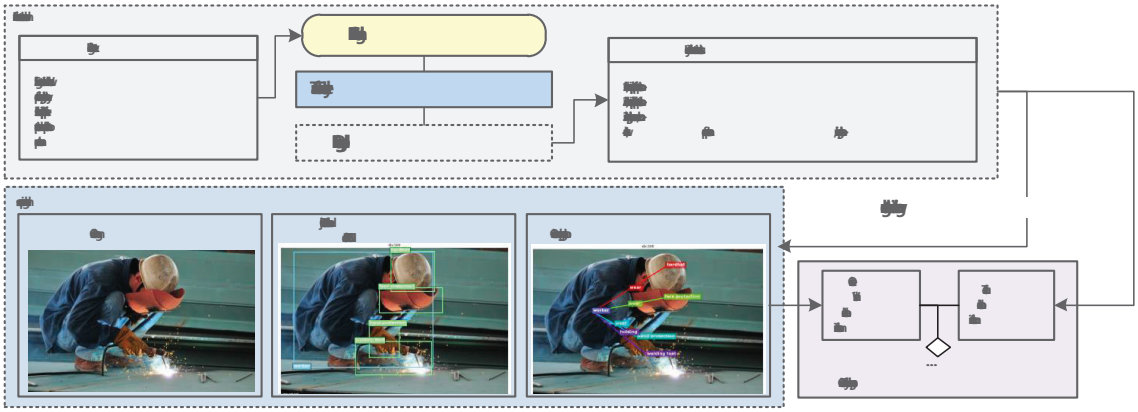

Figure 1. Overall pipeline of the proposed framework.

3.3. BERT-Based Safety Regulations Information Extraction (IE) Module

The outfield test involves vast quantities of text data, including work safety standards, construction regulations, and terminology. The manual extraction of these safety requirements is labor-intensive, costly, and error-prone. This paper proposes a BERT-based IE model for jointly extracting entities and relations from Chinese domain-specific text to address these issues. IE can automatically process regulations text and extract key information to identify the hazards of non-compliance.

Relation extraction is a core task in IE, and its goal is to detect specific types of entities and the relations between entity pairs from unstructured natural language text. It is essential for ontology learning and building knowledge bases [41]. Our proposed IE enables the extraction of multiple relational triples. A relational triple contains a subject entity (S), an object entity (O), and a semantic relation R between entities. A triple is often formalized as $< S, R, O >$. For example, <worker, wear, helmet>. As Figure 2 shows, the proposed model comprises a BERT-based encoding layer and a BIEO tagging decoder.

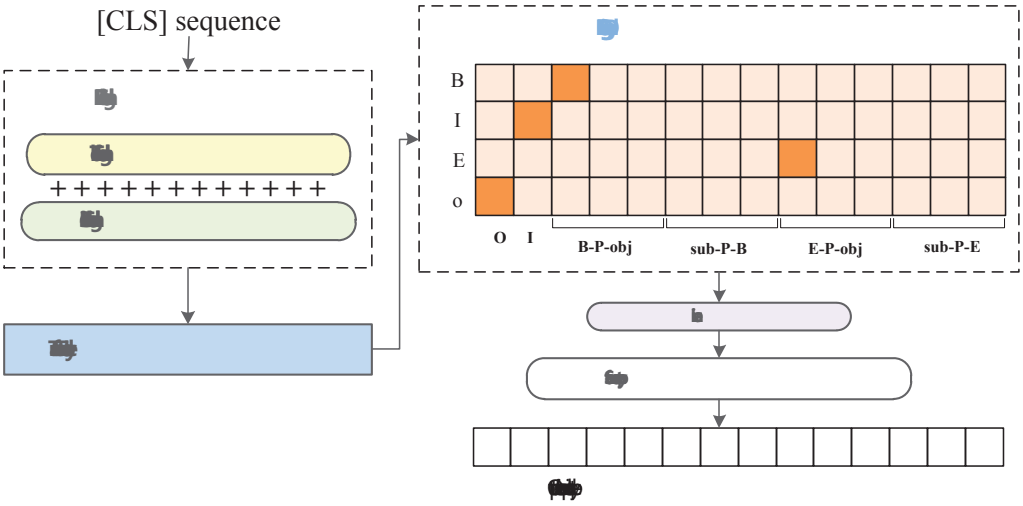

Figure 2. An illustration of proposed regulations information extraction (IE) module.

3.3.1. Pre-Processing

We first design a schema to guide domain knowledge information extraction. Schema can clearly define the entity type and relation in the knowledge bases. A schema is a set of relational triple templates [42]. Each schema contains a subject entity type, a predicate (semantic relation), and an object entity type. Defining the schema is equivalent to constructing ontology in knowledge graphs and helping integrate domain knowledge through formal conceptualization.

The knowledge sources utilized to design the schema include the work safety standardization requirements of the outfield test, the experience of previous scholars, and public information retrieval. We pick the eight most commonly used relation types by analyzing the domain text information. Table 1 shows some examples. Next, we select the related relational triples from the regulations text according to the schema. The relation in a triple should be equivalent to the predicate in the schema, and the subject and object entity correspond to the instances of the pre-defined object types, respectively. These relational triples are used to parse the text and generate annotations for training the IE model.

Table 1. Schema examples in regulations dataset.

Subject Type	Predicate	Object Type	SPO Example
person	be equipped with	PPEs	S: worker, P: be equipped with, O: eye protection
person	perform…operations	working operations	S: worker, P: perform…operations, O: welding operations
working operations	occurrence	occupational injuries	S: welding operations, P: occurrence, O: burns

3.3.2. Encoding Layer

This paper adopts the BERT [43] as the encoding layer for input regulations text. We will briefly review the BERT, a multi-layer transformer-based language representation model, which contains the input embedding space (denoted by W) and N identical Transformers block modules. The input to BERT is a sequence of words where 15% WordPiece tokens are masked. The i-th input word is converted into a one-hot vector $x_{(i)}$, and the positional embedding per input token is denoted as p_i. The original words in the input sentence are translated into N Transformer blocks. The Transformer encoder's hidden dimension is represented as h, and h_j is the hidden state of the input sentence at the j-th layer.

$$h = W * x_i + p_i \tag{1}$$

$$h_j = \text{Transformer}(h_{j-1}), j \in [1, N] \tag{2}$$

The output of the Transformer block is a sequence of contextualized word embeddings (denoted by O), and the output of the pre-training BERT is a word score vector (denoted by y^{x}) for each masked word. The word score vector for i-th position masked word is extracted from the Transformer block output by transposing: $y^{\text{x}} = W^{\top} O^{(i)}$.

For the i-th position masked word, the pre-training BERT's original prediction is to make the normalized exponential for $y^{\,textx}$ infinitely close to the masked word one-hot vector. This is achieved by minimizing the loss between the $y^{\,textx}$ and the masked word one-hot vector. The $0/1$ sequence tagging in this paper is a multi-label classification task. Softmax + cross-entropy is introduced here as a multi-label categorical loss function L_m:

$$L_m = \log\left(1 + \sum_{i \in K} e^{s_i}\right) + \log\left(1 + \sum_{j \in L} e^{-s_j}\right) \tag{3}$$

where s_i and s_j denote the non-target and target class, and K and L denote the category sets of negative and positive samples.

3.3.3. BIEO Tagging Decoder

The semantic representation of regulations text is first output through the BERT-based encoding layer described in the previous section. The BIEO tagging scheme for decoding is introduced in this section. Finally, we explain how to extract key textual relational triples from the text.

We utilize BIEO signs to distinguish entities, drawing inspiration from the classic BIO tagging scheme [44]. The BIEO signs represent the word position within the entity (Begin, Inside, End, Outside). A set of $|N|$ predicates in the pre-defined schema is used to determine the relation type between entities.

Figure 3 is an example to detail our tagger. The input sentence from outfield safety regulations is "Workers should be equipped with face protection during welding operations". It contains two overlapped triples: $<$ worker, perform…operations, welding operations$>$, $<$ worker, be equipped with, face protection$>$, where "perform" and "be equipped with" are the pre-defined predicates in the schema. The words "worker, welding operations, face protection" are all related to the final extracted entities. Next, each word in the input sentence is tagged based on BIEO signs. For example, the word "face" is the first word of the object entity "face protection" and is related to the subject (sub) "worker" and predicate "be equipped with" (P1). So its tag form is $sub - P1 - B$. Similarly, the last word of the object entity "protection" is tagged as $sub - P1 - E$. All words between an entity's first and last word are labeled as "I" (inside). Moreover, the other words unrelated to the final relational triples are labeled as "O". We set B and E signs for each subject entity and object entity, a total of $4|N| + 2$ tags ($2|N|$ "B" tags, $2|N|$ "E" tags, 1 "I" tag, and 1 "O" tag) are generated. The BERT-based encoding layer's output is the tagging decoder's input. Finally, the decoder matches the "B" and "E" signs of the subject and the object to form triples, which is performed by matching tags separated by the total number of predicates. The number of subject and object tags are doubled, so more entities and relational triples can be extracted from a sentence. More comprehensive information can also be obtained from the safety regulations.

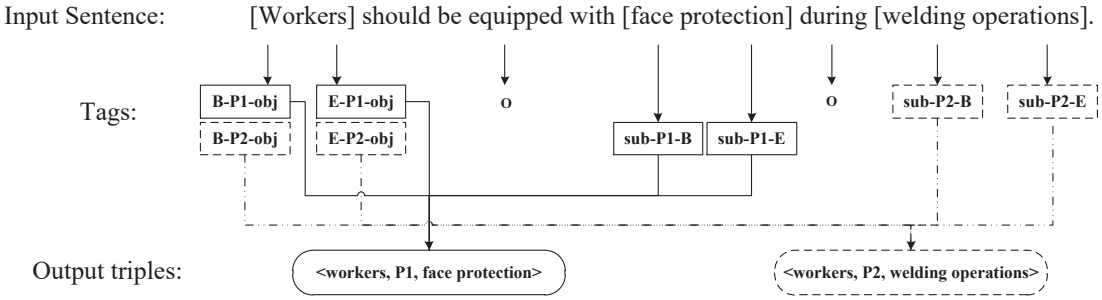

Figure 3. Gold annotation for an example sentence based on our tagger.

3.4. On-Site Scene Parsing Module

The IE module can automatically extract the key information from the regulations. In addition, detecting the visual relations and parsing the on-site scene is necessary for additional research on automatic hazard inspections. The $S, P, O >$ (SPO) relational triples can also express visual relations. An on-site scene graph has the perceptual capacity to recognize the position, class, and interrelationship of entities. The visual relations between entities could be action-based, comparative, or spatial. As shown in Figure 4, the on-site scene parsing module extracts visual and semantic features for each visual relationship proposal.

(1) Semantic modeling: We first introduce a pre-trained fastText model to map the word vectors into an embedding space that preserves higher semantic similarity. (2) Visual network: Based on earlier research [16], we build a separate CNN branch to extract the predicate feature from interactive areas of the subject and object. The structure of the on-site scene parsing network is shown in Figure 4.

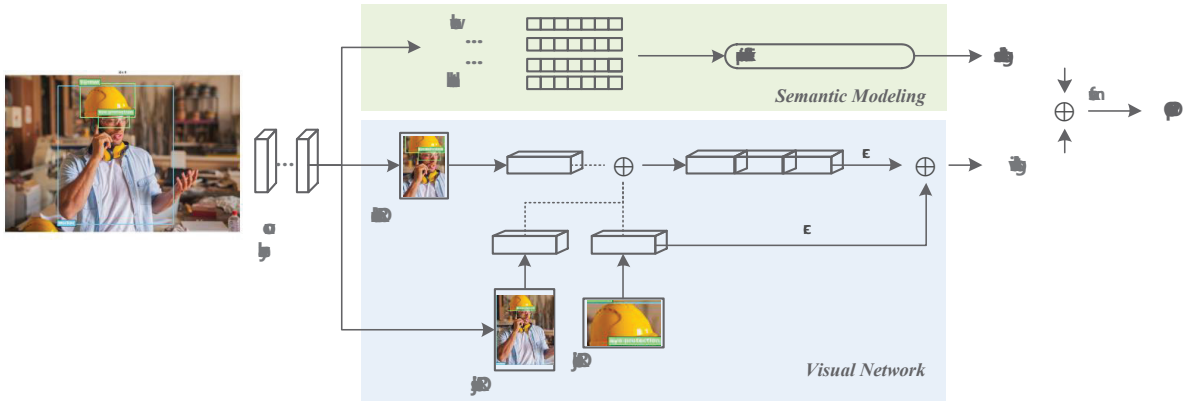

Figure 4. The on-site scene parsing module architecture.

3.4.1. Semantic Modeling

The relation and object labels can be properly initialized using pre-trained word vectors. Here, we introduce pre-trained word embedding, which maps the source words into an embedding space while maintaining higher semantic similarity. We initialize the word embeddings for the objects and predicates with pre-trained two million word vectors fastText learned on Common Crawl [45]. The word embedding is computed by representing a word as a bag of n-grams [46]. This word-internal information-rich embedding outperforms random initialization and the word2vec tool, which consider words as independent representations and disregard the morphological features within words. Inspired by stacked motif networks [47], we condition on objects when predicting predicate class. The statistical analysis shows that object labels are highly predictive of relation labels. For each on-site image, we

compute the empirical distribution $\hat{p}(p \mid s, o)$ over the predicate between the subject and object in the training annotations. We assume the testing set has the same distribution as the training set. In the training annotations set, p represents the instances of the predicate class given the subject class (s) and object class (o).

3.4.2. Visual Network

We employ Mask R-CNN [21] as the object detector that generates the ROI (region of interest) feature map for on-site images. The ROI feature of the subject and object entities (f_{sub}, f_{obj}) are extracted from the convolution layer. The visual features for relations usually come from the interactive areas of subjects and objects, so we build a separate CNN branch to extract the predicate feature from interactive regions of the subject and object based on previous research [16].

The ROI feature of relations (f_{rel}) is also extracted from the CNN branch, the same as the structure of the entity convolution layer. These features are then mapped to the hidden layer nodes through the multilayer perceptron (MLP) and generate hidden features h^s, h^r, and h^o. The entity embeddings of subjects and objects are output:

$$y^s = g(h_k^s) = g(\sum_{i=0}^{M} \mathrm{w_{jk}^s} h_{jk}^s) \tag{4}$$

$$y^o = g(h_k^o) = g(\sum_{i=0}^{M} \mathrm{w_{jk}^o} h_{jk}^o) \tag{5}$$

where i, j, and k denote the node of the input layer, hidden layer, and output layer, respectively; $i \to j \to k$ represents the relative connection among the different layers of the MLP; $g(h)$ is the activation function; and w_{ij} denotes the weight of the current j layer. Then, we generate a fusion relation embedding h_j^r by concatenating the subject and object embeddings with h_{j-1}^r:

$$\begin{aligned} h_j^r =& \mathrm{CONCATENATE}_{(\mathrm{y^s} + \mathrm{h_{j-1}^r} + \mathrm{y^o})} \\ =& y^s \oplus \sum_{i=0}^{M} \mathrm{w_{i(j-1)}^r} h_{i(j-1)}^r \oplus y^o. \end{aligned} \tag{6}$$

Finally, the relation embeddings y^r are output through a fully connected layer:

$$y^r = g(h_k^r) = g(y^s \oplus \sum_{i=0}^{M} \mathrm{w_{jk}^r} h_{jk}^r \oplus y^o). \tag{7}$$

As shown in Figure 4, the logits (the unnormalized class probabilities) from the semantic modeling (logit(p_{sem}) and visual network (logit(p_{vis})) are output through the final fully connected layer. The output logits are added and then perform softmax normalization to obtain the final probability distribution of the predicate class:

$$p(pre) = \mathrm{softmax}(\mathrm{logit}(p_{sem}) + \mathrm{logit}(p_{vis}). \tag{8}$$

3.5. Automatic Work Safety Checking Approach

To conduct on-site safety inspections, we design an automatic control process, as shown in Figure 5. This process aims to identify any deviations between the on-site construction process and the safety regulations by comparing the visual relations output from the on-site scene graphs with the key textual relational information derived from the IE model. The correction instructions are then created and relayed to the safety managers and the on-site workers until the current production step is finished. Integrating on-site visual understanding with plain-text information extraction can make automatic safety inspections effective.

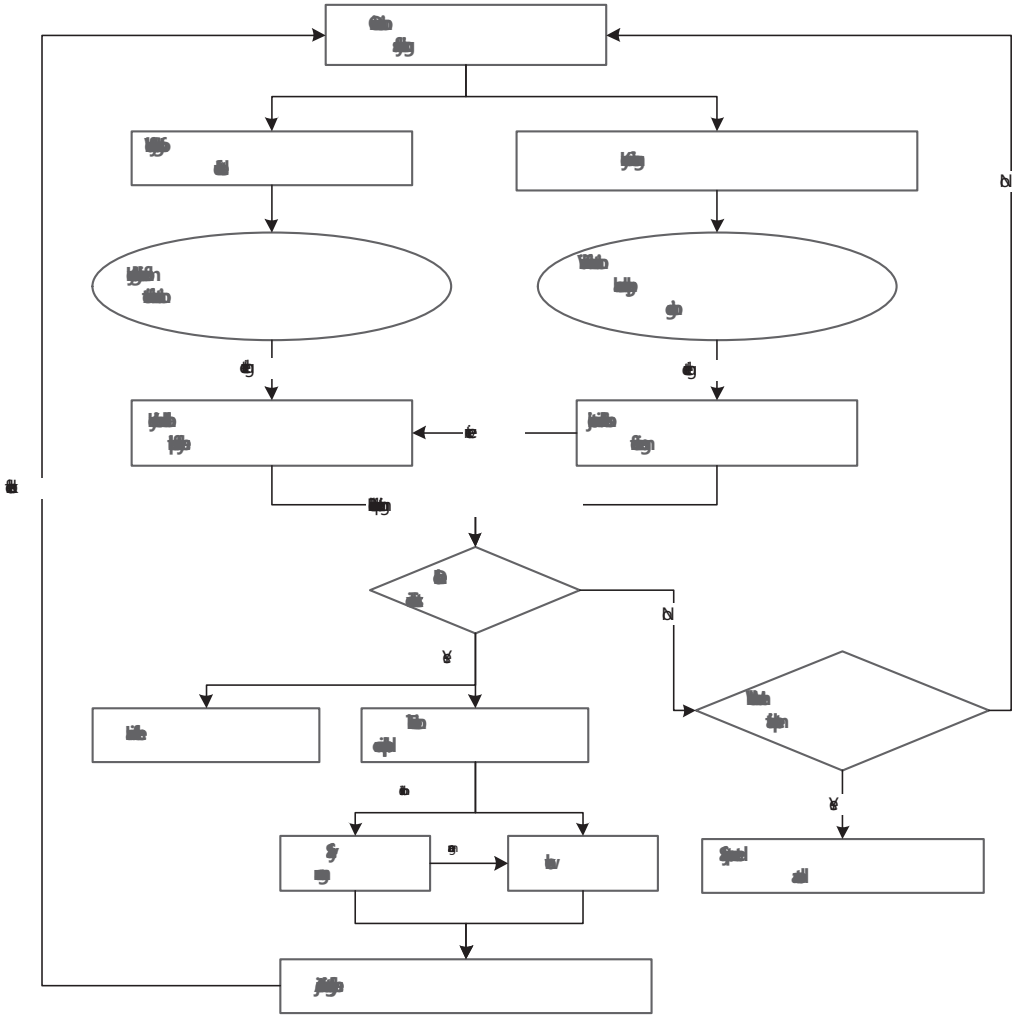

Figure 5. The flowchart of the work safety checking process.

In order to conduct PPE compliance checking, we set a relational information matching procedure to analyze the occurrence of each PPE-related triple ($Y_{ppe} = (P_{ppe}, S_{ppe}, O_{ppe})$) in the set of visual relational triples ($Y_{vis} = (P_{vis}, S_{vis}, O_{vis})$). If no worker is detected in an image from object detection, this scene is not applicable for the PPE compliance checks. Suppose the detected visual relational triple set for each worker contains all the PPE-related textual relational information ($Y_{ppe} \subseteq Y_{vis}$, $Y_{ppe} \neq \varnothing$). In that case, it means that the PPE inspection of workers meets the requirements. If no textual relational triples or part of them exists in the worker-related relationship triples ($Y_{ppe} \cap Y_{vis} = \varnothing$ or $Y_{ppe} \cap Y_{vis} = A$, $A \subsetneqq Y_{ppe}$), it means the PPE checks do not comply with the regulations.

4. Experiments and Results

The experimental details are described in this section. We used key sentences from the outfield work safety standardization to test the feasibility of the IE method. An image dataset for PPE checking was constructed to test the feasibility of our proposed on-site scene parsing method. Our proposed IE and scene parsing approaches on both self-built and public benchmark datasets showed good performance.

4.1. Automated Textual Information Extraction for Outfield Safety Regulations

The proposed IE method was implemented and tested on selected outfield work safety regulations related to hazard identification. Experimental results on the self-built and large-scale public datasets performed well in information extraction.

4.1.1. Schema-Based Dataset Establishment and Model Training

We collected outfield work safety regulations from open network resources and documents released by the enterprises. These texts were first cleaned up by correcting typos and spelling mistakes. The full texts were then split into several separate sentences. Next, we deleted the irrelevant and meaningless sentences. According to the pre-defined schema, we chose 336 key sentences related to on-site hazards for training. We labeled the entities and relations for each candidate sentence using the label studio tool to generate relational triples that follow the schema. Finally, 1218 relational triples in total were obtained.

In the training process, a total of 269 and 67 sentences were, respectively, used for training and validating. We employed the pre-trained BERT-Base Chinese model (12-layer Transformer, 768-hidden, 12-heads, 110 M parameters) [43] for fine-tuning. The learning rate was set to 1×10^{-5}. Additionally, the training process was early stopping when the F1 score on the validation set did not increase for 10 sequential epochs.

4.1.2. Results of Regulations Information Extraction

We followed the evaluation metrics from Fu et al. [48], the standard precision (*Prec.*), recall (*Rec.*), and F1 score (*F*1) were adopted to evaluate our IE model. A relational triple was considered correct only if the two entities and a predicate type were all correct.

To demonstrate the effectiveness, our IE model was first evaluated on the widely used public dataset DuIE [42]. DuIE is a large-scale Chinese dataset built by Baidu Inc for relation extraction, consisting of 210,000 sentences covering 49 predicate types. The proportion of overlapping pattern sentences in the DuIE dataset is higher than the other widely used public datasets, such as NYT[49] and WebNLG [50]. Therefore, extracting information from DuIE is more challenging. As Table 2 shows, our IE model achieved encouraging *Prec.*, *Rec.*, and *F*1 of 77.3%, 82.1%, and 79.6% on the DuIE dataset, respectively, which indicated good performance. For the selected outfield work safety regulations, it achieved *Prec.*, *Rec.*, and *F*1 of 80.1%, 78.6%, and 79.3%, respectively.

Table 2. Results of information extraction on DuIE and Regulations datasets.

Task	DuIE			Regulations Dataset		
	***Prec.* (%)**	***Rec.* (%)**	***F*1 (%)**	***Prec.* (%)**	***Rec.* (%)**	***F*1 (%)**
Information extraction	77.3	82.1	79.6	80.1	78.6	79.3

According to the statistics, most of the sentences in the selected regulations belong to overlapping patterns, making information extraction more difficult. If two relational triples share the same entity pairs or two triples contain at least one overlapping entity but do not share the same entity pairs, the sentence belongs to an overlapping pattern. Next, we conducted a qualitative study to prove the IE approach of extracting key information from work safety regulations. Table 3 shows the results of processing the overlapping sentences in the regulations text.

Table 3. Case study of regulations information extraction.

Instance Examples	Results
(1). Workers should be equipped with face protection and hand protection to prevent burns when performing welding operations.	<workers, be equipped with, face protection> <workers, be equipped with, hand protection> <welding operation, occurrence, burn> <workers, perform…operations, welding operations>
(2). Workers working at height should be equipped with hard hats to prevent head injury from falls.	<workers, be equipped with, hard hats> <working at height, occurrence, head injury from falls> <workers, perform…operations, working at height>

4.2. On-Site Scene Graph Generation for Visual Information Extraction

On-site visual information extraction based on the proposed scene parsing method is described in this section, along with its implementation and performance. We selected several types of PPE objects and built an on-site scene graph dataset to test the feasibility of the proposed method.

4.2.1. On-site Scene Graph Dataset Establishment

First, PPE safety checking rules were selected from outfield work safety regulations based on the IE model in Section 4.1. Second, on-site images were crawled from open image resources based on these safety rules, then manually selected. A total of 829 images were selected as the final candidates and then annotated by the LabelImg annotation tool [51] and PySimpleGUI.

The image annotation process consists of entity annotation and relation annotation. As shown in Figure 6, the bounding box and box labels were used to identify the locations and types of entities. The subject, object, and predicate attributes were adopted to describe the visual relational triple. Finally, the entities and relations annotation results were organized into the Visual Genome [52] format to facilitate model training. As the statistics in Table 4 show, we obtained a total of 2316 visual relations annotations.

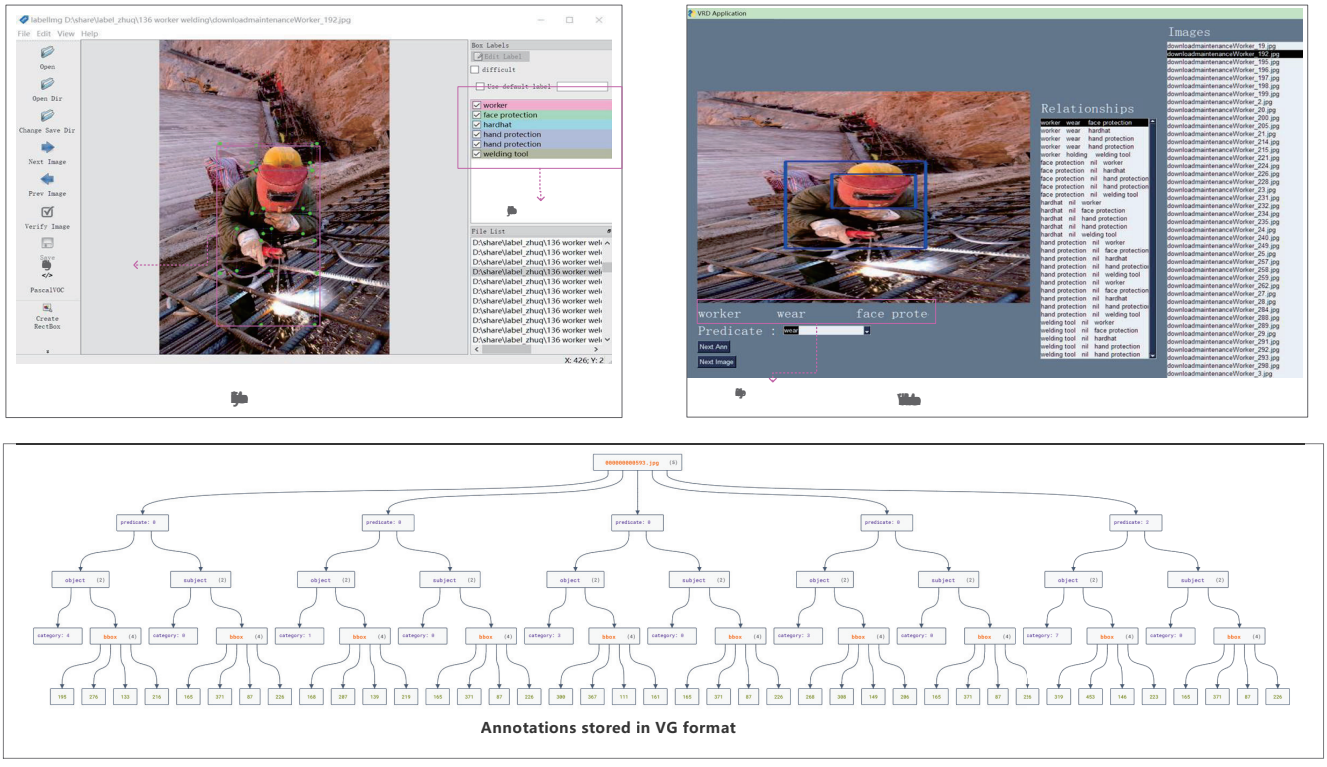

Figure 6. Interface and storage format of the image annotation.

Table 4. Visual relation annotations statistics.

Subject	Predicates	Objects	Instances
Worker	Wearing	Hard hat	1141
		Eye protection	220
		Hand protection	364
		Face protection	169
	Holding	Welding tool	422

4.2.2. Results of On-Site Scene Graph Generation

After the annotations, the dataset was randomly divided into a training set (80%) and a testing set (20%). The basic learning rate was set to 1×10^{-3}, and optimization was performed via momentum SGD. For the object detection task, we used the evaluation metrics from He et al. [21], which include average precision (AP) (averaged over IoU thresholds), AP50, AP75, APS, APM, and APL (AP at different scales). The test results of the object detector using the ResNeXt-101-FPN backbone are displayed in Table 5 .

Table 5. Results of object detection.

Task	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Object detection	0.768	0.841	0.783	0.703	0.652	0.800

Following Zellers et al. [47], we conducted three tasks to evaluate the performance for relation detection: (1) Scene graph generation (SGGen) predicts the subjects and objects appearing in an image and all labels(the class label of subjects, predicates and objects). (2) Scene graph classification (SGCls) predicts the class labels for the subjects and objects given ground-truth bounding boxes and predicts the predicates labels. (3) Predicate classification (PredCls) predicts predicate labels are given ground-truth subjects and object bounding boxes and labels. We use Recall@K (R@20, R@50, and R@100) as the evaluation metrics for these three tasks. Recall (recall $= \frac{TP}{TP+FP}$) is the ratio of the true relationship in the top-K confident relation predictions in an image. Table 6 shows the evaluation results (R@20) of scene graph generation using the ResNeXt-101-FPN backbone.

Table 6. Results of scene graph generation using the ResNeXt-101-FPN backbone.

Task	SGGen	SGCls	PredCls
On-site scene graph generation	0.507	0.855	0.978

Our scene parsing model was also evaluated on a benchmark dataset: Visual Relationship Detection (VRD) dataset [53], which includes 5000 images with 100 object categories and 70 predicate categories.

Figure 7 displays more qualitative results of our model on two datasets. Qualitative results on the self-built on-site dataset mainly showed the visual relationship between workers and PPEs. The results on the VRD dataset showed the ability of our model to extract various types of interactions such as spatial, comparative, or action-based.

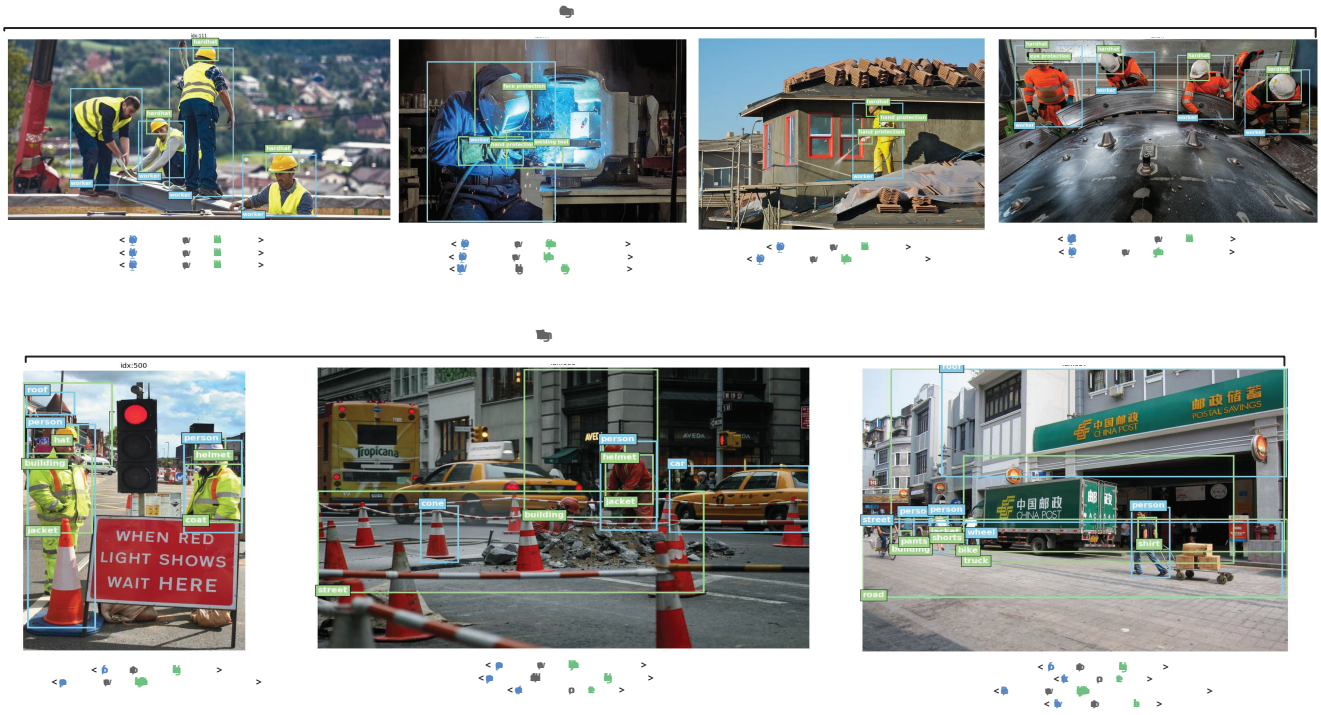

Figure 7. Qualitative examples of scene graph generation on self-built and VRD datasets.

4.3. Case Study of Hazard Identification

As shown in Figure 6, the critical steps of automatic safety inspection are summarized as follows: (1) scene graph generation for the key work sites, (2) key information extraction of safety rules, and (3) matching the two kinds of relational triples output from the visual and linguistic model for PPE compliance checking. Next, we conduct a case study for the specific scenes. We selected two types of work scenes to illustrate the process for safety checking and hazard inference, as shown in Figure 8.

In the outfield working environment, it is required that monitoring equipment be set for the key sites (e.g., working environment of work-at-height). The video surveillance images are utilized as the input of the on-site scene parsing module. As Figure 8 shows, the proposed scene parsing model parsed the images of the work-at-height and welding operation, and the visual relational triples were output. Next, the corresponding safety regulation texts were retrieved from the database according to the specific construction scene, and the key textual relational information was extracted through the IE module.

In order to check PPE compliance, we set up a triple keyword matching procedure to analyze the occurrence of each textual relational triple in the set of visual relational triples and output three types of labels: (1) On-site workers comply with the regulations (Yes). (2) The PPE safety checks do not meet the regulations (No). (3) The scene is not applicable for the PPE compliance checks (N/A). Finally, conditional judgment was utilized to perform hazard inference.

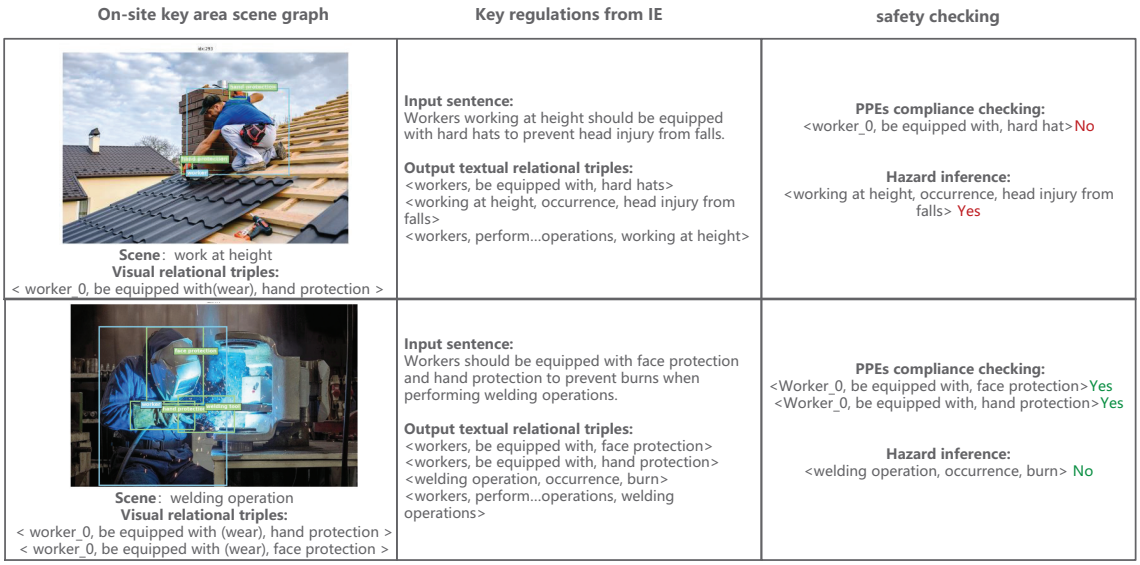

Figure 8. Examples of PPEs compliance checking and hazard inference.

5. Discussion

This paper proposed a framework combining NLP and computer vision to achieve on-site safety inspection and hazard identification in outfield tests. To enable the possibility of compliance checking, the framework extracted information from safety regulations text using a BERT-based IE method. Meanwhile, on-site scene graphs were generated to detect the visual relations between objects to enhance the on-site scene understanding. Compared with previous research of on-site hazard identification based on deep learning, our study has the following advantages. (1) In safety regulations processing, this paper proposed a novel information extraction model based on the BERT encoder and BIEO tagging scheme. Using BIEO to tag sentences can extract more textual relational triples from the regulation texts with more overlapping patterns. The extracted textual relational information can be used for automatic hazard identification against safety rules. (2) In on-site image processing, we designed a scene graph generation method for scene understanding, which can better detect the visual relational information combined with semantic modeling. Additionally, we designed an automatic control process based on matching textual and visual relational information to conduct safety inspections. The proposed automatic work safety checking approach can implement hazard inference for the on-site scene graphs with the regulations retrieved from matching relational triples.

The effectiveness of the proposed method was experimentally evaluated from qualitative and quantitative perspectives. In text processing, the proposed IE model was trained with a self-built Chinese text dataset of outfield safety regulations and a large-scale public Chinese dataset, DuIE. The proposed IE model achieved F1 scores of 79.3% and 79.6% on two datasets, respectively. The high-performance results of F1 scores show that the proposed IE model is competitive among various text information extraction methods and can accurately extract the key information related to hazard identification from safety regulations text. Qualitative results show that the IE model can automatically extract key information from a large amount of text, represented as a textual relational triple such as <workers, be equipped with, hard hats>. In image processing, the scene parsing model was trained with a self-built on-site image dataset and a public benchmark dataset VRD. The average precision (AP50) result for the object detector on the on-site test images is 84.1%. The Recall@20 to evaluate the challenging SGG task is 50.7%. The proposed scene parsing model is competitive in parsing scenes and can extract visual relations between

workers and their interacting objects from images according to the high-performance results of precision and recall@20. Qualitative results show the ability of the scene parsing model to extract various types of relational triples (e.g., spatial, action-based) represented as <worker, wear, hard hats>, <worker, holding, welding tool>, and <person, standing behind, building>. In the safety checking process, the analysis results on different on-site scenes showed that processing the output from the textual and visual modules based on relational information matching enabled automatic safety inspections and hazard inference.

Although our model performed well on various evaluation tasks and metrics, there are still some drawbacks. We next discuss the limitations and future research directions of this paper.

5.1. Limitations

The automatic detection of hazards in outfield tests is facilitated by rules information extraction and scene graph generation. However, there are still some limitations. First, deep-learning-based models require a large amount of training data to perform effectively but establishing the large-scale text and image datasets for the construction site is labor-intensive. We constructed two domain-specific datasets for the information extraction and scene graph generation models, which could be better applied to outfield construction sites. However, the size of the datasets is inadequate to cover most outfield construction hazards, so we need to expand and enrich the annotation data. Second, we mainly detected action-based visual relations in the scene graph generation task without carefully considering geometric and spatial aspects. Due to the impossibility of proximity measurements, utilizing the spatial relations detected by the scene parsing model is hard for safety inspections. The research on pose estimation, which is based on the human skeleton and can provide more precise information about a person for hazard identification, is also not conducted in this paper.

5.2. Future Research

This paper proposed a framework based on deep learning to integrate textual information extraction with scene graph generation and improve construction safety management capabilities. Textual relational triples can be extracted from regulation text using an NLP-based information extraction method. The scene graph generation algorithm detected entities and visual relational triples. An automatic control procedure matched the textual and visual relational triples to enable safety inspection. The primary objective of our ongoing studies is to develop a multi-modal construction knowledge graph. Textual and visual relational triples are continuously integrated into the knowledge graph for data updating. Dynamic prediction of construction site hazard will be achieved using the knowledge graph built on multi-source data fusion. In addition, the personalized safety training recommendation system combining the knowledge graph feature learning can be developed to meet the safety training needs of workers and reduce the occurrence of workers' unsafe behaviors.

6. Conclusions

This work introduced a method to implement outfield construction safety inspections. Unlike previous approaches, we extracted key information from visual and textual modalities and represented them in the same relational triple form to relieve the semantic gap. The IE model employed BERT to encode sentences and extract textual relational triples based on the proposed BIEO tagging scheme. It was evaluated on the self-built outfield regulations dataset and a large-scale Chinese dataset, DuIE. Our method achieved encouraging performance in F1 scores of 79.3% and 79.6%, respectively. The proposed scene parsing method extracted visual relational information by fusing pre-trained fasttext and the deep visual network. It was trained and tested on the self-built on-site image dataset, and our method achieved 84.1% measured by the AP50 metric on object detection. For three tasks widely used to evaluate scene graph generation, our method achieved 50.7%, 85.5%, and

97.8%, respectively, on relation detection measured by SGGen(R@20), SGCls(R@20), and PredCls(R@20). The results demonstrated the effectiveness of the proposed methods in regulation text information extraction and on-site scene parsing. Additionally, an automatic control process was developed by integrating processed visual and textual relational triples to implement the PPE compliance checks. The results of the two scenes showed the feasibility of our method in safety inspection and hazard inference.

Most current studies on safety management using construction images and domain knowledge focus on deep visual–semantic modeling; however, the automatic information extraction from domain texts is not sufficiently explored. The theoretical contribution of this research is to present a deep learning framework for automatically extracting textual and visual relational information, which complements the lack of understanding of the regulation text in previous vision-based models. Outfield tests involve a variety of hazards, and there are numerous construction-related regulations in the outfield work safety requirements. The practical implication of this paper is to provide the possibility of outfield automated safety management by integrating textual and visual information. Further improvements will be to expand multi-source data for building a domain knowledge graph to improve the safety management for multiple types of hazardous operations in outfield tests.

Author Contributions: Conceptualization, X.L.; methodology, X.L.; software, Q.Z.; validation, X.L. and Q.Z.; resources, X.W. and X.J.; writing—original draft preparation, X.L.; writing—review and editing, X.L., W.D., X.W. and X.J. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Aerospace Hongka Intelligent Technology (Beijing) CO., LTD.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used to support the findings of this study are available by contact with the corresponding author.

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

References

1. Zhao, X.; Feng, R.; Wu, Y.; Yu, N.; Meng, X. A complementary filter-based all-parameters estimation for triaxis gyroscopes and optical angular encoders with intrinsic eccentricity. *IEEE Sensors J.* **2020**, *21*, 5060–5069.
2. Wang, L.; Gu, N.; Li, L.; Yu, H.; Sun, Y.; Lu, W. Analysis of capability of the ground target infrared stealth for the space infrared imaging system. In Proceedings of the International Symposium on Optoelectronic Technology and Application 2014: Infrared Technology and Applications, Beijing, China, 13–15 May 2014; SPIE: Bellingham, WA, USA, 2014; Volume 9300, pp. 277–284.
3. China Electronics Standardization Institute. Work Safety Standardization Requirements of Military Electronics Enterprises. SJ 21494. Available online: https://www.cssn.net.cn/cssn/productDetail/e7d1308706e73cb372698593b0868093 (accessed on 9 December 2022).
4. U.S. BUREAU OF LABOR STATISTICS. Bureau of Labor Statistics: Injuries, Illnesses, and Fatalities. Available online: https://www.bls.gov/iif/fatal-injuries-tables/fatal-occupational-injuries-table-a-1-2020.htm (accessed on 9 December 2022).
5. Ministry of Housing and Urban Rural Development of China. Available online: https://www.mohurd.gov.cn/gongkai/fdzdgknr/tzgg/202006/20200624_246031.html (accessed on 9 December 2022).
6. Tang, S.; Shelden, D.R.; Eastman, C.M.; Pishdad-Bozorgi, P.; Gao, X. A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. *Autom. Constr.* **2019**, *101*, 127–139.
7. Luo, H.; Liu, J.; Fang, W.; Love, P.E.; Yu, Q.; Lu, Z. Real-time smart video surveillance to manage safety: A case study of a transport mega-project. *Adv. Eng. Informatics* **2020**, *45*, 101100.
8. Fang, W.; Ding, L.; Love, P.E.; Luo, H.; Li, H.; Pena-Mora, F.; Zhong, B.; Zhou, C. Computer vision applications in construction safety assurance. *Autom. Constr.* **2020**, *110*, 103013.
9. Guo, H.; Yu, Y.; Ding, Q.; Skitmore, M. Image-and-skeleton-based parameterized approach to real-time identification of construction workers' unsafe behaviors. *J. Constr. Eng. Manag* **2018**, *144*, 04018042.
10. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Li, C. Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. *Autom. Constr.* **2018**, *93*, 148–164.
11. Ding, L.; Fang, W.; Luo, H.; Love, P.E.; Zhong, B.; Ouyang, X. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Autom. Constr.* **2018**, *86*, 118–124.

12. Fang, W.; Zhong, B.; Zhao, N.; Love, P.E.; Luo, H.; Xue, J.; Xu, S. A deep-learning-based approach for mitigating falls from height with computer vision: Convolutional neural network. *Adv. Eng. Informatics* **2019**, *39*, 170–177.
13. Nath, N.D.; Behzadan, A.H.; Paal, S.G. Deep learning for site safety: Real-time detection of personal protective equipment. *Autom. Constr.* **2020**, *112*, 103085.
14. Yan, X.; Zhang, H.; Li, H. Computer vision-based recognition of 3D relationship between construction entities for monitoring struck-by accidents. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *35*, 1023–1038.
15. Luo, X.; Li, H.; Yang, X.; Yu, Y.; Cao, D. Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning. *Comput. Aided Civ. Infrastruct. Eng.* **2019**, *34*, 333–351.
16. Zhang, J.; Shih, K.J.; Elgammal, A.; Tao, A.; Catanzaro, B. Graphical contrastive losses for scene graph parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11535–11543.
17. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
18. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring visual relationship for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 684–699.
19. Xiong, R.; Song, Y.; Li, H.; Wang, Y. Onsite video mining for construction hazards identification with visual relationships. *Adv. Eng. Informatics* **2019**, *42*, 100966.
20. Tang, S.; Roberts, D.; Golparvar-Fard, M. Human-object interaction recognition for automatic construction site safety inspection. *Autom. Constr.* **2020**, *120*, 103356.
21. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
22. Kelm, A.; Laußat, L.; Meins-Becker, A.; Platz, D.; Khazaee, M.J.; Costin, A.M.; Helmus, M.; Teizer, J. Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. *Autom. Constr.* **2013**, *36*, 38–52.
23. Yan, X.; Li, H.; Li, A.R.; Zhang, H. Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. *Autom. Constr.* **2017**, *74*, 2–11.
24. Gheisari, M.; Esmaeili, B. Applications and requirements of unmanned aerial systems (UASs) for construction safety. *Saf. Sci.* **2019**, *118*, 230–240.
25. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 1440–1448.
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
27. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
28. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
29. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
30. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Rose, T.M.; An, W. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9.
31. Kim, D.; Liu, M.; Lee, S.; Kamat, V.R. Remote proximity monitoring between mobile construction resources using camera-mounted UAVs. *Autom. Constr.* **2019**, *99*, 168–182.
32. Fang, W.; Ma, L.; Love, P.E.; Luo, H.; Ding, L.; Zhou, A. Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. *Autom. Constr.* **2020**, *119*, 103310.
33. Wu, J.; Cai, N.; Chen, W.; Wang, H.; Wang, G. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. *Autom. Constr.* **2019**, *106*, 102894.
34. Wang, L.; Xie, L.; Yang, P.; Deng, Q.; Du, S.; Xu, L. Hardhat-wearing detection based on a lightweight convolutional neural network with multi-scale features and a top-down module. *Sensors* **2020**, *20*, 1868.
35. Zhang, M.; Zhu, M.; Zhao, X. Recognition of high-risk scenarios in building construction based on image semantics. *J. Comput. Civ. Eng.* **2020**, *34*, 04020019.
36. Wu, H.; Zhong, B.; Li, H.; Love, P.; Pan, X.; Zhao, N. Combining computer vision with semantic reasoning for on-site safety management in construction. *J. Build. Eng.* **2021**, *42*, 103036.
37. Wang, Y.; Xiao, B.; Bouferguene, A.; Al-Hussein, M.; Li, H. Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Adv. Eng. Informatics* **2022**, *53*, 101699.
38. Paneru, S.; Jeelani, I. Computer vision applications in construction: Current state, opportunities & challenges. *Autom. Constr.* **2021**, *132*, 103940.
39. Li, Y.; Wei, H.; Han, Z.; Jiang, N.; Wang, W.; Huang, J. Computer Vision-Based Hazard Identification of Construction Site Using Visual Relationship Detection and Ontology. *Buildings* **2022**, *12*, 857.
40. Chen, S.; Demachi, K.; Dong, F. Graph-based linguistic and visual information integration for on-site occupational hazards identification. *Autom. Constr.* **2022**, *137*, 104191.

41. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 10–12 June 2008; pp. 1247–1250.
42. Li, S.; He, W.; Shi, Y.; Jiang, W.; Liang, H.; Jiang, Y.; Zhang, Y.; Lyu, Y.; Zhu, Y. Duie: A large-scale chinese dataset for information extraction. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Dunhuang, China, 9–14 October 2019; pp. 791–800.
43. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
44. Zheng, S.; Wang, F.; Bao, H.; Hao, Y.; Zhou, P.; Xu, B. Joint extraction of entities and relations based on a novel tagging scheme. *arXiv* **2017**, arXiv:1706.05075.
45. Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; Joulin, A. Advances in Pre-Training Distributed Word Representations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
46. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146.
47. Zellers, R.; Yatskar, M.; Thomson, S.; Choi, Y. Neural motifs: Scene graph parsing with global context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5831–5840.
48. Fu, T.J.; Li, P.H.; Ma, W.Y. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1409–1418.
49. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 5–9 September 2010; pp. 148–163.
50. Gardent, C.; Shimorina, A.; Narayan, S.; Perez-Beltrachini, L. Creating training corpora for nlg micro-planning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017.
51. Tzutalin. LabelImg. Git Code (2015). Available online: https://github.com/tzutalin/labelImg (accessed on 9 January 2023).
52. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73.
53. Lu, C.; Krishna, R.; Bernstein, M.; Fei-Fei, L. Visual relationship detection with language priors. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 852–869.