Computers in Industry 137 (2022) 103610

Contents lists available at ScienceDirect

Computers in Industry

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

Image2Triplets: A computer vision-based explicit relationship extraction framework for updating construction activity knowledge graphs

Zaolin Pan[a,b], Cheng Su[a,c], Yichuan Deng[a,c,*], Jack Cheng[d]

[a] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
[b] Sino-Singapore International Joint Research Institute, Guangzhou, China
[c] State Key Laboratory of Subtropical Building Science, South China University of Technology, Guangzhou, China
[d] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 6 August 2021
Received in revised form 18 December 2021
Accepted 14 January 2022
Available online 31 January 2022

Keywords:
Zero-shot learning
Human-object interaction detection
Computer vision
Explicit relationship extraction
Knowledge graph

ABSTRACT

Knowledge graph (KG) is an effective tool for knowledge management, particularly in the architecture, engineering and construction (AEC) industry, where knowledge is fragmented and complicated. However, research on KG updates in the industry is scarce, with most current research focusing on text-based KG updates. Considering the superiority of visual data over textual data in terms of accuracy and timeliness, the potential of computer vision technology for explicit relationship extraction in KG updates is yet to be explored. This paper combines zero-shot human-object interaction detection techniques with general KGs to propose a novel framework called Image2Triplets that can extract explicit visual relationships from images to update the construction activity KG. Comprehensive experiments on the images of architectural decoration processes have been performed to validate the proposed framework. The results and insights will contribute new knowledge and evidence to human-object interaction detection, KG update and construction informatics from the theoretical perspective.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Efficient knowledge management is essential in the architecture, engineering and construction (AEC) industry (Kamara et al., 2002). However, the AEC industry is fragmented, with the information distributed among numerous and ever-changing stakeholders (Rasmussen et al., 2019), who may never have worked together before and may never again (Xue and Lu, 2020). Knowledge acquisition and reuse in such a fragmented and complex industry are complicated and limited. Therefore, to manage the industry's heterogeneous, discrete and empirical knowledge, an effective tool is needed to enable the structured storage and reuse of knowledge, which is the main focus of the knowledge graph (KG). First proposed by Google in 2012, KGs are essentially semantic networks that reveal the relationships between entities. KG consists of a data layer and a schema layer, with the former organising knowledge in triplets and the latter regulating the representation of knowledge in the data layer through ontologies. KG has made its mark in many fields, such as information retrieval (Li et al., 2020), personalised recommendation (Wang et al., 2018) and automatic Q&A (Liu et al., 2019).

Although researchers have investigated ontologies in many domains of the AEC industry, such as knowledge retrieval (Park et al., 2013), claims management (Niu and Issa, 2012), cost estimation (Ma et al., 2016), risk identification (Zhong et al., 2020), knowledge management (Kamsu-Foguem and Abanda, 2015), structural health monitoring (SHM) (Li et al., 2020) and facility maintenance management (FMM) (Chen et al., 2020), research on KG has been scarce. For example, Leng et al. (Leng et al., 2019) and Zhu et al. (Zhu et al., 2017) constructed a mechanical, electrical and plumbing (MEP) domain KG and a geological data KG using natural language processing (NLP). Pan et al. (Pan et al., 2021) updated construction activity KG using computer vision technology. By applying the KG technology, Rasmussen et al. (Xue and Lu, 2020) managed interrelated project information, Wang et al. (Wang et al., 2020) integrated building information modelling (BIM) for fire drawing review, and Fang et al. (Fang et al., 2020) investigated the knowledge extraction from visual data and developed a hazard identification framework based on object detection. Although these studies mentioned above have explored KG construction and application aspects, few have explored the KG update issue, which is an ongoing and unavoidable issue when structuring KGs.

KG updates consist of instance- and ontology-level updates, the former representing the update of triplet form knowledge in the

* Corresponding author at: School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China.
E-mail address: ctycdeng@scut.edu.cn (Y. Deng).

https://doi.org/10.1016/j.compind.2022.103610
0166-3615/© 2022 Elsevier B.V. All rights reserved.

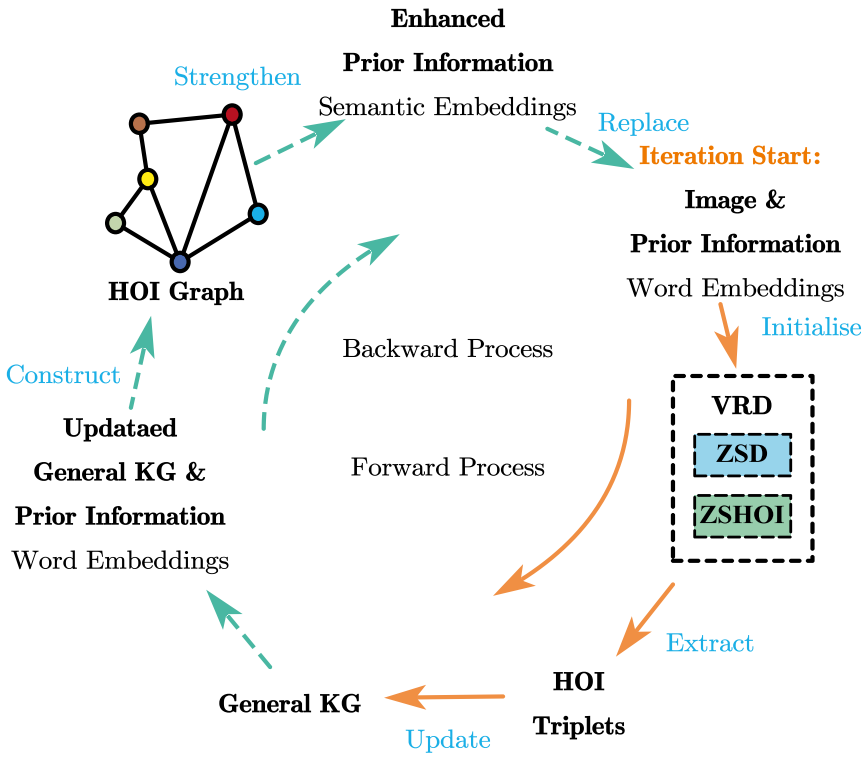


Fig. 1. Schematic diagram of the Image2Triplests framework. The framework can be divided into the forward process and the backward process.

data layer and the latter representing the update of concepts and concept relations in the ontology at the schema level. In general, we expect to implement these two levels of updates automatically. However, ontology-level updates involve conceptual and logical updates that lead to semantic inconsistency problems (Ovchinnikova and Kühnberger, 2006), and this problem has not yet been sufficiently addressed (Ahmeti, 2020), requiring a degree of expert intervention. In contrast, instance-level updates do not involve conceptual and logical updates and can be performed automatically or semi-automatically under the guidance of the ontology. In addition, instance-level updates can provide a more reliable database for ontology updates, facilitating conceptual and logical generalisation. Therefore, we focus on instance-level updates in this paper as opposed to ontology-level updates.

Relationship extraction, a subtask of knowledge extraction, is one of the most critical tasks in instance-level KG updates. Relationships between entities include explicit and implicit relationships. The former are generally visible associations between entities that exist in the dataset at the activity level, such as actional (e.g., holding), spatial (e.g., next to), or comparative (e.g., longer than). The latter generally require inferences from similar properties between objects at the project level, such as associations between documents (Dörk et al., 2011). Unfortunately, both the AEC domain's activity and project level contain many complex implicit relationships, and direct extraction of these implicit relationships is challenging, particularly under data-limited conditions. However, improving understanding at the activity level by extracting and distiling explicit relationships can help discover and infer implicit relationships at both activity and project levels (Li et al., 2021). Therefore, we concentrate on explicit relationships extraction in this study.

Although many industries use textual data for relationship extraction due to their convenience and easy accessibility, we still choose visual data (e.g., images) as our primary source of knowledge for the following reasons. First, this paper focuses on extracting explicit relationships at the activity level in the AEC industry, which contains many complex interactions between humans and objects. While textual descriptions of various construction activities exist, visual data provides a more intuitive and accurate record of these activities (Martinez et al., 2019) and avoids subjective biases and misinterpretations. In the case of onsite monitoring, the relationships extracted from visual data can be seen as visually verified facts, as they record the actual actions of the workers, which may differ from those recorded in textual data. Second, visual data can be updated automatically in a real-time manner through surveillance cameras, which is beneficial for onsite monitoring, while textual data update requires additional manual effort. Additionally, computer vision has made considerable progress in visual relationship detection (VRD) tasks (Lu et al., 2016), particularly in the subtask of human-object interaction (HOI) detection, making it possible to extract human-to-object visual relationships.

As computer vision-based explicit relationship extraction tasks require detecting entities (e.g., people and objects) and relationships between entities (e.g., holds or rides), these tasks are limited by three major difficulties. The first difficulty is the acquisition of training data. As the relationships between entities are fine-grained and related to specific entity classes, quadratic combinations between relationships and entities entail high labelling costs. In addition, these combinations present a long-tailed distribution, making it difficult to obtain enough training data for rare combinations. The second difficulty is the acquisition of new knowledge. KG updates rely heavily on acquiring new knowledge (i.e., unknown relations), whereas both supervised and few-sample learning, in general, rely on training data of known classes and cannot identify unknown classes without the training data of known classes. Third, inconsistencies in action and action labels lead to polysemy problems (i.e., an action may correspond to multiple action labels, and an action label may also correspond to different actions). For example, a person with a steel bar may be described as holding or moving the bar, and the action of installing a window is not the same as installing a ceiling. These phenomena pose additional challenges to the computer vision-based KG update task.

To address the above challenges, a novel framework called Image2Triplets is proposed to extract explicit visual relationships between humans and objects at the activity level. For the first challenge, we believe that compositional methods (Bansal et al., 2020) can address the combinatorial explosion and long-tailed data distribution problem. Compared to non-compositional methods (Chao et al., 2018) that directly predict the entire human-object interaction (HOI) triplet (e.g., < person, ride, bike >), compositional methods that separately predict human, actions and objects and then combine them into triplets require less training data. In addition, transfer learning and data argument techniques are also adopted to further cope with the lack of labelled training data (Smith et al., 2021). As for the second challenge, we argue that new knowledge can be gained via zero-shot learning (ZSL) techniques. This technique transfers the knowledge learned from non-rare classes to rare or unknown classes with additional prior information, thus enabling the unknown class recognition. In particular, humans can imagine unknown HOIs from known HOIs, e.g., from < person, ride, bike > to < person, sit on, elephant > . Therefore, we detect novel HOIs using zero-shot object detection (ZSD) and zero-shot HOI (ZSHOI) models in the proposed framework. For the third challenge, we turn single-label forecasts into multi-label forecasts in the ZSHOI model so that a single action can be associated with multiple action labels. Besides, we introduce additional features in the ZSHOI model to narrow down the candidates for action labels, thus alleviating the problem of one action label corresponding to multiple human poses. In summary, our framework can properly tackle these challenges in the ways described above.

As Pan et al. (Pan et al., 2021) adopted two iterative processes using ZSL technology for new entity extraction and achieved some results, we follow a similar idea for new relation extraction. Our framework consists of two iterative processes: a forward process (solid orange line) and a backward process (dotted green line), as shown in Fig. 1. The forward process initialises the visual relationship detector (VRD), including ZSD and ZSHOI models, which detect novel HOI triplets from images using prior information (e.g., word embedding). These extracted triplets of the form < human, action, object > contain AEC domain-related information that underpins the construction activity KG construction by integrating a general KG. The backward process aims to construct semantic embedding from the HOI graph using graph convolution network (GCN) (Fang et al., 2020) to strengthen the prior information, where the HOI graph integrates the semantic relations between the updated general KG and word embeddings. In the next iteration, the forward process benefits from the better VRD initialised by the semantic embedding and extracts triplets. The framework incorporates the perceptual capabilities of computer vision and the cognitive capabilities of KG that can optimise the visual relationship extraction process.

This paper addresses the computer vision-based KG update task by presenting a framework that can extract novel explicit HOI triplets from images for the data layer of construction activity KG updates. This paper specifically addresses the novel explicit relationship extraction problem in activity level by applying ZSL and HOI detection techniques. This paper first introduces HOI detection techniques into the computer vision-based KG update task to the best of our knowledge. This paper has also performed comprehensive experiments and discussions on the images of architectural decoration processes to validate the proposed framework. The results show that the backward process can enhance the prior information to improve the VRD performance and that the framework can iteratively extract both known and unknown HOI triplets to update construction activity KG. The remainder of this paper is organised as follows. Section 2 reviews the research on knowledge management in construction, relationship extraction, ZSL and HOI. Section 3 introduces the Image2Triplets framework. Section 4 provides comprehensive experiments and detailed discussion, and Section 5 concludes the paper.

2. Related works

2.1. Knowledge management in construction

Up to now, the development and practice of knowledge management in construction has gone through four stages: the emergence period, the expert system and ontology period, the semantic web period and the KG period.

In the first stage, the concept of knowledge management gradually emerged. In 1967, Drucker (Drucker, 2018) pioneered the concept of knowledge workers. In 1977, the Fifth International Conference on Artificial Intelligence first introduced knowledge engineering, and knowledge base systems began to be applied. In 1986, Sveiby published *The Knowledge-Based Enterprise*, which used the term "knowledge management" for the first time and delved into knowledge management's fundamental issues (Sveiby and Risling, 1986).

The second stage is the period of expert systems and ontologies. In the 1980 s, expert systems began to be applied to the construction industry (Mohan, 1990). In 1991, Neches proposed a framework using ontologies to model domain knowledge (Neches et al., 1991), and in the same year, ontologies were applied to building facades monitoring (Fazio et al., 1991). In 1995, Gruber (Gruber, 1995) proposed a widely accepted definition of an ontology: "An ontology is a formal, explicit specification of a shared conceptualisation". However, despite their widespread use, expert systems and ontologies also had their shortcomings. Expert systems rely on the manual acquisition of knowledge by experts, and ontologies focus on the description of concepts and relationships and, as opposed to expert systems, lack a knowledge base consisting of data instances.

The third stage is the period of the semantic web. The semantic web was introduced in 2001 by Berners-Lee (Berners-Lee et al., 2001), the father of the world wide web. Combining the strengths of ontologies and expert systems, the semantic web represents the content on the Internet in the structured semantics form to build a semantically shared knowledge base under the specification of ontologies. In 2002, the semantic web represented by ifcXML (Industry Foundation Classes eXtensible Markup Language) started to be applied to the construction industry (Cheng et al., 2002). In 2008, Anumba and Charles et al. (Anumba et al., 2008) published *Knowledge Management in Construction*, which systematically proposed a knowledge management framework for the construction industry.

The fourth stage is the period of KG, with Google first introducing the concept of KG in 2012. The KG is a typical application of the semantic web, which uses the knowledge base obtained from multiple data sources on the Internet to improve retrieval quality. Since then, as KG embedding techniques evolved (Wang et al., 2014), KGs have been created and successfully applied to many real-world applications in both industry and academia (Wang et al., 2017).

2.2. Relationship extraction

Relationship extraction is one of the most critical subtasks of knowledge extraction. Mostly, the relationship extraction task aims to extract semantic relationships between entities from unstructured text. Thus, relationship extraction is closely related to entity extraction. Relationship extraction generally focuses on extracting possible relationships between entities after identifying them in the text. Currently, relationship extraction methods can be categorised as template-based methods (Flynn et al., 2007), supervised learning-based methods (Miwa and Bansal, 2016; Kambhatla, 2004; dos Santos et al., 2015) and weakly-supervised learning-based methods (Ji et al., 2017; Brin, 1999).

Most early methods for relationship extraction (Flynn et al., 2007) were implemented based on template matching. These methods combine linguistic knowledge and corpus features with

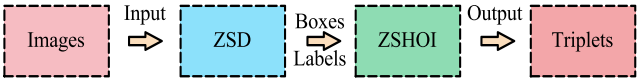


Fig. 2. Schematic diagram of the visual relationship detection (VRD) module. The module follows pipeline-based methods and uses a zero-shot detection (ZSD) and a zero-shot human-object interaction (ZSHOI) detection model.

hand-written templates made by domain experts to match specific textual relationships. Template-based methods can achieve satisfactory results for small-scale and limited domain relationship extraction problems. However, domain experts must have considerable time manually constructing templates when the dataset is large. Besides, the portability of the template-based relationship extraction system is relatively poor (Moreo et al., 2013).

Supervised learning-based relationship extraction methods transform the task into a classification problem. These methods train supervised classification models for relationship extraction based on a large amount of labelled data. Based on the feature extraction differences, these methods are typically classified into traditional methods (Kambhatla, 2004) and deep learning methods (Miwa and Bansal, 2016). While traditional methods rely on feature engineering, deep learning methods do not require manual construction of various features and can be divided into pipeline methods (dos Santos et al., 2015) and joint methods (Miwa and Bansal, 2016). Pipeline methods treat entity extraction and relationship extraction as two separate processes, with relationship extraction based on the entity extraction results. Joint methods combine these two extraction processes and optimise them together in a unified model. Although the joint extraction method can mitigate the pipeline method's error accumulation in the two processes, the pipeline method is more studied and even outperforms the joint method (Liu et al., 2020) in the VRD task. Additionally, the pipeline approach allows for easy replacement of different components, meaning that different components can be selected to adapt to real-world situations. Therefore, a pipeline-based approach is used in this paper.

Weakly supervised learning-based relationship extraction methods primarily include remotely supervised methods (Ji et al., 2017) and bootstrapping methods (Brin, 1999). Remotely supervised methods automatically construct training datasets by aligning the KG with the unstructured text, reducing the reliance on manually labelled data and enhancing the model's cross-domain adaptability. The bootstrapping method learns on an initial seed set consisting of a small number of instances to obtain relationship extraction templates. These templates can extract more instances and iteratively add these instances to the seed set. Via continuous iterating, the bootstrapping method can extract different relationships from the text.

2.3. Zero-shot learning (ZSL)

Most ZSL research has focused on zero-shot image recognition (ZSIR) tasks to recognise unseen classes with additional side information. Depending on the side information, ZSL falls into four types: attributes (Lampert et al., 2014), word embeddings (Romera-Paredes and Torr, 2015), KG (Kampffmeyer et al., 2019) and generative adversarial networks (GAN) (Gao et al., 2020). The early ZSL research (Kambhatla, 2004) connected the seen and unseen classes at the semantic level using attributes. However, considerable time and labour are required for attribute design and data labelling in attribute-based methods. Therefore, word embedding (Romera-Paredes and Torr, 2015) was used for the ZSL to alleviate these limitations. Word embedding, or word vectorisation, trained in large corpora, enables the measurement of semantic distance between different words, bridging the gap between seen and unseen classes at the semantic level. In addition to these methods, researchers distilled the knowledge from KGs (Kampffmeyer et al., 2019), containing structural knowledge of many domains, and generated unseen classes data using GAN (Gao et al., 2020).

Apart from ZSIR, ZSD that aims to detect novel objects with additional side information has attracted increasing attention. ZSD is a crucial component for HOI detection in this study, as we follow the pipeline-based approach. Similarly, ZSD methods can also be divided into different types depending on the side information: attributes (Zhu et al., 2020), word embeddings (Rahman et al., 2020; Rahman et al., 2019), text descriptions (Zhang et al., 2020), KG (Yan et al., 2020) and GAN (Zhu et al., 2020). Early ZSD studies used attributes or word embeddings as prior information to generalise the knowledge learned from seen classes to unseen classes. However, Zhang et al. (Fazio et al., 1991) chose textual descriptions instead of word embeddings because word embeddings are static and may not consider the actual context. Yan et al. (Yan et al., 2020) used GCN to aggregate the prior information in the KG. Additionally, Zhu et al. (Zhu et al., 2020) focus on the adaptability of GANs to the ZSD task.

2.4. Human-object interaction (HOI) Detection

HOI detection, which detects interactions between humans and objects, is critical to human-centric scene understanding (Gupta and Malik, 2015). Early HOI works followed non-compositional methods and benefited from pre-trained object detectors. For example, Chao et al. (Chao et al., 2018) directly predicted HOI labels at the trigram level via non-compositional methods using a pre-train object detection model, thereby simplifying the HOI detection to the HOI classification. However, the combinatorial explosion problem and the long-tail distribution of HOI annotations still limit the non-compositional method. Researchers used compositional methods and focused on verb prediction to mitigate these limitations. For example, Gao et al. (Gao et al., 2018) predicted action labels using humans and objects' respective visual and geometric features. Ulutan et al. (Ulutan et al., 2020) modified human and object features using attention techniques. Xu et al. (Xu et al., 2019) introduced external KGs for verb predicting. As these methods focused on predicting the HOI labels of seen objects, Bansal et al. (Bansal et al., 2020) proposed a framework for detecting the HOI labels of unseen objects. Tang et al. (Xu et al., 2019) and Xiong et al. (Kato et al., 2018) applied HOI detection techniques for construction site safety inspection in the AEC industry.

Additionally, many researchers have devoted their efforts to novel HOI detection. Instead of improving the HOI prediction of seen actions, novel HOI detection aims to scale HOIs with novel objects, actions or combinations using ZSL techniques. For example, Kato et al. (Kato et al., 2018) detected novel combinations of HOI, Liu et al. (Liu et al., 2020) combined external KGs to predict HOIs with novel actions, and Wang et al. (Wang et al., 2020) detected HOIs with novel objects. Nevertheless, none of these studies examined the HOI detection of novel objects and actions, which is critical for KG updates. Our framework integrates ZSD and ZSHOI models based on the pipeline approach that can detect HOIs with novel objects and novel actions.

3. Image2Triplets

As the novel HOI detection requires both the novel object and action detection, this paper uses a ZSD detector (Pan et al., 2021) for novel object detection and focus on novel action detection and construction activity KG updates. The VRD consists of a ZSD and a ZSHOI model designed to detect novel objects and interactions to form HOI triplets, as shown in Fig. 1. Specifically, as the VRD adopts pipeline-based methods, the VRD first utilises a ZSD detector to detect humans and novel objects (i.e., boxes and labels) in the image, then leverages the ZSHOI model to perform novel action detection

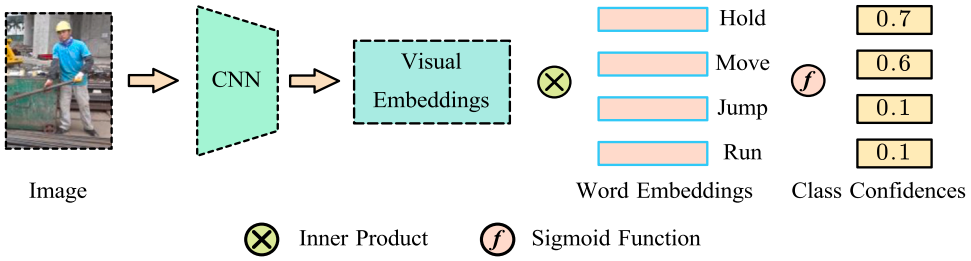


Fig. 3. The architecture of the zero-shot human-object interaction (ZSHOI) detection model in the forward process. This model uses a convolutional neural network (CNN) to extract visual embeddings.

using this information, and finally outputs the triplets, as shown in Fig. 2.

3.1. Problem definition

The VRD, containing ZSD and ZSHOI models, is a critical part of the framework. The formal definitions of the VRD, ZSD and ZSHOI are given below.

3.1.1. VRD

HOI detection detects triplets of the form < human, action, object > in a given image. Formally, HOI can be defined as $<b_h, b_0, a, y>$, where the boundary box $b_h, b_0 \in \mathbb{R}^4$ denotes humans and objects' location, action $a \in A = \{V_1, ..., V_n\}$ denotes the movement performed by a person, and label $y \in Y = \{C_1, ..., C_n\}$ denotes the category of the boundary box (b_h, b_0). The goal of the VRD is to scale the HOI by extending the action and object categories with ZSD and ZSHOI models. Therefore, one can define the VRD as a function $F(\,) = \{f_{ZSHOI}(f_{ZSD}(\,))\}$, where $f_{ZSHOI}(\,)$ and $f_{ZSD}(\,)$ represent the ZSHOI and ZSD models, respectively.

3.1.2. ZSD

ZSD aims to localise and classify seen and unseen objects with additional prior information. The prior information regarding the seen and unseen object categories (i.e., $E_S^O, E_U^O \in \mathbb{R}^q$, where S denotes seen classes, U denotes unseen classes, O denotes objects, and q denotes the length of the prior information), is available in the training and testing phases. Let $Y_S = \{C_1, ..., C_m\}$ and $Y_U = \{C_{m+1}, ..., C_n\}$ denote the seen and unseen categories, respectively, and let B denote the collection of all the human and object boundary boxes. The intersection of seen and unseen category sets is empty (i.e., $Y_S \cap Y_U = \varnothing$). Given an image, the ZSD function $f_{ZSD}(\,)$ aims to recognise seen and unseen object categories $y \in Y = Y_S \cup Y_U$ and localise human and object boundary boxes $b_h, b_0 \in B$ with prior information $E_S^O, E_U^O \in \mathbb{R}^q$.

3.1.3. ZSHOI

Likewise, let $A_S = \{V_1, ..., V_m\}$ and $A_U = \{V_{m+1}, ..., V_n\}$ denote the seen and unseen actions, respectively, and let $E_S^A, E_U^A \in \mathbb{R}^q$ denote the additional prior information of the seen and unseen action categories, respectively. Given an image, the human and object boundary boxes $b_h, b_0 \in B$ and action categories' prior information $E_S^A, E_U^A \in \mathbb{R}^q$, the ZSHOI function $f_{ZSHOI}(\,)$ simultaneously recognises the seen and unseen action categories $a \in A = A_S \cup A_U$.

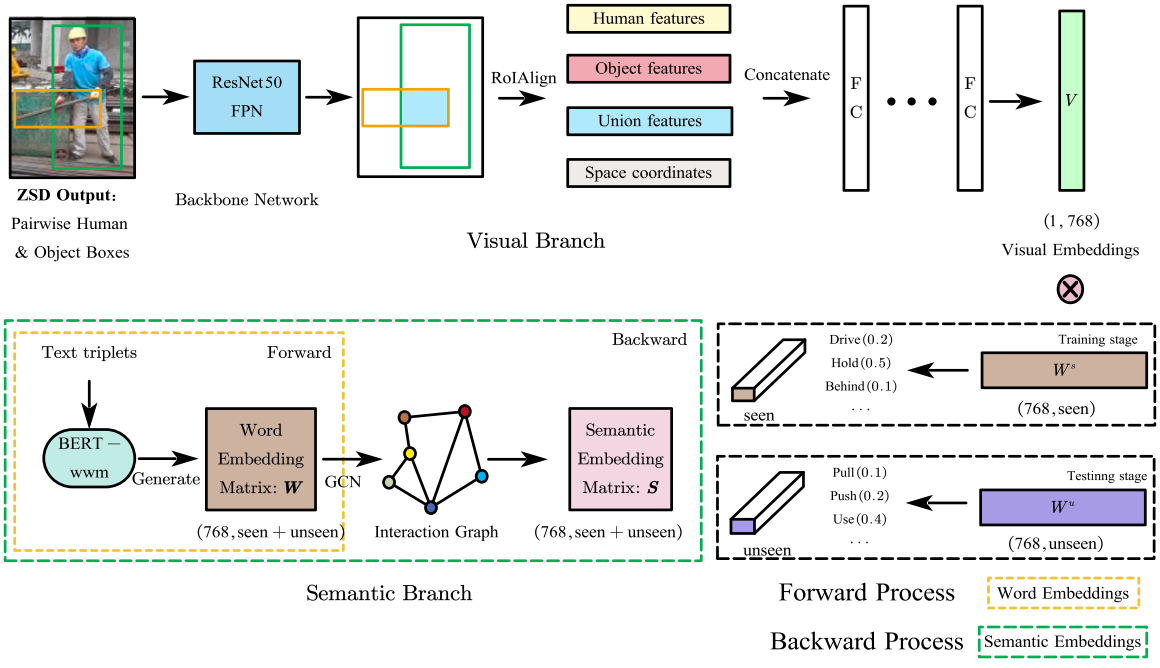


Fig. 4. Two types of ZSHOI models are framed: the word embedding-based ZSHOI model and the semantic embedding-based ZSHOI model. Both models contain two branches: a visual branch and a semantic branch.

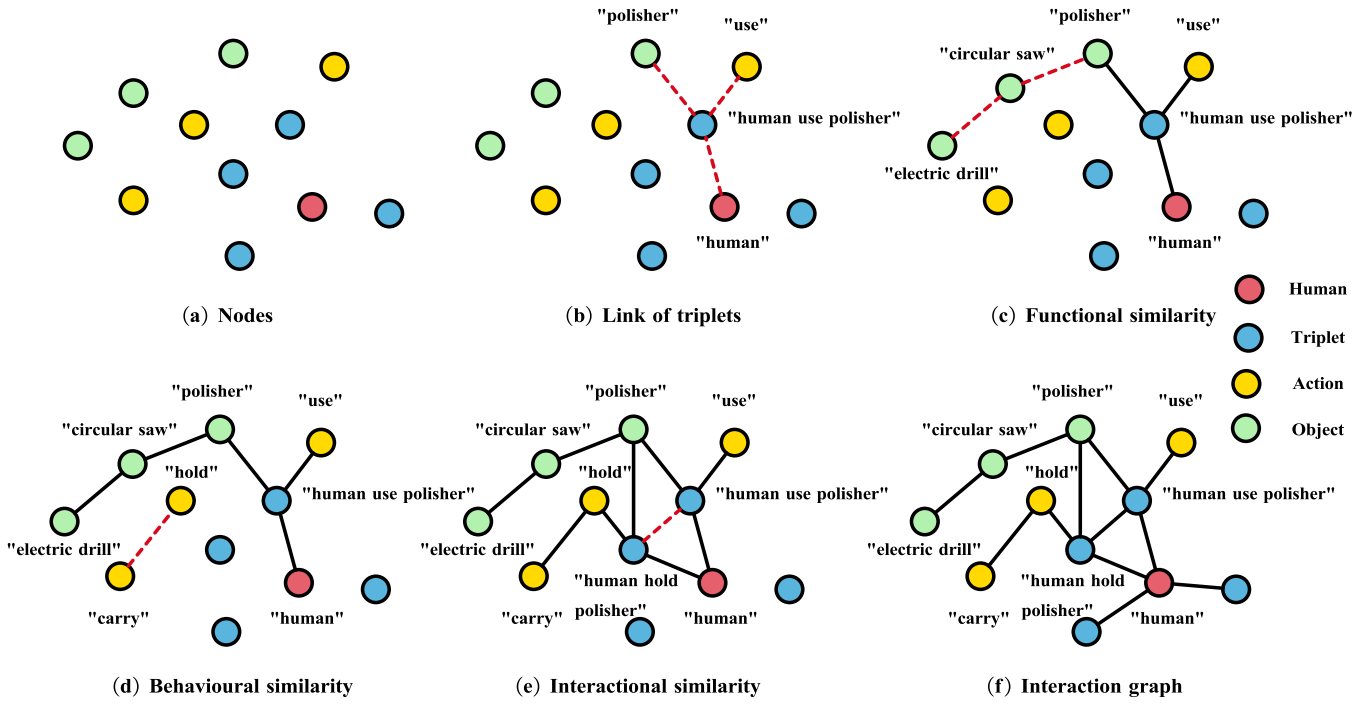


Fig. 5. The pipeline of HOI graph construction. (a) HOI graph contains human, action, object and triplet nodes. (b) Nodes within a triplet are connected. (c) Nodes with functional similarities are connected. (d) Nodes with behavioural similarities are connected. (e) Nodes with interactional similarities are connected. (f) The constructed HOI graph.

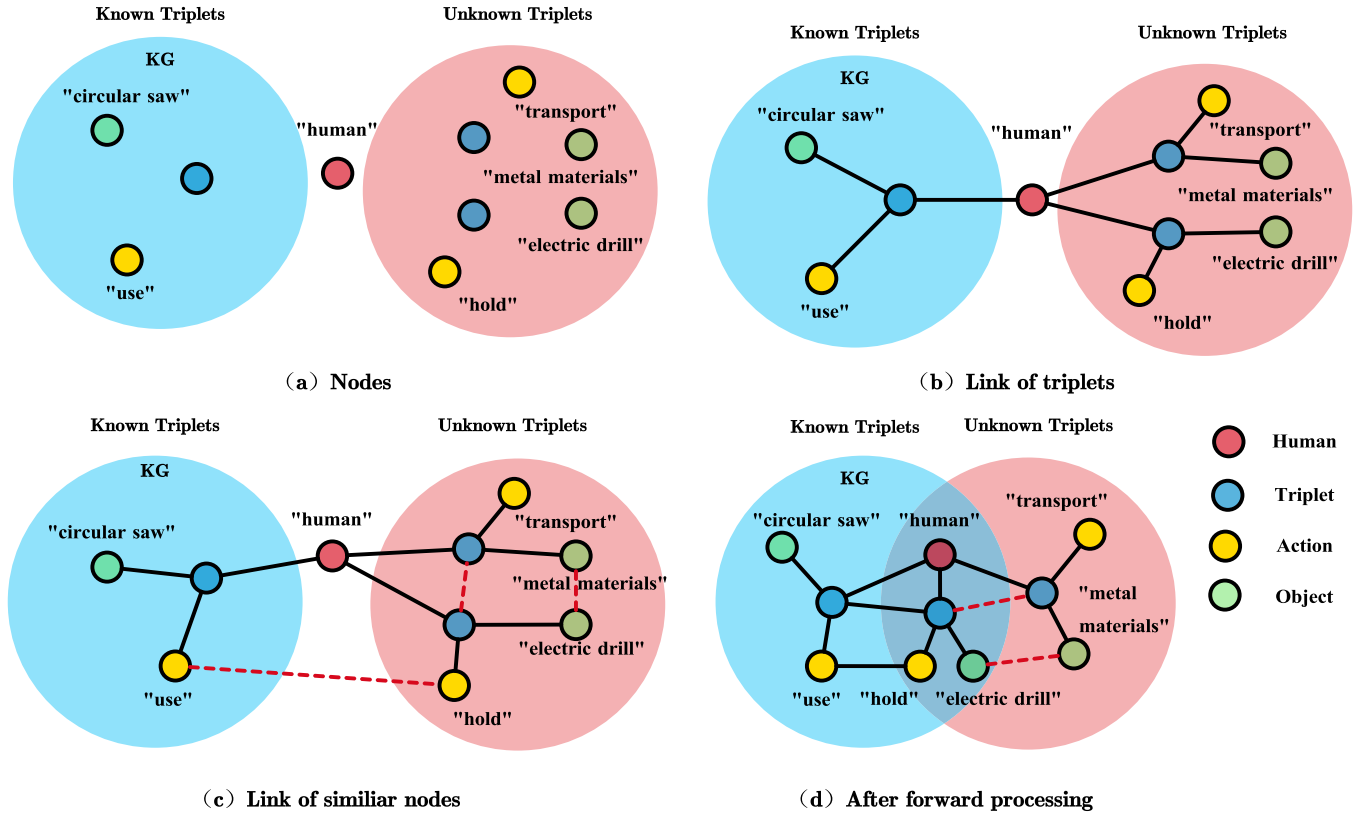


Fig. 6. Iteration of the link between known and unknown triplets. (a) Nodes within the pink circle are known, and nodes within the green circle are unknown. (b) Nodes within the identical triplet are connected. (c) Nodes with functional, behavioural, or interactional similarities are connected. (d) After the forward process, the unknown triplet < human, hold, electric drill > is detected and becomes a known triplet.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

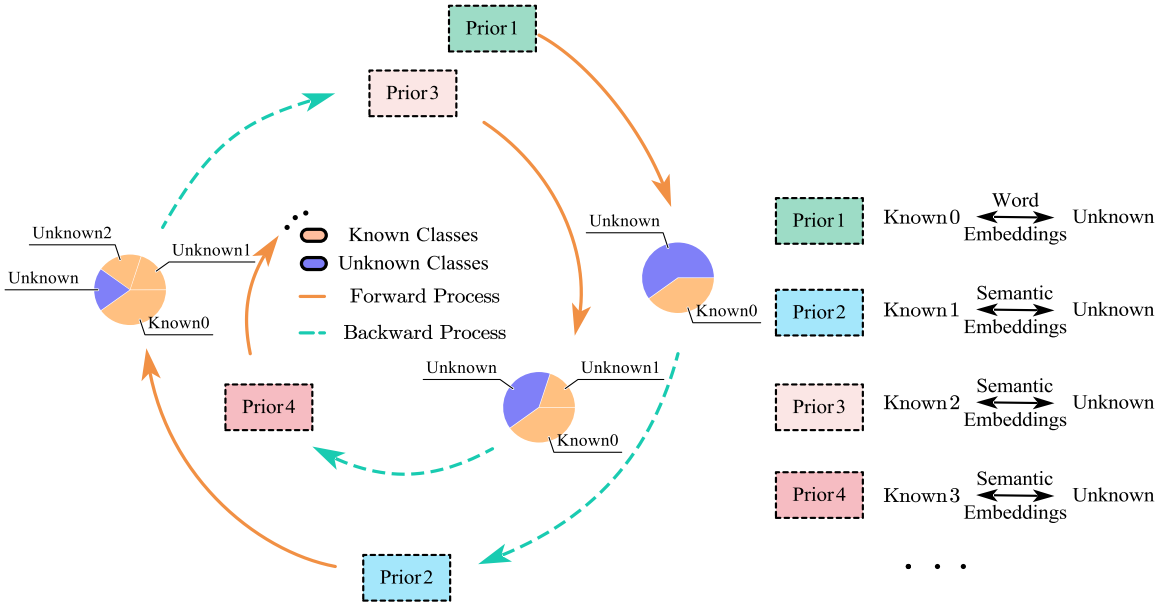


Fig. 7. The overall workflow of the Image2Triplets framework. The purple part of the pipe chart represents known classes, and the orange part of the chart represents unknown classes. The solid orange line represents the forward process, and the dash green line represents the backward process.

3.2. Forward process

3.2.1. The primary idea of the forward process

The ZSHOI model, a critical part of the VRD, is first initialised by prior information and then used to extract the novel HOI. Due to the ability to evaluate semantic links between known and unknown actions, word embeddings are used as prior information in the ZSHOI model, thus enabling the transformation of the known knowledge to the unknown. The ZSHOI model follows a linear mapping approach, mapping visual embeddings directly to the semantic space. Then, the inner product is applied to align features in semantic and visual space, i.e., aligning the word embeddings of all candidate labels and the visual embeddings extracted by a convolutional neural network (CNN), as shown in Fig. 3. We choose the sigmoid activation function instead of the softmax activation function to obtain the final class confidence so that single-label prediction can be turned into multi-label prediction. The predicted class confidence is 1 when the word embedding is compatible with the visual embedding; otherwise, it is 0, and confidences greater than a specific threshold will become the action label for this HOI.

3.2.2. ZSHOI model construction through word embeddings

The ZSHOI model consists of visual and semantic branches, as shown in Fig. 4. The visual branch aims to extract visual embeddings from different human-object pairs, while the semantic branch focuses on the extractions of word embeddings (the orange box). Note that the word embedding-based ZSHOI model uses word embeddings rather than semantic embeddings. The pre-trained ResNet-50-FPN (Lin et al., 2017) model is used as the backbone network for visual feature extraction in the visual branch. Considering that a large amount of background information in the image is not related to HOIs, the regions related to HOIs are mainly where humans and objects are located. Therefore, we focus on the visual features in these regions and use the RoIAlign (He et al., 2020) operation to extract their features. The RoIAlign operation combines the ZSD output (human and object locations) with visual features to obtain human features, object features and union features, where union features are those where the person and object boundary boxes intersect. These features and space coordinates (human and object locations) are then combined and fed into multiple fully connected (FC) layers to obtain the final visual embedding V of length 768. While in the semantic branch, we use pre-trained word embeddings obtained from the BERT-wwm (Bidirectional Transformers-Whole Word Masking) model trained on a large-scale general corpus (Cui et al., 2020) as prior information. The word embedding obtained in this way can gauge the semantic relevance of any two words in the corpus, making it robust prior information. As mentioned before, the ZSHOI model applies a linear mapping approach to align the visual and semantic feature space and derive the class confidence. More specifically, in the training stage, only the word embeddings of seen classes $\boldsymbol{W}^S$ of the shape 768 ×seen are available, as shown below:

$$Seen \quad Class = \text{sigmoid}(V\boldsymbol{W}^S) \tag{1}$$

In the testing stage, only the word embeddings of unseen classes $\boldsymbol{W}^U$ of the shape 768 ×unseen are available. Both seen and unseen word embeddings are available only in the generalised HOI detection (GHOID) task, which simultaneously detects seen and unseen actions.

3.2.3. Relationship extraction and KG update

The ZSD model can detect entities of humans and objects, and the ZSHOI model can identify these interactions. Thus, we can obtain triplets in the form of < human, action, object > combining these results. Considering that general KGs, such as ConceptNet 5.5 (Speer et al., 2017), contain entities and relations from the AEC domain, such as people, bricks, hammers, holding and carrying, it is possible to associate the found unknown triplets with entities or relations that are already present in the general KG, thereby eliminating the need to structure a KG from scratch. For example, we can store the found triplets in a graph database, such as Neo4j (Anon, 2021), automatically or semi-automatic to update the KG. However, there may be an exception where both actions and objects are unknown, which inevitably requires human intervention. In this case, we need to determine the relationship between the unknown object and the most relevant object in the KG so that the detected triplet can be linked to the KG.

3.3. Backward process

Prior information is critical to the performance of the ZSL model. As the nature of ZSL is to transfer known knowledge to unknown,

Fig. 8. Images in the dataset.

ZSL models heavily rely on the semantic link between known and unknown classes. Therefore, the ZSL model requires robust prior information to strengthen the semantic link. Considering that a reasonable KG can lead to more robust prior information than word embeddings (Kampffmeyer et al., 2019), the backward process aims to enhance prior information by distiling knowledge from KGs and word embeddings, as the enhanced prior information should improve the performance of the VRD.

3.3.1. Enhancement of prior information

This section constructs an HOI graph by distiling knowledge in the updated general KG and word embeddings. Fig. 5(f) shows that an HOI category is represented by three entities nodes and one interaction node. For example, < human, use, polisher > and < human, hold, polisher > are represented by four entities nodes "polisher", "human", "use" and "hold", and two interaction nodes "human hold polisher" and "human use polisher", respectively. We follow the method proposed by Liu et al. (Liu et al., 2020) to structure the HOI graph. This method supposes that an unknown HOI triplet can be generalised from the known triplets if functionally, behaviourally, or interactionally consistent with the unknown triplets. For example, a human can recognise the unknown triplets (< human, hold, rebar >) by using their common sense to imagine what it would be based on the known triplets (< human, hold, pipe > and < human, move, rebar >).

The detailed HOI graph construction method involves the following processes. First, all entity and interaction nodes, including known and unknown nodes in the dataset, are added. In addition, we query the updated general KG to find the relevant relationships with those entity nodes. For those entities interconnected in KG, we annotate them and their interconnected relations as triplets and add them as interaction nodes to the graph, as shown in Fig. 5(a). Second, nodes belonging to the identical HOI triplet are linked together, as shown in Fig. 5(b). Third, the entity nodes that have a similar function are connected. More specifically, we use the word embedding generated by the pre-trained language model BERT-wwn, which takes the manually labelled text triplets in the dataset as input, to calculate the cosine similarity of any two object nodes. Since word embedding can measure the semantic distance of entities, two nodes are considered semantically consistent to some extent if their cosine similarity is higher than a given threshold (e.g., 0.5). Then, edges are added to the top-k consistent nodes, as shown in Fig. 5(c). Likewise, the other nodes are linked based on their semantic distances among action and interaction, as shown in Fig. 5(d)(e). The final HOI graph is shown in Fig. 5(f).

The HOI graph consists of the nodes of known and unknown objects, actions and interactions, integrating the knowledge from KG and word embeddings and implying rich semantic connections between known and unknown classes. The HOI graph iteration is shown in Fig. 6. We first add all the known and unknown nodes to the graph, as shown in Fig. 6(a). More specifically, for those nodes interconnected in KG, we annotate them and their interconnected relations as known triplets, and the remaining known triplets are in the training set. Triplets that neither exists in KG nor the training set are annotated as unknown. Then, as shown in Fig. 6(b)(c)), nodes within a triplet and semantically similar nodes are connected. Before the forward process, only a link between "use" to "hold" and a node "human" that links the known and the unknown triplets. After the forward process, the unknown triplet " < human, hold, electric drill > " is detected, and the nodes of this triplet turn into known and are linked to the KG, as shown in Fig. 6(d). Additional two links between the known and unknown triplets are acquired, i.e., from "electric drill" to "metal materials" and from " < human, hold, electric

Table 1
mAP comparison of ZSHOI model with word embeddings and semantic embeddings.

Prior/Task	Seen	Unseen	Seen + Unseen
Word embeddings	0.60	0.41	0.41
Semantic embeddings	**0.61**	**0.58**	**0.53**

Fig. 9. Qualitative results of the ZSD.

drill > " to " < human, transport, metal materials > ". Therefore, as the iterations progress, unknown nodes are gradually identified and transformed into known nodes, thus enhancing the connection between known and unknown nodes.

3.3.2. ZSHOI model construction through semantic embeddings

As illustrated in Fig. 4, the semantic branch is the primary difference between the word embedding-based and semantic embedding-based ZSHOI model. In the forward process, the semantic branch only extracts word embeddings, while in the backward process, the semantic branch aims to extract semantic embeddings. The backward process's semantic branch generates semantic embeddings using the GCN to distil knowledge from the HOI graph and word embeddings, where word embeddings are generated by the BERT-wwm model with text triplets input. Additionally, the shape of the semantic embedding matrix $\boldsymbol{S}$ of the shape 768 × (seen + unseen) is the same as the word embedding matrix $\boldsymbol{W}$ of the shape

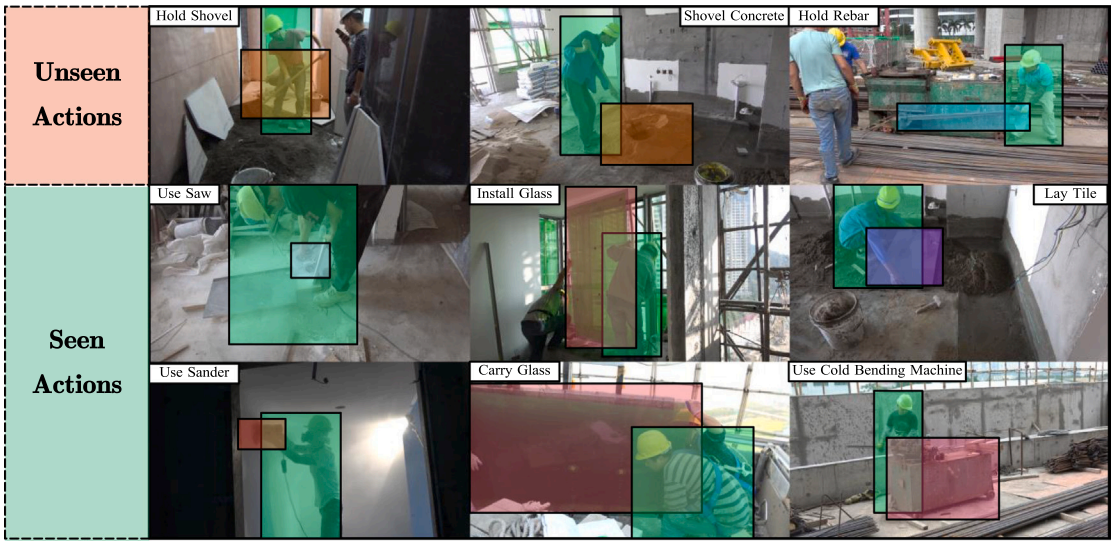


Fig. 10. Qualitative results of the ZSHOI.

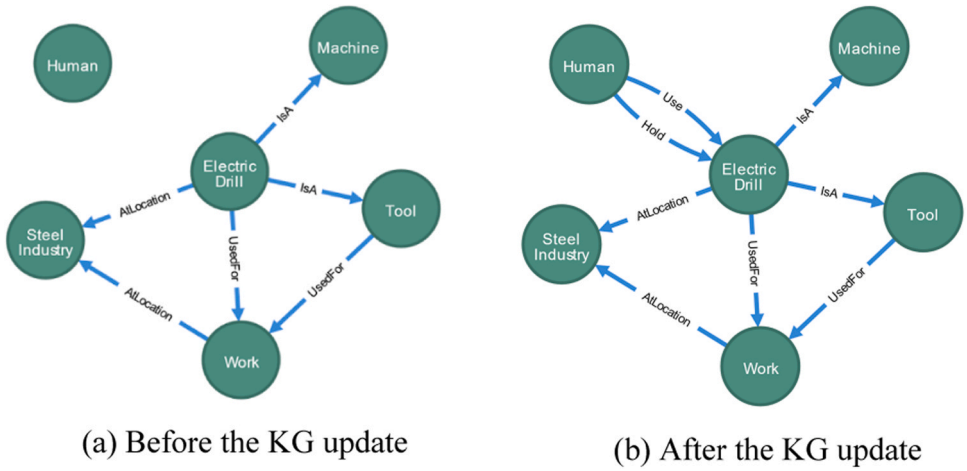


Fig. 11. KG updates. (a) Before updating the KG, the human node was isolated from other nodes. (b) After the KG update, two links are added between the human and electric drill nodes, i.e., use and hold.

768 × (seen + unseen), as these two models share the same visual branch.

3.4. Overall process

The overall process of the framework is shown in Fig. 7. The solid orange line represents the forward process, and the dotted green line represents the backward process. The forward process leverages the VRD initialised by prior information to detect triplets in images, while the backward process augments the prior information. All potential triplets fall into *Known0* in the training dataset or *Unknown* outside the training dataset, where *Unknown* = {*Unknown1*, *Unknown2*, *Unknown3*, …, *UnknownN*}. First, we initialise the VRD by Prior1, connecting *Known0* and *Unknown* via word embeddings. Due to the noise in word embeddings and the limited generalisation ability of *Known0*, the VRD can only recognise a subset of *Unknown* (i.e., *Unknown1*), and the subset *Unknown1* is then converted into known triplets and updated to a general KG (i.e., *Known1* = {*Known0*, *Unknown1*}, *Unknown* = {*Unknown2*, *Unknown3*, …, *UnknownN*}). Next, the backward process constructs Prior2 (i.e., semantic embeddings) to replace Prior1 based on the HOI graph and word embeddings using GCN. The Prior2 then initialises the VRD, which detect triplets from images and recognise a new unknown subset (i.e., *Unknown2*). At this time, *Known2* = {*Known0*, *Unknown1*, *Unknown2*}, *Unknown* = {*Unknown3*, …, *UnknownN*}. As the iteration progresses, the *Known* set grows, and the *Unknown* set shrinks. The iteration should be stopped when the backward process fails to identify new unknown triplets.

4. Experiments

4.1. Experimental setup

4.1.1. Dataset description

As shown in Fig. 8, the ZSHOI dataset contains 13 action classes (10 known and 3 unknown), 42 object or material classes and 73 HOIs. There are approximately 1000 images in total.

4.1.2. Evaluation metric

Mean average precision (mAP) is widely used for object detection and HOI detection tasks, mainly because mAP allows using a single number to compare the performance of different models. Although mAP may penalise those detection results, which are reasonable but not annotated as ground truth (Lu et al., 2016), in this study, we suppose that the ZSD model can detect both the seen and unseen objects correctly and, therefore, can overcome this shortcoming. In addition, in order to fairly compare with existing methods that use mAP as an evaluation metric, we need to evaluate the performance of the ZSHOI model using mAP.

4.1.3. Implementation details

A general KG, ConceptNet 5.5, is used to update construction activity KG. The length of the word embedding and semantic embedding is 768. The GCN model contains four graph convolution layers in the semantic branch with output channels of 2048, 1024, 1024 and 768. Before obtaining the visual embeddings in the visual branch, the concatenated feature is fed into three FC layers, whose output channels are 1024, 512 and 768. The length of the human, object and joint features is 1024, and the length of the coordinate features is 8. The features of known and unknown action nodes are utilised in the training and testing stages, respectively. We use all the action features only in the GHOID task.

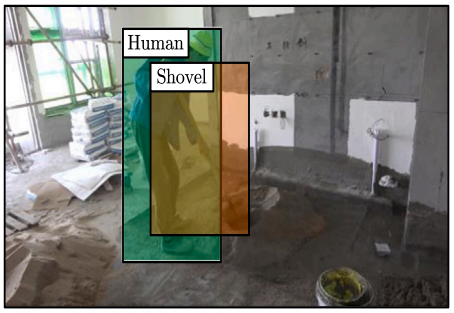
4.2. Quantitative analysis

In this section, we compare the mAP of the word embedding-based and semantic embedding-based ZSHOI model under three different tasks (i.e., Seen, Unseen and Seen + Unseen), which are distinguished by their test categories (i.e., only contains known categories, only contains unknown categories and contains both known and unknown categories). Note that we only perform one forward and one backward process.

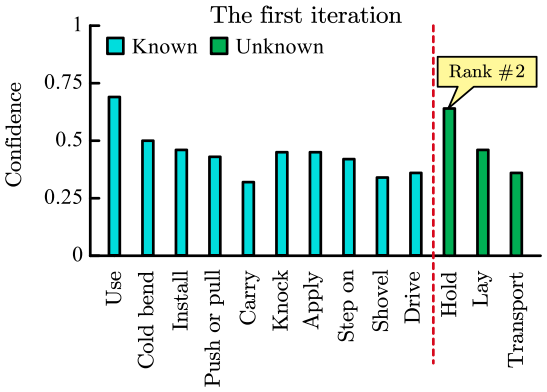
As shown in Table 1, although the mAP performance of these two models is similar in the seen task, the mAP performance of the semantic embedding-based model is better than that of the word embedding-based model in the unseen and seen + unseen tasks, which indicates that the prior information in semantic embeddings is more potent than that in word embeddings.

4.3. Qualitative results

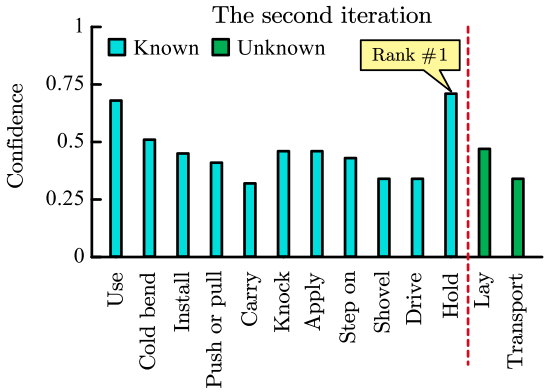
Fig. 9 illustrates the ZSD detection results of the human and object. Our framework can detect known HOIs (the second and third rows) and unknown HOIs (the first row), as shown in Fig. 10. Therefore, we can update the KG after extracting the HOI triplet. For example, new edges can be added between the "human" and "electric drill" nodes, as shown in Fig. 11.



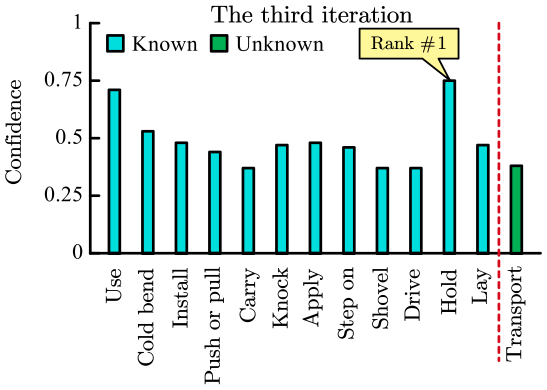
(a.1) Human and a shovel.



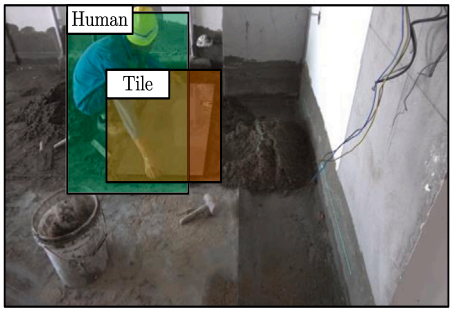
(a.2) At the first iteration, the unseen action "hold" was detected with a confidence level of 0.64, ranking 2nd.



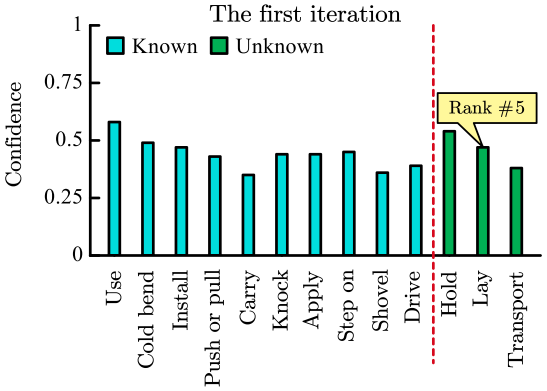
(a.3) At the second iteration, the confidence level for action "hold" was 0.71, ranking 1st.



(a.4) At the third iteration, the confidence level for action "hold" was 0.75, ranking 1st.



(b.1) Human and a tile.

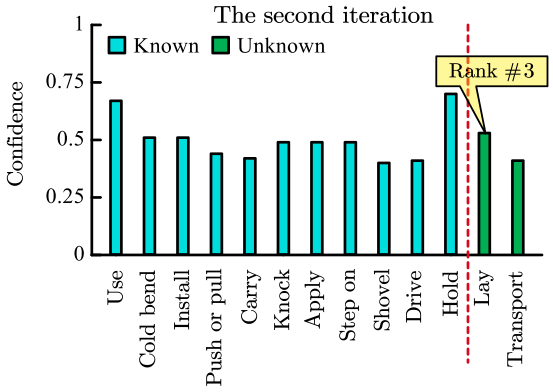


(b.2) At the first iteration, the unseen action "lay" was not detected with a confidence level of 0.47, ranking 5th.

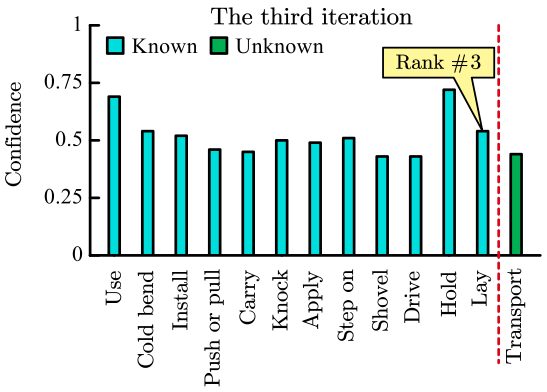
Fig. 12. The detection results of the framework in three iterations in three unseen actions, namely "hold", "transport" and "lay".

4.4. Framework iteration

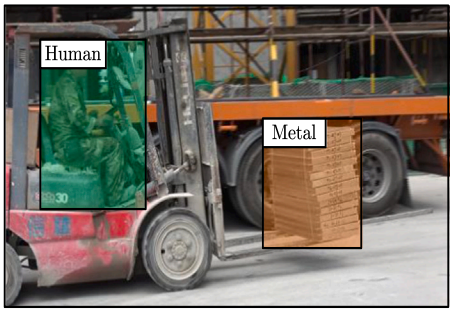
Fig. 12 shows the results of three iterations in three unknown classes, namely "hold", "transport", and "lay". Here, we assume that the results of the top-3 classes will be considered as positive detections. In the first iteration, unknown action "hold" is correctly detected, with a confidence level of 0.64, ranking second out of 13 classes, as shown in Fig. 12(a.2). Then, in the second iteration, we re-train the framework as the unknown class "hold" becomes known class, and the unknown action "lay" is detected with a confidence level of 0.53, ranking third, as shown in Fig. 12(b.3). We also re-train the framework, and detection results are shown in Fig. 12(c.2 to c.4) in the third iteration. Although the unknown action "transport" confidence level slightly increased from 0.47 to 0.54, the framework



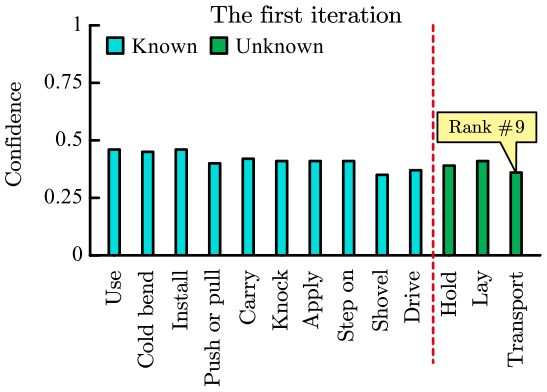
(b.3) At the second iteration, the unseen action "lay" was detected with a confidence level of 0.53, ranking 3rd.



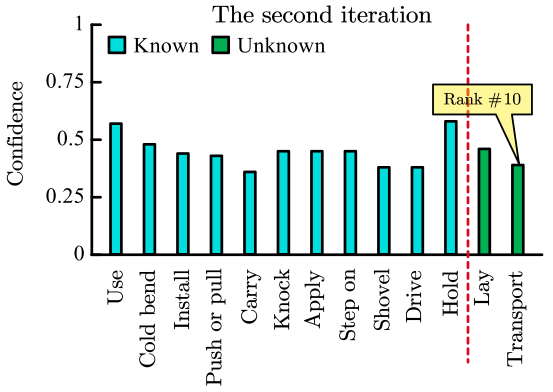
(b.4) At the third iteration, the confidence level for action " lay" was 0.54, ranking 3rd.



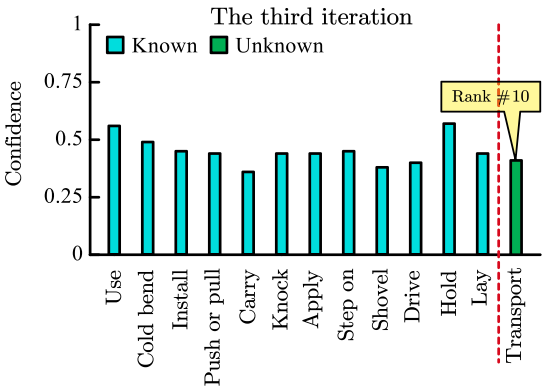
(c.1) Human transports metal materials.



(c.2) At the first iteration, the unseen action "transport" was not detected with a confidence level of 0.36, ranking 9th.



(c.3) At the second iteration, the unseen action "transport" was not detected with a confidence level of 0.39, ranking 10th.



(c.4) At the third iteration, the unseen action "transport" was not detected with a confidence level of 0.41, ranking 10th.

Fig. 12. (*continued*)

still failed to detect this action. These results indicate that although the generalisation capability of the framework can be enhanced by adding known class data, this enhancement is not limitless.

4.5. Case study

In this section, an onsite monitoring scenario is used to reveal the potential application value of the proposed method. In construction sites, abnormal, unregulated or unsafe human behaviours are usually associated with specific objects that may distract workers or pose safety issues. The interactions between humans and these objects form the basis for detecting abnormal, unregulated or unsafe human behaviours. In order to simplify the task, we suppose these objects can be accurately detected using ZSD or general object detection models, and, therefore, the proposed method can be applied to the known and unknown HOI detection. In this scenario, a welding

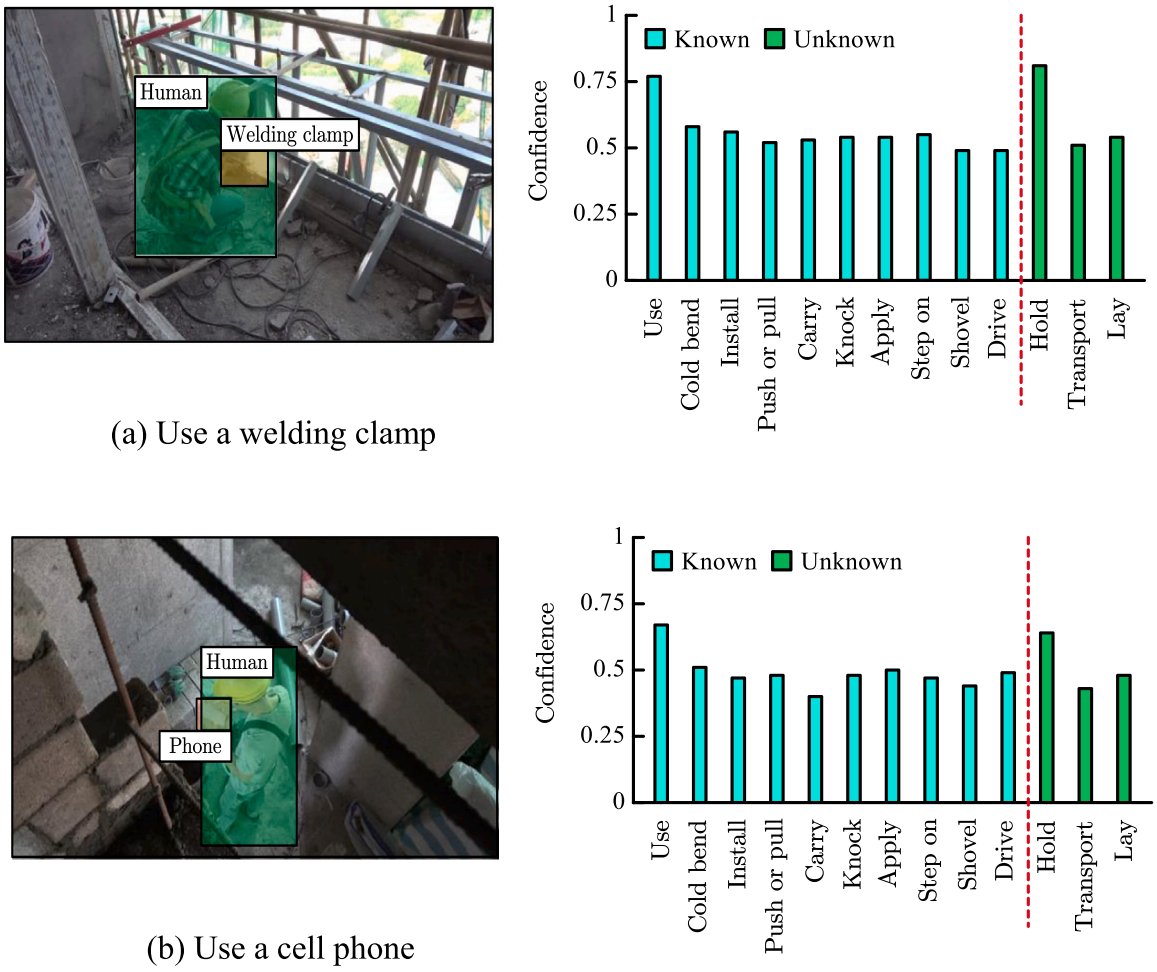


Fig. 13. Two cases in construction site monitoring scenarios.

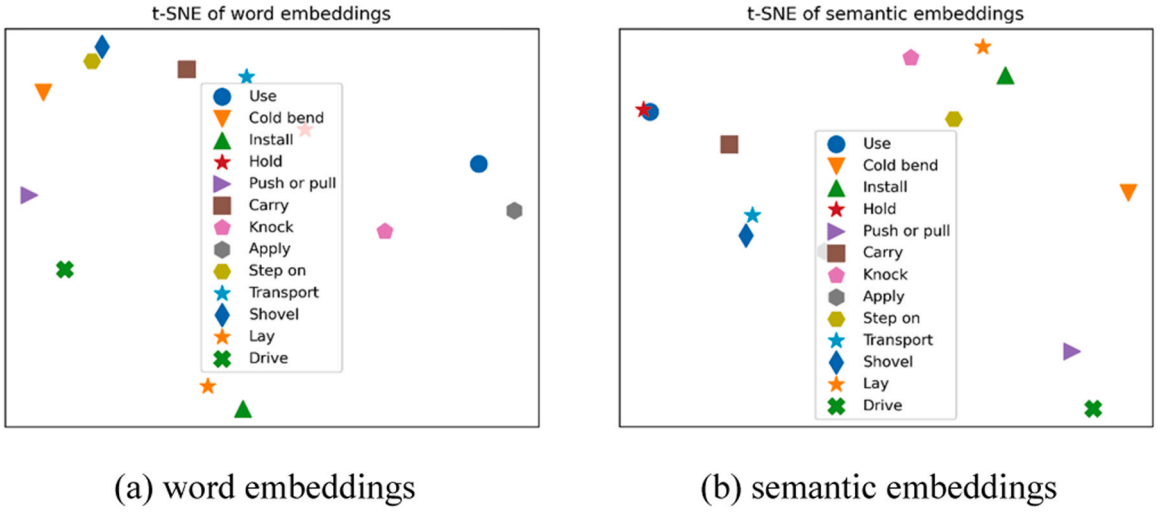
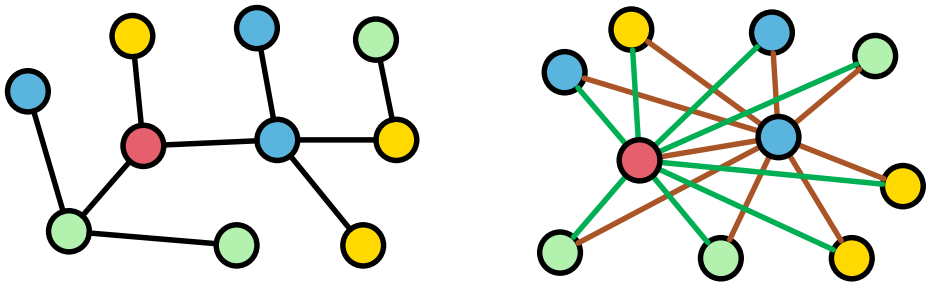


Fig. 14. t-SNE visualisations of the word embedding and semantic embedding. Different shapes of graphics represent different types of actions.

clamp and a cell phone are considered sensitive objects, and the backward process is utilised to detect HOIs. As shown in Fig. 13, the detection results are plotted as histograms. The horizontal axis represents the interactions, where the first ten interactions represent the known, and the last three represent the unknown, and the vertical axis represents the confidence level (between 0 and 1). In the case of a worker using a welding clamp, the highest confidence level is an unknown action "hold" of more than 0.75, followed by a known action "use". In the case of a worker using a cell phone, the top two confidence levels are the known action "use" and the unknown action "hold", both of which exceeded 0.6. These results indicate that although training data is limited, the backward process still can



（a）Random Graph （b）Partial fully connected（PFC）graph

Fig. 15. Two different graph structures. (a) Nodes in this graph are randomly connected. (b) Partial nodes in this graph are fully connected to all the other nodes.

Table 2
mAP comparison of ZSHOI model with different graph structures. Three types of graphs are used: random connective graph (random), partial fully connective graph (PFC) and HOI graph.

Method/Task	Seen	Unseen	Seen + Unseen
Random	0.22	0.40	0.20
PFC	0.16	0.38	0.16
HOI	**0.61**	**0.58**	**0.53**

detect interactions regarding these sensitive objects, both known and unknown, which will help reduce the amount of training data and the effort for data labelling when there is a need to detect unknown human-object interactions.

5. Discussion

5.1. A straightforward comparison between word embeddings and semantic embeddings

Although we have already performed a performance comparison between these two types of prior information in Section 4.2, we expected to demonstrate the difference in their generalisation capabilities straightforwardly. Therefore, we used a dimensionality reduction algorithm t-SNE (van der Maaten and Hinton, 2008) to visualise these two types of prior information, as shown in Fig. 14. Each marker in the figure represents an action, where the star-shaped marker represents an unknown action, and all others are known actions. It is clear that the distance between known and unknown actions is closer in Fig. 14(b) than in Fig. 14(a). For example, 'hold' becomes closer to 'carry', and 'transport' becomes closer to 'shovel', indicating that the semantic word embedding can better generalise the known knowledge to the unknown.

5.1.1. Different structures of HOI graphs

Although the backward process, which uses the HOI graph to generate semantic embeddings, showed better performance than the forward process that uses word embeddings in three different tasks, we would like further to explore the impact of graph structure on model performance. Therefore, two different graph structures were adopted: the random graph and the partial fully connected (PFC) graph. As shown in Fig. 15, nodes in the random graph are randomly connected, while the PFC graph has a few nodes connected to all other nodes. The early stopping (Prechelt, 1998) training technique is adapted to avoid overfitting. The main implication of early stopping is to stop training when the model's performance on the validation set (labelled unseen class instances) starts to degrade so that overfitting due to continuous training can be avoided. Let the random graph serve as a baseline, as its mAP is close to random guesses, as shown in Table 2. In the unseen task, the random graph and PFC graph have the similar mAP, which is lower than that of the HOI graph, indicating that the random and PFC graphs have weaker generalisation ability than the HOI graph and do not well transfer the knowledge learned from the known class to the unknown class. The HOI graph also has the highest mAP in the seen task and the seen + unseen task. These results suggest that a well-structured HOI graph leads to better generalisation performance.

5.1.2. Effect of different seen and unseen splits on the generalisation capability of the ZSHOI model

The key idea of ZSL is to generalise the known knowledge to the unknown. However, model performance is limited by the amount of seen classes. Therefore, it is necessary to analyse the effect of the seen and unseen split on the model generalisation ability. Table 3 shows the percentage of different category instances in the dataset, with 45.40%, 50.71% and 53.89% of the seen categories in the three different seen and unseen splits. This division in Table 3 ensures that

Table 3
The proportion of instances of seen categories in different seen and unseen splits.

Class / S:U	Use (35.48)	Cold bend (0.41)	Install (1.64)	Hold (43.67)	Push or pull (0.31)	Carry (9.51)	Knock (1.23)	Apply (0.92)	Step on (1.84)	Transport (1.94)	Shovel (1.43)	Lay (0.51)	Drive (1.12)	Seen (%)
2:8	S	S	U	U	U	S	U	U	U	U	U	U	U	45.40
5:5	S	S	U	U	U	S	U	S	S	U	S	U	S	50.71
8:2	S	S	S	U	S	S	S	S	S	U	S	U	S	53.89

Note that the capital "S" represents the seen class and the capital "U" represents the unseen class. The number in the bracket represents the percentage of instances of that class in the dataset, e.g., "Use (35.48)" indicates that the seen class instances of "Use" account for 35.48% of all instances in the dataset.

Table 4
mAP comparison of ZSHOI model with different seen/unseen split.

Seen/unseen split/Task	Seen	Unseen	Seen + Unseen
2:8	**0.63**	0.10	0.18
5:5	0.61	0.19	0.38
8:2	0.61	**0.58**	**0.53**

Table 5
mAP comparison of ZSHOI with different feature combinations. Four types of features are used: human features (H), object features (O), union features (U) and space coordinate features (S).

Method/Task	Seen	Unseen	Seen + Unseen
H	0.36	0.45	0.29
O	0.38	0.43	0.35
U	0.45	0.41	0.40
H+O	0.52	0.48	0.47
H+O+U	0.52	0.50	0.48
H+O+U+S	**0.61**	**0.58**	**0.53**

the number of seen and unseen instances is close to 1:1, thus reducing the interference in the results due to unbalanced training and testing instances and helping to mitigate the over-fitting issue. As shown in Table 4, as the seen and unseen split grows, the mAP in the unseen task and seen+unseen task also increases, which indicates that the increase in seen classes improves the model generalisation capability. The three splits have similar mAP in the seen task, mainly because the seen class instances used for training are pretty similar.

5.1.3. Effect of different feature fusion styles on the performance of the ZSHOI model

As mentioned in Sections 1 and 2, we introduced additional features to mitigate the action polysemy problem. In this section, we use ablation experiments to analyse this issue. As shown in Table 5, four types of features are used: human features (H), object features (O), union features (U) and space coordinate features (S). As the numbers of fused features increased, the mAP of these three tasks also increased, suggesting that multi-feature fusion improves model performance and may alleviate the action polysemy issue. Notably, when these three features (i.e., H, O, and U) were used individually, the human feature obtained the highest mAP in the seen task and had a similar mAP to the other features in the unseen task and seen + unseen task, indicating that the human region may contain richer information.

Table 6
mAP comparison of ZSHOI with other methods.

Method/Task	Seen	Unseen	Seen + Unseen
LDHOI	0.14	0.39	0.22
ConSE	0.59	0.40	0.16
CLHOI	0.22	0.49	0.22
Word embeddings	0.60	0.41	0.41
Semantic embeddings	**0.61**	**0.58**	**0.53**

5.1.4. Effect of ZSD on the scope of unknown triplets

The HOI detection of novel objects relies heavily on ZSD to detect novel objects. However, the ZSD, pre-trained in a public domain dataset, may not perform well when applied to the construction domain. For example, we choose the PL-ZSD model (Rahman et al., 2020), which was proposed in 2020, to detect the novel object in the construction industry. More specifically, the PL-ZSD model is pre-trained in the Microsoft COCO dataset (Lin et al., 2014), and the word embeddings of unseen classes are replaced by the word embeddings of the construction industry objects, such as bricks, glasses and drills, to simulate the unseen object detection in construction contexts. As shown in Fig. 16, the PL-ZSD model cannot detect these three kinds of unseen objects, indicating that the scope of training data limits the generalisation ability of this model. Apart from the scope of training data, according to Pan (Wang et al., 2020), ZSD is also sensitive to prior information, and robust prior information can improve the performance of ZSD. Therefore, we suggest enhancing the prior information in the construction context to improve the ZSD performance, as a better ZSD can detect a more extensive range of novel objects and thus expand the scope of triplets.

5.1.5. Comparison with other methods

In this part, we compare the word embedding-based and semantic embedding-based ZSHOI models with the following approaches: LDHOI (Xu et al., 2019), CLHOI (Kato et al., 2018) and adjusted ConSE (Norouzi et al., 2013). As shown in Table 6, the semantic embedding-based ZSHOI model achieves the best performance in these three tasks, demonstrating our approach's superiority.

Fig. 16. The detection results of the PL-ZSD method in the construction context.

6. Conclusions

As an efficient knowledge management tool, KG using its structural storage and logical inference capabilities, can help improve the efficiency in AEC industries, such as indoor scene design, project management and construction site monitoring. This paper is dedicated to the research of KG updates based on explicit HOI extraction at the activity level in the AEC industry with the application to the interior decoration construction process. A novel computer vision-based explicit relationship extraction framework, called Image2Triplets, is proposed to update the data layer of the construction activity KG. We introduce the ZSHOI technique in the visual-based KG updating to address the new relationships extraction issue. Considering the lack of comprehensive and large-scale datasets in the industry and the complexity of construction sites, performing HOI detections with limited data is challenging. The framework alleviates this issue by introducing the ZSL technique and incorporating the prior information from the general KG to enable the migration of knowledge learnt on known classes to unknown classes.

More specifically, we combine ZSD and ZSHOI detection techniques to develop two iterative processes (i.e., a forward process and a backward process) to extract explicit HOI triplets from images. We use prior information to initialise the ZSD and the ZSHOI detection model in VRD, and these two models are used together for known and unknown HOI extraction. The proposed framework is tested using construction images of architectural decoration processes. The results show that the framework can detect both known and unknown triplets and that the extracted triplets can be used to update the data layer of the construction activity KG. In addition, the experiments verify that the HOI relationship graph can enhance the prior information, thus improving the performance of VRD and that the backward process achieves the best results in the Unseen task compared to existing methods.

However, several issues still need to be addressed. For example, we assumed that the ZSD model could correctly detect the seen and unseen entities, while the actual performance of ZSD is generally poor, and at this stage, the performance of the ZSHOI model is not yet satisfactory. Furthermore, some manual intervention is still required in KG updating, and we have not yet tested our framework on large-scale datasets. In the future, we will further address these issues and focus on the visual-based multimodal KG construction, ontology updates and automatic KG completion.

CRediT authorship contribution statement

Zaolin Pan: Methodology, Formal analysis, Investigation, Writing – original draft. **Cheng Su:** Resources, Writing – review & editing, Funding acquisition. **Yichuan Deng:** Conceptualization, Methodology, Writing – review & editing. **Jack Cheng:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the support of the Guangdong Science Foundation, Grant No. 2018A030310363; the support of the Science and Technology Program of Guangzhou, Grant No. 201804020069; and the support of the China-Singapore International Joint Research Institute, Project No. 205-D119001.

Code availability

The code of the proposed framework is available on GitHub (https://github.com/CrossStyle/Image2Triplets).

References

A. Ahmeti, Updates in the Context of Ontology-Based Data Management, Technische Universität Wien, 2020.
Anonl. Neo4j, Neo4j, 2021.
Anumba, P., Chimay, J., Egbu, 2008. Charles and Carrillo, Knowledge management in construction. John Wiley & Sons.
Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R., 2020. Detecting human-object interactions via functional generalisation. The Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press, pp. 10460–10469. https://doi.org/10.1609/aaai.v34i07.6616
Berners-Lee, T., Hendler, J., Lassila, O.R.A., 2001. The semantic web. Sci. Am. 284, 34–43.
Brin, S., 1999. Extracting patterns and relations from the world wide web. In: Atzeni, P., Mendelzon, A., Mecca, G. (Eds.), The World Wide Web and Databases. Springer Berlin Heidelberg, Berlin, Germany, pp. 172–183. https://doi.org/10.1007/10704656_11
Chao, Y., Liu, Y., Liu, X., Zeng, H., Deng, J., 2018. Learning to detect human-object interactions. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, Lake Tahoe, NV, USA, pp. 381–389. https://doi.org/10.1109/WACV.2018.00048
Chen, W., Das, M., Chen, K., Cheng, J.C.P., 2020. Ontology-based data integration and sharing for facility maintenance management. Construction Research Congress 2020:Computer Applications. American Society of Civil Engineers (ASCE), Reston, VA, USA, pp. 1353–1362. https://doi.org/10.1061/9780784482865.143
Cheng, J., Trivedi, P., Law, K., 2002. Ontology mapping between Psl And Xml- based standards for project scheduling. Int. Conf. Concurr. Eng. Constr. 3.
Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, G. Hu, Revisiting Pre-Trained Models for Chinese Natural Language Processing, CoRR. abs/2004.1 (2020).⟨https://arxiv.org/abs/2004.13922⟩. Model link: ⟨https://github.com/ymcui/MacBERT⟩.
Dörk, M., Carpendale, S., Williamson, C., 2011. EdgeMaps: visualising explicit and implicit relations. 78680G Chung Wong, P., Park, J., Hao, M.C., Chen, C., Börner, K., Kao, D.L., Roberts, J.C. (Eds.), Vis. Data Anal. 2011. https://doi.org/10.1117/12.872578
dos Santos, C.N., Xiang, B., Zhou, B., 2015. Classifying relations by ranking with convolutional neural networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. The Association for Computer Linguistics (ACL), Beijing, China, pp. 626–634. https://doi.org/10.3115/v1/p15-1061
P. Drucker, The effective executive, Routledge, 2018. https://doi.org/https://doi.org/10.4324/9780080549354.
Fang, W., Ma, L., Love, P.E.D., Luo, H., Ding, L., Zhou, A., 2020. Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. Autom. Constr. 119, 103310. https://doi.org/10.1016/j.autcon.2020.103310
Fazio, P., Bédard, C., Gowri, K., 1991. Building envelope design query language. Computing in Civil Engineering and Symposium on Data Bases: Proceedings of the Seventh Conferenc. American Society of Civil Engineers (ASCE), New York, NY, USA, pp. 719–728.
Flynn, P., Zhou, L., Maly, K., Zeil, S., Zubair, M., 2007. Automated Template-Based Metadata Extraction Architecture. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (Eds.), International Conference on Asian Digital Libraries. Springer Berlin Heidelberg, Berlin, Germany, pp. 327–336. https://doi.org/10.1007/978-3-540-77094-7_42
Gao, C., Zou, Y., Huang, J.-B., 2018. iCAN: instance-centric attention network for human-object interaction detection. ArXiv E-Prints arXiv:1808.10437.
Gao, R., Hou, X., Qin, J., Chen, J., Liu, L., Zhu, F., Zhang, Z., Shao, L., 2020. Zero-VAE-GAN: generating unseen features for generalised and transductive zero-shot learning. IEEE Trans. Image Process. 29, 3665–3680. https://doi.org/10.1109/TIP.2020.2964429
Gruber, T.R., 1995. Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum. Comput. Stud. 43, 907–928. https://doi.org/10.1006/ijhc.1995.1081
Gupta, S., Malik, J., 2015. Visual semantic role labeling. ArXiv E-Prints arXiv:1505.04474.
He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2020. Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. 42, 386–397. https://doi.org/10.1109/TPAMI.2018.2844175
Ji, G., Liu, K., He, S., Zhao, J., 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Singh, S.P., Markovitch, S. (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017. AAAI Press, San Francisco, California, USA, pp. 3060–3066 (San Francisco, California, USA).
Kamara, J.M., Augenbroe, G., Anumba, C.J., Carrillo, P.M., 2002. Knowledge management in the architecture, engineering and construction industry. Constr. Innov. 2, 53–67. https://doi.org/10.1108/14714170210814685
Kambhatla, N., 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL), Barcelona, Spain. https://doi.org/10.3115/1219044.1219066

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., Xing, E.P., 2019. Rethinking knowledge graph propagation for zero-shot learning. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Long Beach, CA, USA, pp. 11479–11488. https://doi.org/10.1109/CVPR.2019.01175

Kamsu-Foguem, B., Abanda, F.H., 2015. Experience modeling with graphs encoded knowledge for construction industry. Comput. Ind. 70, 79–88. https://doi.org/10.1016/j.compind.2015.02.004

Kato, K., Li, Y., Gupta, A., 2018. Compositional learning for human object interaction. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, Switzerland, pp. 247–264. https://doi.org/10.1007/978-3-030-01264-9_15

Lampert, C.H., Nickisch, H., Harmeling, S., 2014. Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Mach. Intell. 36, 453–465. https://doi.org/10.1109/TPAMI.2013.140

Leng, S., Hu, Z.Z., Luo, Z., Zhang, J.P., Lin, J.R., 2019. Automatic MEP knowledge acquisition based on documents and natural language processing. Proceedings of the 36th International Conference of CIB W78. International Council for Research and Innovation in Building and Construction, Newcastle-upon-Tyne, UK, pp. 800–809.

Li, J., Wang, Z., Wang, Y., Hua, Z., Jing, W., 2020. Research on distributed search technology of multiple data sources intelligent information based on knowledge graph. J. Signal Process. Syst. 1–10. https://doi.org/10.1007/s11265-020-01592-5

Li, R., Mo, T., Yang, J., Jiang, S., Li, T., Liu, Y., 2020. Ontologies-Based Domain Knowledge Modeling and Heterogeneous Sensor Data Integration for Bridge Health Monitoring Systems. IEEE Trans. Ind. Inform. 17, 321–332. https://doi.org/10.1109/tii.2020.2967561

Li, X., Lyu, M., Wang, Z., Chen, C.-H., Zheng, P., 2021. Exploiting knowledge graphs in industrial products and services: a survey of key aspects, challenges, and future perspectives. Comput. Ind. 129, 103449. https://doi.org/10.1016/j.compind.2021.103449

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, USA, pp. 936–944 https://doi.org/10.1109/CVPR.2017.106. Model link: https://github.com/pytorch/vision/blob/main/torchvision/models/detection/backbone_utils.py.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 740–755.

Liu, Y., Xu, B., Yang, Y., Chung, T., Zhang, P., 2019. Constructing a hybrid automatic Q&A system integrating knowledge graph and information retrieval technologies. Foundations and Trends in Smart Learning. Springer, Singapore, Singapore, pp. 67–76. https://doi.org/10.1007/978-981-13-6908-7_9

Liu, Y., Yuan, J., Chen, C.W., 2020. ConsNet: learning consistency graph for zero-shot human-object interaction detection. Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, pp. 4235–4243. https://doi.org/10.1145/3394171.3413600

Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L., 2016. Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016. Springer International Publishing, Cham, Switzerland, pp. 852–869. https://doi.org/10.1007/978-3-319-46448-0_51

Ma, Z., Liu, Z., Wei, Z., 2016. Formalized representation of specifications for construction cost estimation by using ontology. Comput. -Aided Civ. Infrastruct. Eng. 31, 4–17. https://doi.org/10.1111/mice.12175

Martinez, P., Al-Hussein, M., Ahmad, R., 2019. A scientometric analysis and critical review of computer vision applications for construction. Autom. Constr. 107, 102947. https://doi.org/10.1016/j.autcon.2019.102947

Miwa, M., Bansal, M., 2016. End-to-end relation extraction using lstms on sequences and tree structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. The Association for Computer Linguistics (ACL), Berlin, Germany, pp. 1105–1116. https://doi.org/10.18653/v1/p16-1105

Mohan, S., 1990. Expert systems applications in construction management and engineering. J. Constr. Eng. Manag. 116, 89–99. https://doi.org/10.1061/(ASCE)0733-9364(1990)116:1(87)

Moreo, A., Eisman, E.M., Castro, J.L., Zurita, J.M., 2013. Learning regular expressions to template-based FAQ retrieval systems. Knowl. Based Syst. 53, 108–128. https://doi.org/10.1016/j.knosys.2013.08.018

Neches, R., Fikes, R.E., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.R., 1991. Enabling technology for knowledge sharing. AI Mag. 12, 36. https://doi.org/10.1609/aimag.v12i3.902

Niu, J., Issa, R.R.A., 2012. Framework for production of ontology-based construction claim documents. Computing in Civil Engineering 2012. American Society of Civil Engineers (ASCE), Reston, VA, USA, pp. 9–16. https://doi.org/10.1061/9780784412343.0002

M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, J. Dean, Zero-Shot Learning by Convex Combination of Semantic Embeddings, ArXiv E-Prints. (2013) arXiv:1312.5650.

Ovchinnikova, E., Kühnberger, K.-U., 2006. Aspects of automatic ontology extension: adapting and regeneralizing dynamic updates. Proceedings of the Second Australasian Workshop on Advances in Ontologies - Volume 72. Australian Computer Society Inc.,, AUS, pp. 51–60.

Pan, Z., Su, C., Deng, Y., Cheng, J., 2021. Video2Entities: a computer vision-based entity extraction framework for updating the architecture, engineering and construction industry knowledge graphs. Autom. Constr. 125, 103617. https://doi.org/10.1016/j.autcon.2021.103617

Park, M., Lee, K., Lee, H., Jiayi, P., Engineering, J.Y., 2013. Ontology-based construction knowledge retrieval system. KSCE J. Civ. Eng. 17, 1654–1663. https://doi.org/10.1007/s12205-013-1155-6

Prechelt, L., 1998. Early stopping - but when? In: Orr, G.B., Müller, K.-R. (Eds.), Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 55–69. https://doi.org/10.1007/3-540-49430-8_3

Rahman, S., Khan, S., Porikli, F., 2019. Zero-shot object detection: learning to simultaneously recognise and localise novel concepts. Computer Vision – ACCV 2018. Springer International Publishing, Cham, Switzerland, pp. 547–563. https://doi.org/10.1007/978-3-030-20887-5_34

Rahman, S., Khan, S., Barnes, N., 2020. Improved visual-semantic alignment for zero-shot object detection. Proc. AAAI Conf. Artif. Intell. 34, 11932–11939. https://doi.org/10.1609/aaai.v34i07.6868

Rasmussen, M.H., Lefrançois, M., Pauwels, P., Hviid, C.A., Karlshøj, J., 2019. Managing interrelated project information in AEC knowledge graphs. Autom. Constr. 108, 102956. https://doi.org/10.1016/j.autcon.2019.102956

Romera-Paredes, B., Torr, P.H.S., 2015. An embarrassingly simple approach to zero-shot learning. Proceedings of the 32nd International Conference on Machine Learning. PMLR, Lille, France, pp. 2152–2161.

Smith, M.L., Smith, L.N., Hansen, M.F., 2021. The quiet revolution in machine vision - a state-of-the-art survey paper, including historical review, perspectives, and future directions. Comput. Ind. 130, 103472. https://doi.org/10.1016/j.compind.2021.103472

Speer, R., Chin, J., Havasi, C., 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. In: Singh, S.P., Markovitch, S. (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017. AAAI Press, San Francisco, California, USA, pp. 4444–4451.

K.E. Sveiby, A. Risling, Kunskapsföretaget: seklets viktigaste ledarutmaning?, LiberFörlag, 1986.

Ulutan, O., Iftekhar, A.S.M., Manjunath, B.S., 2020. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Seattle, WA, USA, pp. 13614–13623. https://doi.org/10.1109/CVPR42600.2020.01363

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., Guo, M., 2018. RippleNet: propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, pp. 417–426. https://doi.org/10.1145/3269206.3271739

Wang, J., Mu, L., Zhang, J., Zhou, X., Li, J., 2020. On intelligent fire drawings review based on building information modeling and knowledge graph. Constr. Res. Congr. 2020 812–820. https://doi.org/10.1061/9780784482865.086

Wang, Q., Mao, Z., Wang, B., Guo, L., 2017. Knowledge graph embedding: a survey of approaches and applications. IEEE Trans. Knowl. Data Eng. 29, 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499

Wang, S., Yap, K.-H., Yuan, J., Tan, Y.-P., 2020. Discovering human interactions with novel objects via zero-shot learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, pp. 11649–11658. https://doi.org/10.1109/CVPR42600.2020.01167

Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014. Knowledge graph embedding by translating on hyperplanes. Proc. AAAI Conf. Artif. Intell. 28.

Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S., 2019. Learning to detect human-object interactions with knowledge. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Long Beach, CA, USA, pp. 2019–2028. https://doi.org/10.1109/CVPR.2019.00212

Xue, F., Lu, W., 2020. A semantic differential transaction approach to minimising information redundancy for BIM and blockchain integration. Autom. Constr. 118, 103270. https://doi.org/10.1016/j.autcon.2020.103270

Yan, C., Zheng, Q., Chang, X., Luo, M., Yeh, C.-H., Hauptman, A.G., 2020. Semantics-preserving graph propagation for zero-shot object detection. IEEE Trans. Image Process. 29, 8163–8176. https://doi.org/10.1109/TIP.2020.3011807

Zhang, L., Wang, X., Yao, L., Zheng, F., 2020. Zero-shot object detection with textual descriptions using convolutional neural networks. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, Glasgow, UK, pp. 1–6. https://doi.org/10.1109/IJCNN48605.2020.9207417

Zhong, B., Li, H., Luo, H., Zhou, J., Fang, W., Xing, X., 2020. Ontology-based semantic modeling of knowledge in construction: classification and identification of hazards implied in images. J. Constr. Eng. Manag. 146. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001767

Zhu, P., Wang, H., Saligrama, V., 2020. Don't even look once: synthesising features for zero-shot detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, pp. 11690–11699. https://doi.org/10.1109/CVPR42600.2020.01171

Zhu, P., Wang, H., Saligrama, V., 2020. Zero shot detection. IEEE Trans. Circuits Syst. Video Technol. 30, 998–1010. https://doi.org/10.1109/TCSVT.2019.2899569

Zhu, Y., Zhou, W., Xu, Y., Liu, J., Tan, Y., 2017. Intelligent learning for knowledge graph towards geological data. Sci. Program.(2017). https://doi.org/10.1155/2017/5072427