Computer Vision and Deep Learning to Manage Safety in Construction: Matching Images of Unsafe Behavior and Semantic Rules

Weili Fang, Peter E.D. Love, Lieyun Ding, Shuangjie Xu, Ting Kong, and Heng Li

***Abstract*—The determination of people.s unsafe behavior from images in construction has been typically based on hand-made rule approaches, which renders it difficult to identify multiple acts of unsafe behavior within an image and accordingly apply safety rules. This article aims to develop a computer vision and deep learning method that can match images of people's unsafe behavior with semantic safety rules. Our proposed method consists of: 1) image feature representation; 2) safety rule feature representation; and 3) feature fusion similarity whereby unsafe behavior extracted from an image is matched with safety rules. We validate the effectiveness of our method using an image database of people's unsafe behavior from different sites associated with the construction of the Wuhan Metro Project (China). The results of our research explicitly demonstrate that our method is robust and can accurately recognize people's unsafe behavior and the corresponding safety rule that has been contravened. To this end, we suggest that construction organizations can use our method to manage safety better as part of a behavior-based safety strategy and thus prevent accidents.**

***Index Terms*—Computer vision, deep learning, image-text matching, safety, unsafe behaviour.**

I. Introduction

DESPITE the ongoing effort being made by construction organizations to improve safety in their projects, accidents continue to occur on-sites worldwide [3], [30], [39], [44]. In China, for example, over 2850 people were killed on construction sites from 2012 to 2016 [33]. However, such deaths are preventable, particularly as 88% of accidents that materialize on-site are attributable to people's unsafe behavior [15]. Hence, it is necessary to mitigate people's unsafe behavior to reduce accidents and improve safety performance in construction.

Behavior-based safety (BBS) is an effective approach to modify people's unsafe behavior to undertake work more safely [4], [6], [9], [29]. Traditionally the process of observing unsafe behavior has been reliant on a construction organization's staff physically walking around a site to monitor and control subcontractors to ensure they are adhering to safety regulations. Indeed, this can be a very time-consuming process. Recognizing the need to improve the efficiency of detecting and identifying people's unsafe behavior on construction sites, computer vision, and deep learning have been used to automatically perform such tasks [8], [10], [12], [13], [26], [42], [49], [50]. For example, Fang *et al*. [11] employed a computer vision approach juxtaposed with a Mask R-CNN to identify people traversing over structural supports above deep foundation pits [12]. Likewise, Fang *et al*. [11] sought to determine whether people wore their safety harness while working at heights [10]. Despite the contribution of these studies, the reliance on the use of hand-made rule-based approaches to detect unsafe behavior from images renders it difficult to identify multiple acts of unsafe behavior within an image and accordingly apply safety. Considering these limitations, we develop a computer vision and deep learning approach that can match people's unsafe behavior with semantic image-safety rules on construction sites. We employ a stacked cross attention network (SCAN), which is an image-text matching approach developed by Lee *et al*. [21] to determine the visual semantic similarities between an imagine (i.e., unsafe behavior) and a sentence (i.e., safety rules). The rationale for drawing on the work of Lee *et al*. [21] is based on the ability of a SCAN to retrieve text accurately from Flickr30K and Microsoft Common Objects in Context (MS-COCO) datasets. An attention mechanism is provided for both images and safety rules to determine possible alignments simultaneously. Furthermore, to achieve higher accuracy results, the hyper-parameters (e.g., learning rate and optimizer) of the SCAN are adjusted accordingly. Extensive datasets (ImageNet, COCO, etc.) are first pretrained to improve feature learning and representation. As a result, people's unsafe behavior implied in images will be determined by computing their semantic similarity with safety rules.

Manuscript received 19 October 2020; revised 31 May 2021; accepted 6 June 2021. Date of publication 2 September 2021; date of current version 14 September 2023. This work was supported by the National Natural Science Foundation of China under Grant 71732001, Grant 51978302, and Grant 51878311. Review of this manuscript was arranged by Department Editor T. Daim. *(Corresponding author: Shuangjie Xu.)*

Weili Fang is with the Department of Building School of Design and Environment, National University of Singapore, 117565, Singapore (e-mail: bdgfw@nus.edu.sg).

Peter E.D. Love is with the School of Civil and Mechanical Engineering, Curtin University, Perth, WA 6845, Australia (e-mail: plove@iinet.net.au).

Lieyun Ding is with the Department of Construction Management, School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China (e-mail: dly@hust.edu.cn).

Shuangjie Xu is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: shuangjiexu@gmail.com).

Ting Kong and Heng Li are with the Department of Building and Real Estate, Faculty of Construction and Environment, Hong Kong Polytechnic University, Hong Kong, China (e-mail: ting.kong@connect.polyu.hk; heng.li@polyu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TEM.2021.3093166.

Digital Object Identifier 10.1109/TEM.2021.3093166

TABLE I
PRIOR WORKS ON COMPUTER VISION AND DEEP LEARNING APPROACHES FOR UNSAFE BEHAVIOR MONITORING

Approach	Description	Author
Yolov2 and affine transformation method	Detect people who enter extractor's working dangerous area	Luo *et al.* [27]
Mask R-CNN and Ontology	Detect people's unsafe behavior by reconstructing knowledge graph from images	Fang *et al.* [13]
Spatial and temporal attention pooling network	Detect people's identity for unsafe behavior management	Wei *et al.* [47]
Mask R-CNN	Detect people who traverse structural supports over a deep foundation pit	Fang *et al.* [12]
Faster R-CNN and Classification deep network	Detect people not wearing safety harness working on height	Fang *et al.* [10]
Faster R-CNN	Detect people not wearing their hardhat in construction site	Fang *et al.* [11]
CNN and long short-term memory	Detect people's abnormal unsafe behavior	Ding *et al.* [8]

The rest of this article is organized as follows. Section I commences our article by reviewing the literature that has examined unsafe behavior using computer vision and deep learning, and image-text matching. Section II introduces and describes our method for matching people's unsafe behavior with safety rules. Section IV uses a case study to test and validate our method, which is followed by a discussion of our results in Section V by particularly placing emphasis on the study's contribution and limitations. Finally Section VI concludes this article.

II. COMPUTER VISION AND DEEP LEARNING

Computer vision techniques provide an ability to rapidly collate and recognize unsafe behavior on construction site cost-effectively and efficiently. However, the performance of vision-based unsafe behavior identification is reliant on the capture and representation of a large number of images/videos [40]. With advances in deep learning, we have seen it increasingly used in conjunction with computer vision to create and adopt systems to monitor and manage people's unsafe behavior on-sites [8], [10], [12].

Table I summarizes key studies that have used deep learning and computer vision as part of a BBS strategy. For example, Fang *et al*. [11] applied a computer vision approach with a Mask R-CNN to identify people traversing structural supports and their spatial relationship to recognize unsafe actions [12]. Likewise, Ding *et al*. [8] applied a hybrid deep learning approach by integrating a convolutional neural network (CNN) and long-term short memory (LSTM) to detect unsafe behavior. The CNN in Ding *et al*.'s work is used to extract spatial features, and the LSTM is used to remove temporal features.

There has been a proclivity for studies such as those identified in Table I to rely on hand-made rule approaches to identify people's unsafe behavior. The orthodoxy of such a hand-made approach uses machine learning models to recognize object and scene, or processes spatial and temporal relations, and then compares the extracted information with rules or existing knowledge to determine people's unsafe behavior. In this case, the scalability of the developed models jeopardizes their reliability, rendering them difficult to recognize a new unsafe behavior. For example, a model designed to detect people wearing their hardhat cannot identify someone wearing their safety harness while working at heights. So, if we want to detect a person not wearing their harness, a new model needs to be developed and trained using a dataset with the specific unsafe behavior. What is more, such a hand-made rule approach is challenging to detect when an image contains more than one unsafe behavior being committed.

A. Image-Text Matching

The task of image-text matching has received considerable attention in the field of computer vision [1], [22], [48]. This task aims to accurately measure the visual semantic similarity between an image visual and a text. Deep learning has been combined with image-text matching approaches to form two categories: 1) joint embedding learning and 2) pairwise similarity learning.

Joint embedding learning focuses on finding latent space and then compute the similarity of images and texts in this found latent space. The approach usually associates features from two modalities (e.g., image, text) with correlation loss using deep canonical correlation analysis (DCCA) [49] and bi-directional ranking loss [24], [45], [46]. The DCCA used a deep learning model to learn nonlinear transformations of image and text, and it achieved a higher accuracy. However, the major issue of DCCA is an eigenvalue [31], [32], [46]. The bidirectional ranking loss [24], [45], [46] extends the triplet loss [41], which needs the similarity distance of matched samples to be much smaller than unmatched samples for the task of image and text matching.

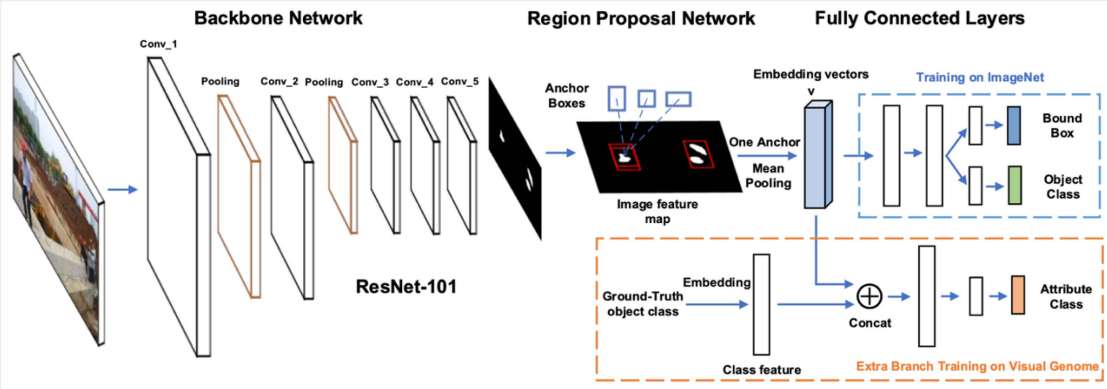

Fig. 1. Network of faster R-CN.

Moreover, the former shared the same disadvantage with the latter in selecting negative samples and margins.

Pairwise similarity learning aims at developing a similarity network to predict the matching score of image and text. Several studies have sought to design similarity score for image and text matching [45], [46], while others have aimed to maximize alignments between regions in images and word in texts [23], [31], [32], [35], [37]. However, this strategy of designing a similarity network may lack efficiency in preparing pairs of images and predicting the matching score of images and texts.

We can observe from the literature that attention has focused on designing embedding networks based on an attention mechanism that attempts to capture the correspondences between the detected visual objects and the textual items (e.g., words) [16]–[18], [21], [25]. With the inclusion of an attention mechanism, research has shown that it can improve the reliability and accuracy of the model's ability to demonstrate a relationship and match of an image and text [21], [35].

III. Research Method

We use a SCAN to achieve higher accuracy on unsafe behavior identification by using an image-text matching method. Our applied method consists of the following three key steps: 1) image features representation; 2) safety rule features representation; and 3) feature fusion similarity for unsafe behavior identification.

A. Image Features Representation

In our article, we firstly extract and encode images regions into features. Then, following Anderson *et al*. [2], a Faster R-CNN represents images with bottom-up attention. The Faster R-CNN was proposed by Ren *et al*. [38] and consists of three key parts: 1) ResNet-101 backbone network; 2) region proposal network (RPN); and 3) fully connected layers for classification and bounding box regression. The Faster R-CNN adopted in this article for feature representation utilizes the ResNet-101, as noted in Fig. 1.

The input image is first fed into the ResNet-101 network to extract image feature maps. Then, the obtained feature maps are input to RPN and then generate a region of interest (RoI) with high objectiveness scores. Next, a small feature map for each proposal box is extracted by RoI pooling. Finally, the fully connected layers are used to refine and output results for classification and bounding box regression, pretrained for object classification on ImageNet datasets. It was also pretrained to predict attribute classes utilizing the database of Visual Genomes, which was adopted from Anderson *et al*. [2] to learn a good representation of semantic feature. A detailed description of the faster R-CNN can be found in Ren *et al*. [38].

To learn high-level semantic representations from images, the faster R-CNN predicts attributes and instances instead of object classes. Thus, the predicted instance contains objects and other salient ones that are not easy to localize (i.e., objects such as "deep excavation" and attributes like "near"). As shown in Fig. 2, we add an extra branch to predict the "attribute class".

As for each selected image region, i, f_i is defined as the mean-pooled convolutional feature. An h-dimensional vector is hence generated from f_i by adopting a fully-connect layer:

$$v_i = W_v f_i + b_v. \tag{1}$$

Therefore, the feature representation extracted from an image is embedding vectors $v = \{v_1, v_2 \cdots v_k\}, v_i \in \mathbb{R}^D$, where each v_i encodes a region.

B. Safety Rules Features Representation

Safety rules need to be mapped on to a dimensional vector space as the regions of images need to connect safety rules and images. A recurrent neural network (RNN) is employed in this research to embed words within the context of safety rule and its semantics [21]. A one-hot vector is used to present the i-th word in a safety rule. Then, an embedding matrix is used to map the one-hot vector of each word to a 300-dimensional vector. Finally, a bidirectional gated recurrent unit (GRU) is used to map the vector and context of the safety rule to the final word's features by summarizing information in both directions.

The GRU is a simplified yet enhanced variant of an RNN, extracting the mapping relationship among time series data with lower complexity and faster computation [5]. The structure of the GRU contains an update and rest gate. Here, the update gate aims to ensure that part of the current hidden state should be

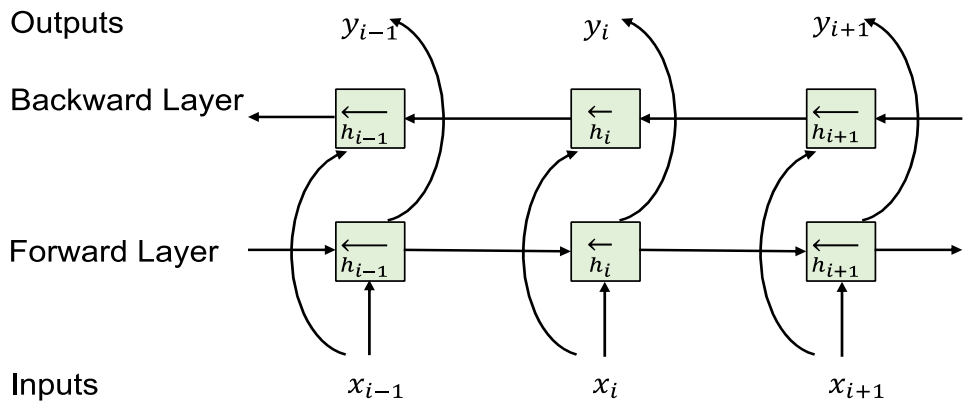

Fig. 2. Structure of bidirectional GRU.

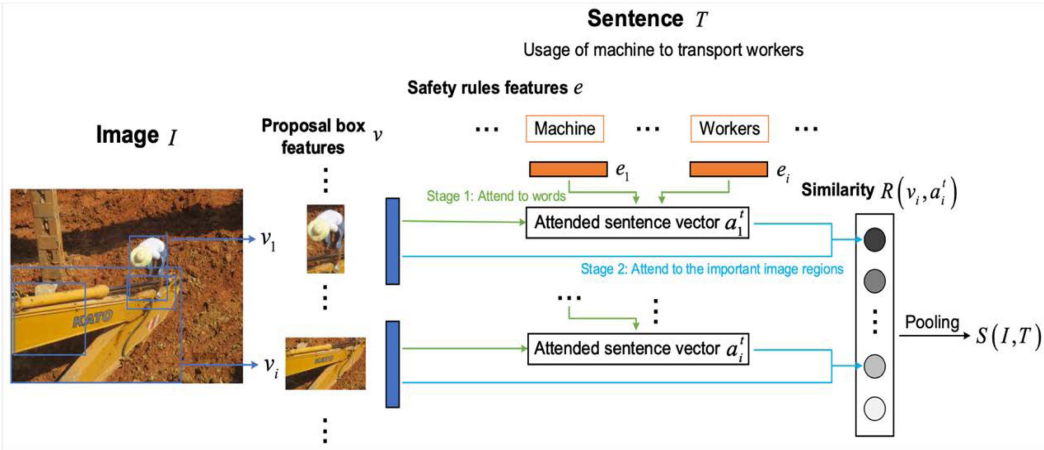

Fig. 3. Workflow of image-safety rules for the SCAN.

updated through the candidate's hidden state. Likewise, the rest gate seeks to ensure that hidden states are partly ignored. The structure of the Bi-directional GRU is denoted in Fig. 2.

The bidirectional GRU contains a forward GRU and a backward GRU, computed by (2) and (3)

$$\overrightarrow{hi} = \overrightarrow{GRU}\,(x_i) \quad i = [1, n] \tag{2}$$

$$\overleftarrow{hi} = \overleftarrow{GRU}\,(x_i) \quad i = [1, n]. \tag{3}$$

The final word feature is computed using (4).

$$e_i = \frac{\left(\overrightarrow{hi} + \overleftarrow{hi}\right)}{2} \quad i = [1, n]. \tag{4}$$

C. Feature Fusion Similarity

A stacked cross attention is used to identify people's unsafe behavior by using both regions in images and words in safety rules as context while computing their similarly [21]. The SCAN inputs consist of two parts: 1) a set of image features; and 2) a set of safety rule features. The output of the SCAN is a similarity score, which is used to measure the similarity of an image-rule pair. Thus, the SCAN is defined by two complementary formulations: 1) image- safety rule; and 2) safety rule-image.

1) Image-Safety Rule Stacked Cross Attention: Fig. 3 presents the workflow of image-safety rules for the SCAN, which consists of two stages: 1) attends to words in the safety rules; 2) determines the importance of the image regions for the safety rules by comparing each image region to the corresponding attended safety rules vector.

Given an image *I* and safety rule *T*, the cosine similarity matrix is used to compute the similarity of all the possible pairs, as presented in (5)

$$s_{ij} = \frac{v_i^T e_j}{\|v_i\| \, \|e_j\|} \quad i \in [1, k] \qquad j \in [1, n] \tag{5}$$

where *k* is the number of detected regions from an image, and *n* is number of words in a safety rule. According to Karpathy *et al.* [19], the threshold of the similarities is set to 0 in this research, and the similarity matrix is normalized as:

$$\overrightarrow{s_{ij}} = [s_{ij}] + / \sqrt{\sum_{k}^{i=1} [s_{ij}]_+^2}. \tag{6}$$

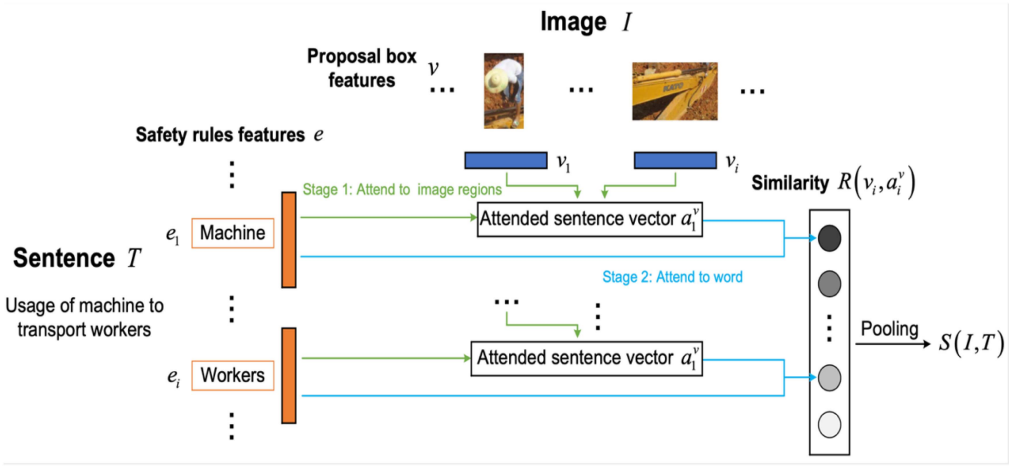

Fig. 4. Workflow of the safety rules-image stacked cross attention.

Following Chorowski *et al*. [7], a weighted combination of word representations is introduced to compute the attended safety rule vector, which can be calculated using (7). More details on attention weight can be found in Luong *et al*. [28]

$$a_i^t = \sum_{n}^{j=1} a_{ij}\, e_{ij} \tag{7}$$

$$a_{ij} = \frac{\exp\left(\lambda_1 \overline{s_{ij}}\right)}{\sum_{n}^{j=1} \exp\left(\lambda_1 \overline{s_{ij}}\right)}. \tag{8}$$

Here, the cosine similarity is used to compute the similarity between a vector of each image region and a vector of an attended safety rule. In this instance, the importance of the territory in each image given the safety rule can be determined. The cosine similarity can be computed by (9)

$$R\left(v_i, a_i^t\right) = \frac{v_i^T a_i^t}{\|v_i\| \left\|a_i^t\right\|}. \tag{9}$$

The semantic similarity of image and safety rule is calculated by LogSumExp (LSE) pooling and average (AVG) pooling [18].

$$S_{\mathrm{LSE}}(I,T) = \log\left(\sum_{i=1}^{k} \exp\left(\lambda_2 R\left(v_i, a_i^t\right)\right)\right)^{(1/\lambda_2)} \tag{10}$$

$$S_{\mathrm{AVG}}(I,T) = \frac{\sum_{i=1}^{k} R\left(v_i, a_i^t\right)}{k} \tag{11}$$

where, λ_2 is a factor determining the magnification of importance for the most relevant pairs of image region featurev_i and attended sentence vector a_i^t.

2) Safety Rules-Image Stacked Cross Attention: Fig. 4 presents the workflow of the safety rules and image SCAN. In this formulation, we focused on image regions referring to each word in safety rules and further exploring the importance of each word by comparing the corresponding image vector of each word. The cosine similarity between the *i*-th region in an image and the *j*-th word in a safety rule is normalized in (12)

$$\overline{s'_{i,j}} = [s_{i,j}] + / \sqrt{\sum_{j=1}^{n} [s_{i,j}]_+^2}. \tag{12}$$

Drawings on the work Lee *et al*. [21] the weighted combination of image region features is defined using the following (13):

$$a_j^v = \sum_{i=1}^{k} a'_{ij} v_i \tag{13}$$

$$a'_{ij} = \exp\left(\lambda_1 \overline{s'_{i,j}}\right) / \sum_{i=1}^{k} \exp\left(\lambda_1 \overline{s'_{i,j}}\right). \tag{14}$$

The relevance between the *j*-th word in the safety rule and an image of people's unsafe behavior is measured as follows:

$$R'\left(e_j, a_j^v\right) = \frac{e_j^T a_j^v}{\|e_j\| \left\|a_j^v\right\|}. \tag{15}$$

Two indices, including LSE pooling and AVGs pooling, are adopted here to determine the final similarity matching score of a given image *I* and shown safety rule *T*

$$S'_{\mathrm{LSE}}(I,T) = \log\left(\sum_{i=1}^{k} \exp\left(\lambda_2 R'\left(e_j, a_j^v\right)\right)\right)^{(1/\lambda_2)} \tag{16}$$

$$S'_{\mathrm{AVG}}(I,T) = \frac{\sum_{j=1}^{n} R'\left(e_j, a_j^v\right)}{n}. \tag{17}$$

We use triplet loss as the objective for the task of matching images of unsafe behavior and semantic rules. In our task, we pay more attention to the hardest negatives, and design our loss function following SCAN[21].

IV. Experiment

Our research focuses on detecting people's unsafe behavior as they approach hazardous work areas on a construction site. We

TABLE II
SELECTED PEOPLE'S UNSAFE BEHAVIOR RELATED TO HAZARDOUS AREA IN CONSTRUCTION SITES

Category	Safety Rules
1	Traversing structural supports (concrete supports or steel supports) without railings over foundation pit
2	Approaching dangerous zone without the usage of edge protection (i.e., roof, scaffold, holes, stairwell, elevator, and foundation pit)
3	Non-usage of safety harness
4	Usage of machine to transport workers
5	Non-usage of safety net
6	Not taking warning during lifting unwanted worker get into the dangerous areas
7	Appearance of workers within the operating radius of excavators during working
8	Absence of warning signage in dangerous zone

Fig. 5. Examples of unsafe behavior extracted from our database.

obtain rules from Chinese safety standards and operating instructions found in the Standard for Construction Safety Assessment of Metro Engineering (GB 50715-2011) [51] and Quality and Safety Check Points of Urban Rail Transit Engineering (2011) [52]. Eight types of unsafe behavior related to hazardous work areas that occur on Chinese construction sites, identified in Fang *et al*. [12], are used as examples to validate the effectiveness and feasibility of our approach, as shown in Table II.

A. Data Collection

An image database is used to train and test our model to validate the feasibility of our computer vision approach (Table II). The photographs of people's unsafe are collected from different sites being used to construct the Wuhan Metro, China. The images in the dataset are acquired from different viewpoints, at varying scales, poses, occlusions and under changing lighting conditions to avoid potential bias. Fig. 5 presents examples of the developed database. Our developed database consists of about 1000 images, which are divided into training and testing database according to ratio of 7:3.

B. Model Training

A pretraining strategy is used to train our proposed multimodal fusion approach to improve the accuracy of recognizing people's unsafe behavior. The MS-COCO dataset is used for pretraining as it contains a rich source of scenarios and objects.

TABLE III
PEOPLE'S UNSAFE BEHAVIOR DETECTION RESULTS (PRECISION AND RECALL)

Category	#1	#2	#3	#4	#5	#6	#7	#8
Precision (Average)	0.97	0.88	0.97	0	0.09	0.5	0.4	0.25
Recall (Average)	0.90	0.85	0.42	0.20	0.40	0.22	0.5	0.40

The sentences are tokenized using the Punkt Sentence Tokenizer [20], provided by NLTK Software Toolkit [34]. A one-layer GRU network is used for safety rule embedding. The features extracted by the GRU and the final joint embedding space are set to 1024, which is the same as the output of the image network. The input dimension of the GRU network is 300. The final dimension of the image embedding is set to 2048. The Faster R-CNN implementation uses an intersection over union threshold of 0.7 for region proposal suppression and 0.3 for object class suppression. The top 36 RoIs with the highest-class detection confidence scores are selected. The final results used both image-text and text-image attention.

The triplet loss is set to 0.2. The mini-batch size is set to 128, and the gradient is clipped by the threshold of maximum gradient norm to 2.0. In the pretraining strategy, the learning rate for the Adam optimizer is set to 0.0002 for 15 epochs, and then we decay the learning rate by 0.1 in the training of our dataset for 15 epochs. The mini-batch is set to 128. The gradient clip is set to 2. We use 1 Nvidia Titan RTX 2080Ti to train our proposed approach, and the whole training process took approximately two days.

C. Evaluation Performance Metrics

Three key performance indicators (KPIs) are used to evaluate the performance of hazard detection: 1) precision; (2) recall; 3) recall rate@k. recall rate@k is a percentage of top-k relevance detected hazard to all items. Here, precision and recall are defined as follows:

$$\begin{cases} \text{precision} = \frac{\text{TP}}{\text{TP+FP}} \\ \text{recall} = \frac{\text{TP}}{\text{TP+FN}} \end{cases} \tag{18}$$

where "precision" referred to the ratio of correctly detected positive unsafe behavior to the total classified positive unsafe behavior. "recall" referred to the ratio of correctly detected positive hazard to all unsafe behavior in an actual class. A "true positive" (TP) refers to an unsafe behavior where the approach makes a correct detection. A "false positive" (FP) occurs when the developed approach detects people's unsafe behavior (A) as another unsafe behavior (B). A "false negative" (FN) means that the developed approach detects a false sample as a positive sample.

D. Results

Due to the limited size of the training sets, a k-fold cross-validation was used to evaluate our approach. The k is set to 5 to balance the computing source and accuracy. The database was randomly divided into five parts. Four parts are used for training. The fifth part is used for testing.

We run our model five times and then present average precision and average recall results in Table III. The R@1, R@2, and R@3 were 81.1%, 87.8%, and 96.4%, respectively, as shown in Table IV. In Table III, we can see that the proposed detection approach has achieved better performance on category #1 than other categories. Here, the precision and recall for the proposed approach on category #1 were 97% and 90%, respectively.

Fig. 6 presents an example of "usage of machine to transport workers," which is used to illustrate how the attention mechanism works in our article. In this example, the faster R-CNN approach is used to detect the interest regions and generate the bounding boxes, as shown in Fig. 6 (left). Then, our attention mechanism is used to associate the obtained bounding box with each word in the safety codes, as noted in Fig. 6 (right). Finally, output the matching results.

Fig. 7 presents a confusion matrix for our detection results for each category of unsafe behavior. Notably, some hazard categories (namely categories #3 to #8) were unable to be detected as the datasets required additional training. However, our method can improve the ability to train data under varying conditions. Examples of the correct detection results are presented in Fig. 8, and examples of error detection are shown in Fig. 9.

Table V presents our ablation experiments on different components of the network. Here, six different experiments are implemented to verify the five basic components of the network.

1) Hard negatives, a usual method in the loss function, need to search for the complex samples in the loss calculation.
2) N-directional GRU indicates the use of a One-directional or Bi-directional GRU network.
3) Image-text SCAN determines whether image attention is used or not.
4) Text-image SCAN, which determines whether text attention is used or not.
5) Pooling indicates whether average or max pooling is used between image and text features.

The second line of Table V is used as our basic network. The five components are not loaded, as we use the base image and text network to extract features. We also compare the similarity of image-text features to obtain a sentence or image recall. The following groups of experiments introduced a new network component, respectively. From Table V, we can conclude that the five components (e.g., hard negatives and N-directional GRU) help in improving performance (i.e., R@1 with 3.3

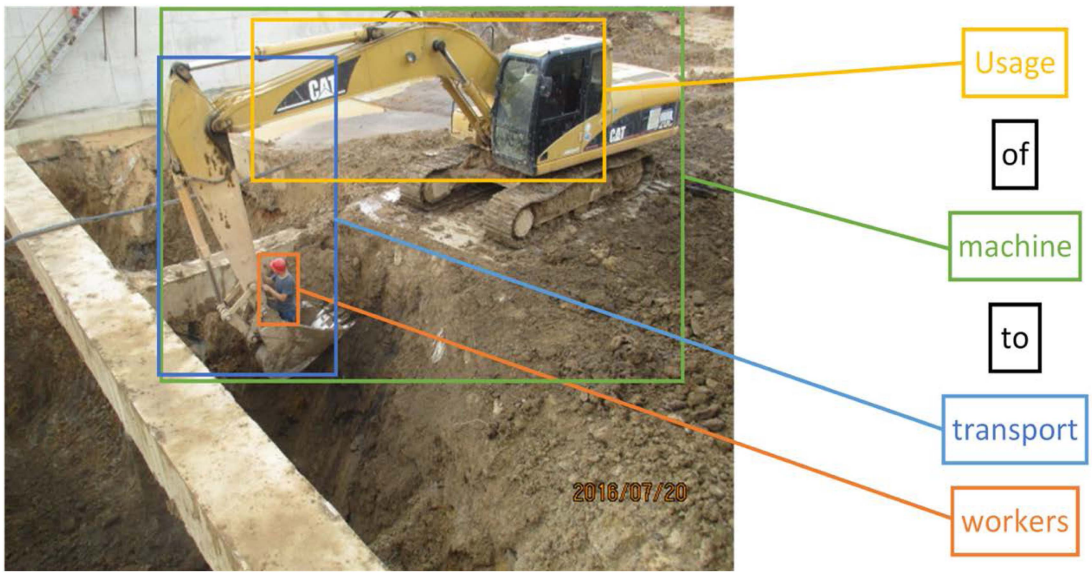

Fig. 6. Example of attention mechanism on the proposed approach.

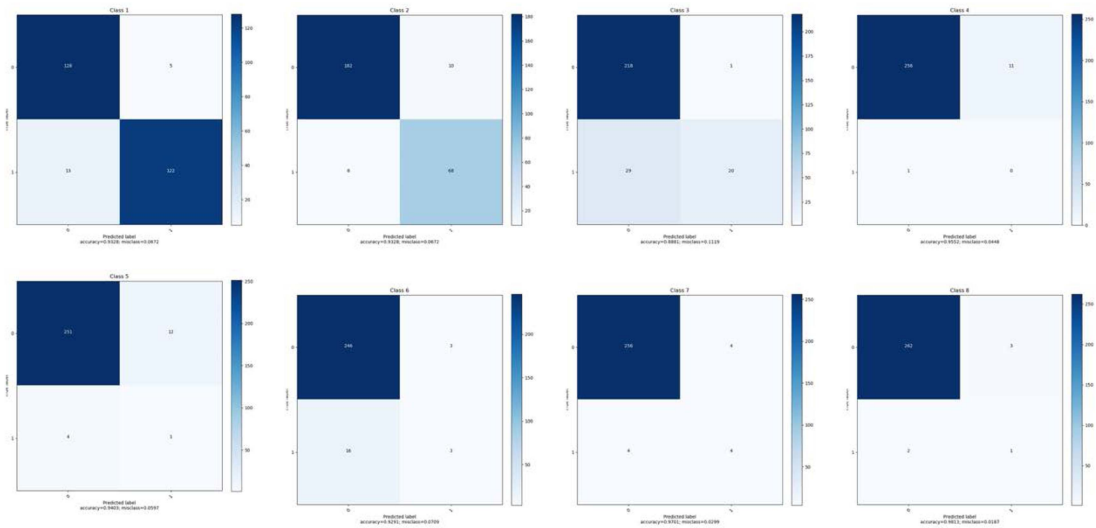

Fig. 7. Confusion Matrix of detection results for each class.

Fig. 8. Examples of correct detection of people's unsafe behavior.

TABLE IV
PEOPLE'S UNSAFE BEHAVIOR DETECTION RESULTS (R@1, R@2, AND R@3)

Folds	1	2	3	4	5	Average
R@1(%)	82.1	80.8	81.5	81.3	79.9	81.1
R@2(%)	88.6	87.9	88.5	87.5	86.4	87.8
R@3(%)	97.3	96.5	96.9	96.1	95.3	96.4

Fig. 9. Examples of error detection of people's unsafe behavior.

TABLE V
ABLATION STUDY OF THE EFFICACY FOR DIFFERENT NETWORK COMPONENTS

Hard Negatives	N-directional GRU	Image-Text Stacked Cross Attention	Text-Image Stacked Cross Attention	Pooling	Results		
					R@1	R@2	R@3
×	One- directional	×	×	MAX	64.3	70.9	78.4
×	One- directional	×	×	MEAN	65.7	72.5	80.1
√	One- directional	×	×	MEAN	69.3	76.1	83.8
√	Bi-directional	×	×	MEAN	72.8	79.5	86.2
√	Bi-directional	√	×	MEAN	76.1	83.6	91.4
√	Bi-directional	√	√	MEAN	81.1	87.8	96.4

and 5.0), particularly for the image-text SCAN and image-text SCAN.

E. Evaluation

Six representative methods for image-text matching in computer science are selected to compare with our approach to evaluate the model's performance. These six types of methods include: 1) deep visual semantic alignments (DVSA), 2) hierarchical multimodal LSTM (HM-LSTM), 3) order-embeddings, 4) SM-LSTM, 5) visual semantic embedding (VSE)++, and 6) semantic concepts and order (SCO), which determines whether image attention is used or not. Tables VI and VII introduce these six approaches.

We use the same training and testing dataset with our applied SCAN approach to compare our method with the others. The hyperparameters (e.g., learning rate) of each model are followed with the original references, respectively. Tables VI and VII presents the detection results, showing our developed model is more accurate than the other six methods in automatically identifying people's unsafe behavior. Additionally, a *t*-test to determine if our model is statistically superior to the results obtained from the six other image-text matching methods. According to Johnson [55], if $p < 0.05$, the performance is significantly different. From Tables VI and VII, we can see that SCO achieved better performance than other methods (e.g., DAVS, HM-LSTM). Thus, we perform a *t*-test between our approach and SCO. The result is presented in Table VIII. The *t*-value for R@1, R@2, and R@3 were 46.687, 27.883, and

TABLE VI
BRIEF INTRODUCTION OF OUR SELECTED SIX APPROACH

Method	Description	Author (Year)
DVSA	A novel combination of CNN and bidirectional RNNs is used.	Andrey and Li [1]
HM-LSTM	Exploited the hierarchical relations (e.g., sentences and phrases, image and regions) for feature representation jointly.	Niu *et al*. [36]
Order-embedding	Encode order structure of visual-semantic hierarchy into learned distributed representations.	Vendrov *et al*. [43]
SM-LSTM	A multimodal context-modulated attention mechanism is used at each timestep.	Huang *et al*. [18]
VSE++	A hard negative is incorporated in the loss function to improve accuracy on retrieval.	Faghri *et al*. [14]
SCO	Image features are correctly represented and semantic order by learning semantic concepts predicted.	Huang *et al*. [16]

TABLE VII
COMPARISON OF STATE-OF-THE-ART APPROACHES

Algorithms	Our approach	DVSA	HM-LSTM	Order-embeddings	SM-LSTM	VSE++	SCO
R@1	**81.1**	40.8	45.3	48.8	56.0	56.6	60.4
R@2	**87.8**	49.6	56.2	60.2	66.8	67.3	76.3
R@3	**96.4**	68.1	73.6	77.1	83.9	84.2	88.2
Training speed (s/image-text pair)	**1.46**	2.13	0.63	0.92	0.84	0.96	1.72
Testing speed (ms)	**417**	523	176	324	295	345	562

TABLE VIII
PERFORMANCE COMPARISON OF METHODS USING *T*-TEST

Indicator	Our method	SCO						*t* value	*p-value*
	Mean	1	2	3	4	5	Mean		
R@1	81.1	59.7	61.1	60.4	59.3	61.4	60.4	46.687	4.899e-13
R@2	87.8	75.9	76.8	76.1	75.5	77.2	76.3	27.883	8.171e-11
R@3	96.4	87.6	88.9	87.9	87.3	89.4	88.2	19.053	3.446e-09

19.053, respectively. The *p*-value for R@1, R@2, and R@3 was 4.899e-13, 8.171e-11, and 3.446e-09, respectively, which were much lower than 0.05, implying that our approach is significantly better than the SCO.

V. DISCUSSION

Ensuring the safety of people in construction is a challenge. Digital technology has an enabling role to play in helping managers to provide a safe workplace [56]. However, the responsibility for safety not only resides with a construction organization and its managers but also everyone on-site. Establishing a culture psychological safety, where people are able to "speak-up" about acts of unsafe behavior without fear of reprisal, has been shown to be difficult to cultivate as error prevention mindset pervades practice [57]–[59]. However, our computer vision and deep learning approach can explicitly help support a BBS strategy as it provides site management with a mechanism to identify unsafe behavior in real-time proactively. In doing so, they can take immediate action to modify people's future intentions by engaging in the process of persuasive communication and education can bring to the fore the woes and whereabouts of safety.

A. Research Contributions

We introduce an attentive matching strategy to acquire a context from both image and safety rules. The similarity between the regions images and words associated with safety rules is obtained using a cross-attention strategy. Our proposed approach can match safety rules with images and is also scalable.

Current hand-made rule-based approaches can only detect people's unsafe behavior using predefined safety rules. However, our approach not only checks safety rules from knowledge learned but also those contained with the dataset. The results from our experiment (Table III) demonstrate we can detect several types of unsafe behavior within an image (Fig. 8).

Our proposed model was initially pretrained using the COCO dataset which contains several scenarios that can helpful to detect of people's unsafe behavior. Furthermore, to achieve levels of accuracy, the hyper-parameters (e.g., learning rate, optimizer) of our model were adjusted in our dataset.

Publicly available datasets have been made public within the computer vision community, such as PASCAL and MS COCO, but no specific unsafe behavior image database is available for use in construction. Recognizing this knowledge gap, we created a new database of people's unsafe behavior derived from real-life project surveillance videos. Access to the training dataset together with the Python code is available upon the authors upon request. Our database based on the taxonomy of eight unsafe behavior provides a springboard to aid in developing more advanced detection techniques in construction.

B. Limitations

It should be acknowledged that our article has several limitations. First, we used the pretrained deep learning model, which was limited in size, to extract features, affecting detection accuracy. However, this limitation can be addressed in our future studies by creating a more extensive database of people's unsafe behavior. Another limitation affecting the accuracy of our method is the class imbalance within the database. Additionally, the class-based accuracies of the proposed solution have a high variance. We suggest that this limitation can be addressed by using active learning, transfer learning approaches [53], [54], or creating a larger unsafe behavior database with model training.

Second, given the detection speed, the attention mechanism in our proposed approach takes a considerably long time to compute features and, therefore, the detection of unsafe behavior. We acknowledge that detection speed is a problem in this article, and therefore, is something we needs consideration future studies. Semantic web technology can allow various sources of information to be made available in a format that can be searched and retrieved from the Internet. In doing so, semantic web principles can be used as a resource that can take advantage of its important advantages in this regard in our future work. We suggest that semantic web principles can be integrated with computer vision to create a large-sized database for the model training.

Third, when we computed the similarity of features extracted from images and words, the verb and adverb in words cannot be matched with activities in images, such as "walking." In this instance, there will be an influence on the ability to detect unsafe behavior accurately. Finally, the developed image-text matching approach was applied to detect a limited number of people's unsafe behavior related to a dangerous area. Extend the feasibility of our approach to different type's unsafe behavior in construction would be a fertile line of inquiry for our future studies.

VI. Conclusion

This article developed a semantic image-safety rules matching approach with an attention mechanism to identify people's unsafe behavior in construction sites. The experimental results we presented demonstrated that our developed approach can recognize people's unsafe behavior with a high level of accuracy automatically. The attention mechanism we used enabled attention with context from both people's unsafe behavior image and safety rules to be discovered simultaneously. As a result, we identified people's unsafe behavior accurately. The model's accuracy (e.g., R@1, R@2, and R@3) to detect people's unsafe behavior was 81.1, 87.8, 96.4, respectively, which exceeded the current state-of-the-art methods (e.g., DVSA, HM-LSTM).

To enable our developed approach to be used in practice by construction organizations, our future research will focus on creating a: 1) large database of unsafe behavior images for model training to improve detection accuracy; and 2) real-time video question answering system so that site managers can retrieve information for safety in construction site automatically.

References

[1] K. Andrej and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[2] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[3] L. Ben-Alon and R. Sacks, "Simulating the behavior of trade crews in construction using agents and building information modeling," *Automat. Construction*, vol. 74, pp. 12–27, 2017.

[4] W. Fang, P. E. D. Love, H. Luo, and L. Ding, "Computer vision for behavior-based safety in construction: A review and future directions," *Adv. Eng. Inform.*, vol. 43, 2020, Art. no. 100980.

[5] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014.

[6] R. M. Choudhry, "Behavior-based safety on construction sites: A case study," *Accident Anal. Prevention*, vol. 70, pp. 14–23, 2014.

[7] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[8] L. Ding, W. Fang, H. Luo, P. E. D. Love, B. Zhong, and X. Ouyang, "A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory," *Automat. Construction*, vol. 86, pp. 118–124, 2018.

[9] A. R. Duff, I. T. Robertson, R. A. Phillips, and M. D. Cooper, "Improving safety by the modification of unsafe behavior," *Construction Manage. Econ.*, vol. 12, no. 1, pp. 67–78, 1994.

[10] W. Fang, L. Ding, H. Luo, and P. E. D. Love, "Falls from heights: A computer vision-based approach for safety harness detection," *Automat. Construction*, vol. 91, pp. 53–61, 2018.

[11] Q. Fang *et al.*, "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automat. Construction*, vol. 85, pp. 1–9, 2018.

[12] W. Fang *et al.*, "A deep learning-based approach for mitigating falls from height with computer vision: Convolutional neural network," *Adv. Eng. Inform.*, vol. 39, pp. 170–177, 2019.

[13] W. Fang, L. Ma, P. E. D. Love, H. Luo, L. Ding, and A. Zhou, "Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology," *Automat. Construction*, vol. 119, 2020, Art. no. 103310.

[14] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *Proc. British Mach. Vis. Conf.*, 2018.

[15] H. W. Henrich, *Industrial Accident Prevention: A Scientific Approach*, New York, NY, USA: Wiley, 1959.

[16] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6163–6171.
[17] F. Huang, X. Zhang, Z. Zhao, and Z. Li, "Bi-directional spatial-semantic attention networks for image-text matching," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2008–2020, Apr. 2019.
[18] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2310–2318.
[19] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.
[20] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Comput. Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.
[21] K. H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.
[22] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4654–4662.
[23] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1890–1899.
[24] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4107–4116.
[25] Q. Liu, B. Wang, and Y. Zhu, "Short-term traffic speed forecasting based on attention convolutional neural network for arterials," *Comput.-Aided Civil Infrastructure Eng.*, vol. 33, no. 11, pp. 999–1016, 2018.
[26] X. Luo, H. Li, Y. Yu, C. Zhou, and D. Cao, "Combining deep features and activity context to improve recognition of activities of workers in groups," *Comput.-Aided Civil Infrastructure Eng.*, vol. 35, no. 9, pp. 965–978, 2020.
[27] H. Luo, J. Liu, W. Fang, P. E. D. Love, Q. Yu, and Z. Lu, "Real-time smart video surveillance to manage safety: A case study of a transport mega-project," *Adv. Eng. Inform.*, vol. 45, 2020, Art. no. 101100.
[28] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Assoc. Comput. Linguist.*, 2015.
[29] P. E. D. Love, S. Veli, P. Davis, P. Teo, and J. Morrison, "See the difference in a precast facility: Changing mindsets with an experiential safety program," *J. Construction Eng. Manage.*, vol. 143, no. 2, 2017, Art. no. 05016021.
[30] P. E. D. Love, P. Teo, J. Smith, F. Ackermann, and Y. Zhou, "The nature and severity of workplace injuries in construction: Engendering operational benchmarking," *Ergonomics*, vol. 62, no. 10, pp. 1273–1288, 2019.
[31] Z. Ma, Y. Lu, and D. Foster, "Finding linear structure in large datasets with scalable canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 169–178.
[32] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2623–2631.
[33] Ministry of Housing and Urban-Rural Development, "The short reports of fatal accidents in china's building construction activities," 2017. Accessed: Aug. 15, 2017. [Online]. Available: http://sgxxxt.mohurd.gov.cn/Public/AccidentList.aspx
[34] "Natural language toolkit," [Online]. Available: https://www.nltk.org/
[35] H. Nam, J. W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 299–307.
[36] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal LSTM for dense visual-semantic embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1881–1889.
[37] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 49–58.
[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[39] R. Sacks, O. Rozenfeld, and Y. Rosenfeld, "Spatial and temporal exposure to safety hazards in construction," *J. Construction Eng. Manage.*, vol. 135, no. 8, pp. 726–736, 2009.
[40] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Appl. Sci.*, vol. 7, no. 1, 2017, Art. no. 110.
[41] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.
[42] J. Shen, X. Xiong, Y. Li, W. He, P. Li, and X. Zheng, "Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning," *Comput.-Aided Civil Infrastructure Eng.*, vol. 36, pp. 180–196, 2020.
[43] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Inter. Conf. Learn. Represent.*, 2016.
[44] G. M. Waehrer, X. S. Dong, T. Miller, E. Haile, and Y. Men, "Costs of occupational injuries in construction in the United States," *Accident Anal. Prevention*, vol. 39, no. 6, pp. 1258–1266, 2007.
[45] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5005–5013.
[46] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.
[47] R. Wei, P. E. D. Love, W. Fang, H. Luo, and S. Xu, "Recognizing people's identity in construction sites with computer vision: A spatial and temporal attention pooling network," *Adv. Eng. Inform.*, vol. 42, 2019, Art. no. 100981.
[48] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image-text matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 21, pp. 5412–5425, Dec. 2020.
[49] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3441–3450.
[50] X. Yan, H. Zhang, and H. Li, "Computer vision-based recognition of 3D relationship between construction entities for monitoring struck-by accidents," *Comput.-Aided Civil Infrastructure Eng.*, vol. 35, pp. 1023–1038, 2020.
[51] Ministry of Housing and Urban-Rural Development of the People's Republic of China, "Quality and safety check points of urban rail transit engineering," 2011. [Online]. Available: http://www.zgjsjl.org.cn/uploadfile/2011 12/temp 11121215128737.pdf
[52] Ministry of Housing and Urban-Rural Development of the People's Republic of China, "Standard for construction safety assessment of metro engineering (GB 50715-2011)," 2011b. [Online]. Available: http://www.spsp.gov.cn/page/CN/2011/GB%2050715-2011.shtml
[53] U. Aggarwal, A. Popescu, and C. Hudelot, "Active learning for imbalanced datasets," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1428–1437.
[54] S. Al-Stouhi and C. K. Reddy, "Transfer learning for class imbalance problems with inadequate data," *Knowl. Inf. Syst.*, vol. 48, no. 1, pp. 201–228, 2016.
[55] D. H. Johnson, "The insignificance of statistical significance testing," *J. Wildlife Manage.*, vol. 63, pp. 763–772, 1999.
[56] P. E. D. Love and J. Matthews, "The 'How' of benefits management for digital technology: From engineering to asset management," *Automat. Construction,* vol. 107, 2019, Art. no. 10293.
[57] P. E. D. Love, "Creating a mindfulness to learn from errors: Enablers of rework containment and reduction in construction," *Develop. Built Environ.,* vol. 1, no. 1, 2020, Art. no. 100001.
[58] P. E. D. Love *et al.*, "Houston we have a problem: A view of quality and safety tensions in projects," *Prod. Plan. Control,* vol. 30, no. 16, pp. 1354–1365, 2019.
[59] P. E. D. Love, L. Ika, H. Luo, Y. Zhou, B. Zhong, and W. Fang, "Rework, failure and unsafe behaviour: Moving toward an error management mindset in construction," *IEEE Trans. Eng. Manage.,* to be published, doi: 10.1109/TEM.2020.2982463.

Weili Fang received the Ph.D. degree in civil engineering-construction engineering and management from the School of Civil Engineering and Mechanics, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2019.

He is currently a Research Fellow with National University of Singapore, Singapore. His research has appeared in several leading international journals such as *ASCE Journal of Construction Engineering and Management, Automation in Construction, Advanced Engineering Informatics*, and IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT AND PRODUCTION PLANNING AND CONTROL.

Dr. Fang was the recipient of numerous national and international awards for his research including the prestigious International CIC Construction Innovation Award in 2017.

Peter E.D. Love received the Ph.D. degree in operations management from Monash University, Melbourne, Australia, in 2013 and the higher doctorate of science degree for his contributions in the field of civil and construction engineering from Curtin University, Perth, Australia.

He is a John Curtin Distinguished Professor with the School of Civil and Mechanical Engineering, Curtin University, Perth, Australia. He has authored or coauthored more than 450 scholarly journal papers, which have appeared in leading journals such as the *European Journal of Operations Research, Journal of Management Information Systems, Journal of Management Studies*, IEEE TRANSACTIONS IN ENGINEERING MANAGEMENT, and *International Journal of Operations and Production Management, Production Planning and Control, and Transportation Research A: Policy and Practice*. His research interests include operations and production management, resilience engineering, infrastructure development, and digitization in construction.

Ding Lieyun received Ph.D. degree in management science and engineering from Tongji University, Shanghai, China, in 2002. He is a Professor of Construction Management with the School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, China, and an Academician with the Chinese Academy Engineering, Beijing, China. His research interests include virtual, safe, and automated construction. His research findings have been successfully applied in many metro and infrastructure construction projects in China. His scholarly works have been published in journals such as *Advanced Engineering Informatics, Automation in Construction, Reliability Engineering and System Safety,* and *Safety Science*.

Dr. Lieyun is the Executive Editor-in-Chief of Frontiers of Engineering Management and the Editor-in-Chief of *Journal of Civil Engineering and Management (Chinese)*.

Shuangjie Xu received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2016 and 2019, respectively. He is currently working toward the Ph.D. degree at the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong.

He has authored or coauthored several papers about computer vision. His research interests include computer vision, machine learning, and deep learning as they apply to image and video analysis and understanding.

Mr. Xu has been a Reviewer for several top conferences, including IEEE International Conference on Computer Vision (ICCV), Conference on Computer Vision and Pattern Recognition (CVPR), and Association for the Advancement of Artificial Intelligence (AAAI).

Ting Kong received the master's degree from the Department of Digital Construction and Engineering Management, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2020. She is currently working toward the Ph.D. degree at the Department of Building and Real Estate, Hong Kong Polytechnic University, Hong Kong.

Her research interests include construction safety management and construction informatics. Her research has appeared in several leading international journals such as *Automation in Construction*, *Tunneling and Underground Space Technology*.

Heng Li received the B.S. and M.S. degrees in civil engineering from Tongji University, Shanghai, China, in 1984 and 1987, respectively, and the Ph.D. degree in architectural science from the University of Sydney, Sydney, NSW, Australia, in 1993.

He is a Chair Professor of Construction Informatics with Hong Kong Polytechnic University, Hong Kong. He has authored two books and more than 500 articles. His research interests include building information modeling, robotics, functional materials, and Internet of Things.

Dr. Li is a Reviews Editor of *Automation in Construction*. He is also an Editorial Board Member of *Advanced Engineering Informatics*. He was a recipient of the National Award from Chinese Ministry of Education in 2015, and the Gold Prize of Geneva Innovation in 2019.