Vision-assisted behavior-based construction safety: Integrating computer vision and natural language processing

by

Yiheng Wang

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in
Construction Engineering and Management

Department of Civil and Environmental Engineering
University of Alberta

Abstract

Background: Construction sites can be hazardous places. Behavior-based safety is a method to optimize workers' behaviors and improve site safety. Previous behavior-based safety has been criticized for their low efficiency because of manual observation. The community has conducted enormous studies about applying advanced computer vision-based methods to automate the monitoring and observation of construction sites. However, the lack of methods for extracting semantic information and identifying safety hazards from construction imagery still poses a significant challenge for the development of sophisticated vision-assisted behavior-based safety programs.

Objectives: This research aims to automate the processes where the manual observation and inspection is needed in the traditional construction safety management by (1) enrich the information could be extracted from construction images, supporting safety hazard identification, (2) automate the safety hazard identification on site, and enable reasoning about the hazard identification according to safety regulations, and (3) automate the image records management and retrieval for efficient safety analysis.

Methods: Firstly, this research proposes a method to extract objects, activities, and interaction information from construction images. This method utilizes image captioning techniques to generate image captions for construction images containing semantic information. Secondly, this research proposes a novel visual–text semantic similarity method to compare construction image captions with safety regulation rules, enabling automatic safety hazard identification and reasoning. Finally, this research proposed a novel content-based image retrieval method for construction image repositories-

based object detection. This will help safety managers query and retrieve similar cases from monitoring image records, and conduct behavior analysis.

Outcomes: This research will improve the current vision-based construction management applications in the following ways: (1) it helps automate the monitoring and observation of construction sites; (2) it provides an automated method to identify potential safety hazards on construction sites and give reasoning of their violation about safety rules; and (3) it provides an information retrieval system for construction image repositories, enabling fast image retrieval and case-based reasoning and analysis.

Preface

This thesis is an original work by Yiheng Wang. This thesis is organized in a monograph format. Three journal papers related to this thesis have been published or submitted for reviewing, which are listed below.

Chapter 3 of this thesis has been published as **Wang, Y.**, Xiao, B., Bouferguene, A., Al-Hussein, M., & Li, H. (2022). Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Advanced Engineering Informatics*, 53, 101699. I was responsible for the data collection and analysis, method development and implementation, and the manuscript composition. Xiao, B assisted with the data collection and contributed to manuscript edits. B., Bouferguene, A., Al-Hussein, M., & Li, H. were the supervisory authors and were involved with concept formation and manuscript revising.

Chapter 4 of this thesis has been submitted and is under reviewing as **Wang, Y.**, Xiao, B., Bouferguene, A., & Al-Hussein, M. (2023). Proactive Safety Hazard Identification Using Visual–Text Semantic Similarity for Construction Safety Management. I was responsible for the data annotation and analysis, method development and implementation, and the manuscript composition. Xiao, B assisted with the data annotation and contributed to manuscript edits. B., Bouferguene, A., & Al-Hussein, M. were the supervisory authors and were involved with concept formation and manuscript revising.

Chapter 5 of this thesis has been submitted and is under reviewing as **Wang, Y.**, Xiao, B., Bouferguene, A., Al-Hussein, M.. (2023). Content-based Image Retrieval for Construction Site Images: Leveraging Deep Learning-based Object Detection. I was responsible for the data collection and analysis, method development and implementation, and the manuscript composition. Xiao, B assisted with the data collection and contributed to manuscript edits. B., Bouferguene, A., & Al-Hussein, M. were the supervisory authors and were involved with concept formation and manuscript revising.

Acknowledgements

The journey of my Ph.D. studies at the University of Alberta has been a challenging but rewarding experience, made even more so by the unprecedented circumstances of the COVID-19 pandemic. Despite the disruptions and uncertainties caused by the pandemic, I am grateful for the opportunity to have pursued my research and completed my degree. I would like to express my deepest appreciation to my family, supervisors, exam committees, and friends for their unwavering support and encouragement throughout this difficult time. Their guidance, feedback, and collaboration were invaluable in helping me navigate the challenges of learning and research.

I would like to express my deepest gratitude to my supervisors, Dr. Ahmed Bouferguene and Dr. Mohamed Al-Hussein, for their invaluable guidance, support, and encouragement throughout my research. Their unwavering commitment to my project and their insightful comments and suggestions have been instrumental in shaping my ideas and refining my arguments. I also want to thank Dr. Geoffrey Shen, Dr. Farook Hamzeh, Dr. Karim El-Basyouny, Dr. Ying Hei Chui, Dr. Xinming Li and Dr. Qipei Mei for their helpful feedback and suggestions, which have greatly improved the quality of my work. In addition, I want to thank Dr. Danbing Long, for her precious suggestions and guidance.

I am also indebted to the staff, colleagues, and students of the Department of Civil and Environmental Engineering at the University of Alberta, who provided me with valuable resources and support during my research, especially Dr. Bo Xiao, Ms. Tzu-jan Tung, Ms. Xue Chen, and Mr. Zicong Huang. Their expertise, knowledge, and enthusiasm have been a constant source of inspiration and motivation.

I would like to express my heartfelt thanks to my friends, Ms. Peipei Feng, Mr. Hangyu Li, Mr. Xin Wang, Ms. Shan Sun, and Mr. Kai Wang, for their unwavering support, encouragement, and understanding throughout both my academic journey and personal life. Their presence and friendship have been a constant source of comfort and strength, especially during times of difficulty, depression, and low mood. I am grateful for their willingness to lend an ear, offer a kind word, or simply be there

for me when I needed it most. Their support has been an integral part of my success, and I am honored to have such wonderful friends in my life.

I would like to express my special gratitude to my parents, Ms. Lingli Lu and Mr. Peng Wang, who have been my biggest supporters throughout my life. Their love, sacrifice, and unwavering belief in me have been the driving force behind my success. I am grateful for their endless support, whether it was providing financial assistance or offering words of advice and comfort during times of stress and difficulty. Their unwavering love and commitment to our family have set a powerful example for me to follow. I could not have achieved my academic goals without them, and I will always cherish their love, guidance, and support.

In conclusion, I am humbled and grateful for the difficulties, challenges, opportunities, and support that have enabled me to complete this Ph.D. program.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BBS	Behavior-based Safety
V-BBS	Vision-assisted Behavior-based Safety
CV	Computer Vision
NLP	Natural Language Processing
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
R-CNN	Region-Based Convolutional Neural Networks
AEC	Architecture-Engineering-Construction
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
CBIR	Content-based Image Retrieval
IR	Information Retrieval
IRS	Information Retrieval System
SOR	Safety Observation and Reporting
FC	Fully Connected
YOLO	You Only Look Once
RPN	Region Proposal Network
FCN	Fully Convolutional Network
GNN	Graph Neural Network
UAV	Unmanned Aerial Vehicle
SVM	Support Vector Machine
BERT	Bidirectional Encoder Representations from Transformers
MOCS	Moving Objects in Construction Sites Dataset
ACID	Alberta Construction Image Dataset

CPU	Central Processing Unit
GPU	Graphic Processing Unit
CUDA	Compute Unified Device Architecture
LOESS	Locally Estimated Scatterplot Smoothing
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
CIDEr	Consensus-based Image Description Evaluation
SPICE	Semantic Propositional Image Caption Evaluation
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MS COCO	Microsoft Common Objects in Context Dataset
NLG	Natural Language Generation
AP	Average Precision
mAP	Mean Average Precision
IoU	Intersection of Union
PPE	Personal Protective Equipment
RMAC	Regional Maximum Activation of Convolutions
VG	Visual Genome
ROI	Region of Interest
KNN	K Nearest Neighbor

Chapter 1
Introduction

1.1 Background

The construction industry is a crucial sector in North America, contributing 4.3% to the US GDP in 2021[1]. Unfortunately, construction sites are hazardous places with various safety risks that pose significant threats to workers. In the US alone, the construction industry reported over 1,000 deaths and 75,400 nonfatal injuries in 2020 [2]. The top three categories of fatal hazards are falls and slips, transportation incidents, and inappropriate contact with objects and equipment [2]. The frequency and severity of these accidents highlight the urgent need to improve safety measures in the construction industry.

Previous research has shown that many of these hazards can be prevented by enhancing safety management and minimizing exposures that may contribute to dangers and health impacts for construction workers. Behavior-based safety (BBS) is a widely studied approach that has been shown to be effective in promoting safe behavior. However, previous BBS methods in the construction industry have limitations such as being manual, time-consuming, and subject to observer bias, making them inefficient and error-prone [3–5]. These disadvantages can be attributed to the tedious and labor-intensive nature of visual observation and the difficulties in monitoring all workers continuously [6].

Recent technological advancements, such as cameras, drones, and smartphones, have become standard equipment in construction engineering, enabling professionals to record construction site imagery, including images and videos [7–9]. Therefore, the number of digital images and videos captured on-site has exponentially increased, with more than 400,000 images being captured from one typical project site during the construction phase [10]. Traditionally, construction sites have been documented and tracked using images and videos taken on the project site. These images and videos have been used for various purposes, including documenting and tracking the status of the project

[11], keeping a visual record of safety and quality inspections [12], maintaining a visual timeline of site progress, and providing evidence against damage claims [13].

The availability of construction site image repositories provides an opportunity to develop tools that can automate the inspection and observation of construction sites. Researchers have increasingly adopted computer vision (CV) technologies in the construction industry [5] to recognize critical information in monitoring images. Computer vision is a subfield of artificial intelligence that enables computers to process, analyze, and understand images and videos, allowing for the recognition and classification of objects, people, scenes, and events [14]. By leveraging computer vision, researchers have developed methods to recognize hazardous postures and actions, detect missing personal protective equipment (PPE), and automate BBS programs.

Despite the promising results of computer vision, it has its limitations when used alone in BBS. For example, it can only detect simple repeating objects or activities, which may not be the focus of typical BBS programs [5]. Furthermore, computer vision cannot leverage domain knowledge about safety regulations and guidelines, limiting the ability to infer whether the behaviors and interactions presented in the image scene follow the safety regulations. Fortunately, recent advances in natural language processing (NLP) technologies have enabled computers to process, understand, and infer from natural text languages [15]. Therefore, NLP can help extract and evaluate semantic meanings from safety regulations and guidelines, providing valuable domain knowledge for assessing the safety of complex activities.

The construction industry plays a significant role in the economy, but it also poses a considerable threat to workers due to safety hazards. Previous studies have shown that many accidents are preventable by enhancing safety management and minimizing exposures. BBS is a widely studied approach to promoting safe behavior, but previous methods have limitations. Recent advancements in CV technologies have enabled researchers to automate the inspection and observation of construction sites. However, CV alone has limitations, and NLP can help extract and evaluate semantic meanings from safety regulations and guidelines, providing valuable domain knowledge for assessing the safety of complex activities.

1.2 Behavior-based Safety

There is a considerable risk of accidents and injuries in the construction industry, which can result in large financial expenses and personal tragedies. Traditional safety programs have historically focused on compliance with regulations and PPE. However, there has been growing recognition of the importance of BBS in preventing accidents and injuries. BBS is a proactive approach to safety that emphasizes identifying and changing unsafe behaviors and reinforcing safe ones. In the construction industry, BBS has the potential to reduce accidents and injuries by encouraging workers to take responsibility for their safety and promoting a safety culture throughout the organization.

Construction work is inherently hazardous, with workers often exposed to multiple risks simultaneously. BBS programs can help identify and address the underlying causes of unsafe behaviors, such as lack of training, unclear instructions, fatigue, stress, or pressure to meet deadlines. The basic principles of BBS are observation, feedback, and reinforcement. BBS programs encourage workers to observe and report unsafe behaviors and provide feedback to their colleagues on correcting them. Positive reinforcement is used to reward safe behaviors and create a safety culture within the organization.

The basic steps of a BBS program include: (1) identifying unsafe behaviors; (2) observing or sampling identified behaviors over a time period; (3) providing feedback to increase desired behaviors and decrease undesirable ones through coaching and mentoring; and (4) presenting feedback regarding performance to the relevant audiences within the organization [16].

The effectiveness of BBS in the construction industry has been the subject of several studies and has shown promising results in reducing accidents and injuries. However, implementing a successful BBS program in construction requires careful planning and execution, which involves training workers and management on the principles of BBS, selecting appropriate observation and feedback methods, and establishing a system for tracking and reporting data.

Overall, behavior-based safety is an important approach to preventing accidents and injuries in the construction industry. Its emphasis on changing unsafe behaviors and promoting a safety culture can help reduce risks and create a safer work environment for all workers.

1.3 Computer Vision for BBS

Implementing a behavior-based safety (BBS) program in construction can be a challenging and labor-intensive process, particularly when identifying and highlighting people's unsafe behaviors. This is because the initial steps of BBS involve intensive manual monitoring and observation to identify unsafe behaviors and observe them over time. However, this process is crucial for workers to reflect on and learn about how their unsafe actions can jeopardize not only their safety but also that of their co-workers. Recent advancements in computer vision have enabled the automatic capture and identification of unsafe behavior and hazards in real time from two-dimensional (2D) digital images and videos. These developments have generated considerable interest in the construction industry, leading to extensive research on the potential application of computer vision in practice [17].

Drawing on a synthesis of existing knowledge [7,9,18], Figure 1 outlines the contents, main techniques, deployment process, and challenges involved in effectively implementing CV for BBS. Vision-based construction management involves four main steps: imagery data collection, feature processing, information recognition of the imagery, and specific management application.

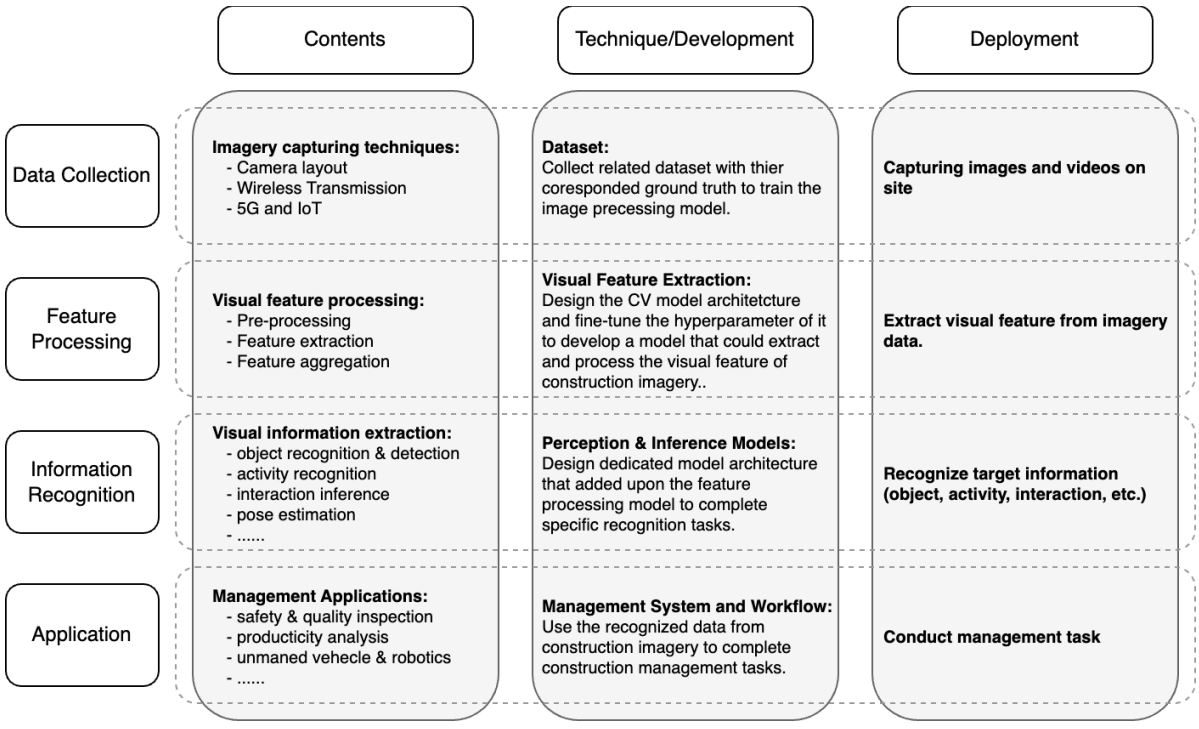


Figure 1. Contents, Techniques, Deployment and Challenges corresponding to the four aspects of vision-based application workflow.

The first step of vision-based construction management is visual data collection. This involves designing the construction site's camera layout and data transfer plan. The visual data includes images, videos, and laser point cloud data. Imagery data collected by a single camera is adopted for monocular vision analysis [19], while data collected by multiple cameras are employed for stereo analysis [20]. Currently, the main target subjects of the visual data are heavy equipment [21], human labor [22], and construction materials [23]. The visual data serves as the training data for deep learning models in the technology development process. In the deployment process, the visual data is the input data of the management system.

The raw imagery data must be processed for deep learning models. Unlike simple processing like image cropping [24], video clipping [25,26], and keyframe selection [27], which are commonly used in traditional computer graphics methods, the essential processing step in computer vision methods is feature processing [18]. It outputs a sequence of vectors representing the raw imagery's most significant visual feature. Current technology commonly utilizes convolutional neural network (CNN) as the feature extractor. CNN has several architectures, such as VGG [28], Inception [29], ResNet [30], etc. Based on the architecture and parameters of the CNN, the extracted feature represents the image's edges, patterns, and objects. The processed feature vector is then fed into a subsequent process to generate a high-level perception of the captured imagery.

The information recognition process relates to essential computer vision tasks that mimic basic human visual perception abilities. The recognition could be categorized based on the information extraction target. Object detection recognizes its pixel location and category. Activity recognition recognizes the behavior of the target object. Interaction inference recognizes the relationship between objects. Pose estimation recognizes the human key point and then infers the human pose.

The last step of vision-based construction management is to combine the above-described basic information from site imagery with certain rulesets, logic, and workflow to complete specific management tasks. For example, by combining the object detection result of workers and scaffold with the safety rule, the safety of workers' behavior could be determined [31].

1.4 Knowledge Gaps

Despite the promising benefits provided by the computer vision technologies for the BBS program, there are several challenges when automating the observation and inspection on construction sites:

(1) Most applied computer vision-based construction recognition models learn from correlations and recurring patterns in the input data (specifically between the features extracted from images). Compared to humans, these methods can rarely draw causality or extract higher-level semantic understanding (activities, interactions, and attributes) from imagery. Construction imagery contains a large amount of inexplicit information that needs an advanced recognition model to extract the related information, like interaction and attributes. Current construction studies merely deal with this semantic information or only extract a small portion of it, which cannot form a comprehensive understanding of the visual scene presented in the construction imagery. Hence, many vision-based applications have a low level of understanding of construction imagery and cannot be applied practically to complex safety management applications, such as the BBS program.

(2) Most computer vision-based safety hazard identification models merely deal with safety regulations and guidelines. These models typically rely on pre-annotated safety labels to learn the visual patterns related to safety hazards. Therefore, the ability of these models is limited to the pre-defined hazard types and needs to address the dynamic and complex nature of safety hazards on the construction site. Moreover, since these models output safety labels, they cannot provide insight into which safety rule is being violated. This is a critical limitation, as identifying the violation of safety rules is crucial to preventing future safety hazards and developing effective safety interventions.

(3) Even though an extensive repository of construction monitoring images assisted in observing a target behavior over time, the image data are manually retrieved and analyzed when conducting safety analysis on past records and similar cases. Construction monitoring images are essential for analyzing and observing target behaviors over time, studying the behavior patterns, and relating issues in the BBS program. However, manual sorting of construction monitoring images can be a time-consuming and labor-intensive process, and it can also be prone to errors or inconsistencies. Furthermore, content analysis, which involves manually identifying and tagging images with relevant keywords or labels, can also be subjective and may not capture all relevant information. Moreover, without an adequate information retrieval system, retrieving specific photos from a vast archive of construction monitoring images can be problematic.

1.5 Research Objectives and Scope

1.5.1 Research Objectives

This research aims to streamline on-site safety management efforts by automating observation and inspection on construction site including construction imagery processing, inspection, and management, thereby improving the capability and performance of vision-assisted BBS solutions. With the adoption of the proposed methods, construction engineers can automatically extract more semantic information from the construction imagery, enabling automatic safety hazard identification and reasoning. Additionally, the proposed method includes an information retrieval system for construction monitoring images that facilitates faster observation over time and behavior analysis. The objectives of this research and their relationship with BBS are illustrated in Figure 2.

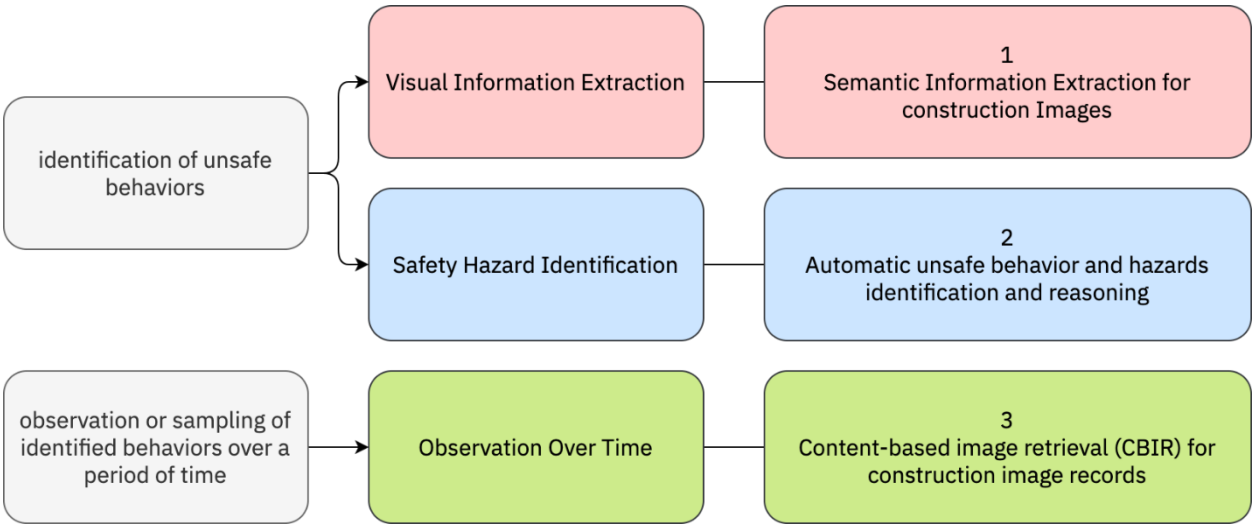


Figure 2. The three objectives of this research and their relationship with BBS

This research focuses on the first two steps of BBS: (1) the identification of unsafe behaviors; (2) the observation or sampling of identified behaviors over a time period. When using CV to automate the identification of unsafe behaviors, two steps are needed: firstly, extracting the visual information from the construction images, and secondly, identifying the safety hazards. These two steps correspond to the first two objectives of this research. In addition, an information retrieval system is required for observation over time in order to retrieve the relevant image records from the construction image repository. This is related to the third objective of this research. The details of the three objectives are as follows:

(1) *To enrich the information could be extracted from construction images, supporting safety hazard identification* : This objective focuses on improving the capabilities of computer vision techniques in understanding and interpreting visual data from construction sites. The goal is to automate the observation of related information required by safety management (e.g., safety hazard identification).

In achieving the first objective, and resolving the first challenge identified in former section, this research proposed a method for *semantic information extraction for construction images*: semantic information extraction for construction images involves identifying and analyzing the content of images to extract information such as objects, activities, and relationships as text descriptions. In this way, this method extracts rich semantic information from construction images and enable downstream safety hazard identification with the extracted information.

(2) *To automate the safety hazard identification on site, and enable reasoning about the hazard identification according to safety regulations*: This objective aims to automate the potential safety hazard identification in common safety management practices. Furthermore, it seeks to enable reasoning about these identified hazards in accordance with safety regulations, providing a comprehensive understanding of the risks involved.

In achieving the second objective, and resolving the second challenge identified in former section, this research proposed a method for *automatic unsafe behavior and hazards identification and reasoning*: developing an unsafe behavior and hazards identification and reasoning using CV and NLP techniques to extract information from images and text related to safety hazards in the construction industry. This method analyzes the meaning of the rule text and image captions using NLP techniques to identify and match relevant keywords and phrases. The similarity between the extracted information can then be quantified using semantic similarity measures, which assess the degree of similarity between the text and image captions. In this way, this method can automatically comparing the image content with the related safety regulations, and generate output about safety hazard identification and reasoning.

(3) *To automate the image records management and retrieval*: This objective is about automating and improving the efficiency of organizing and retrieving of construction site images, thereby

facilitating the extraction of semantic information and the identification of safety hazards on a large construction image collections.

In achieving the third objective, and resolving the third challenge identified in former section, this research proposed a method for *content-based image retrieval for construction image records*: Developing an information retrieval system for construction image records based on the visual content of the images. This objective addresses the challenge of lacking a dedicated IRS for construction imagery and how to distinguish and balance the background and object features of construction imagery. With the help of this method, the construction manager could easily retrieve similar images to observe a target behavior over time and perform behavior pattern identification and analysis.

1.5.2 Connections Between Objectives

Objective 3, the creation of a content-based image retrieval method, serves as a foundational database that supports the first two objectives. This image retrieval system will store and organize construction images in a manner that allows for efficient and accurate extraction of semantic information (Objective 1) and hazard identification and reasoning (Objective 2).

The image retrieval system is designed to be content-based, meaning that it uses the actual content of the images (such as the presence of construction equipment, workers, safety measures, etc.) to index and retrieve images. This feature is crucial for the success of the first two objectives. For Objective 1, the semantic information extraction method relies on the ability to access relevant images from the database. Similarly, for Objective 2, the automatic safety hazard identification and reasoning method will utilize the images stored in the database to analyze the behaviors over time. The image retrieval system will provide a comprehensive visual context that aids in the identification and understanding of safety hazards.

In summary, the first two objectives could be used either on real time monitoring and observation of construction site, or work with Objective 3 to analyze past image records. Objective 3 is not just a standalone goal, but a critical component that enables and enhances the success of Objectives 1 and 2. It ensures that the methods developed for semantic information extraction and hazard identification have a robust, content-rich image database to work with, ultimately leading to a more effective and accurate system for improving construction site safety.

1.5.3 Relation to Existing Safety Models and Frameworks

This research, focusing on the integration of computer vision and natural language processing for construction site safety management, aligns with and complements existing safety models and frameworks in the construction industry.

Cognitive Systems Engineering (CSE): Jens Rasmussen's CSE [32] emphasized the understanding of cognitive processes in complex work domains. This research aligns with CSE principles by designing a system that supports human decision-making in the complex domain of construction safety. The system aids in understanding the work environment and potential safety hazards that comes from violations of safety rules, thereby supporting informed decision-making.

Working Near the Edge: Howell et al. [33] focused on managing risks inherent in high-risk construction environments. Our research complements this approach by providing a tool that automates the identification of such risks, allowing for more efficient and accurate risk management.

Other Safety Models: Various other safety models in construction emphasize the importance of hazard identification, risk assessment, and effective safety management. This research contributes to these goals by automating the observation and hazard identification process based on safety rules, and by improving image records management.

It's important to note that this research is not intended to replace any existing safety models or frameworks. Instead, it provides a tool that enhances and supports these models. The system automates the observation of construction sites, identifies potential safety hazards based on established safety rules, and improves the management of image records. This allows for more efficient and accurate safety management, while still relying on human expertise for final decision-making.

1.5.4 Research Scope and Hypothesis

This research does not include construction videos, but a further study could extend this framework to construction videos since videos consist of a sequence of images. This research hypothesis is that the camera layout on the construction site is well planned, and the image data can be successfully transferred and stored. It should also be noted that the proposed framework targets the basic methodological challenges in the current technology development process of vision-assisted BBS and validates the feasibility of the proposed methods. The main scope of the present research is to provide an essential workflow for the steps that involve construction site monitoring and observation

in the BBS program, which could extract semantic information from construction imagery, compare the extracted information with safety rules, and retrieve target images from the construction image repository.

Chapter 2
Literature Review

2.1 Computer Vision Techniques

As workers' unsafe behavior is a major contributor to site hazards [34], BBS programs have been implemented for decades to improve the safety performance of construction projects and organizations [35,36]. BBS is an approach that creates a safety partnership between management and employees by (1) identifying and observing unsafe actions, (2) providing direct feedback to individuals who committed unsafe actions, and (3) improving future safety awareness and performance through coaching and training [16,17]. Among these steps, observation is an essential one. However, manual observation could be labor-intensive, subjective, and budget-challenging. In addition, the corresponding manual safety observation and reporting (SOR) method also has drawbacks, such as increased administration time and data [37].

Recently, deep learning-based CV techniques have shown visual recognition ability comparable to or even better than human ability [38,39]. It has also been identified as a reliable method for automatically recognizing and capturing hazardous activities made by individuals during construction [17]. The BBS programs have been automated to automatically recognize objects and activities based on CV techniques, such as image classification, object detection, and image captioning. Figure 3 visually illustrates these tasks and sorts them based on label density and semantic richness.

2.1.1 Image Classification

Image classification is a task that labels the category of a whole image. An image classification algorithm, for example, can take images of various construction equipment as input and assign a class label such as "excavator," "dump truck," "forklift," and so on. A typical image classification model utilizes several CNN layers to extract the image feature map and then feed the feature map into a fully connected (FC) layer to get the feature vector of the input image. A classification layer, usually with a

SoftMax activation function, is the final layer of the model to predict the classification probability. There are many image datasets for image classification, such as ImageNet, which includes over one million images. Many well-known CNN architectures are proposed for image classification tasks, such as LeNet [40], Inception [29], and ResNet [30]. However, the image classification model has an obvious shortcoming: it can only recognize one single object in each image and cannot detect its location. So, it may fail when multiple objects are presented in the image to classify. In image classification, an entire image is analyzed, and one label is generated for this image; as such, the label density of image classification is low. And the predicted label often contains a single keyword; as such, the semantic richness of image classification is low.

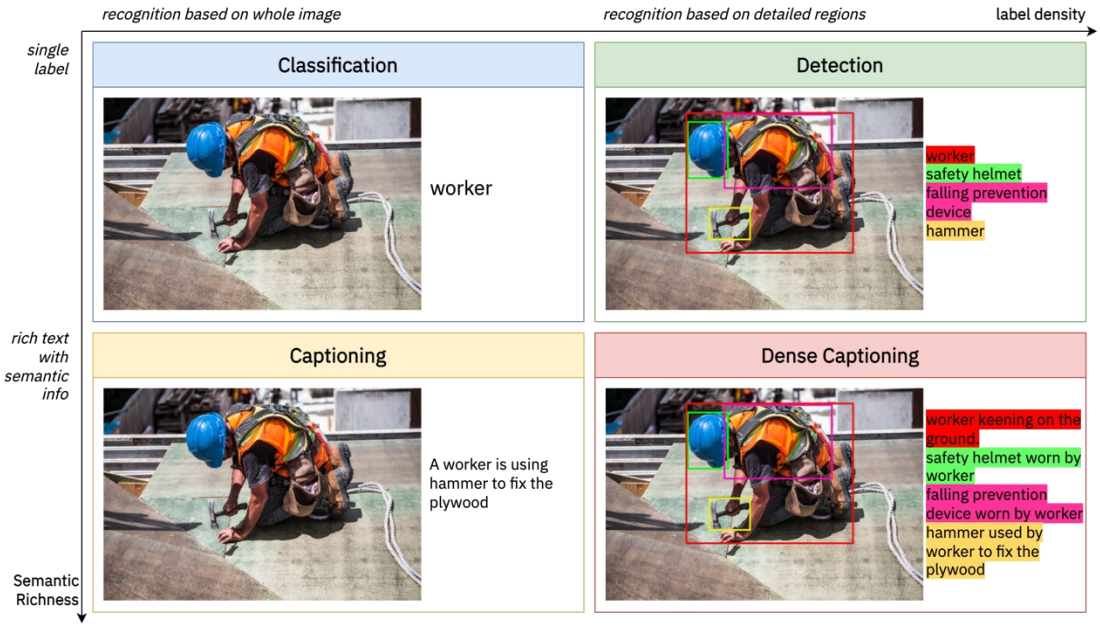


Figure 3. Comparing the label density and semantic richness of various visual recognition tasks.

2.1.2 Object Detection

Object detection is a task that extracts the category and localization information of the objects in an image. Object detection models often rely on a CNN pre-trained on an image classification dataset. Based on the type of object region generation, there are two types of object detection models:

(1) The first type is named a two-stage detector based on the region proposal. This type's most commonly used model architecture is Faster R-CNN [41]. Faster R-CNN utilized a CNN-based backbone to extract the feature map of the input image. Then a Region Proposal Network (RPN) will

search the feature map and find region zones of the feature map that potentially contain objects. The proposed region is named the region of interest (RoI). Each RoI will be further processed by classification to predict the category label and regression to indicate the object's location.

(2) The second type of detector is named a one-stage detector based on regression. This model type depends on a fixed number of region anchors or proposals for an input image. The regression and classification will directly process these regions, mapping the image features to bounding box coordinates and class probabilities. An example of this model is You Only Look Once (YOLO) [42].

Object detection yields higher label density than image classification. However, because it predicts single keyword labels, the semantic richness is low.

2.1.3 Instance Segmentation and Semantic Segmentation

Instance segmentation is very similar to object detection. Unlike the object detection model, which locates the bounding box of the target object, the instance segmentation model finds all the pixels in the image that belongs to the target object. The widely utilized instance segmentation model is Mask R-CNN [43], which is similar to Faster R-CNN. Compared to Faster R-CNN, Mask R-CNN adds a mask generation head after the RPN. The mask generation head utilizes a fully convolutional network (FCN) [44] to predict if the pixel belongs to the object in each RoI.

2.1.4 Visual Relationship Detection

Visual relationship detection captures interactions between pairs of objects in images or aims to reason over relationships among salient objects in images. Visual relationship detectors are often built upon object detectors since the object detector can predict object classification and localization. For example, Lu et. al. [45] trained another CNN to classify the relationship between the objects according to the union of both object regions. More recent research [46] utilized a graph neural network (GNN) to detect the relationship.

2.1.5 Image Captioning

Image captioning is a task that generates a natural language description for an input image. An image captioning model could extract semantic information from the images. The widely used image captioning methods were template-based, which first detected a specific set of visual concepts from images. Then, the detected visual concepts are connected by sentence templates or specific grammar

rules to form a sentence [47–49]. Recent studies have implemented deep learning techniques in image captioning to provide more accurate and natural captions. The base architecture for neural image captioning models is the encoder-decoder architecture. This base architecture combines CNN as the image extraction module and RNN as the sentence generation module [50,51]. The early encoder-decoder model approaches have obvious limitations. Firstly, the image feature vector has a fixed dimension and easily loses or ignores some important visual information when decoding in the language generation process [52]. Secondly, such a structure cannot model the correspondence between the input visual features and the output sequence, and this correspondence is significant for the tasks in image captions.

The "show, attend and tell" model [53] introduces the attention mechanism into the classic encoder-decoder image caption model. The image coding part of image captioning no longer uses the top-level representation of CNN. Still, it extracts the bottom-level vector representation from CNN (previously a one-dimensional vector, now a three-dimensional grid). The attention distribution on each grid is calculated using the attention mechanism in the decoding process. Then, when the image is formed, each word will be noticed in the visual grid area. Meanwhile, other types of attention mechanisms are proposed, such as adaptive attention [54], bottom-up attention [55], and transformers [56].

Compared with image classification, image captioning provides more semantic information regarding activities and interactions and generates labels with more semantic richness. However, because the output of captioning is based on the whole image, the label density remains low.

2.2 Computer Vision Applications in AEC industry

Many studies have been conducted to retrieve useful information from heavy equipment operation images for management purposes such as: improving the machine's health and condition [57], improving environmental performance [58], monitoring progress [59,60], and identifying potential improvement areas [61]. The target information that researchers are interested in retrieving from the construction operation images could be categorized into three types: (1) construction object detection, (2) operation activity recognition, and (3) interaction and scene analysis.

2.2.1 Construction Object Detection

Construction object detection identifies and locates the construction objects from images. Early studies have implemented background subtraction algorithms [24] and machine learning models [62] as the detection method. Recent studies have utilized more sophisticated deep learning methods based on CNNs and recurrent neural networks (RNNs) to conduct the object detection task. Deep learning methods such as Faster R-CNN have been adopted to detect construction objects, including workers, excavators, and sewer pipes [63,64], which have improved performance over non-deep learning studies. For example, Fang et al. [65] proposed a real-time detection model of workers and excavators on the construction site, which provides managers with information about workers and equipment to improve their decision-making. Cheng and Wang [63] proposed an automatic detection model for sewer pipes, laying the foundation for applying deep learning technology to detect sewage pipe defects. Kim et al. [66] utilized YOLO-V3 as the object detector to process the UAV-captured site images and localize the object, followed by image rectification and distance measurement. The proposed method can detect the danger of being injured around workers and provide early warning to ensure that workers are in a safe working environment.

Although much research has been conducted on construction object detection, the detection results cannot represent all the information in the image scene, nor can they express the complex activity scene of the relationships between different objects.

2.2.2 Activity Recognition

The second direction of vision-based operation monitoring recognizes the construction activities of the identified construction objects. The recognized actions of on-site equipment (e.g., working, idling) and workers (e.g., unsafe behaviors) can be further used for operational performance analysis and safety assurance. Early studies have implemented rule-based and thresholding algorithms [67,68] and support vector machines (SVM) [69,70] as activity recognition methods. Recent studies adopted deep learning networks as the image processor to recognize workers' operations and mechanical equipment gestures [71,72]. Deep learning models have shown better image feature extraction and processing capabilities [38]. For example, Kim and Chi [73] proposed a hybrid model integrating CNN and long short-term memory (LSTM) that performs visual feature extraction and sequential pattern analysis to recognize earthmoving equipment activities. Their proposed method is used for automated cycle time and productivity analysis, which is beneficial for automated excavators' cycle time and

productivity monitoring. Moreover, Luo et al. [74] utilized deep learning to automatically recognize and track excavators' position, posture, and movement on construction sites, reducing the incidence of safety mishaps and preventing injury to workers within the operating range of on-site machinery and equipment.

2.2.3 Interaction and Scene Analysis

Object interaction and scene analysis received less interest than the former two research areas. Existing studies mainly utilized LSTM as an essential tool to identify the interaction through sequential pattern recognition for the object interaction. For example, Cai et al. [75] utilized the corresponding positional relationships through LSTM to classify the working group of identified entities, which further organized the interaction activity. The construction site's safety and productivity can be improved by analyzing the interaction between workers and equipment. Another LSTM-based model [76] is proposed to predict the workers' trajectories while considering the interaction with the contextual objects. It can automatically detect and collect data on workers' movements and building sites, assisting in proactively preventing engineering accidents. Other computer vision techniques are also utilized to analyze the interaction and scene. For example, Kim et al. [77] utilized a scene parsing method to identify whole image areas. Ham and Kamari [78] used semantic segmentation to investigate the UAV-captured image for objects' location and analyze the spatial composition by detecting all the object zones in the image and analyzing the relative locations between different zones. Moreover, Tang et al. [79] utilized scene graph techniques [80] to analyze the human-object interaction in the site image for a safety inspection.

Most of the scene analysis models are proposed separately from the object detection models, which means that they lack visual grounding of the recognition results. For example, when a model recognizes an activity, it cannot tell which object is doing this activity. And this visual grounding is important when the input image contains multiple objects.

2.2.4 Need for Label Density and Semantic Richness

A sophisticated vision-based safety management program requires CV technologies that can generate semantic richness and high label density. Semantic richness is required to understand what is happening in a given scene as well as how people and objects are interacting in it; information regarding the activity and interaction is paramount for understanding the context of a scene because

the majority of safety risks stem from inappropriate actions and complex spatial and temporal object structures [5]. Finally, the need for a procedure that can generate a high label density can be linked to its accuracy because the underlying algorithm is not likely to ignore important information in the scene. Several studies have been conducted to generate dense and semantically rich labels. For example, Zhang et al. [81] added interaction labels into the object detection model to generate a scene graph. However, these labels are based on keywords, not on natural language. Dense captioning [82] is a promising approach that jointly generates dense and rich annotations in a single model but has not been adopted in vision-based safety management programs yet.

2.3 Natural Language Processing in Construction Safety

Data regarding construction safety, such as safety regulations, safety guidelines, construction plans, safety reports, incident logs, and worker communications, are generally stored in various electronic text formats and used to track the progress of construction projects, identify, and address safety hazards and improve communication between workers and managers. NLP can be used in construction safety management to analyze large amounts of text-based data, such as safety reports and incident logs, to identify patterns and potential hazards [83], enabling construction companies to identify and address potential safety risks more effectively and efficiently, thus improving safety at construction sites [84]. Like CV, NLP provides several techniques for text data processing, such as document classification, knowledge extraction, and factor analysis. These techniques can be employed for improving construction safety management.

Document classification: Document classification is the process of arranging documents into predefined categories or labels and can be used to categorize construction records in the construction industry, such as safety reports, incident logs, and progress reports to aid in pattern identification and decision-making. For example, by using key attributes from injury reports, Tixier et al. [85] created an undirected network and then used multiple graph clustering algorithms to identify potential safety conflicts. Zhang [86] proposed a hybrid structured deep neural network incorporating (learned) word embeddings for the automatic classification of the causes of construction accidents. Fang et al. [87] developed a DL-based text classification method based on BERT, a DL-based NLP network, to automate the classification of near-miss data in safety reports.

Knowledge extraction: Knowledge extraction is the method of automatically extracting structured information from unstructured or semi-structured texts and involves extracting entities, relationships, and events from text as well as recognizing concepts, themes, and sentiments. Knowledge extraction has many applications in construction safety management. For example, Yeung et al. [88] developed a knowledge extraction and representation system to restructure a complex safety accident narrative text and created a narrative map to help safety project trainees comprehend and remember the important elements and events that led to the accident. Martnez-Rojas et al. [89] used NLP to extract seven types of contents from safety and health plans documents by using manually established rules to check if the plans satisfy safety requirements at an early stage. Wang and El-Gohary [90] proposed a DL-based method to automatically extract and represent relations that describe fall protection requirements and represented the extracted information in the form of knowledge graph-based queries to aid in the automatic checking of safety requirements.

Factor analysis: Factor analysis is a statistical approach used for determining underlying correlations and patterns between data samples. Mathematically, it reduces the dimensionality of large datasets, retaining a small subset of the original variables, known as factors, that explain most of the variance in the dataset. In the construction sector, factor analysis can be used to determine the underlying elements that lead to safety incidents, such as identifying primary causes of accidents and injuries or discovering patterns in worker communication that may indicate a possible safety hazard. For example, Kim and Kim [91] discovered the elements of a construction fire accident from public news items by using multiple morphological analyses and then assessed the primary causes leading to fire accidents in different seasons by using principal component analysis. Xu et al. [92] extracted 37 safety risk factors from 221 metro construction accident reports by using text mining technology. Pan et al. [93] developed a graph-based model to reveal the interdependency between body part factors and accident types..

Most NLP applications employed in construction safety management operate using only text data. However, in recent years, the utilization of digital cameras has gained tremendous popularity in the construction industry, making videos and images an important data format for monitoring construction sites. Combining CV and NLP in construction safety management can provide a powerful tool for managers to obtain a comprehensive understanding of the safety conditions on site, which, in turn, will help address any identified safety hazards more effectively and efficiently and improve overall safety at construction sites.

Zhong et al. [94] proposed an integrated method that accepts image and text data. They first used NLP to extract potential hazard ontologies from project documents and then manually tagged construction image scenes with predefined categories. Finally, they obtained similar pictures by calculating the degree of similarity between image annotation (text) vectors. However, they did not utilize CV techniques to process the images automatically. Zhang et al. [81] proposed a more advanced hazard identification system by combining CV and NLP. This system first detects keywords regarding objects and interactions from the image and then uses a supervised BERT model to calculate the probability of hazards by inputting the keywords and safety regulations. Although this study achieved promising results, it has two limitations: (1) The semantic information in the image is represented as keyword pairs, which have less semantic richness than a complete sentence. As such, this method requires additional processing steps. (2) This hazard identification model is supervised learning-based and thus requires extensive data collection and model training in the case of a large number of safety rules.

Chapter 3
Semantic Information Extraction from Construction Images

3.1 Introduction

The construction industry is one of the largest industry sectors in North America, which contributes to 4.3% of the GDP of the United States in 2021 [1]. Cameras have recently become standard equipment in construction engineering, allowing construction professionals to monitor their sites remotely [7–9]. Analyzing construction images/videos by vision-based methods is beneficial to construction management in terms of improving crew productivities [73], improving the machine's well-being [57], improving environmental performance [58], monitoring progress [59,60], reducing safety risks [95], and enhancing construction logistics [65]. Extracting semantic information (e.g., objects, activities, and interactions between objects) from construction images is the fundamental step for many vision-based applications in construction management [7,96–99].

Object detection is a vision-based technology that can extract pre-defined classes of objects and their location information from construction images, which has been applied to defect detection [100,101], equipment classification [102], and safety monitoring [103] in construction engineering. However, object detection can only provide information of object category and localization, which may not be sufficient for advanced construction applications (e.g., activity recognition and interaction analysis) [99]. Therefore, other technologies have been employed for extracting semantic information from images or videos. For example, Kim and Chi [73] utilized an additional model based on the Recurrent Neural Network (RNN) upon the object detector to recognize the activity of excavators. Liu et al. [104] proposed a method to extract the semantic information from the site image as natural language descriptions. Tang et al. [9] combined the object detection and human-object-interactions recognition module to ground the interaction information of workers onto the image. Moreover, NLP technologies also play an important role in conducting similar roles. When current techniques could not directly extract target information from the images, NLP techniques extract information from

human observation reports. For example, researcher utilized Named Entity Recognition techniques to extract information about safety accident [91,105], equipment and labor information [106,107], and relations [108,109].

Currently, executing separate dedicated models on the site image could extract the object, activity, and interaction information, achieving the goal of semantic information extraction [31]. However, executing separate models is time-consuming, and separate models may lack the consistency of the entity label since they are trained on different datasets. Also, extracted semantic information by separate models lacks visual connections between the recognized labels with image regions. Visual connection is vital because the semantic information extracted in labels cannot provide enough information for analysis or decision-making [5,22,95,110]. For example, to track the activity of equipment or labor, it is required to know the object's location so that its trajectory can be generated and analyzed [61,111,112]. Sometimes, an activity cannot be identified as unsafe for safety management unless this activity happens in some restricted areas [31]. This means providing the location of the object and its activity is required [97,113]. Therefore, combining object detection and semantic information extraction into an integrated model is a promising way to provide richer and more integrated information for downstream analysis and decision making.

Based on the above analysis, this section proposed a novel vision-based method by integrating object detection, image captioning, and data post-processing to extract semantic information for the construction machine images with the visual connection. This method contains a novel process to integrate object detection and image captioning to extract information about object categorization and location, object activity, and interactions between objects. A novel attention mechanism added to the integrated model achieves the visual connection between the object detection results and the extracted semantic information. This study provides a novel integrated model that will extract semantic information with visual connection for construction images and videos. The extracted information could enhance the visualization ability of current methods by providing object categorization, location, and activity information. It could also facilitate object tracking and safety management. For example, provided the location and activity information, the activity trajectory of the equipment could be generated, and the unsafety behavior could also be analyzed and determined.

3.2 Methodology

Aiming to extract the related semantic information (objects, activities, and interactions), This section combined an image object detector and a language decoder as an integrated method. Figure 4 presents the overall architecture of this method. The proposed method is an extension of a typical encoder-decoder-based image captioning method. In the typical method, the encoder is a CNN that extracts the useful semantic feature from the whole image. The decoder then predicts the description words of the image according to the image features.

The input is the original image for the method proposed in this study. The object detector – encoder – will output the object location, category, and feature maps for the whole image and object regions. The feature maps are used as the inputs of the language decoder. The decoder is a Recurrent Neural Network (RNN) built upon Long Short-Term Memory (LSTM) cells and the attention mechanism. In essence, the LSTM cells utilize the feature maps of the image to predict the caption words, with the help of an attention mechanism providing the correspondence between the input object features and the output words (i.e., the Attention Maps). In the last step, this study post-processed outputs, including the object location, object category, attention maps, and caption words, to make the outputs more intuitive and integrated, which is presented in the right panel of Figure 4. The detailed introductions of each module are described later in the following subsections.

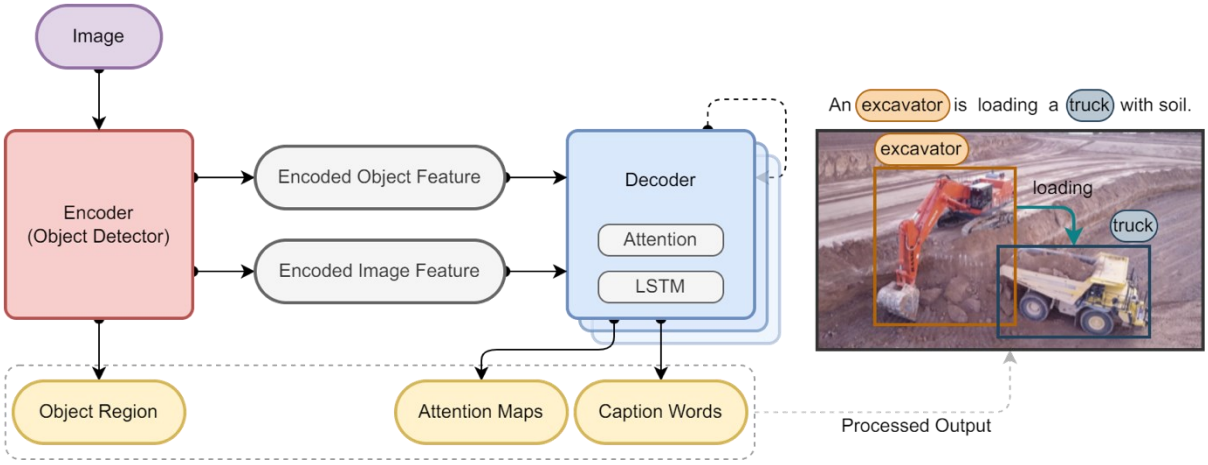


Figure 4. Overall flowchart of the proposed method.

3.2.1 Feature Extraction

This study utilized the Mask R-CNN [43] as the image decoder to complete two tasks: (1) recognizing the objects in the image by performing object detection and instance segmentation; and (2) extracting the image feature and object feature based on recognized objects in step one. The reason for choosing this architecture is that based on our prototype experiments, this study found that the Mask-RCNN provides a better performance boost for the whole model than other object detection architectures.

Mask R-CNN is an extended version of the original Faster R-CNN [41]. To complete the object detection task, the original Faster R-CNN model has a category classification head and a bounding box regression head for each Region of Interest (RoI). The Mask R-CNN adds a head parallel to the existing heads to predict the instance segmentation using a Fully Convolution Network (FCN) [44].

Figure 5 presents the detailed architecture of the proposed encoder. The encoder utilizes the ResNet-101 network as its backbone to extract the feature map from the input image. ResNet-101 is an image classification network originally trained on the ImageNet dataset [114] and is wildly utilized for other tasks. This study also utilized the five convolution stages in the ResNet ('Res1' to 'Res5') as the feature extractor in this encoder. In panel A of Figure 5, the "7x7, 64" denotes the filter size and depth of the convolution process. "Res2" denotes ResNet's second stage, and so forth. "x3" denotes a stack of three consecutive convolution layers, and so forth. The backbone could get the image features after the final convolutional layer of the 4-th stage and 5-th stage, which we call C4 and C5, respectively. The C5 feature is sent directly to an average pooling layer and a fully connected (FC) layer to get the feature vector v for the whole image, which will be sent to the decoder later.

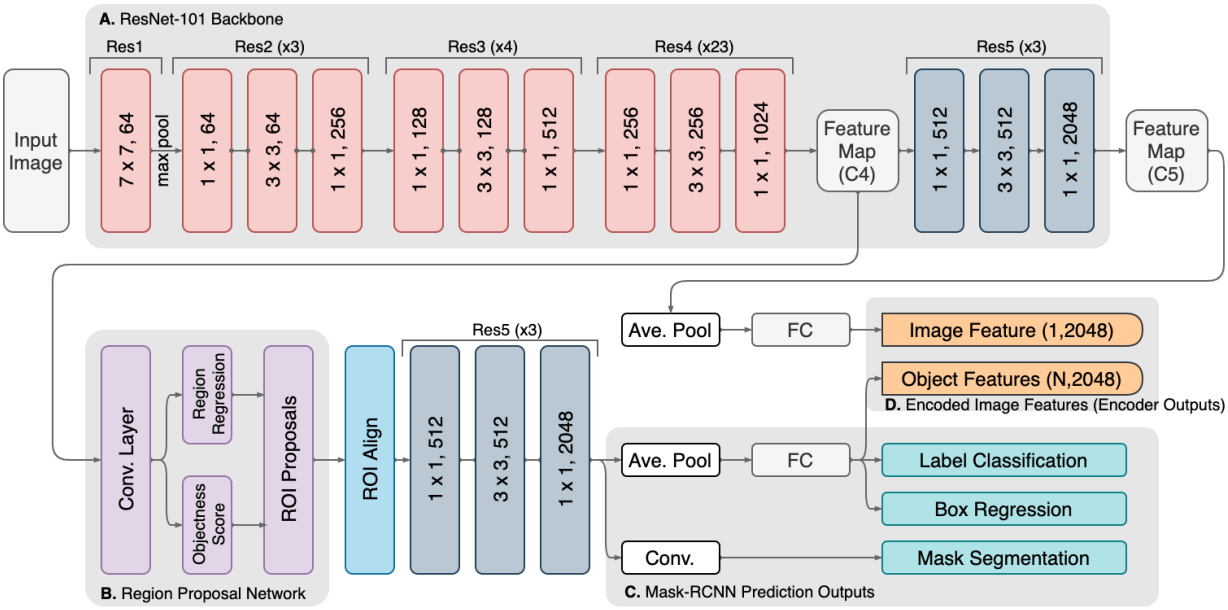


Figure 5. The architecture of the Mask R-CNN-based Encoder.

The C4 feature is sent to the Region Proposal Network (RPN) on the other path. The RPN utilizes the C4 feature map to predict a set of object region proposals with a broad range of scales and ratios. The RPN is a small network that slides over the convolutional feature map. RPN has a classifier and a regressor. At each sliding location, the classifier calculates the objectiveness score – the probability of a region having a target object; and the regressor calculates the proposal coordinates. Regions with high objectiveness scores will be selected as the Regions of Interest (RoIs) and transferred to subsequent processes with coordinate values.

After cropping the object features on the C4 feature map based on the coordinates of each RoI using the RoI Align layer, another 'Res5' convolution stage further processes the object features to shrink the feature size and extract more semantic meaningful features. At this point, the output heads, labeled as panel C in Figure 5, could provide the outputs as a usual Mask R-CNN models after some simple pooling, deconvolution, and fully-connect layers.

Moreover, the loss function of the Mask-RCNN-based Encoder is defined as:

$$L_{Mask-RCNN} = L_{cls} + L_{box} + L_{mask} \tag{3-1}$$

$$L_{cls} = -\frac{1}{N}\sum_{i} c_i^* p(c_i) \tag{3-2}$$

$$L_{box} = \lambda \sum_{i \in \{x,y,w,h\}} \text{smooth}_{\text{L1}}(t_i - t_i^*) \quad \text{3-3}$$

$$L_{mask} = -\frac{1}{M} \sum_i [m_i^* \log p(m_i) - (1 - m_i^*) \log(1 - p(m_i))] \quad \text{3-4}$$

where:

N, M are the total number of classes, and mask pixels,

c is the predicted and ground truth class,

c^* is the binary indicator if the label c is correct,

λ is the balancing weight, whose default value is 10,

$\text{smooth}_{\text{L1}}$ is the smooth L1 loss function,

t, t^* are the predicted and ground truth bounding box coordinates for class u,

x, y, w, h are the coordinates values,

m are the predicted and ground truth mask pixels,

m^* is the binary indicator if the label m is correct

3.2.2 Image Captioning-based Decoder

In this method, the object detector will be the encoder to extract the image features. The encoder will extract both image features of the whole image and local objects:

$$\boldsymbol{v}, \boldsymbol{v}_i = \text{encoder}(image), \qquad \boldsymbol{v}, \boldsymbol{v}_i \in \mathbb{R}^d \quad \text{3-5}$$

where the $\boldsymbol{v}$ is a d-dimension feature vector for the whole image and the $\boldsymbol{v}_i$ is a d-dimension feature vector for the $i - \text{th}$ object region proposal.

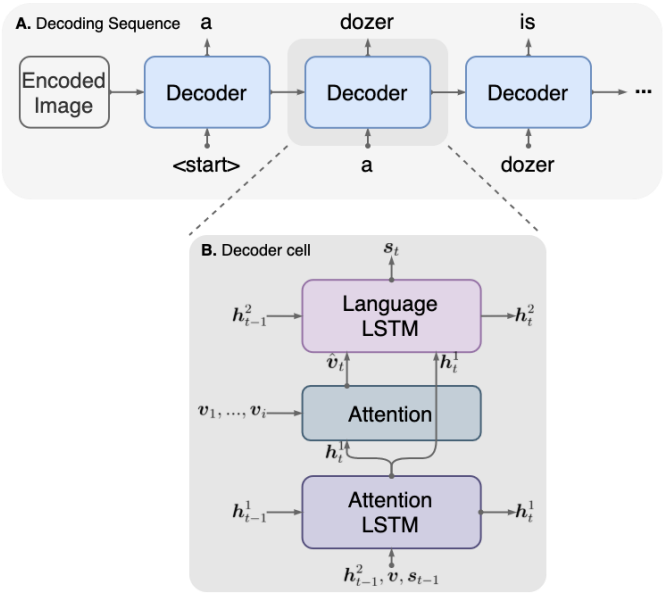


Figure 6. The sequence of the decoding process and the detail of a decoder cell.

As shown in Figure 6, two-layer LSTMs with an attention cell have been adopted as the decoder, namely the Attention LSTM and the Language LSTM. The Attention LSTM layer further processes the image feature. It also helps the attention cell calculate the visual attention on the object zones, which later becomes the connection between the image captioning words and detected object zones. The input of the Attention LSTM layer at each time step is a concatenation of the output of the previous Language LSTM layer $\boldsymbol{h}^2_{t-1}$, the feature vector of the whole image $\boldsymbol{v}$, and the embedding vector of the previous predicted word $\boldsymbol{s}_{t-1}$(the word embedding is learned by random initialization without any pre-training):

$$\boldsymbol{x}^1_t = [\boldsymbol{h}^2_{t-1}, \boldsymbol{v}, \boldsymbol{s}_{t-1}] \tag{3-6}$$

These inputs give the Attention LSTM layer the related information about the Language LSTM's current state, the feature map of the whole image, and the memory of the generated caption words so far.

After getting the hidden state of the Attention LSTM by:

$$\boldsymbol{h}^1_t = \mathrm{LSTM}(\boldsymbol{x}^1_t, \boldsymbol{h}^1_{t-1}) \tag{3-7}$$

an attention cell is followed to calculate a normalized attention weight $\alpha_{i,t}$ on each of the k object zone features v_i as follows:

$$c_t = \boldsymbol{w}_c^T \tanh(W_v \boldsymbol{v}_i + W_h \boldsymbol{h}_t^1) \quad \text{3-8}$$

$$\alpha_t = \text{softmax}(c_t) \quad \text{3-9}$$

where the $\boldsymbol{w}_c^T$, W_v, and W_h are trainable parameters. Then the attended object features, which is the weighted summation of each object feature, could be obtained:

$$\boldsymbol{v}_t = \sum_{i=1}^{K} \alpha_{i,t} \boldsymbol{v}_i \quad \text{3-10}$$

The Language LSTM layer takes the concatenation of the attended object feature and the output of the Attention LSTM layer as the input. It outputs the hidden state of the language model:

$$\boldsymbol{x}_t^2 = [\boldsymbol{v}_t, \boldsymbol{h}_t^1] \quad \text{3-11}$$

$$\boldsymbol{h}_t^2 = \text{LSTM}(\boldsymbol{x}_t^2, \boldsymbol{h}_{t-1}^2) \quad \text{3-12}$$

A linear layer with a SoftMax activation function added upon the Language LSTM takes these outputs and predicts the conditional distribution over possible output words at time step t, given the generated sequence of words $(s_1, \ldots, s_{t-1})$ so far:

$$p(s_t | s_{1:t-1}) = \text{softmax}(W_p \boldsymbol{h}_t^2 + \boldsymbol{b}_p) \quad \text{3-13}$$

Provided with the ground truth caption words $s_{1:T}^*$, this study updates all the trainable parameters (noted as θ) in the captioning model by minimizing the following cross-entropy (XE) loss with stochastic gradient descent learning:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log\big(p_\theta(s_t^* | s_{1:t-1}^*)\big) \quad \text{3-14}$$

3.3 Implementations, Experiments, and Results

3.3.1 Experimental Setup

This study trained the detecting encoder and the captioning decoder using two datasets. The metadata and examples of datasets are provided in Table 1. To ensure the robustness of the trained model, this study include as many as possible data instances from different scenarios (environment, view of angle, weather, etc.). For training the encoder, this study utilized the dataset for Moving Objects in Construction Sites (MOCS) [115]. The MOCS dataset contains 41,668 images collected from

174 different construction sites and 13 object categories. This study utilized the training set (19,404 images) and the validation set (4,000 images) to pre-train the encoder.

For the captioning decoder, this section annotated 6000 images selected from the Alberta Construction Image Dataset (ACID) with natural language descriptions [116]. This annotated dataset is referred to as the ACID-C dataset. For each image, this study collected two to three captions. Captions corresponding to the images are then pre-processed by tokenization. Tokenization splits the sentence into a list of single words and drops all the punctuations. Next, at the beginning and the end of the token list, this method adds '<start>' and '<end>' words to indicate the start and end of a sentence. Finally, a word dictionary is built that includes all the words that appear in the caption corpus and indexes them by numbers. These processed 6000 data instances are randomly divided into training and validation sets according to an 8:2 ratio.

Table 1. Data sample for the training datasets.

Module	Training Image	Visualized Labels	Labels	Categories	Total Images
Encoder			{… 'bbox': [279.0, 398.0, 252.0, 124.0], 'category_id': 10, 'segmentation':[…] imgid:608, …}	Worker Static Crane Hanging Head Crane Roller Bulldozer Pump Truck Concrete Mixer Pile Driving	19404(Train) 4000(Val)
Decoder		N/A	{… raw: "grader is driving on the road by a man" imgid:6, sentid:10, …}	N/A	4800(Train) 1200(Val)

3.3.2 Implementations

The experiments ran on a workstation with Xeon E5-2678 CPU and GeForce RTX 2080 Ti GPU. The software environment is configured as ubuntu 18.04, Python 3.8, PyTorch 1.7.1, CUDA 11.0, Detectron2, and other related packages.

This study trained the encoder with an initial learning rate of 0.02 and a batch size of 8 images per batch. The learning rate decayed at step 90000 to 2×10^{-3}, and then decayed to 2×10^{-5} at step 125,000. The point in this figure is the raw record, and the lines are Locally Estimated Scatterplot Smoothing (LOESS) of the raw values. The training process was terminated after 155,000 steps since there was no significant loss value reduction and validation performance improvement. This study then trained the decoder with this pre-trained encoder. The models are trained using the cross-entropy loss. The number of training epochs is set to 30. The learning curve about the loss value and learning rate for both encoder and decoder are shown in Figure 7. The total loss of the encoder is calculated as Equation 3-1 to 3-4, and the total loss of the decoder is calculated as Equation 3-14.

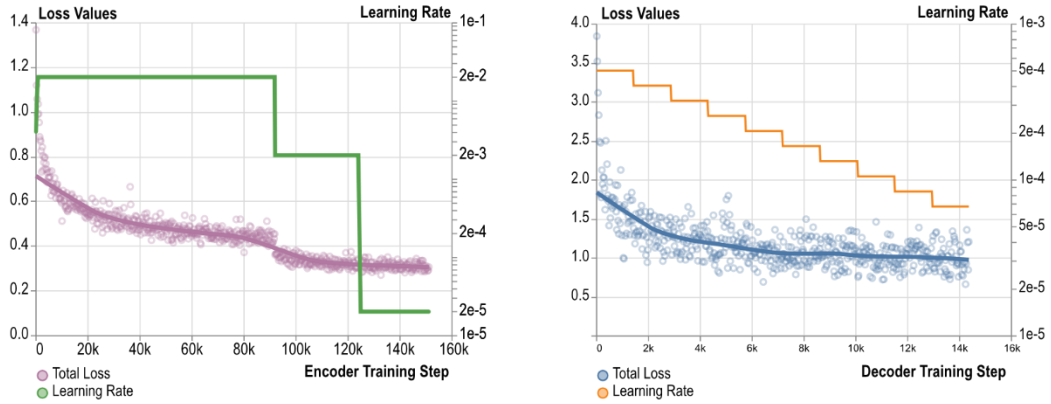


Figure 7. The learning curve of the decoder.

As for data post-processing, This study built a parser based on Python and utilized the NLP package spaCy and Sng_Parser [117] to accelerate the coding process.

3.3.3 Evaluation Metrics

This study uses the following automatic NLP evaluation metrics to evaluate the performance of the semantic information extraction as an image captioning task and the language quality of the generated captions. For all these metrics, a higher score indicates better performance.

Bilingual Evaluation Understudy (BLEU)

The overlap between the predicted single word or n-gram (sequence of n-adjacent words) and a collection of reference sentences is measured by BLEU [118]. The semantic meaning of the words is not taken into consideration by BLEU, which solely assesses word and sentence length matches. The

BLEU is divided into four types: BLEU-1, BLEU-2, BLEU-3, and BLEU-4, with the number indicating the number of words used up to n-grams.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

The recall score of the produced sentences corresponding to the reference phrases is measured using n-grams in ROUGE [119]. ROUGE-L, which computes the Recall and Precision of the longest common subsequences between the candidate and reference phrases, is the most often used variant in picture captioning.

Consensus-based Image Description Evaluation (CIDEr)

CIDEr [120] measures the co-existence frequency of the n-grams in both candidate and reference sentences after converting the terms in both phrases to their root forms. The term-frequency-inverse-document-frequency (TF-IDF) method was used for the measurement. In NLP, the TF-IDF is a commonly used statistical approach. It assesses the significance of a phrase inside a text in the context of a collection of documents.

Semantic Propositional Image Caption Evaluation (SPICE)

SPICE [121] score is measured by comparing the scene graph tuples of the candidate sentences to those of the reference sentences and calculating the degree of similarity. The scene graph contains the various objects, their qualities, and the relationships that were derived from the text.

3.3.4 Experimental Results

Table 2 provides the evaluation results of the image captioning task of the decoder and other benchmark models. The inference process speeds around 4.25 frames per second on a single RTX 2080 Ti GPU. Managers could utilize better hardware or conduct parallel computing to achieve real-time operation. The proposed model has a validation performance on the evaluation metrics with BLEU-4 of 0.36, ROUGE of 0.57, CIDEr of 1.84, and SPICE of 0.41. The performance of the image captioning implementation in the construction community [104] has also been reported in Table 2 as benchmarks 1, 2, and 3. These models are in similar encoder-decoder structures but without attention mechanism and the visual connections to the original input image. These benchmark models are trained on a customized dataset for on-site workers and have the highest CIDEr score of 1.61 and SPICE of 0.36.

MS COCO [122] is a large computer-vision community dataset that provides an image captioning dataset for real-life scene images. It also provides the performance of leading implementations in the computer vision community. On its leaderboard [123], the top two implementations are TencentVision [124] and panderson@MSR/ACRV [55]. The evaluation results of TencentVision are CIDEr-1.12 and SPICE-0.21; and the evaluation results of panderson@MSR/ACRV are CIDEr-1.18 and SPICE-0.22.

For all the metrics, a higher score means better performance. The output range for BLEU, ROUGE, and SPICE is $[0,1]$, and range for CIDEr is $[0,10]$. A score close to zero indicates poor overlap between predictions and references for BLEU and ROUGE. A score close to one indicates a strong overlap between predictions and reference words. CIDEr and SPICE are originally designed for image captioning tasks with semantic match; a higher score indicates higher semantic similarity between predictions and references. Though the dataset used in our study is different from other benchmarks, the results suggest that our model has a comparable ability to extract semantic information from images as other state-of-arts equivalents.

Table 2. Model performance of the image captioning evaluation metrics.

Model (Dataset)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	SPICE
Proposed * (ACID-C)	0.61	0.52	0.44	0.36	0.57	1.84	0.41
Benchmark 1 (a dedicated dataset [104])	0.52	0.43	0.36	0.32	0.50	1.09	0.36
Benchmark 2 (a dedicated dataset [104])	0.66	0.58	0.52	0.48	0.65	1.55	0.35
Benchmark 3 (a dedicated dataset [104])	0.68	0.60	0.54	0.49	0.62	1.61	0.35
TencentVision [124] (MS COCO[122])	0.79	0.64	0.49	0.36	0.57	1.12	0.21
panderson@MSR/ACRV [55] (MS COCO[122])	0.80	0.64	0.49	0.37	0.57	1.18	0.22

The evaluation of the decoder also demonstrated slightly lower BLEU scores and better CIDEr and SPICE scores. The BLEU metric is based on strict matches of the words and sentence length

between the predicted and reference captions. Previous investigations have reported that BLEU algorithmic variations "don't reflect either grammaticality or meaning preservation very well" [125] and "don't map well to human judgments in evaluating NLG (natural language generation) tasks" [126]. CIDEr and SPICE are originally designed for image captioning tasks, adding semantic match and word mapping. Thus, CIDEr and SPICE are more suitable for evaluating the ability to extract semantic information from images. Our decoder performs well on CIDEr and SPICE, and its results suggest that it has a good ability to extract semantic information from the site images.

3.3.5 Image Results Demonstration

Figure 8 provides samples of the retrieved semantic information as captions for the images. It also provides a visualization of the parsed model output for three test site images in Figure 9. These three sample images are never used in the training process of this model. The object detection and instance segmentation results are provided in the first row. The second row provides the predicted image description of the image. The third row provides the visualized attention mapping for the first object word in the caption with the image object zone. The connected object zone is drawn as a red square in the image. The fourth row shows the post-processed semantic information in tabular format. And the fifth row provides the scene graph constructed for the semantic information.

The first column of Figure 9 shows that the encoder has successfully recognized the excavator and truck in the image. The decoder has extracted the semantic information as a caption: "an excavator is excavating dirt while a dump truck is parking beside". The attention map has connected the object word "excavator" with the object region of the excavator recognized by the encoder. The data post-processing has identified the entities "excavator", "dirt", and "dump truck" in the caption. The rich semantic information represented as triplet "excavator-excavating-dirt" is also extracted successfully. Finally, the graph formatted semantic information could be extracted as shown in the fifth row. The second and third columns also show that our system has successfully extracted the semantic information in other test images.

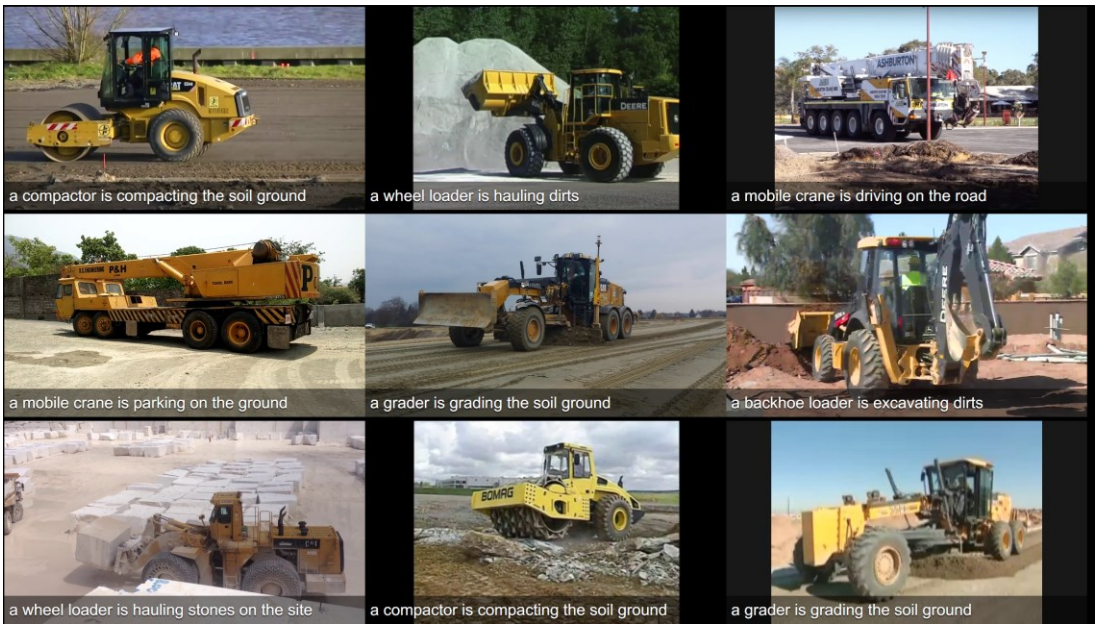


Figure 8. Examples of the semantic information extracted as a caption for construction images.

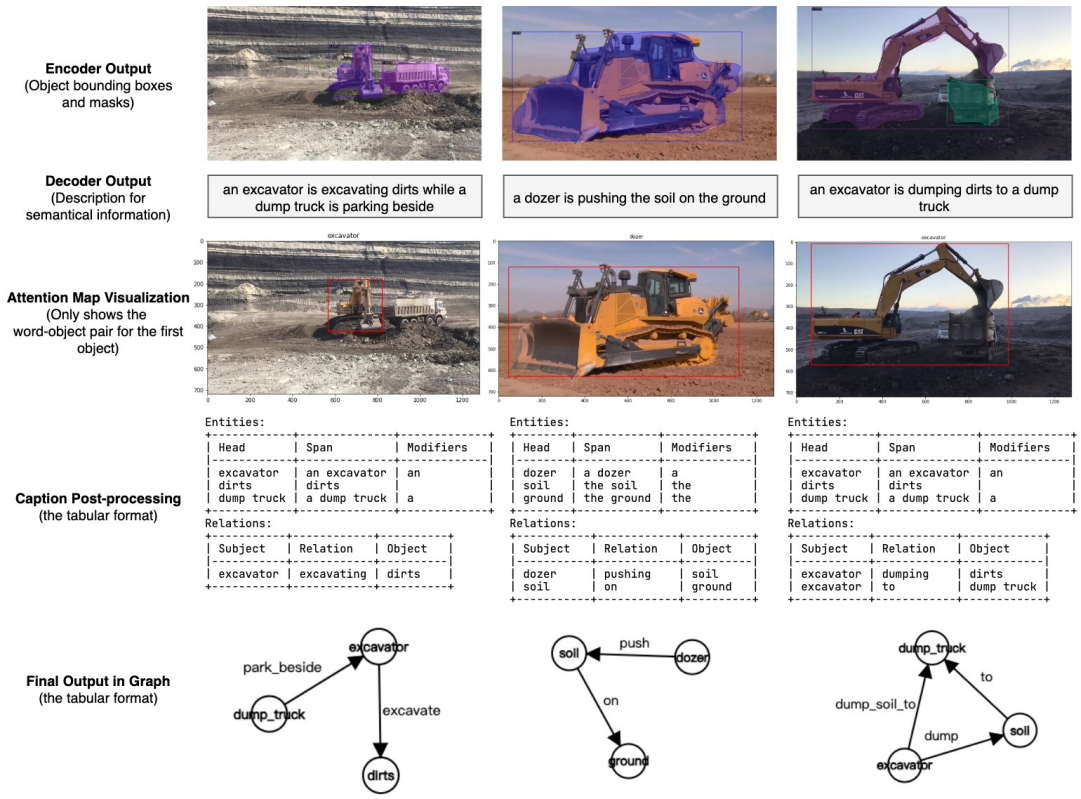


Figure 9. Visualization of the model output, including the encoder and decoder outputs, the visualized attention map, and the final output in tabular and graph formats.

3.4 Discussion

3.4.1 Feasibility of Encoder

This study evaluates the object detection and instance segmentation performance on the MOCS validation set of the encoder to indicate the feasibility of the encoder. This study utilizes Mean Average Precision (mAP) to evaluate the performance of the object detection and instance segmentation [54].

The evaluation metric is based on average precision (AP). Given a certain level of confidence value α, the AP_α integrates the precision scores at different recall levels r:

$$\mathrm{AP}_\alpha = \frac{1}{11} \sum_{r \in \{0.0, 0.1, \ldots, 1.0\}} Precision(r) \tag{3-15}$$

and the mAP_α is average of AP_a values over all N classes:

$$\mathrm{mAP}_\alpha = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_{\alpha,i} \tag{3-16}$$

Finally, this study utilized the following metrics to conduct the evaluation:

$$\mathrm{mAP} = \frac{1}{10} \sum_{\alpha \in \{0.5, 0.55, \ldots, 0.95\}} \mathrm{mAP}_\alpha \tag{3-17}$$

The mAP performance of the encoder can be found in Table 3. The bounding box mAP of the encoder on the MOCS validation dataset is 54.1%. It also reports the evaluation scores from the benchmark models from [115]. The benchmark models for object detection are Mask R-CNN, Faster R-CNN, YOLO-v3, and RetinaNet, whose mAP scores are 50.8%, 50.6%, 39.0%, 50.0%, respectively. The mAP score of the proposed encoder for instance segmentation is 40.6%. The benchmark models for instance segmentation tasks are Mask R-CNN, MS R-CNN, and SOLO-v3, whose mAP scores are 43.18%, 42.86%, and 43.64%, respectively.

Table 3. Evaluating the mAP performance for the object encoder.

Type	Model	mAP(%)
Bounding Box	*Proposed Model*	54.1
	(Mask R-CNN based)	
	Faster R-CNN	50.6
	YOLO-v3	39.0
	RetinaNet	50.0
Segmentation Mask	*Proposed Model*	40.6
	(Mask R-CNN based)	
	MS RCNN	42.9
	SOLO-v3	43.6

The results show that our encoder has a slightly better performance on object detection and a slightly lower performance on instance segmentation than other benchmark models. The evaluation indicates that our encoder could correctly extract the visual features of the object regions.

3.4.2 Failure Cases

Table 4 provides several false inference samples in the validation process. There are two main types of error. The first type of error is the false recognition of construction materials. For example, in the first sample in Table 4, the model recognizes the material in the backhoe loader's bucket as dirt instead of rock. In the second sample, the construction material here is snow, but the model recognized it as soil. The second type of error is false recognition of the construction equipment. For example, in the third sample in Table 4, the model recognized the object as a wheel loader by mistake, which should be a dozer. In the fourth sample, the model mistakenly recognized a steel box as a dump truck. In the fifth sample, the model made two types of mistakes; it recognized a pump truck as a mobile crane by mistake and failed to recognize the construction material that the pump truck was working on.

Table 4. Examples of failure cases in the validation process.

Index	Image	Caption
1		a backhoe loader is excavating dirt
2		a grader is grading the soil ground
3		a wheel loader is driving on the ground
4		an excavator is dumping dirt into a dump truck
5		a concrete mixer truck is dumping cement while a mobile crane is transferring

The first type of error may come from the absence of regional material features. In the implementation, the encoder was trained on a dataset that only labels the heavy equipment. So, the detailed visual feature is only for equipment. When inferencing the construction material, the decoder has to look at the whole image feature, which is less detailed. To solve this problem, further studies could prepare a dataset that labels the construction material. This will improve the accuracy of material recognition. The second type of error may come from a confusing angle of view or over-similar objects. Improving the number of data instances of the training dataset with more variations should solve this problem.

3.4.3 Discussion of Explainable AI

Though deep learning has demonstrated its potential to automate industries, achieving or surpassing human performance in recognition precision for certain tasks, it is often identified as a "black box" or opaque system. As the "black box" paradigm remains in deep learning, existing deep learning-based methods merely provide clues or explanations for why specific results are generated or the decisions are made. [127] This limitation occasionally generates absurd results and undermines users' faith in using deep learning models to make important decisions. Thus, interpretability is essential for users to understand, trust, and effectively manage the deep learning models for many crucial applications [128], especially for making life and death decisions such as construction safety monitoring and management.

While the proposed method tries to extract semantic information from construction imagery with adequate accuracy, it also extends the interpretability toward the explainable artificial intelligence (XAI) goal. In the image encoding process, the proposed method did encode not only the entire image feature but also the image feature based on detected object zones. This makes the decoder could utilize a specific region of data from the whole image. By monitoring which region of data is utilized, users could understand what object data is utilized. During training, the attention mechanism in the decoder generates an object-word relevance loss, which digitizes whether the generated word considers the object zones in the input image well. With more and more training steps executed, the model could generate better outputs considering the object regions. During the testing or implementation, the attention mechanism generates weights for each encoded object zones, indicating the relevance between the object zone and the extracted word in the pair. By providing the weight for the object-word pair, this model provides insight into the utilization of inputs towards the decision-making, at least on some intuitive level.

This method did not solve all the interpretability problems of the deep learning models. For example, it cannot provide an intuitive explanation of the weights in the CNN-based visual feature extractor. More studies are needed to improve the interoperability of deep learning models.

3.4.4 Methodological and Practical Contributions

This proposed method has demonstrated its ability to extract semantic information with visual connections from construction imagery. It has several methodological contributions:

(1) This method proposed to utilize regional visual features as the input for the image captioning-based decoder. The regional visual feature is extracted by the object detector-based encoder. This regional visual feature integrates object detection and image captioning. At the same time, it also improves the performance of the information extracting precision.

(2) This method proposed an attention mechanism in the decoder. This module achieves an explicit visual connection between the extracted information and the image region. The visual connection extends the extracted semantic information with location information. Thus, it could improve the visualization and enable more sophisticated management tasks.

(3) This method improves the explainability of the vision-based models. The explainability may enable an intuitive representation of the calculation mechanics inside the deep learning model and improve the utilization of deep learning models in construction management.

Besides the methodological contributions, this proposed method also has the following practical implications:

(1) This method could improve the visualization and documentation in construction management. The visual connection ability enables displaying the related image zone for extracted information. This provides an intuitive visualization of the extracted semantic information. The extracted semantic information could serve as enriched metadata to simplify the construction image documentation process.

(2) This method could improve the current practice of vision-based monitoring and management by providing richer semantic information and visual connections. The extracted semantic information could provide an integral information package for downstream construction management applications. This information enables more complex decision-making processes and management tasks. For example, to monitor the usage of equipment on-site, managers could simply retrieve and analyze the activity and combine it with the time tag in the image metadata.

3.4.5 Limitations of the Proposed Method

This study proposed an integrated system architecture that extracts semantic information with visual connections from site images for the construction community. The limitations of the proposed method are illustrated as follows:

(1) This study implemented and validated the method for construction images. Though the feasibility of this method has been confirmed, applying the proposed method to construction videos still needs minor modifications to datasets and the input module of the proposed architecture.

(2) Though our model performs better than existing methods, it may fail on several conditions, such as rare view-of-angle and confusing construction materials. Moreover, there may be a mismatch between the image regions and caption words. Since there is no ground truth of the visual connection in the training dataset, fine-tuning the visual connection is tricky and difficult. Extending the data instances of the training dataset should solve this problem.

(3) The proposed model is not fully interpretable. For example, the explainability of the CNN-based encoder is still low. Future research in both the computer science and construction management communities is still needed to further extend the explainability of deep learning models.

3.5 Conclusion

This section presented an integrated information extraction system for on-site images, extracting semantic information such as objects, activities, and interactions. It contains three modules: (1) the object detector-based encoder, which detects the object in the image and extracts the feature maps of the image and objects; (2) the image captioning decoder, which extracts the semantic information as a natural language sentence according to feature maps; (3) the post-processing module, which parses the output sentence into scene graph and maps the recognized objects with the object zones of the image. The extracted information has both advantages of rich information and visual connections. Our implementation's evaluation results show good performance on various tasks such as object detection and semantic information extraction via captioning.

The contributions of this study are: practically, (1) this method could improve the visualization and documentation in construction management; (2) this method could improve the current practice of vision-based monitoring and management by providing richer semantic information and visual connections; methodologically, (1) this method uses regional visual feature for better recognition and information extraction performance, (2) this method designs an architecture which could provide a visual connection between the image and extracted information, (3) the architecture of the proposed method provides a certain level of explainability.

Chapter 4
Automatic Safety Hazards Identification and Reasoning

4.1 Introduction

Construction sites are hazardous places with various safety risks that pose a serious threat to workers. In the United States, the construction industry reported over 1000 deaths and 75,400 nonfatal injuries in 2020 [2]. The top three categories of fatal hazards are falls and slips, transportation incidents, and inappropriate contact with objects and equipment [2]. The frequency and severity of these accidents highlight the urgent need to improve safety measures in the construction industry.

Many of these hazards can be prevented by enhancing safety management and minimizing exposures that may contribute to hazards and affect the health of construction workers. Behavior-based safety (BBS) is effective in promoting safe behavior. BBS is an approach to occupational safety that focuses on changing individual behaviors to reduce workplace accidents and injuries. However, existing BBS methods in the construction industry have limitations such as being manual, time-consuming, and subject to observer bias, thus making them inefficient and error-prone [3–5]. These disadvantages can be attributed to the tedious and labor-intensive nature of manual observation and the difficulties in monitoring all workers continuously [6].

To address these limitations and automate the inspection and observation of construction sites, computer vision (CV) technologies have been increasingly adopted in the construction industry [5]. CV is a subfield of artificial intelligence that enables computers to process, analyze, and understand images and videos, thus allowing for the recognition and classification of objects, people, scenes, and events [14]. By leveraging CV, methods have been developed to recognize hazardous postures and actions [8], detect missing personal protective equipment (PPE) [129], and automate construction safety management [17].

Although existing methods yield satisfactory results, there are limitations to using CV alone in safety management. For example, CV methods can detect only simple repeating objects or activities

[5]. In contrast, the traditional method of BBS prefers detecting complex patterns and relationships of worker behavior and then identifying potential safety hazards. Furthermore, CV methods cannot leverage domain knowledge regarding safety regulations and guidelines, thus limiting their ability to infer if the behaviors and interactions presented in the image follow the safety regulations. Recent advancements in natural language processing (NLP) technologies have enabled computers to process, understand, and infer natural text languages [15]. NLP can aid in extracting and evaluating semantic meanings from safety regulations and guidelines, thereby providing valuable domain knowledge for assessing the safety of complex construction activities.

In this section, the author proposed a framework for construction safety management by integrating CV and NLP technologies to automate the safety hazard identification and reasoning in construction sites. The safety hazard identification could generate classifications of a behavior about whether it is safe or unsafe, while the hazard reasoning provides the reason of an identified unsafe behavior. The framework comprises two modules: (1) an image processing module based on CV and dense image captioning technologies to recognize behaviors and interactions in images, and (2) a text processing module based on NLP technologies to extract and evaluate the semantic similarities of safety regulations and guidelines. The proposed framework can improve safety management in the construction industry and minimize the occurrence of fatal and nonfatal injuries.

4.2 Methodology

To extract semantic information regarding objects, activities, and interactions and link it to the domain knowledge base for hazard causes and identification, this study combined dense image captioning and NLP within a single framework. A graphical overview of the modules and submodules used in the proposed framework is presented in Figure 10. The first module transforms image data into text data. It first recognizes several target regions inside the site image. These image regions often contain important objects for safety management, such as workers, PPE, and tools. Subsequently, this module recognizes the semantic information inside each region and describes the information in a natural language caption format. The second module performs safety hazard reasoning and identification by measuring the semantic similarity between the region captions obtained from the first module and domain knowledge base. This module first groups the captions belonging to a worker

and then uses semantic similarity to determine whether the actions performed in the image follow the safety rules.

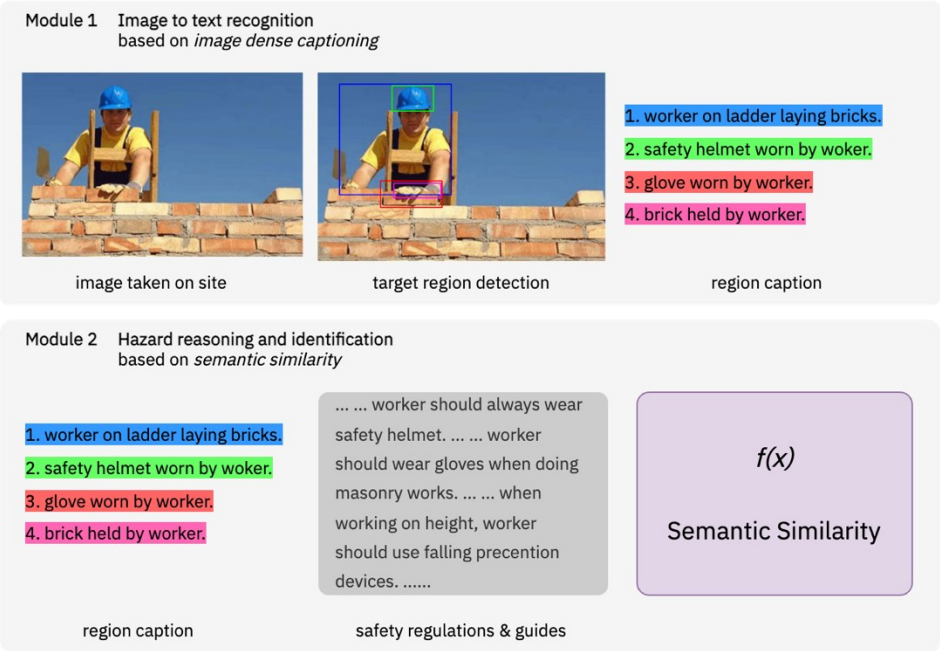


Figure 10. Main modules of the proposed method.

The computational flow of the proposed framework is illustrated in Figure 11. Figure 11a and 3b correspond to the first module and second module, respectively. The details regarding each part and the calculation processes are presented in the following subsections.

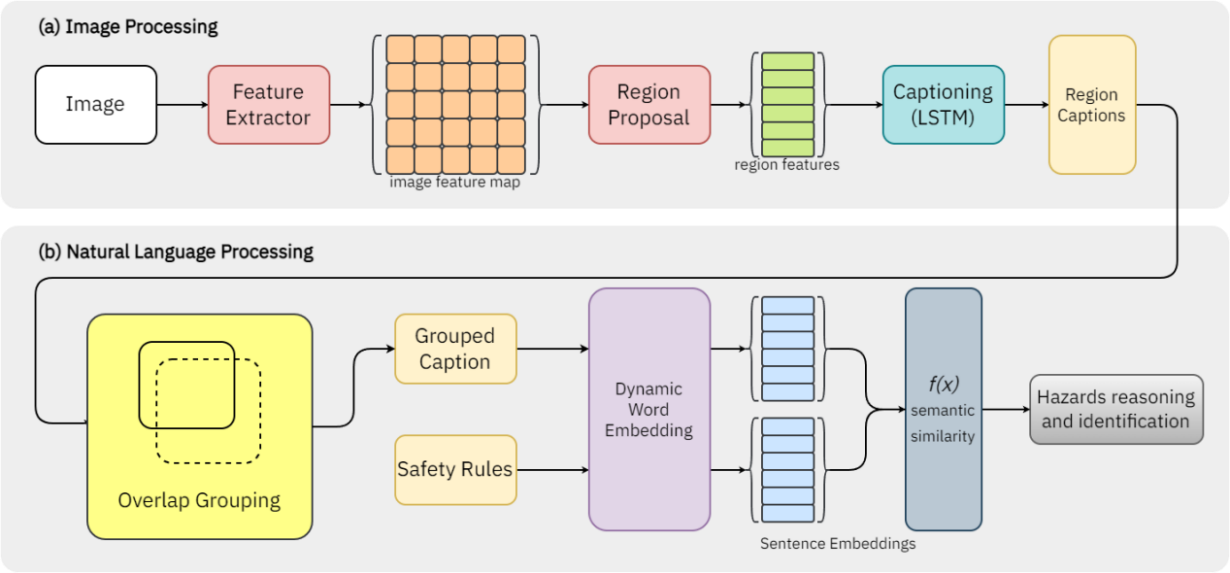


Figure 11. Computational workflow of the proposed method.

4.2.1 Visual Recognition and Description Generation

In this subsection, the calculation process illustrated in Figure 11a is described. An image of the construction site is taken as input. After processing this image, semantic information is extracted as text descriptors of multiple subregions of the image. The image data is generally represented using a three-dimensional (3D) tensor:

$$I \in \mathbb{R}^{3\times W\times H} \tag{4-1}$$

Feature extraction: The image data are first forwarded to a convolutional neural network (CNN)-based feature extraction network to obtain the image feature map. CNNs are a type of neural network designed to process grid-like input and have been used to process image data [38]. The CNN breaks down the grid data into smaller feature maps to reflect the visual properties of the image. The deeper the layer in the CNN, the higher the representation level of the image feature in the layer's output. For instance, shallow layers of the CNN identify lines and edges in the image when processing images, whereas deeper layers identify shapes, such as the overall shape of a safety helmet.

$$\boldsymbol{V} = \mathrm{CNN}(I), \boldsymbol{V} \in \mathbb{R}^{C\times H'\times W'} \tag{4-2}$$

where $\boldsymbol{V}$ is the visual feature map, C is the number of channels of the feature map, H' and W' are the height and width of the feature map.

Region proposal: The extracted image feature map is forwarded to a region proposal network (RPN) to generate target regions in the image. RPN is a CNN proposed by Ren et al. [41] and has been used for object detection. RPN uses the sliding window approach, wherein the network generates multiple subregions of the input image with different locations, sizes, and aspect ratios. These subregions have their counterparts called anchor boxes on the extracted feature map. The RPN then uses these anchor boxes to predict whether each anchor box contains an object of interest and, if so, adjusts the size and position of the bounding box to fit the object more accurately. This is accomplished using a set of regression coefficients learned during training:

$$\boldsymbol{B}, \boldsymbol{s} = \mathrm{RPN}(\boldsymbol{V}), \boldsymbol{B} \in \mathbb{R}^{n\times 4}, \boldsymbol{s} \in \mathbb{R}^{n} \tag{4-3}$$

After obtaining the subregions' coordinates $\boldsymbol{B}$ and scores $\boldsymbol{s}$, the final target subregions are filtered out by applying a threshold score, whereas the image feature vectors of the subregions are obtained using another small CNN:

$$\boldsymbol{v}_i = \mathrm{CNN}(\boldsymbol{V}, \boldsymbol{B}_i), \boldsymbol{v}_i \in \mathbb{R}^{d} \tag{4-4}$$

Dense captioning: The region features are forwarded to a long short-term memory (LSTM)-based recurrent neural network (RNN) to realize semantic information recognition and extraction. Figure 12 illustrates how the semantic information is recognized and how the region caption is generated in the LSTM cells. The LSTM cell performs several calculations to encode the necessary information as the hidden state. The hidden state transfers the information between cells, meaning that the previous cell's hidden state is used to calculate the hidden state of the current time step.

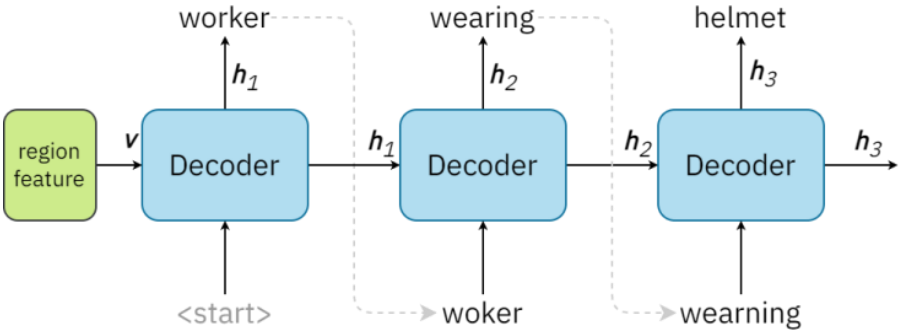


Figure 12. The region caption generation process.

The previous hidden state and previous output word are inputted into the LSTM cell at the current time step. Caption words are represented by one-hot vectors. The initial hidden state is the region feature vector, and the initial input word is a special word token "<start>" indicating the start of the captioning process:

$$\begin{aligned} \boldsymbol{h}_t, \boldsymbol{c}_t &= \mathrm{LSTM}(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t), \\ \boldsymbol{x}_t &= \boldsymbol{c}_{t-1} \text{ for } t > 0, \\ \boldsymbol{x}_0 &= < \text{start} >, \\ \boldsymbol{h}_0 &= \boldsymbol{v}, \end{aligned} \tag{4-5}$$

where t is the timestep in the LSTM decoding process, $\boldsymbol{h}_t$ is the hidden state of the decoder at time step t, $\boldsymbol{x}_t$ is the input of the LSTM at time step t, $\boldsymbol{c}_t$ is the vector representation of the output word at time step t, and $\boldsymbol{v}$ is the feature vector of an image subregion.

As this method is a dense captioning application, all the region features are forwarded to the LSTM cells to obtain the caption text for each region.

4.2.2 Caption Grouping

In the grouping process, captions are organized into separate groups on the basis of which worker the caption is describing. Because multiple workers may be present in an image, the captions may be disordered and not properly grouped by the workers. To resolve this issue, a grouping method is used to assign captions belonging to the same worker into a single group. This helps ensure that all captions describing a particular worker are organized together, thus making it easier to analyze and identify the safety status. In this study, overlap measurement was used as the grouping method.

First, the main captions that describe the worker are extracted. The data annotator is required to describe the worker in the first word of the main captions; this is achieved by filtering out captions that begin with the word "worker." Subsequently, to calculate the intersection area, each of the bounding boxes related to the remaining captions is compared with the main bounding boxes. This allows the calculation of the intersection ratio as follows:

$$ratio = \frac{area(box_1 \cap box_2)}{\min\ (\text{area}(box_1), \text{area}(box_2))} \quad \text{4-6}$$

An overlap ratio threshold λ is used to determine which main caption group the current caption belongs to. In this study, the ratio threshold was set as 0.5. Accordingly, when a caption box had an overlap ratio larger than 0.5 with multiple main boxes, the caption was assigned to the main group that has the highest overlap ratio. In contrast, when a caption box had no overlap ratio larger than 0.5 with any main boxes, the caption was regarded as redundant.

4.2.3 Word Embedding and Sentence Embedding

As illustrated in Figure 11b, the NLP-based hazard reasoning and identification process includes steps called "static word embedding" and "dynamic word embedding." The word embedding steps are essential for the second module as they enable understanding the semantic meaning.

One-hot word vectorization: Traditionally, machine learning (ML) algorithms utilize one-hot vectorization to represent words. One-hot vectorization is a method used for encoding categorical data as numerical data that can be used as input to ML models, such as words in a lexicon. This method creates a new dimension for each word in the vocabulary and assigns a binary value (1 or 0 to respectively indicate the presence or absence of the word. This method produces a high-dimensional and sparse representation of the words, with most values being zero. The one-hot vectorization of

several words and the two-dimensional and 3D visualization of some of the word vectors are illustrated in Figure 13a. This method is simple and efficient for encoding categorical data but cannot represent the semantic relationship between words. As shown in Figure 13a, all the word vectors are orthogonal to each other, providing no clue regarding the relationship between words such as synonyms and antonyms.

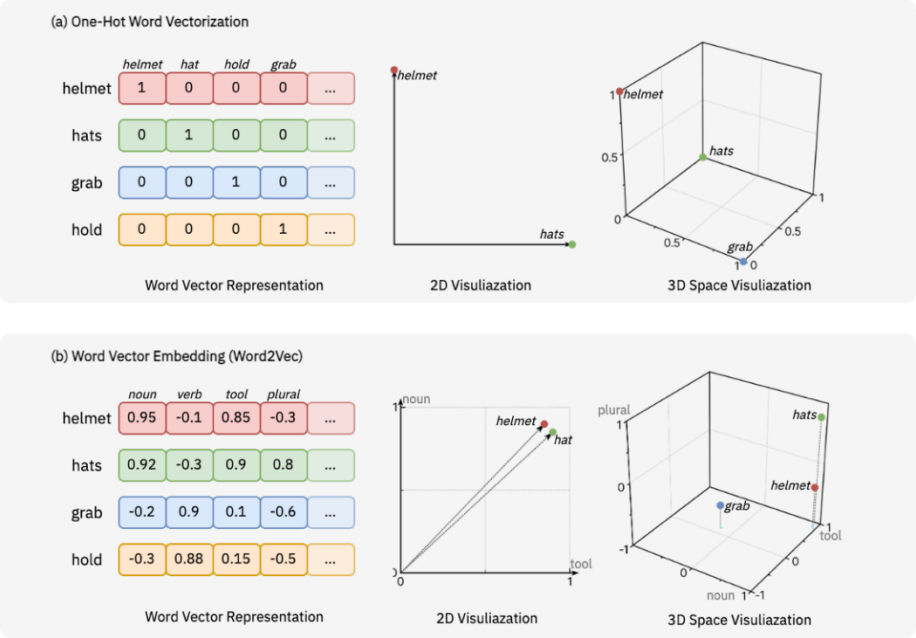


Figure 13. Comparison between one-hot word vectorization and word embedding.

Word embedding: In contrast to one-hot vectorization, where word vectors are orthogonal, the word embedding method represents words as continuous-valued vectors in a low-dimensional space where words that have similar meanings are spatially close together. The vector representation for words is typically trained using large datasets and can capture word context as well as word relationships. A common word embedding training method is Word2Vec [130], which produces a vector space with a dimensionality of several hundred; each unique word in the corpus is allocated a matching vector in the space, as illustrated in Figure 13b. Thus, each dimension in the vector space can be used to measure a semantic or syntactic property of the word, resulting in semantically similar words having close representation vectors.

Static word embedding: Word embedding based on Word2Vec is a type of static word embedding. Word vectors do not change when used with a different language corpus, and the same word is always represented by the same vector no matter which context it appears in, hence the name

"static." Word vectors are typically pretrained on a large corpus of text and then used as a fixed input for downstream tasks. The use of this type of word embedding is illustrated in Figure 14a. The word embedding lookup table is obtained through the training process. Each time a word vector is required, the program indexes the lookup table to query the corresponding word vector. In other words, embeddings do not adapt to the specificity of sentences or datasets for which the model is being called. Thus, static word embedding represents the semantic characteristics of a word in a large context scenario and may not provide an accurate representation of the semantic relationship in the current context:

$$\boldsymbol{S} = \mathrm{Word2Vec}(\boldsymbol{C}), \boldsymbol{C} = [\boldsymbol{c}_1 , \boldsymbol{c}_2 , \ldots , \boldsymbol{c}_n]^\top \tag{4-7}$$

where $\boldsymbol{c}_i$ is the one-hot vectorization of the $i-\mathrm{th}$ word in a sentence, and $\boldsymbol{S} = [\boldsymbol{s}_1 , \boldsymbol{s}_2 , \ldots , \boldsymbol{s}_n]^\top$ is the static word embedding matrix containing static word vectors in it.

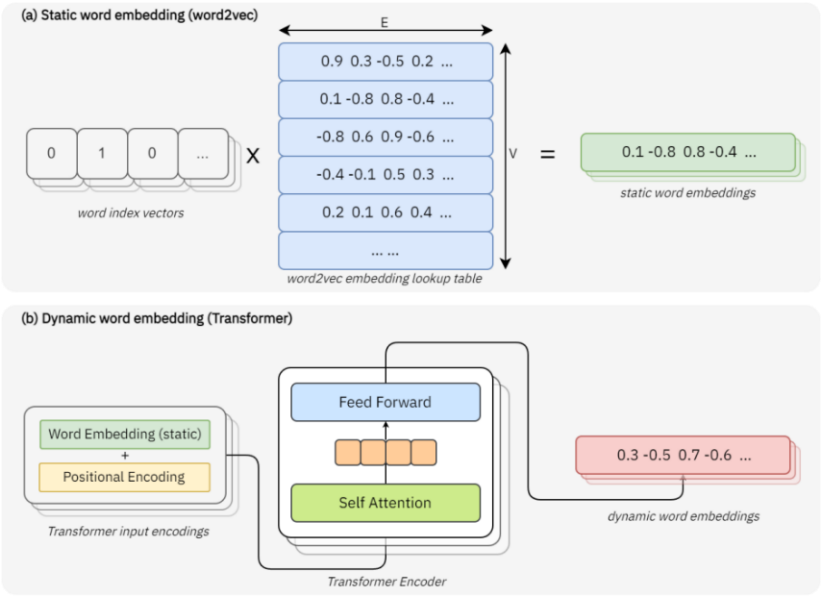


Figure 14. Comparison between the static word embedding method based on Word2Vec and the dynamic word embedding method based on the Transformer network.

Dynamic word embedding: Dynamic word embedding provides contextualized word vectors that adapt to specific sentences. Dynamic embedding uses more advanced models to create representations that change depending on the context. Several dynamic word embedding models are available, such as BERT [131], universal sentence encoder (USE) [132], and Sentence-BERT [133]. These models are based on the Transformer architecture, which is a type of neural network proposed by Vaswani et al. [134]. As illustrated in Figure 14b, dynamic word embedding is achieved using a

Transformer-based language encoder. Generally, the input of the Transformer-based encoder is the static word embedding together with the positional encoding that gives the position of a word in a sentence. The Transformer-based encoder is trained on a large language corpus to learn the parameters for encoding word embeddings. During encoding, the self-attention mechanism allows the model to focus on certain parts of the input while processing it, thus enabling the model to understand the context in which words appear and generate better context-sensitive embeddings for a given sentence input:

$$\boldsymbol{D} = \mathrm{Transformer}(\boldsymbol{C}), \boldsymbol{C} = [\boldsymbol{c}_1\ , \boldsymbol{c}_2\ , \ldots, \boldsymbol{c}_n]^\top \quad \text{4-8}$$

where $\boldsymbol{s}_i$ is the static word embedding of the $i-th$ word in a sentence, and $\boldsymbol{D} = [\boldsymbol{d}_1\ , \boldsymbol{d}_2\ , \ldots, \boldsymbol{d}_n]^\top$ is the static word embedding matrix containing static word vectors in it.

After word embeddings for the words in a sentence are generated, the vector representation of a given sentence can be obtained by applying average pooling on the word vectors:

$$\boldsymbol{s} = \boldsymbol{S}, \boldsymbol{d} = \overline{\boldsymbol{D}} \quad \text{4-9}$$

4.2.4 Universal Sentence Encoder

In this study, the USE [132] was used as the dynamic sentence embedding tool because the USE exhibited better performance on the data and implementation in this study. However, the USE does not outperform other methods such as BERT and Sentence-BERT in all scenarios; future research is required to explore different embedding methods.

The USE is a pretrained DL model that can generate dynamic numerical embeddings of sentences. The USE is based on a deep neural network trained on various NLP tasks, such as sentiment analysis, paraphrase identification, and natural language inference. One of the key features of the USE is its ability to encode the semantic meaning of a sentence rather than just its surface-level syntax. This means that sentences with similar meanings are mapped to similar vectors even if they use different words or structures.

The USE is based on a deep neural network architecture called the Transformer, which was proposed by Vaswani et al. [134] for machine translation tasks and has since become a popular choice for various NLP tasks due to its ability to model long-range dependencies and capture global context.

In the USE architecture, the main component is the encoder, which is a multilayer Transformer network that processes the input sentence and generates a sequence of hidden states, one for each token in the sentence. The pooling layer then aggregates the hidden states into a fixed-length vector, which serves as the sentence embedding. The Transformer-based encoder is illustrated in Figure 15.

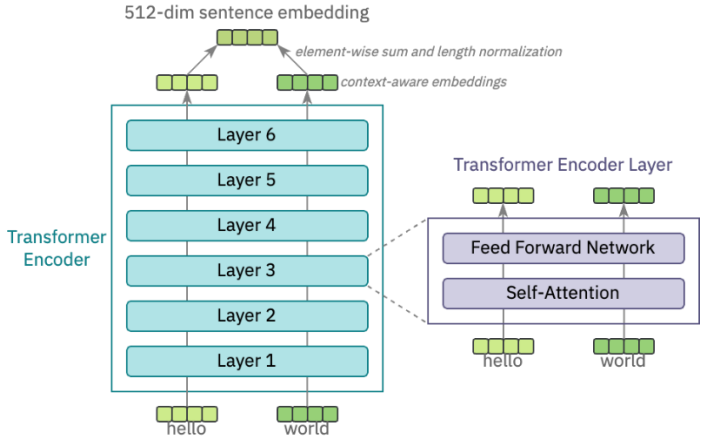


Figure 15. Architecture of the Transformer-based encoder.

The USE can generate text embeddings that can be used for various natural language understanding tasks. To achieve this, the model is trained using a multitask learning framework, where it learns to perform several tasks simultaneously. This approach helps the model learn a more general understanding of the input text, resulting in embeddings that can be used for multiple tasks. The multitask training process is illustrated in Figure 16.

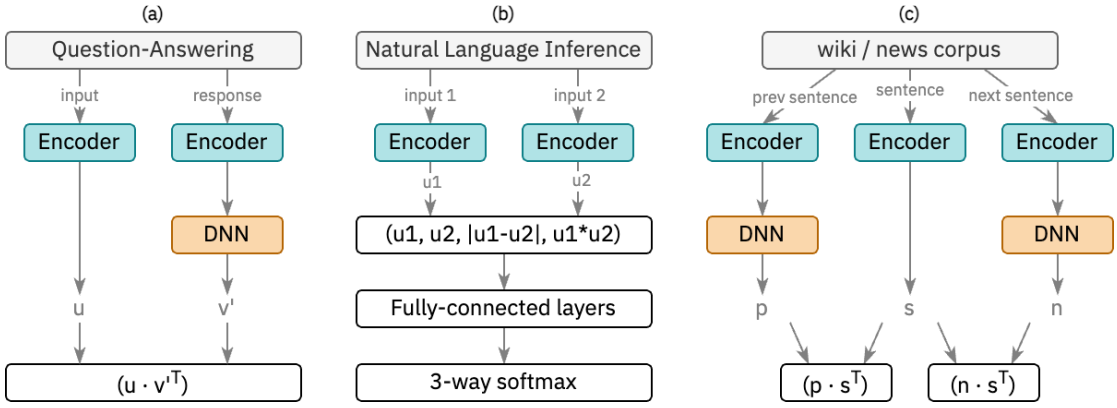


Figure 16. Multitask training in the universal sentence encoder [132]; the tasks and task structures share the same encoder layers and parameters.

In the question-answering task (Figure 16a), the model is trained to predict the correct answer to a given question according to the context provided. This task helps the model learn to understand the semantic relationship between questions and answers; this can be useful for other tasks that involve understanding the meaning of the text.

The Stanford Natural Language Inference dataset [135] comprises sentence pairs, where each pair is labeled as either entailment, contradiction, or neutral (Figure 16b). The model is trained to predict the correct label for each sentence pair; this helps the model learn to recognize semantic relationships between sentences. This task enhances the model's ability to understand the meaning of the text at a sentence level; this can be useful for tasks that involve comparing sentences.

Furthermore, the model is trained on large-scale unsupervised data from sources such as Wikipedia and news articles (Figure 16c). This training helps the model learn general language understanding and construct a broader knowledge base. Pretraining on such data exposes the model to diverse topics, writing styles, and vocabulary, thereby enabling it to generate embeddings that can be used for a wide range of tasks.

The combination of the aforementioned tasks helps the USE learn a rich understanding of text semantics, thereby allowing it to generate embeddings that capture the meaning of the input text. The final pretraining of the USE benefits from this multitask learning approach, resulting in embeddings that can be used for tasks such as text classification, sentiment analysis, and semantic textual similarity. This is also the reason why a pretrained USE model was utilized as the dynamic sentence embedding model in this study.

4.2.5 Rule Compliance Checking

Rule compliance checking is performed by measuring the semantic similarity of the embedding vectors of captions and rules. As discussed in section 3.3, word embedding is a technique that represents words in a continuous, dense vector space, where each dimension represents some semantic or syntactic feature of the word. These vectors, which in the context of this study describe the region captions and the safety rule corpus, constitute a vector space that enables the evaluation of the semantic similarity between these two textual pieces of information. This process is illustrated in Figure 17 intuitively.

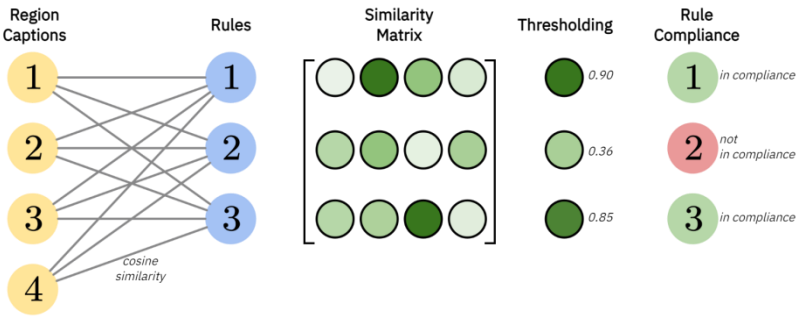


Figure 17. Graphical illustration of rule compliance checking based on semantic similarity match.

Semantic similarity is a measure of how closely related two words, phrases, or sentences are in meaning. The most common approach used for calculating semantic similarity is to calculate the cosine similarity of the word embeddings of the words or phrases in query and target texts. Given the dynamic embedding vectors of both image captions set $\mathcal{C} = (\boldsymbol{d}_1^{\mathcal{C}}, \boldsymbol{d}_2^{\mathcal{C}}, \ldots, \boldsymbol{d}_m^{\mathcal{C}})$ and rule sets $\mathcal{R} = (\boldsymbol{d}_1^{\mathcal{R}}, \boldsymbol{d}_2^{\mathcal{R}}, \ldots, \boldsymbol{d}_n^{\mathcal{R}})$, a similarity matrix:

$$\boldsymbol{S}_{m\times n} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{bmatrix} = [s_{ij}] \tag{4-10}$$

can be obtained by calculating the cosine similarity:

$$s_{ij} = \text{cosin\_sim}(\boldsymbol{d}_i^{\mathcal{C}}, \boldsymbol{d}_j^{\mathcal{R}})_{i=1,\ldots,m;j=1,\ldots,n} = \frac{\boldsymbol{d}_i^{\mathcal{C}} \cdot \boldsymbol{d}_j^{\mathcal{R}}}{\|\boldsymbol{d}_i^{\mathcal{C}}\|\|\boldsymbol{d}_j^{\mathcal{R}}\|} \tag{4-11}$$

Rule compliance checking can be regarded as evaluating the semantic similarity between image captions $\mathcal{C}$ and related safety rules $\mathcal{R}$. The workflow of checking the compliance between region captions and related safety rules is displayed in Figure 17. First, the semantic similarity s between each of the region captions and safety rules is calculated and a similarity matrix $\boldsymbol{S}_{m\times n}$ is obtained:

$$\boldsymbol{S} = \text{cosin\_sim}(\boldsymbol{d}^{\mathcal{C}}, \boldsymbol{d}^{\mathcal{R}}). \tag{4-12}$$

Next, the semantic similarity vector $\boldsymbol{r}$ is calculated by applying maximum pooling on the similarity matrix:

$$\begin{gathered} \boldsymbol{r} = [r_j], \\ r_j = \max(s_{ij})_{i=1,2,\ldots,m}, s_{ij} \in \boldsymbol{S} \end{gathered} \tag{4-13}$$

After obtaining the similarity vector, the safety rules that must be complied with can be selected according to a given threshold $\alpha \in [0\ ,\ 1]$. If the similarity score is larger than α, then it can be selected as a safety rule that must be complied to:

$$\mathcal{R}' = \{\mathcal{R}_j\}, \text{if } r_{j,j=1,2,\ldots,n} \geqslant \alpha. \tag{4-14}$$

4.3 Experiments and Implementation

4.3.1 Data Preparation

This study prepared a dataset containing 2000 site images shared by Zhang et al. [81]. This study labeled dense image caption labels on these images. Labeling guidelines and examples are provided in Figure 18. The target image subregions were categorized into four main classes: human, PPE, tool, and construction materials. For each region caption, this method developed a language template to guide the caption labeling process, which required subject, operation, and object described in the caption; attributes (e.g., color and number) and complement (e.g., environment information) were also encouraged to be described.

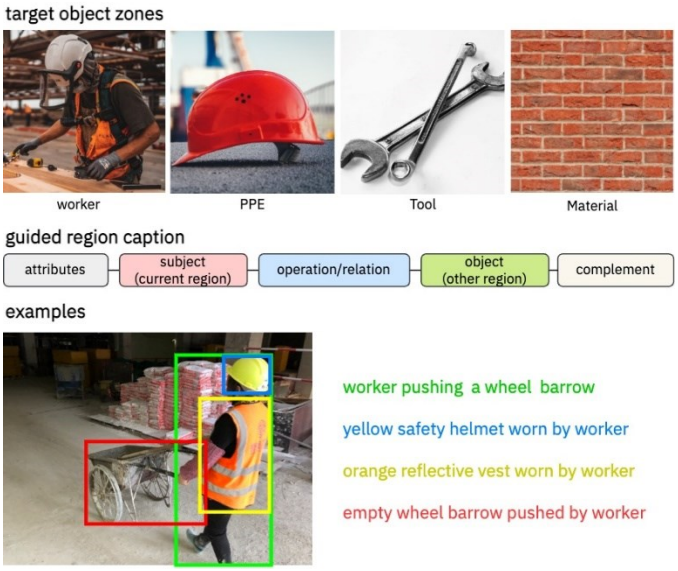


Figure 18. Dense captioning labeling guidelines and examples.

In addition, this study developed an online annotation task schema to assist the two-step labeling process, as shown in Figure 19. In the first step, the labelers were asked to draw bounding boxes for a set of target object categories. In the second step, after each bounding box was annotated on the image,

a query popped up asking the labeler to provide a text description for the drawn bounding box, providing semantic information for the subregion.

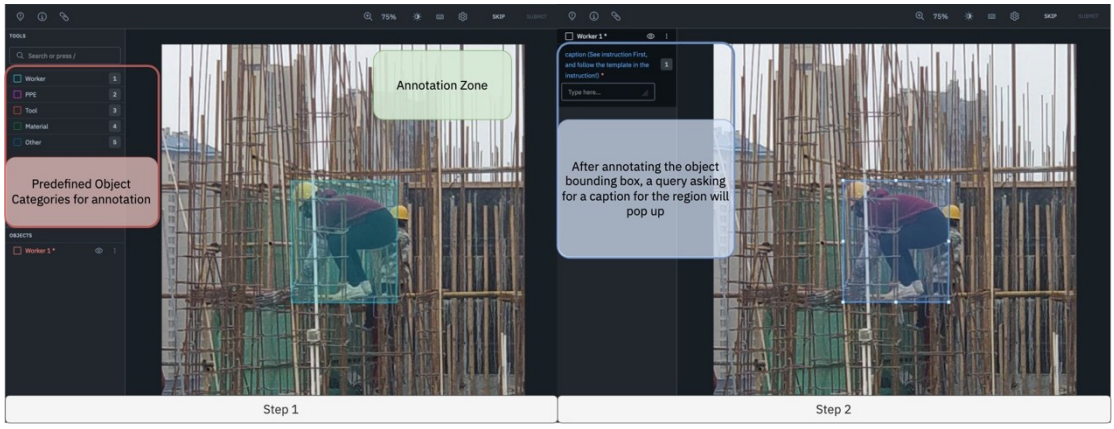


Figure 19. The annotation platform and schema.

As a result, more than 10,000 image subregions and captions were labeled. As illustrated in the "example" section of Figure 18, each image in the dataset has an $n \times 4$ bounding box annotation matrix indicating the subregion coordinates, where n is the number of subregions. Each image also contains a list of strings, which contains n captions indicating the captions of each of the subregions.

This study implemented the proposed method in a Python environment by using the workflow shown in Figure 20. In addition to the prepared and annotated image dense captions, this study prepared several safety rule sets as presented in Table 5. In preparing these safety rules, we selected rules from OSHA codes related to the construction tasks in the dataset. The author also interpretate the rules in simple sentences in according to the requirement of current model. More complex rules need additional processing that will be discussed in the discussion part.

After labeling the caption dataset for the construction images, the dataset was randomly split into the training set and test set in the 8:2 ratio. The dense captioning model was then trained and fine-tuned. This study constructed this model based on the proposed method by using Python libraries such as PyTorch [136] and TorchVision [137]. This method used a pretrained ResNet-50 network [138] as the image feature extracting network since ResNet is a commonly utilized feature extractor. The captions and related rules sets were then fed into the NLP part for sentence embedding and semantic similarity evaluation. For the NLP, this study developed the module by using Python libraries such as HuggingFace Transformer [139] and spaCy [140]. Furthermore, this study used Word2Vec as the static word embedding method and the USE as the dynamic word embedding model.

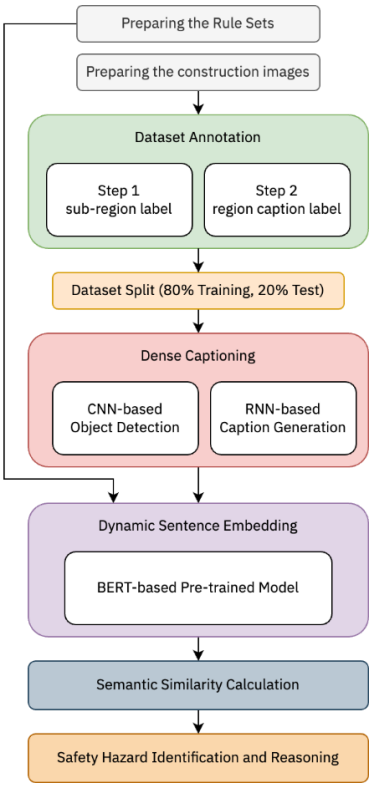


Figure 20. The implementation workflow.

Table 5. The safety rules used in this study.

Ruleset	Activity Type	Regulations
1	Height working	1. Worker should wear hardhat.
		2. Worker should wear falling prevention device.
2	concrete	1. Worker should wear foot protection boots.
		2. Worker should wear gloves.
		3. Worker should wear safety hats.
3	bricking	1. Worker should wear safety helmet.
		2. Worker should wear gloves.

4.3.2 Evaluation Metrics

Two objectives must be achieved when generating regional captions by using the proposed module: (1) generating well-localized target regions (in object detection tasks), and (2) generating accurate descriptions (in image captioning tasks).

Mean average precision (mAP) is a common metric used to evaluate the performance of object detection models. It is calculated as the mean of the average precision (AP) for different threshold levels. The threshold is defined as intersection of union (IoU), which quantifies the overlap between the ground-truth (GT) detection bounding box and the predicted (PD) box:

$$\text{IoU} = \frac{\text{area}(GT \cap PD)}{\text{area}(GT \cup PD)} \tag{4-15}$$

An IoU threshold $\alpha \in [0,1]$ is used to distinguish the true positive (TP), false positive (FP), and false negative (FN) detections. For example, if $\alpha = 0.5$, then a detected bounding box should have an IoU larger than 0.5 with the ground truth to be considered as a correct detection (i.e., TP). Based on the IoU threshold, for each image in the validation dataset, the model's predicted bounding boxes are compared with the ground truth annotations to create a set of true positive (TP), false positive (FP), and false negative (FN) detections, and the precision and recall are calculated accordingly:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \end{aligned} \tag{4-16}$$

The AP is the mean of the precision at different recall levels:

$$AP_\alpha = \frac{1}{11} \sum_{r \in \{0.0, 0.1, \ldots, 1.0\}} \text{Precision}_\alpha(r) \tag{4-17}$$

mAP is the average of AP at different IoU thresholds:

$$mAP = \frac{1}{5} \sum_{\alpha \in \{.3, .4, .5, .6, .7\}} AP_\alpha \tag{4-18}$$

Metric for Evaluation of Translation with Explicit Ordering (METEOR) [141] is a metric used to evaluate the quality of image captioning models. It is a variant of the BLEU [118] metric, which is commonly used in machine translation. METEOR is based on n-gram overlap between the predicted and reference captions but also considers synonyms and paraphrases by using a stemming algorithm and WordNet, a large lexical database of English. METEOR is a more sophisticated and robust method than BLEU for evaluating image captioning models as it considers synonyms, paraphrases, and word order. This study also used a METEOR threshold $\varepsilon \in \{0, .05, .1, .15, .2, .25\}$ to distinguish TP, FP, and FN. In addition, this study used another mAP score mAP_{cap} to measure the AP across all pairwise settings of IoU thresholds α and METEOR thresholds ε.

When conducting manual checking of the result of activity classification, hazard identification, and hazard reasoning, accuracy metric is used:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predections} = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{4-19}$$

4.4 Results and Discussion

4.4.1 Model Predictions and Evaluation Results

Examples of the dense captioning output obtained using the model are shown in Figure 21; only the target subregions with top 10 confidence are shown, and the regional captions are sorted according to the confidence rate. The evaluation results for mAP are presented in Table 6. The inference speed of the model was around 15 images/s on a single RTX 3080 Laptop GPU. The proposed model exhibited an mAP of 51.3% when evaluating only the target-region localization. An et al. [115] reported a benchmark mAP of 50.64% for a Faster R-CNN model (ResNet50) on the MOCS dataset. This means that the object detection performance of the model is feasible. When the METEOR threshold was considered, the proposed model yielded an mAP_{cap} of 41.2%.

The accuracies of activity classification, hazard identification, and rule reasoning were mannually evaluated. Because the dataset has no label about hazard identification and reasoning, and the these processes in the proposed method is unsupervised, the author manually checked the output and compared them to human judgement using common Accuracy metric. As can be seen from the results presented in Table 7, the accuracy scores of the activity classification were 75%, 84%, and 93% for bricking, height working, and spreading concrete, whereas for hazard identification, the accuracy scores were 99%, 84%, and 64%, respectively. The rule reasoning accuracy was evaluated to determine whether the model correctly predicted the safety rule that had been violated. As can be seen from the analysis results, the corresponding score was almost identical to hazard identification accuracy. In addition, some examples of safety hazard identification and rule reasoning are provided in Table 8.

Figure 21. Examples of the dense captioning outputs of the site images.

Table 6. Evaluation scores for the mean average precision of the dense captioning model.

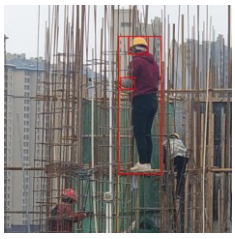
Model	mAP (%)	mAP_{cap} (%)
This Study	51.3	41.2
An et al. [115]	50.6	

Table 7. Activity classification, hazard identification, and reasoning accuracy.

Task Type	Activity Classification (%)	Hazard Identification Accuracy (%)	Rule Reasoning Accuracy (%)
Bricking	75	99	99
Height working	84	84	84
Spreading concrete	93	64	63

Table 8. Examples of the image captioning and safety hazard identification.

Image	Captions	Unsafe?	Violated Rule
	["worker laying clay bricks on the wall," "gloves worn by the worker," "brick held by the worker"]	TRUE	["worker should wear a safety helmet"]

	[“worker laying clay bricks on the wall,” “yellow safety helmet worn by the worker,” “gloves worn by the worker,” “brick held by the worker”]	FALSE	[]
	[“worker spreading the concrete mix,” “yellow safety helmet worn by the worker,” “rubber boots worn by the worker,” “rubber boots worn by the worker,” “concrete spreader held by the worker,” “gloves worn by the worker,” “rubber boots worn by the worker”]	FALSE	[]
	[“worker spreading the concrete mix,” “wheelbarrow pushed by the worker,” “gloves worn by the worker”]	TRUE	[“worker should wear a safety helmet,” “worker should wear safety boots”]
	[“worker creating a scaffold at a height,” “yellow safety helmet worn by the worker,” “gloves worn by the worker”]	TRUE	[“worker should wear fall-prevention device”]
	[“worker creating a scaffold at a height,” “yellow safety helmet worn by the worker,” “fall-prevention device worn by the worker”]	FALSE	[]

In summary, the integration of image captioning and semantic similarity for the automatic identification of safety hazards in construction images has immense potential; however, there are some limitations. Failure cases and major error causes are presented in the subsequent subsection. Further investigation and refinement are required to enhance the proposed method’s accuracy and efficiency. Nevertheless, the evaluation results demonstrate that the proposed method can feasibly integrate image captioning and semantic similarity techniques for identifying safety hazards. The

system successfully identified a considerable proportion of the hazards in the images, highlighting the potential of this method to enhance safety on construction sites.

4.4.2 Failure Cases

This study identified three major causes for the errors: (1) false image captioning, which occurs when the system generates a caption that does not accurately describe the contents of the image; (2) missing image captioning, where the system fails to provide a caption for an object; and (3) over-similar word embedding, which occurs when the system cannot distinguish between words in the construction domain. Examples of failure cases along with the associated errors that caused these failures are presented in Table 9.

Table 9. Examples of the failure cases.

Cause	Image	Captions	Unsafe?	Violated Rule
False image captioning (False caption indicated by underline)		['falling prevention device worn by worker', 'worker forming a scaffold on the height', 'yellow safety helmet worn by worker']	False	N/A
Missing captioning (Missing caption for the rubber boots)		['glove worn by worker', 'worker spreading concrete mix', 'yellow safety helmet worn by worker', 'glove worn by worker', 'concrete mix vibrator held by worker']	TRUE	['worker should wear safety boots']

The first cause of error identified in the study is false image captioning, which occurs when the system generates a caption that does not accurately describe the contents of the image. This error can have different consequences depending on which part of the image the false caption is describing. When the caption refers to the main worker behavior, it can lead to the model matching the wrong safety ruleset on it. This type of mistake contributes the most to the activity classification error. In

contrast, when a false caption is describing objects or other regions in the image, it can cause the semantic similarity matching to yield wrong results. This error may lead to the identification of irrelevant safety hazards or missing actual ones, undermining the accuracy of the safety hazard identification and safety rule reasoning.

The second cause of error identified in the study is missing image captioning, which occurs when the system fails to provide a caption for an important object (e.g., PPE) or an important subregion. This is problematic when the worker is following safety guidance, but the system identifies the worker's behavior as a safety hazard due to the lack of image captioning. This type of mistake is more significant when describing images related to the activity of "spreading concrete" where it mainly contributes to the sudden drop in the accuracy of hazard identification. In the case of the error presented in Table 9, the image captioning module overlooked the safety boots worn by the worker, and the hazard identification module falsely predicted a safety hazard due to the lack of foot protection.

The two aforementioned causes of error identified in this study, that is, false image captioning and missing image captioning, are limitations of the image captioning module. The occurrence of false image captioning and missing image captioning imply that the image captioning module may not have been trained on a sufficiently diverse set of construction site images containing numerous potential safety hazards. These limitations can be addressed by improving the efficiency of the image captioning module through the optimization of the model architecture or by increasing the size and diversity of the training dataset. Incorporating more data instances into the training dataset will improve the model's ability to accurately describe safety hazards in various contexts.

The third cause of error identified in this study is over-similar word embedding, which occurs when the system cannot distinguish between semantically similar words in the construction domain. This problem was observed several times in this study when generating word embeddings of PPE words. Because many PPE items share similar semantic features and functions, the model's word embedding module may have assigned similar word vectors to these words, leading to confusion during semantic similarity matching. To address this issue, this study suggests fine-tuning the similarity threshold to eliminate this type of error. However, a more effective solution would be to train the word embedding model on a corpus specific to the construction domain as it would enable the model to learn to distinguish between semantically similar words in the construction context, such

as different types of PPE, thereby resulting in more accurate semantic similarity matching and a reduction in the occurrence of over-similar word embedding errors.

4.4.3 Feasibility Discussion

Image captioning is a highly researched and developed field of CV and NLP. Image captioning can be employed for generating descriptions of construction sites, buildings, and other structures from construction images. The application of DL algorithms and pretrained models, such as CNNs and RNNs, can aid in the high-precision analysis and comprehension of the visual information present in construction images. This data can then be used to create descriptive captions that provide useful information regarding the structures depicted in the images. The viability of using image captioning for this purpose depends on several factors, such as the quality and resolution of the images, the diversity of the structures and scenes within the dataset, and the availability of vast quantities of annotated training data. Nevertheless, with the advancement of CV and NLP technologies, the potential for using image captioning to generate descriptions for construction images is substantial and warrants further investigation.

Semantic similarity is the method of measuring the degree of relatedness between two pieces of text, such as image captions and safety regulations. This method can be used to compare image captions generated by image captioning algorithms with safety regulations in the construction industry. The feasibility of using semantic similarity for this purpose depends on several factors, such as the quality of the image captions, the specificity and coverage of the safety regulations, and the performance of the semantic similarity algorithms.

If the image captions adequately represent the visual information contained in the images and the safety requirements are extensive and well-defined, semantic similarity can be an effective method for detecting potential safety violations at construction sites. By matching image captions with safety requirements, instances in which the represented structures or activities violate safety regulations can be feasibly detected. This information can then be used to take corrective action and improve the overall safety of the construction site.

4.4.4 Methodological and Practical Contributions

The proposed method can integrate visual and textual semantic information for safety hazard identification and provides the following advantages:

1) The proposed method utilizes dense captioning to extract regional descriptions from site images. Compared with existing visual recognition models, the proposed method provides outputs with higher label density and semantic richness; in addition, it provides richer semantic information regarding objects, actions, and interaction with localization for safety management.

2) The proposed method can integrate construction image data and text data by using visual–text semantic similarity. First, the image data are transformed into text-based regional captions, which contain semantic information. Next, word embeddings of the region captions are used to understand and process the text data. The proposed workflow minimizes the gap between image data and text data in the development of vision-based safety management programs.

3) Compared with existing safety hazard identification methods that use visual and text data, the proposed method decouples visual recognition from safety hazard identification, which means that the developer can optimize the visual module or language module separately without changing the overall architecture. In addition, the language module of the proposed method is an unsupervised method, which means that the language module retains its robustness when the safety rule changes, or new safety rules are added.

In addition, the proposed method offers the following practical advantages:

1) The proposed method can automate traditional construction safety management, which requires observations to be collected by dedicated personnel and is thus expensive and less efficient. The proposed method employs CV technologies to help observe construction sites through an automated process, thereby improving the efficiency and lowering the cost.

2) The proposed method can be used to improve existing vision-based safety management programs by automating hazard identification. The integration of automatic semantic information extraction with dense captioning and visual–text semantic similarity techniques enables the proposed method to make inferences regarding complex safety hazards, enabling managers to make safety-related decisions efficiently.

3) Though not presented in this study, the proposed method can be employed for other applications in construction management. For example, the regional descriptions

generated by the dense captioning module can be utilized to generate daily reports on site automatically. The descriptions can also be utilized as metadata for site image archives; this would aid in querying the image database in an information retrieval system. Moreover, the visual–text semantic similarity technique can be used to check if manual observation reports are in agreement with image monitoring records.

4.4.5 Limitations and Recommendations

The limitations of the proposed method and areas for improvement are described as follows:

1) In this study, safety rulesets comprising simple and clear sentences were used, which are not necessarily the same as the original safety regulations. Other NLP techniques must be explored to parse the safety regulations into usable rulesets. This type of technology relates to the sentence simplification task (split-and-rephrase) in NLP. Researchers in the construction industry are suggested to adopt the existing tools in NLP (e.g., [142–145]) to help in this task.

2) In real safety management practices, there are complex rules that require make decision on multiple parameters/rules. In this scenario, combining more advanced rule matching (such as decision tree, or other multi-criteria decision-making methods) with cosine similarity to determine the rule compliance. It is also feasible to utilizing large language models (LLM) like GPT-4 for better reasoning capability on complex rule matching.

3) The word embedding models used in this study, namely the Word2Vec-based static model and the BERT-based dynamic model, were developed for general purposes, and may ignore important semantic meanings and relationships for some words in the construction safety context and may not provide the best performance for construction related applications. Word embedding models trained on construction-related corpus may improve the NLP performance in construction scenarios.

4.5 Conclusion

This section proposed a novel approach for the automatic identification of safety hazards in construction images by integrating image captioning and semantic similarity techniques. The evaluation results demonstrated that the proposed method is feasible and has potential for further

investigation. The proposed method can accurately identify safety hazards in construction images by using natural language descriptions and semantic similarity measures. The results indicated that the integration of image captioning and semantic similarity has promise for improving safety in construction environments; however, the accuracy and efficiency need to be improved.

The contributions of this study are as follows: (1) It proposed a method that utilizes dense captioning to extract rich semantic information from site images and visual–text semantic similarity to integrate image data and text data. (2) It can be used to automate the observation process of traditional construction safety management programs. (3) It is advantageous for vision-based safety management as it enables automatic safety hazard identification. (4) It can be used for automatic report generation and information retrieval from image databases.

In future studies, the author will extend the dataset to include more on-site activities and interactions to further increase the information that can be extracted from the site images. In addition, the author will develop more NLP-based technologies to make automatic safety hazard identification on more complex safety regulations.

Chapter 5
Content-based Image Retrieval for Construction Image Management

5.1 Introduction

An information retrieval system (IRS) is critical in construction management because it facilitates the effective organization and management of information related to construction projects. The complexity and large volume of technical documents require the use of an IRS to organize and categorize information, making it easier to access and retrieve relevant information when needed [84]. Additionally, an IRS can reduce the effort required to gather information, potentially assisting construction managers in making informed decisions based on the most up-to-date information available [83].

Recently, cameras have become standard equipment for monitoring construction projects and improving management [18,39,66,96]. However, the increasing volume of digital images and videos captured on-site has created a challenge for construction management. For instance, a typical construction project during the construction phase may capture more than 400,000 images [10]. However, such construction visual data are manually sorted in most cases, content-analyzed, and then preserved [146]. The existing IRS in the construction industry focuses on retrieving text data in a text corpus. Image data, unlike text data, cannot be directly queried and retrieved using text keywords without. Therefore, developing new visual data management and retrieval methods for the construction industry with the ability to perceive, associate, and analyze image records within a specific site, zone, and even a precise angle of view with geographical and temporal boundaries is essential [147–150].

One such method is content-based image retrieval (CBIR), which retrieves images based on their content or characteristics. This method uses algorithms to extract attributes from images, such as color, texture, and shape, to search for, and retrieve, similar images. CBIR has several advantages, including eliminating the requirement for manual annotations or metadata, which allows for retrieving images

without keyword input. Additionally, it can be applied to managing large image collections and suits the expanding needs of image repositories. CBIR is also robust to changes in lighting, orientation, and other factors, making it possible to retrieve similar images even if they are not identical [151–153]. CBIR is intuitive and efficient; it facilitates relevance-based image retrieval and helps users quickly find the most relevant images based on the features they are interested in.

Deep learning-based computer vision technologies have demonstrated their superiority in extracting visual features from images, with their ability being comparable to that of humans [38]. Incorporating computer vision in CBIR can greatly enhance the capability of image retrieval systems, particularly in construction scenarios. Deep learning-based computer vision algorithms can be trained to recognize and extract specific visual features, making CBIR a powerful tool for image retrieval applications.

However, current deep learning-based CBIR methods may not be as efficient in construction scenarios as in general scenarios, as they tend to gather images based on the visual feature of the whole image or large-scale image regions. This approach may not be effective for construction images, which often present complex scenes with many similar objects, such as workers, equipment, and materials, making the extracted visual features indistinguishable. This characteristic of construction images presents unique challenges for CBIR. Therefore, there is a pressing need for a more detailed granularity method to extract nuanced visual information for construction image retrieval for improving construction management, tracking progress, and monitoring safety management activity, among others.

This section proposes a feature aggregation process based on object detection for retrieving construction images via the CBIR method. The method is specifically designed for construction images and consists of three main processes. First, a deep learning-based feature extractor extracts visual features from the construction image using a convolutional neural network (CNN). Second, a feature aggregator based on object detection aggregates regional visual features for both the object of interest regions and background object regions. Finally, an indexing process enhances the feature representation using mathematical methods, enabling effective retrieval of construction images based on detailed visual features. The proposed retrieval method can effectively retrieve construction images, aiding the management and retrieval of construction visual data, as well as other applications, such as

localization, behavior recognition, and tracking workers and equipment across different cameras and sites.

5.2 Methodology

As shown in Figure 22, the proposed image retrieval method comprises two primary steps: feature extraction aggregation and indexing. In the first step, the CNN obtains the convolution feature map of the entire image when a new construction image is added to the image database. Additionally, the feature aggregation module aggregates the feature vectors of specific regions on the feature map, and the feature representation of the image content is stored in the feature set. In the second step, the feature vectors undergo normalization processing during indexing to simplify the feature vectors, and the image similarity is calculated by measuring the distance between feature vectors. Thereafter, the images are retrieved based on their similarity.

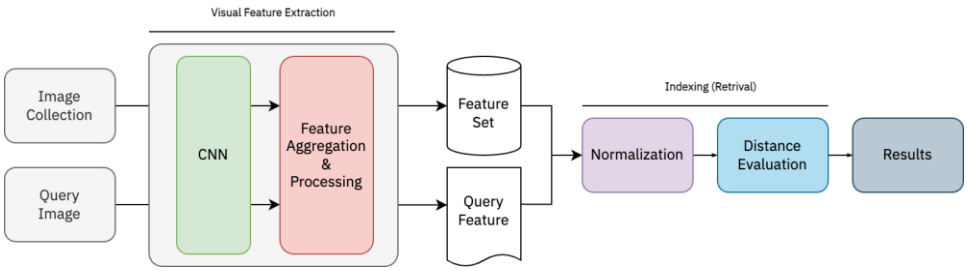


Figure 22. Overall framework of the proposed content-based image retrieval method.

The proposed feature extractor includes a feature aggregator based on object detection, which sets it apart from typical solutions. This feature aggregator conducts refinement work on the feature map, obtaining more detailed visual features about the subregions of the construction images. Figure 23 illustrates the detailed computational workflow. In summary, the proposed feature extractor employs CNNs to obtain a feature map of the input image. After acquiring the feature map, the feature aggregator “crops” the feature map of the entire image based on object zones to obtain feature maps for each object zone. This object-level feature map allows for refined control of visual features in construction images. The process details of the proposed method are elaborated in the following subsections.

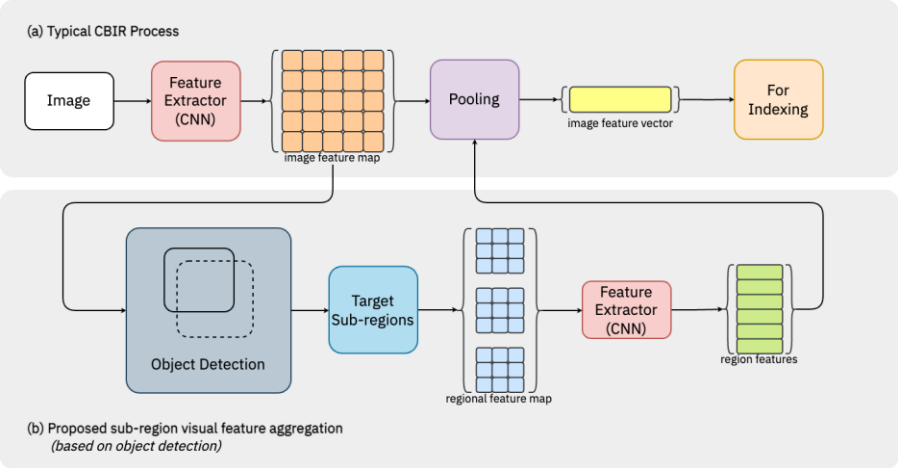


Figure 23. Computational workflow of the typical method and the proposed method.

5.2.1 Feature Extraction

The first module of the proposed engine is the CNN backbone, which conducts convolution calculations on the image to obtain the feature map. CNNs are specialized neural networks that process grid-like data and have been highly effective in processing image data [38]. In essence, CNNs transform the image data into increasingly smaller feature maps that represent the visual characteristics of the image. The deeper the layer in the CNN, the higher the level of representation of the image feature in the output of the layer. For instance, when using a CNN to process excavator images, the shallow layers of the CNN recognize lines and edges in the image, whereas deeper layers recognize shapes, such as the excavator bucket.

Several recent studies have proposed numerous CNN architectures with good performance in extracting image features, including VGG [28], Inception [29], MobileNet [154] and ResNet [155]. Each architecture has its own strengths and optimized applications. For this study, ResNet was chosen as the image feature extractor because it maintains a low training error rate while enabling very deep neural networks. It also enables skip connections, which facilitate easier training and reuse of the features learned in previous layers on deep CNNs. ResNet is chosen because it is the standard feature extraction network for the object detection module used in this study, which will be discussed in the following sections.

Mathematically, given a colored image, the CNN processes and transforms it into a three-dimensional feature map $\boldsymbol{V}$:

$$\boldsymbol{V} = \mathrm{CNN}(I), I \in \mathbb{R}^{3\times W\times H}, \boldsymbol{V} \in \mathbb{R}^{C\times W\prime\times H\prime} \tag{5-1}$$

where I denotes the input image having a width W, a height H, and three-color channels (red, green, and blue).

5.2.2 Baseline Feature Aggregator

Although the feature map output by the CNN backbone already represents the image content, a feature aggregator module is also required to extract significant image features, filter out noise on the feature map, and transform the feature map into a vector. Therefore, the feature aggregator processes the image feature map into meaningful feature vectors. This section introduces two commonly used aggregation methods that serve as baselines for comparison with the proposed method.

Generalized-mean pooling (baseline aggregator)

Considering that the output of the CNN backbone $\boldsymbol{V}$ denotes the feature map of the image, the GeM features are given by

$$\mathrm{GeM}(\boldsymbol{v}, p) = (\frac{1}{C}\sum_{i=0}^{C}\boldsymbol{v}_i^p)^{\frac{1}{p}} \tag{5-2}$$

$$\boldsymbol{v} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_C]^\top \tag{5-3}$$

The pooling parameter p can be manually set or learned. Because this study are using this method as a baseline, the p is set as three, following common practices [156].

Regional maximum activation of convolutions (baseline aggregator)

Regional maximum activation of convolutions (RMAC) [157] is a feature aggregator proposed for general CBIR methods. RMAC extracts regional visual features of an image and adds details to the output feature vector. It extracts regional features from a CNN by computing the maximum activation of each feature map within each region. The regions are defined by a set of fixed-size rectangular boxes that are densely sampled across the image. The maximum activations within each region are then pooled together using an integral operator to obtain a fixed-size vector representation of the image.

In R-MAC, the regions are defined by square boxes, and a scale parameter L is used to determine the size of the box. For an image feature map with a width of W' and height of H', when $L = 1$, the side length of the box is determined as

$$L_{box} = \min\,(W', H') \tag{5-4}$$

The boxes are sampled along the longer side of the image, ensuring that the intersection of consecutive boxes is as close as possible to 40%. For every other scale, the box's side length is determined as:

$$L_{box} = \frac{2 \times \min(W', H')}{L + 1} \tag{5-5}$$

Figure 24 illustrates the sample regions of R-MAC at different scales. The gray box represents the top-left region, and the dashed boxes represent the neighboring regions. The region features $\boldsymbol{f}_{R,L}$ are extracted by region-of-interest pooling (RoI Pooling) [158] to obtain a set of regional feature maps. The final feature vector is

$$\mathbf{F}_{\mathrm{RMAC}} = \sum_{R=1}^{N} \sum_{L=1}^{M} \boldsymbol{f}_{R,L} \tag{5-6}$$

where N is the total number of regions and M is the total number of levels L.

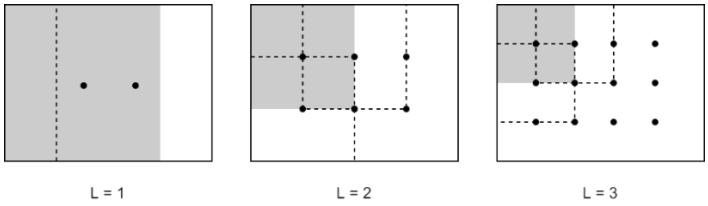


Figure 24. Sample regions extracted in the R-MAC aggregator at three scales (L = 1, 2, 3).

5.2.3 Proposed Feature Aggregator based on Object Detection

While RMAC is a powerful feature aggregation method, it has several shortcomings when applied in construction image retrieval scenarios. RMAC samples rectangular boxes across the entire image, causing some subregions to contain little or no meaningful visual content. This makes it sensitive to noise and irrelevant information in construction image retrieval. Moreover, the sampling subregions are not tightly related to the objects in the construction image, resulting in it lacking detailed object-level information and being difficult to interpret.

To address these issues, this study proposes a feature aggregation method that utilizes object detection to detect object subregions, thus improving the CBIR performance for construction images. The new method selectively chooses relevant subregions and provides interpretable information about the image, including the location and type of detected objects. Furthermore, the new method is less sensitive to noise and clutter in the image, making it effective in detecting the presence and absence of objects and filtering out irrelevant subregions.

This study utilizes a Faster R-CNN [41] based model as the regional visual feature extractor and aggregator to complete two tasks: (1) recognizing the objects in the image by performing object detection; and (2) extracting the object feature based on the recognized objects in step one. This method chose this architecture because, based on their prototype experiments, they found that Faster R-CNN provides a better performance boost for the whole model than other object detection architectures.

Figure 25 presents the detailed architecture of the proposed feature extractor and aggregator. The module uses the ResNet network as its backbone to extract the feature map from the input image. ResNet is an image classification network originally trained on the ImageNet dataset [114] and is widely utilized for other tasks. The authors also use the five convolution stages in the ResNet (Res1–Res5) as the feature extractor in this module. In panel A of Figure 25, "7x7, 64" denotes the filter size and depth of the convolution process. Res2 denotes the second stage, and so forth. "x3" denotes a stack of three consecutive convolution layers, and so forth. The backbone extracts the image features after the final convolutional layer of the 4-th stage, which is coded as C4 features.

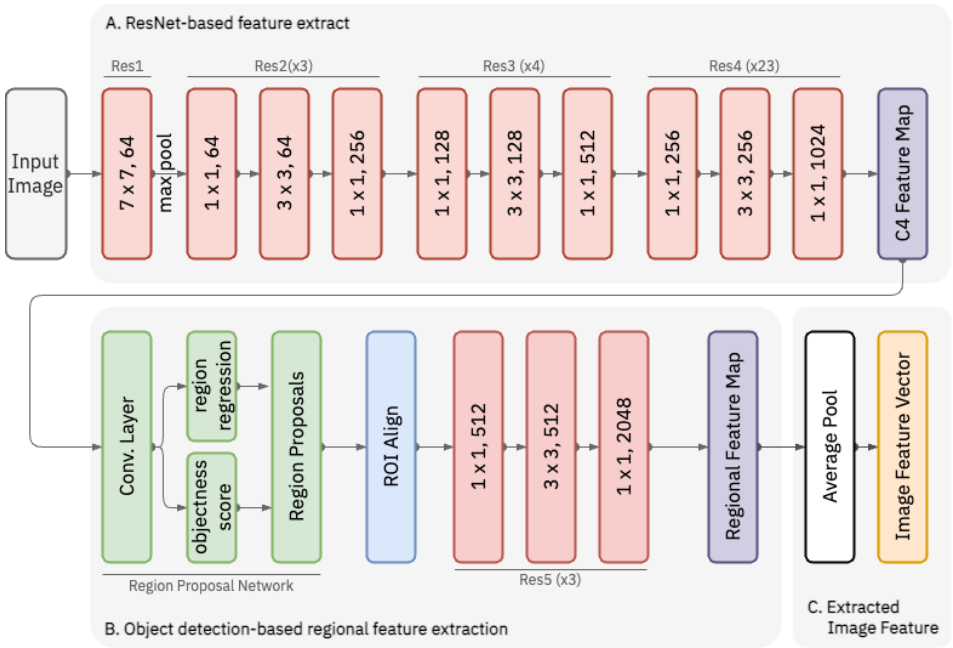


Figure 25. The architecture of the Faster R-CNN-based regional feature extraction and aggregation.

The detected regions come from the Regional Proposal Network (RPN), as shown in Figure 25b. Figure 26 illustrates the basic architecture of an RPN. Its main function is finding image regions that

potentially contain objects of interest. To achieve this, the RPN utilizes a sliding window that iterates over the feature map outputted by the CNN backbone. The window slides over all the feature maps like a convolution filter. At each sliding position on the feature map, the sliding window generates several anchor boxes with different sizes and aspect ratios (typically 0.5, 1, or 2). The features remaining in the anchor box are extracted and processed by a simple convolution stage to obtain a feature vector of the anchor box. The network then calculates the object score of this anchor box to determine whether it contains an object of interest and resizes it to fit the object via box-coordinate regression. The region with an object score higher than a given threshold is selected as the object region to extract the visual object feature:

$$\boldsymbol{B}, \boldsymbol{s} = \mathrm{RPN}(\boldsymbol{V}), \boldsymbol{B} \in \mathbb{R}^{n \times 4}, \boldsymbol{s} \in \mathbb{R}^{n} \tag{5-7}$$

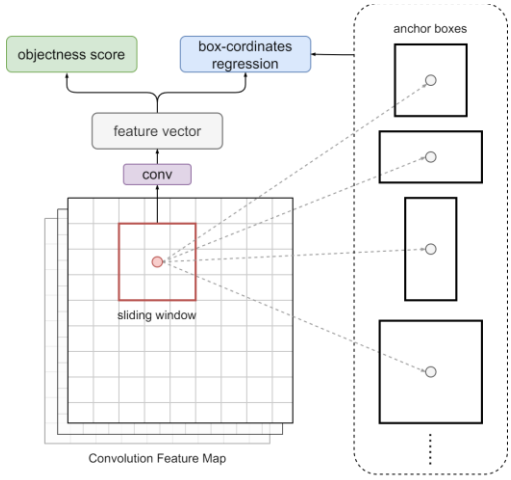


Figure 26. Illustration of the calculation process of regional proposal network.

The proposed object detection-based aggregator utilizes a different approach from RMAC for extracting regional features from the image feature map: specifically, it utilizes region of interest (ROI) Align [43] during the process. It works by dividing an ROI into a grid of uniformly sized rectangular cells and generating four regularly spaced sampling points inside each cell. The feature map values for each of these sampling points are obtained using bilinear interpolation to reduce quantization errors and improve alignment accuracy. The final feature value for each cell is computed as the maximum of the four interpolated values. ROI Align helps to extract accurate and fine-grained features from image sub-regions. Additionally, it can handle ROIs of varying sizes and aspect ratios:

$$\boldsymbol{v}_i = \mathrm{ROI\_Align}(\boldsymbol{V}, \boldsymbol{B}_i) \tag{5-8}$$

After obtaining the ROI features of each region generated by the RPN, the ROI features are further processed by the fifth convolution stage of the ResNet to obtain feature vectors. The feature maps of the regions are transformed into vectors:

$$\boldsymbol{f}_i = \mathrm{Res5}\ (\boldsymbol{v}_i) \tag{5-9}$$

The final vector representation of these regions is calculated as the mean pooling of these vectors:

$$\boldsymbol{F}_{object-based} = \mathrm{Mean}\ (\boldsymbol{f}_i) \tag{5-10}$$

Notably, the proposed object detection-based feature aggregator could be trained to yield various types of object regions. An RPN pretrained on a dataset that covers construction objects will propose the significant object regions for workers, heavy equipment, PPE, and tools, among others, on site images. On the other hand, an RPN pretrained on a general object dataset will propose the general object regions (often presented on the construction images as background objects), such as trees, soils, and roads, of the site images. By balancing those object regions, the proposed aggregator can generate image feature vectors that contain more detailed content information about the site images.

5.2.4 Indexing

After extracting visual feature vectors from construction images, an indexing method for effective retrieval of construction images is implemented. The indexing process involves normalization, distance measurement, and ranking. The process guarantees the retrieval and ranking of the most relevant construction images at the top, resulting in an efficient and effective content-based image retrieval system for construction management.

Normalization

Vector normalization is an important step in information retrieval systems because it helps to ensure that the similarity between two vectors is based on the direction of the vectors rather than their magnitude. Therefore, it can help in eliminating the influence of vector length on the similarity calculation [159], thereby improving the performance of the retrieval system by enhancing the discrimination of the feature vectors. In other words, it helps to make the feature vectors more comparable and distinguishable, which leads to more accurate retrieval results.

This study utilized L2-normalization on the image feature vectors. L2-normalization is a widely used technique in information retrieval systems that normalizes the feature vectors to have a unit length. The L2-norm, which is related to the Euclidean norm, is defined by:

$$\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} |\boldsymbol{x}_k|^2} \tag{5-11}$$

Distance Metric

Table 10 lists some of the most common distance metrics used to compare the similarity of real-valued vectors in machine learning-based applications. The *Euclidean Distance* is the straight-line distance between two points in Euclidean space. It is calculated by taking the square root of the sum of the squared differences between the corresponding elements of the two vectors. The *Manhattan Distance* is the distance between two points in a grid-like path. It is calculated by summing the absolute differences between corresponding elements of the two vectors. The *Minkowski Distance* is a generalization of both the Euclidean and Manhattan distances. It is calculated by taking the $p-th$ root of the sum of the absolute differences between corresponding elements of the two vectors, where p is a positive integer. In addition, the *Cosine Distance* is defined as the distance obtained by considering the angle between two vectors. It is calculated by subtracting 1 from the dot product of the two vectors divided by the product of their magnitudes.

Table 10. Common distance metrics for comparing the similarity of real valued vectors.

Distance Metric	Calculation	
Euclidean Distance	$d(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\sum_{i=1}^{k} (\boldsymbol{u}_i - \boldsymbol{v}_i)^2}$	(5-12)
Manhattan Distance	$d(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{k} \lvert \boldsymbol{u}_i - \boldsymbol{v}_i \rvert$	(5-13)
Minkowski Distance	$d(\boldsymbol{u}, \boldsymbol{v}) = \left(\sum_{i=1}^{k} \lvert \boldsymbol{u}_i - \boldsymbol{v}_i \rvert^p \right)^{\frac{1}{p}}$	(5-14)
Cosine Distance	$d(\boldsymbol{u}, \boldsymbol{v}) = 1 - \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}$	(5-15)

Each distance metric has its own advantages and disadvantages, and the choice of distance metric may depend on the specific scenario. The present study tested these distance metrics and observed that the differences in evaluation results on the normalized feature vectors were subtle. Because the feature vectors were normalized to unit vectors, the rank ordering produced by both Euclidean Distance and Cosine Distance in the experiments conducted were identical. The experiments demonstrated that the Euclidean Distance performed the best. Notably, it is the most commonly used

metric for k-nearest neighbors (k-NN) algorithms, and this study used it as the distance metric in this study.

Ranking Algorithm

This study used the k-NN ranking algorithm to retrieve similar images from the image dataset based on a given query image. The k-NN algorithm is widely used in image retrieval tasks due to its simplicity and effectiveness [84,160]. Moreover, the k-NN algorithm is nonparametric and does not require model training, which makes it computationally less expensive than other machine learning-based ranking algorithms. Furthermore, our feature vector normalization reduces the dimensionality of the data, making the k-NN algorithm more efficient in retrieving the most similar images.

The k-NN algorithm operates by calculating the distance between the feature vector of the query image and the feature vectors of all images in the dataset. The algorithm then retrieves the k images with the smallest distance to the query image, which are ranked as the top k similar images.

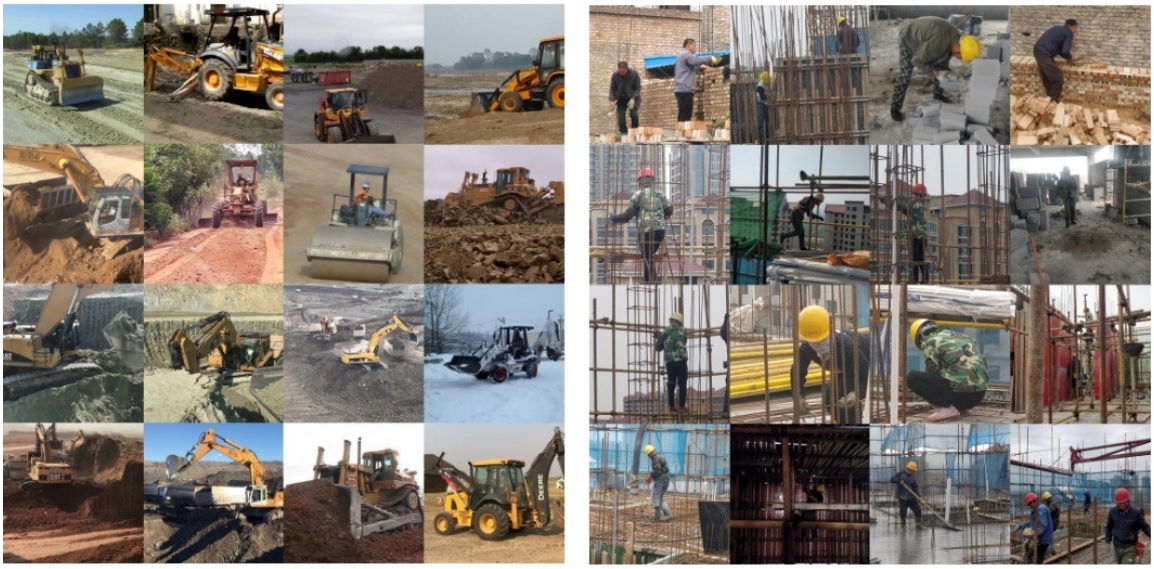
5.3 Implementations, Experiments, and Results

5.3.1 Image Collections and Retrieval Scenario Setup

This study utilized two construction image collections for the experiments, as listed in Table 11. The equipment collection comprised heavy equipment operations, with heavy equipment and materials being the primary objects in the images. The collection was sourced from the Alberta Construction Image Dataset (ACID) [116]. The worker collection, on the other hand, comprised worker activities, with workers, personal protective equipment (PPE), tools, and materials being the primary objects in the images. This collection was obtained from the work of Zhang et al. [161]. Figure 27 provides sample images from both collections.

Table 11. Details of the construction image collections.

Label	Target Scene	Main Objects in Scene
Equipment	heavy equipment operations	Heavy Equipment *(excavator, truck, loader, etc.)* Materials *(soil, rocks, bricks)*
Worker	worker activities	Worker PPE *(helmet, glove, harness, etc.)* Tool *(hammer, ladder, hoe, etc.)* Material *(concrete, brick, wood, rebar, etc.)*



(a) Heavy Equipment-Centric Image Collections (b) Worker-Centric Image Collections

Figure 27. Sample images from the construction image collections used in this study.

This study considers two scenarios in construction image retrieval:

1) The first scenario is retrieving images taken from the *same construction site or view angle*. This scenario allows users to quickly retrieve images that are relevant to a specific construction project or view. For example, a construction manager could retrieve images of a specific construction phase, such as foundation laying or beam installation, from different dates or time points to compare progress and identify any problems. This can improve project documentation and decision-making, ultimately improving project outcomes.

2) The second scenario is retrieving images that have the *same construction activity*. This scenario allows users to quickly retrieve images that are relevant to a specific construction activity, such as excavation or scaffolding. This can be helpful for productivity analysis and safety management. For example, a construction manager could retrieve images of excavation activities to monitor progress and ensure that the excavation is being done correctly and safely. This scenario can also aid in identifying potential hazards and improving overall safety in construction sites.

After categorizing the construction images based on the established scenarios, the images were assigned to different categories and labeled with a unique ID. Images taken from the same construction site or the same construction activity were assigned to the same category. Subsequently, the images were divided into two sets: the query and gallery set. The query set comprised images used as input for the image retrieval system, whereas the gallery set served as the image database to be retrieved.

5.3.2 Object Detection Model Developing and Training

To enable the proposed feature aggregator to detect important sub-regions of construction images, an object detection model was utilized in this study. The Faster R-CNN architecture was selected as it is a state-of-the-art object detection method based on deep CNNs [41]. The Detectron2 deep learning framework was used to build and train the detection models [162].

To train the object detection model, we used a subset of our construction image dataset that contains annotated bounding boxes for the relevant objects. The objects in the image were labeled by bounding boxes with the object category based on the scene presented in the construction image. The annotated dataset was then randomly split into training and validation sets in an 8:2 ratio, and data augmentation was performed during training, including randomly flipping and rotating the images, as well as performing brightness and contrast adjustments. The Adam optimizer was used during training [163], and the loss function used was a combination of classification and regression losses.

After training, the object detection model was fine-tuned on the construction image dataset used in the experiments by using the same hyperparameters as those in the original training. The best validation performance model was used, and the object detection model was employed to generate the relevant object proposals for the proposed object detection-based feature aggregator.

To verify that the object detection models developed in this study could accurately establish subregions from the construction images, the mean average precision (mAP) evaluation for object detection was conducted on each object detection model, which is consistent with previous object detection-related studies [115,116,164]. The evaluation results verified the feasibility of using these models to aggregate sub-regional image features in construction images. Table 12 provides the details and evaluation performance of the object detection models used in this study.

Table 12. Details and evaluations of the object detection models utilized in this study.

Label	Dataset	Target Object	Object Kind	Architecture	$AP_{0.5-0.95}$(%)	$AP_{0.5}$(%)
DET-Equipment	ACID	Heavy Machine	Foreground	Faster R-CNN	50.6	74.6
DET-Worker	Newly developed dataset	Worker, PPE, Tool, Material	Foreground	Faster R-CNN	61.6	87.0

In addition to the two object detection models trained in this study to detect construction-related objects, a pretrained Visual Genome model was utilized to detect general and background objects. It is a widely used object detection model pretrained on a large-scale visual dataset [165]. By incorporating this pretrained model, the proposed method could detect and balance foreground and background objects in the construction images.

5.3.3 Retrieval Model Development

The construction image retrieval system was built in a Python environment using the relevant package [166] to expedite the development process. The proposed object detection-based feature aggregator was developed and implemented, and the object detection module was connected to the retrieval system.

To compare the performance of the proposed method with existing CBIR methods, several benchmark models proposed in prior studies were implemented. The first benchmark model directly utilized the feature maps output by VGG without any detailed object feature, as proposed by Ha et al. [167]. The second benchmark model used the same CNN feature extractor (ResNet) as the proposed method, making the comparison fair. The third benchmark model adopted the RMAC feature

aggregator [157]. Instead of using the visual features of the whole image, R-MAC extracted several regions on the image based on a fixed grid of regions, regardless of where the object was located in the image.

The proposed models included two different methods based on the proposed feature aggregator, with one method using only foreground object region feature maps and the other using both foreground and background object region feature maps. The references for each model are provided in Table 13.

Table 13. The details about the benchmark models and the proposed models implemented in this study.

Label	CNN	Aggregation	Object Detector	Note
Benchmark 1	VGG	GeM	--	Uses the final feature map outputted by VGG feature extractor.
Benchmark 2	ResNet	GeM	--	Uses the final feature map outputted by ResNet feature extractor.
Benchmark 3	ResNet	RMAC	--	Uses a combination of subregion feature maps. The subregion features are aggregated based on a fixed grid of the image.
Proposed - Foreground	ResNet	DET	Equipment/Worker	The proposed method that only uses the feature maps of foreground object regions.
Proposed - Combined	ResNet	DET	Equipment/Worker & VG	The proposed method that uses the feature maps of both foreground and background object regions.

5.3.4 Retrieval Evaluation Metrics

To evaluate the performance of the proposed construction image retrieval method, mAP and recall at k metrics were utilized. These are common evaluation metrics for image retrieval methods and are based on precision and recall:

$$\text{Precision} = \frac{|\{relevant\ images\} \cap \{retrieved\ images\}|}{|\{retrieved\ images\}|} \tag{5-16}$$

$$\text{Recall} = \frac{|\{relevant\ images\} \cap \{retrieved\ images\}|}{|\{relevant\ images\}|} \tag{5-17}$$

The Average Precision measures the performance of a set of retrieval results based on the Precision (p) and Recall (r) metrics. If a submitted result has N rows sorted by its confidence score, then the Average Precision is computed using the following formula:

$$\text{AP} = \sum_{k=1}^{N} p(k)\Delta r(k) \tag{5-18}$$

The mAP is the mean of all the AP scores for all queries:

$$\text{mAP} = \frac{1}{|Q_R|} \sum_{q \in Q_R} \text{AP}(q) \tag{5-19}$$

This study also utilized $Recall@k$ for *metric learning* to evaluate the image retrieval performance. $Recall@k$ in this context defined as the percentage of queries with at least one neighbor retrieved in the first k results [168]. This type of evaluation has commonly been used in recent image retrieval competitions [169–171] and related studies [172–174].

In the context of this work, the proposed image retrieval method retrieves top-k items as the results. If at least one correct retrieval result is obtained, this process is labeled positive and will be given a score of one. If no correct result is obtained, the score is zero. The recall at k metric is the mean of scores among N times of retrieval attempts:

$$score = \begin{cases} 1, & result\ contains\ at\ least\ one\ correct\ instance \\ 0, & result\ contains\ none\ correct\ instance \end{cases} \tag{5-20}$$

$$Recall_k = \frac{1}{N} \sum_{i=1}^{N} (score) \tag{5-21}$$

This study reports the mAP and the recall at 1, 2, or 4 for metric learning to verify the performance and feasibility of the proposed method.

5.4 Results and Discussion

5.4.1 Experimental Results

Table 14. Evaluation results of the models for same-site retrieval.

Model	mAP(%)	$Recall_1$(%)	$Recall_2$(%)	$Recall_4$(%)
Benchmark 1	52.4	71.2	76.2	83.8
Benchmark 2	62.1	81.2	85.0	91.2
Benchmark 3	69.2	90.0	96.2	98.8
Proposed - Foreground	59.4	80.0	88.8	95.0
Proposed - Combined	86.4	97.5	98.8	100

Table 14 presents the evaluation results of the models on the same site retrieval scenario, including the mAP and recall scores for each model at different levels. Among the benchmark models, Benchmark 1 achieved the lowest mAP of 52.4%, whereas Benchmark 3 achieved an mAP of 69.2%. Both proposed models, Proposed–Foreground and Proposed–Combined, achieved higher mAP scores of 59.4% and 86.4%, respectively, than the benchmark models. The Proposed–Combined model achieved the highest recall scores of 97.5%, 98.8%, and 100% for recall at 1, 2, and 4, respectively, indicating that it can accurately retrieve images from the same construction site. This model also outperformed the other models in terms of mAP, demonstrating the effectiveness of the proposed object detection-based feature aggregator for construction image retrieval.

Table 15. Evaluation results of the models on same-activity retrieval.

Model	mAP(%)	$Recall_1$(%)	$Recall_2$(%)	$Recall_4$(%)
Benchmark 1	40.6	50	60	75
Benchmark 2	38.6	50	60	80
Benchmark 3	38.2	60	65	80
Proposed - Foreground	39.7	70	85	90
Proposed - Combined	77.3	95	100	100

Table 15 presents the evaluation results of the models on the same-activity retrieval scenario. The baseline models achieved lower mAP scores than those presented in Table 5, with Benchmark 1 achieving the lowest mAP of 40.6%. This performance drop is unsurprising considering the same-

activity retrieval is stricter on the desired retrieval results. The proposed models, Proposed–Foreground and Proposed–Combined, achieved mAP scores of 39.7% and 77.3%, respectively. The Proposed–Combined model achieved the highest recall scores of 95%, 100%, and 100% for recall at 1, 2, and 4, respectively, indicating that it can accurately retrieve images with the same-construction activity.

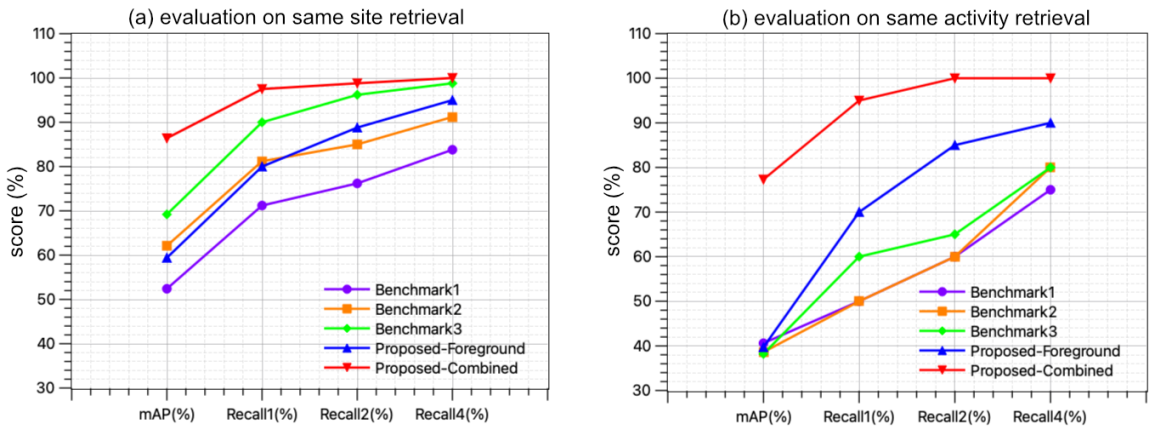


Figure 28. Curve graph of the evaluation results.

This study also illustrated the evaluation results graphically in Figure 28. As presented in the evaluation results, the proposed method performs better in both the same-site and same-activity retrieval scenario. Among the several model variants, the model architecture that combines the ResNet backbone with foreground and background object feature aggregators had the best evaluation score, indicating that combining and balancing the foreground object and background features would significantly improve retrieval precision.

5.4.2 Results Visualization

Scenario 1 – Same Site Retrieval

Figure 29 presents a visualization of the top five retrieval results for an image query on the same site retrieval scenario. The first row of the figure represents the ground truth, i.e., images taken from the same construction site that the models are expected to retrieve. The query image is enclosed in a purple bounding box, and the retrieval results are presented in each subsequent row. The correct retrieval results are highlighted with a green bounding box, whereas the false retrieval results are indicated with a red bounding box.

Figure 29. Visualized example of site image retrieval on the equipment image collection.

Scenario 2 – Same Activity Retrieval

Figure 30 illustrates the results of the same activity retrieval on the worker image collection using two query examples. In the first query, the input image shows a worker laying concrete bricks on the ground. The benchmark models did not perform well in this query because images containing workers were received but without considering the activity or other important construction objects such as bricks. In contrast, the proposed models could correctly retrieve the target construction images. In the second query, the input image shows a worker tying netting to the scaffold. This query is more challenging because many images contain scaffolds, but the workers are engaged in different tasks. Moreover, the benchmark models did not perform well in this query. The proposed method using only the foreground object features correctly retrieved three instances, with two being incorrect. The proposed method using both foreground and background object features correctly retrieved four instances, with one being false.

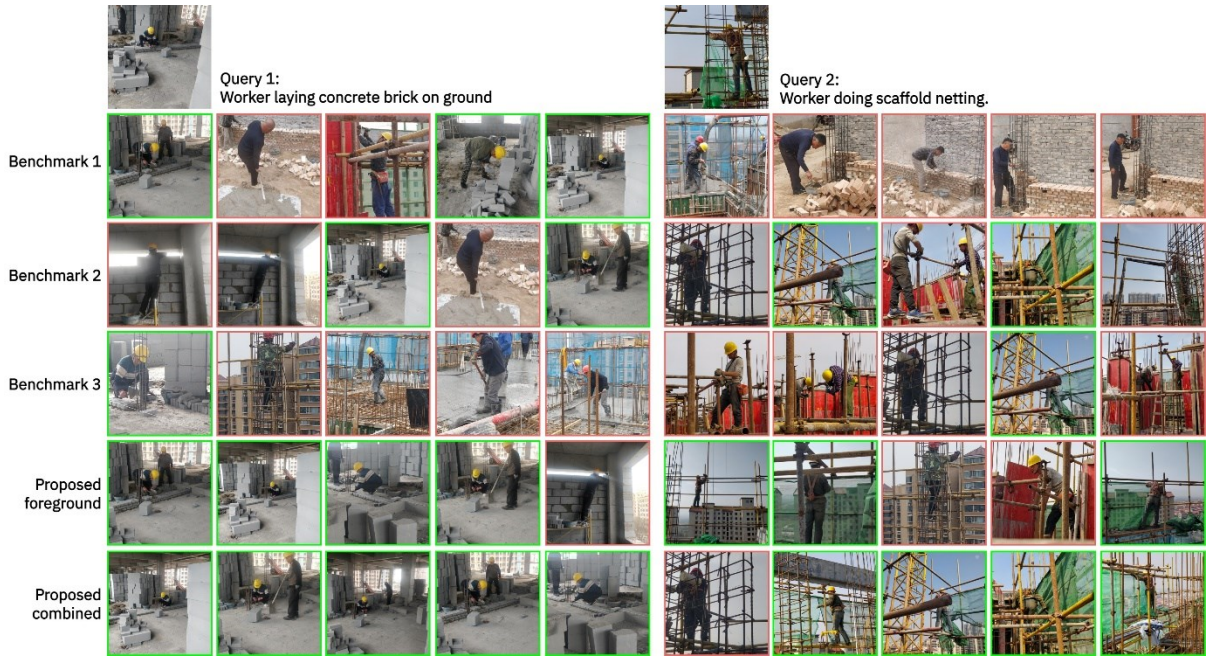


Figure 30. Visualized example of site activity retrieval on the worker image collection.

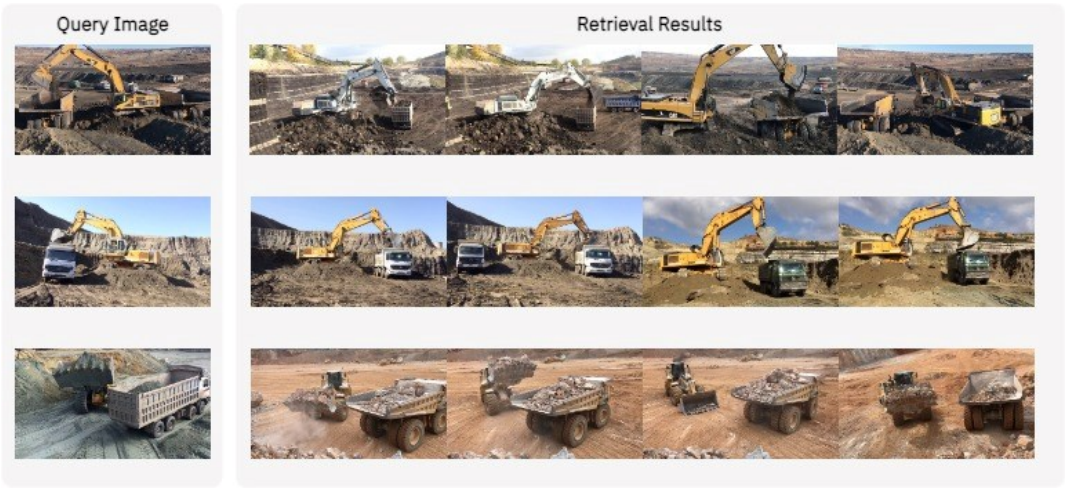


Figure 31. Visualized example of site activity retrieval on the equipment image collection (using the proposed-foreground model).

The proposed method for construction image retrieval can also be applied to retrieving same activity images for equipment images. Figure 31 presents examples of the Proposed–Foreground model for same-activity retrieval for the equipment image collection. The model could retrieve same activity images across different construction site images, demonstrating its effectiveness in retrieving construction images based on activity.

5.4.3 Content-based Method Compared to Label-based Method

Label-based methods for image retrieval often have several limitations, and using metadata or automatic tagging methods may not accurately capture the actual image content. For instance, an image taken on a construction site and tagged with location and time may not provide any information about what objects are present in the image or what is happening in the image. Therefore, retrieving specific images that contain certain objects or events can be challenging. In contrast, CBIR allows users to search for images based on their actual visual content. The system can retrieve images that contain specific objects or events, even if they are not tagged with specific metadata. This makes it easier to find relevant images, saving construction managers time and effort.

Moreover, metadata or automatic tagging methods may not be able to capture all the relevant information in an image. An image may contain several different objects, and automatic tagging may only capture information about one or two of those objects. In contrast, CBIR can capture information about all the objects in an image, making it easier to retrieve images that contain specific object combinations.

Overall, while metadata or automatic tagging based on location/time may provide basic information about the construction site, they may not capture the detailed visual content of construction images, which can be critical for construction management. CBIR methods can efficiently capture the visual content of images, making it easier for managers to retrieve and analyze the images based on specific visual features. However, there may be certain scenarios where label-based methods are preferred, such as when specific objects or actions are irrelevant in the retrieval. In such cases, label-based methods may be preferable. Therefore, both CBIR and label-based methods have their advantages and limitations, and the choice between them should be based on the specific needs and scenarios of construction management.

5.4.4 Method Efficiency and Granularity

Previous CBIR methods, such as RMAC, may have limitations in construction-related image retrieval due to the entire image or rectangular boxes being sampled to define the visual features, resulting in some subregions with little or no meaningful image content. It is sensitive to noise and clutter in the construction image because regions that contain noisy or irrelevant information can contribute to the feature representation. Moreover, previous methods do not provide object-level

information about the image, including the presence or absence of specific objects or object categories. This can limit its usefulness for tasks such as object recognition and localization.

To address these shortcomings, this study proposes an improved feature aggregation method that selectively chooses sub-regions containing relevant objects or regions of interest through object detection. This method can provide targeted and interpretable information about the image, such as the location and type of detected objects, improving CBIR performance for construction images. Additionally, this method is less sensitive to noise and clutter as it can filter out irrelevant subregions.

However, this proposed method may require additional training and computational resources, which can be achieved by upgrading the computation hardware. The benefits of increased interpretability and improved performance make it a worthwhile tradeoff for construction image retrieval scenarios.

5.4.5 Methodological and Practical Contributions

This study makes several methodological contributions. First, it proposes a content-based image retrieval method based on visual features extracted via CNN, a deep learning-based computer vision technology. By extracting higher-level visual features about objects, rather than only low-level visual features about shapes and patterns, this study demonstrates the feasibility and performance of computer vision techniques in constructing CBIR. This could provide better retrieval performance in construction image management. Second, this method introduces a new feature aggregator based on object detection that extracts detailed object visual features. This module is carefully designed to aggregate and balance the visual features of both common foreground objects and background objects in the construction image. This allows the module to represent more detailed and richer visual features of the construction images, resulting in accurate retrieval of target images, even when construction images have similar visual properties.

This study also has practical implications. First, the construction image retrieval system can assist construction managers to gather images from the same construction site or view of angle, providing a structured visual monitoring repository that represents the changes in the site over time. This can help construction managers to identify and track the progress of working zones over time. Second, the construction image retrieval system can help construction managers in querying same-activity images from a large monitoring image collection, providing visual reference for productivity analysis and

safety management. For example, if an unsafe behavior is captured in the monitoring image, the construction manager could retrieve similar behavior images, analyze the behavior patterns, and develop a solution for this unsafe behavior. This could ultimately improve safety in the construction industry.

5.5 Conclusion

This study presents an improved content-based construction image retrieval method using object detection to extract detailed visual features from construction images. The proposed method incorporates a feature extraction approach that balances the visual features of foreground objects and background objects in construction images. The study prepared two construction image collections including heavy equipment images and worker activity images, and evaluates the mAP and recall at k on same-site retrieval and same-activity retrieval scenarios. The evaluation results report that it outperforms existing methods.

Methodologically, this study proposes a novel feature aggregator that extracts detailed object visual features and configurations that fit different construction image retrieval scenarios. Practically, the proposed method is useful in documenting and tracking progress and improving productivity and safety for construction management application. The authors plan to improve the proposed method by extending the dataset to include more objects and combining the foreground and background object detectors.

Chapter 6
Conclusions, Contributions, and Future Works

6.1 Conclusions

A construction site can be a hazardous place, suffering from fatal and non-fatal injuries. Safety managers have adopted BBS programs for decades to improve site safety. Recently, with the help of construction monitoring imagery containing important visual information about worker behaviors and interactions and CV technology that could automatically recognize target information on the image, the manual observation of the construction site has been automated to a certain level. However, the current practice of using CV technology in BBS programs is challenged in three notable respects:

(1) The recognized information from construction images is limited either in label density or semantic richness, so the current visual recognition method does not provide the information required for further data analysis and safety hazard identification.

(2) CV-only methods need help understanding safety regulations and guidelines. Safety hazard identification and reasoning still need manual work.

(3) Existing vision-based applications are inefficient because, although a large quantity of monitored image data is acquired and stored, significant time is consumed in processing and retrieving related information during similar case analysis and behavior pattern identification.

This research aims to automate the processes involves manual observation and inspection in traditional BBS and construction safety management. In addressing the gaps above, three objectives are included in this research:

(1) To enrich the information could be extracted from construction images, supporting safety hazard identification.

(2) To automate the safety hazard identification on site, and enable reasoning about the hazard identification according to safety regulations.

(3) To automate the image records management and retrieval for safety analysis.

This research proposes to integrate CV and NLP technologies into image description generation, safety hazard identification, and content-based image retrieval for a sophisticated vision-assisted BBS program. The main idea here is to convert the image data into structured text captions, thereby extracting important semantic information about objects, activities, and interactions from the site monitoring images. In this way, the image scene information can be compared with the safety regulations using semantic similarity comparison, which is an NLP technique, since they are in the same data format. In addition, this research also proposes an image retrieval method dedicated to construction images, providing fast and accurate image query that helps the observation of target behavior over time and similar case analysis. To accomplish the goal and objctives of this research, the methods are summarized below:

(1) *Development of a semantic information extraction method for construction images*: Recently, vision-based monitoring has been widely adopted in construction management to improve crew productivity, reduce safety risks, and facilitate site planning. However, automated retrieval of semantic information (e.g., objects, activities, and interactions between objects) from construction images remains challenging due to the complex nature of construction sites. This research proposes a novel semantic information extraction method by integrating deep learning object detection and image captioning, which aims to explore more salient information from construction images or videos. In the proposed method, object detection has been employed as an encoder to extract the feature maps of construction object zones and the holistic image. Image captioning has been selected as the decoder to extract the semantic information. Furthermore, a post-processing method has been proposed to parse the semantic information into tabular and graph formats for better accessibility and visualization. In experiments, the proposed method has achieved a consensus-based image description evaluation (CIDEr) of 1.84. In addition, more salient and hidden information behind construction images can be presented to construction managers to assist their decision-making.

(2) *Development of an automatic safety hazard identification and reasoning method*: The construction industry has a high rate of fatal and nonfatal injuries, which can be prevented by optimizing safety management and promoting safe behavior. One approach, BBS, has been

widely studied, but traditional manual methods are time-consuming and error-prone. To automate BBS, researchers have adopted CV technologies, which can recognize hazardous postures and actions and detect missing PPE. However, existing methods have significant limitations, such as only detecting simple repeating objects or activities. This research proposes a framework for a vision-based BBS program that bridges the gap between construction monitoring images and safety knowledge bases. The framework includes two modules: an image processing module that uses computer vision and dense image captioning technologies and a text processing module that uses natural language processing technologies to evaluate semantic similarities. The proposed framework was tested on a dense image captioning dataset and achieved a mean average precision of 50.64% and an average safety hazard identification accuracy of 82.3%. The results suggest that the proposed framework has the potential to automatically identify potential safety hazards on monitoring images and improve safety management in the construction industry.

(3) *Development of a content-based image retrieval method*: Visual data (i.e., images and videos) has become necessary documentation in construction management, potentially replacing traditional paper-based site documentation. However, retrieval of construction images containing specific contents of interest, such as images from the same angle view across different capturing devices, is still challenging due to the large volume of images accumulated in construction projects. This research proposes a content-based image retrieval method to accurately retrieve interested construction images by inputting a query image. The proposed method was validated in the experiment to retrieve target images from construction images of 60 sites. The proposed method achieved the best mean average precision of 86.4%. This technology contributes to decision-making applications in construction management by providing a quick information retrieval system.

6.2 Contributions

This research makes several notable contributions to the knowledge of vision-assisted behavior-based safety for construction. The academic and industrial contributions are outlined in this section.

6.2.1 Methodological contribution

(1) The proposed method for semantic information extraction from construction images utilizes regional visual features extracted by an object detector-based encoder as input for the decoder, which integrates object detection and image captioning to improve information extraction precision. Additionally, the method incorporates an attention mechanism in the decoder, which establishes a visual connection between the extracted information and the image region, enhancing the extracted semantic information with location information. This explicit visual connection improves visualization and enables more sophisticated management tasks, leading to more accurate and effective image captioning. Combining these two methods presents a promising approach to improving image captioning performance.

(2) This method proposes a novel approach based on dense captioning to extract regional descriptions from construction site images, offering outputs with higher label density and semantic richness than existing visual recognition models for construction images. The method enhances BBS programs by providing richer semantic information about objects, actions, and interactions with localization, contributing to a more accurate and effective analysis of construction site images. The method presents a promising approach for improving the quality of regional descriptions for construction site images, providing a valuable contribution to the field.

(3) The proposed method presents a novel workflow that integrates construction image data and text data through visual-text semantic similarity, offering a valuable contribution to developing the vision-assisted BBS program. The workflow transforms image data into text-based regional captions with semantic information, providing a more accurate representation of construction site images. Additionally, the method utilizes word embeddings of the region captions and other text in construction safety regulations to understand and process text data, allowing for a more comprehensive safety hazard analysis. By bridging the gap between image and text data, the proposed workflow enhances the capability of vision-assisted BBS, contributing to a more efficient safety hazard identification. The integration of image and text data through visual-text semantic similarity presents a promising approach for improving the analysis of construction site data.

(4) The study proposes a novel CBIR method for construction image management that is based on visual features extracted by CNN. The study verifies the feasibility and performance of using

deep learning-based CV techniques in constructing CBIR, with CNN being able to extract higher-level visual features about objects beyond just low-level features about shapes and patterns. The proposed method introduces a new feature aggregator that extracts detailed object visual features from the CNN feature map, carefully designed to balance the features of significant objects and environmental objects in construction images. The method presents detailed and richer visual features of the construction images, resulting in improved retrieval precision and recall compared to existing methods, even when images have similar visual properties. The proposed method provides a valuable contribution to the field of CBIR for construction image management, enhancing the accuracy and effectiveness of image retrieval through integrating deep learning-based computer vision technologies.

6.2.2 Practical Implication

(1) This method could improve the visualization and documentation of construction management. The visual connection ability enables displaying the related image zone for extracted information. This provides an intuitive visualization of the extracted semantic information. The extracted semantic information could serve as enriched metadata to simplify the construction image documentation process. This method could improve the current practice of vision-based BBS and management by providing richer semantic information and visual connections. Furthermore, the extracted semantic information could provide an integral information package for downstream safety management applications such as safety hazard identification.

(2) The proposed method presents a valuable contribution to automating traditional BBS programs, offering a more cost-efficient and productive solution to construction site observation. Traditional BBS programs rely on manual observation, which can be time-consuming and labor-intensive. By utilizing computer vision technologies, the proposed method automates the observation of construction sites, enhancing the efficiency and productivity of traditional BBS programs. Furthermore, the automation of BBS through computer vision technology provides a more accurate and comprehensive analysis of construction site data, contributing to improved safety management practices and decision-making processes. The proposed method presents a promising approach to the automation of traditional BBS programs, contributing to a safer working environment for

construction workers and increased efficiency in construction site management. Furthermore, the proposed method can reduce the risk of human error and bias in traditional BBS programs, making it a valuable addition to the field of construction safety management.

(3) This method improves existing vision-assisted BBS programs by providing automatic hazard identification ability. The proposed method could use the extracted semantic information to infer potential hazards automatically. Furthermore, the dense captioning technique and visual-text semantic similarity technique enables the proposed method to identify more complex safety hazards. By utilizing the extracted semantic information and the novel workflow for integrating image and text data, the proposed method can automatically infer potential safety hazards. The dense captioning technique and visual-text semantic similarity technique employed in this method enable it to identify more complex safety hazards, making it suitable for more complex decision-making processes and safety management tasks compared to other vision-assisted BBS programs. With its ability to automatically identify potential safety hazards, the proposed method can enhance the efficiency and accuracy of safety hazard identification in construction sites, leading to improved safety management practices and a safer working environment for construction workers.

(4) Developing a CBIR system for construction images is valuable to its management. With the ability to quickly query and retrieve related image records, the CBIR system enhances the efficiency and productivity of construction site management. By providing a more accurate and detailed representation of construction site images, the CBIR system can help identify and analyze similar cases and behaviors in the image repository, leading to more effective behavior pattern analysis for BBS programs. This analysis can contribute to a better understanding of safety hazards and risks on construction sites, leading to more effective safety management practices. The CBIR system provides a valuable tool for construction site managers to improve their decision-making processes and identify potential safety hazards. Overall, the contribution of a CBIR system for construction images can enhance the efficiency and effectiveness of construction site management and contribute to a safer working environment for construction workers.

6.3 Practice and Implementation Considerations

6.3.1 Potential Safety Applications

There are several applications that this proposed methods could fit in current safety management workflow:

- *Safety Training*: The model could be used to generate examples of potential hazards for safety training purposes. By providing real-world examples from actual construction sites, the training could be more effective and engaging for workers.
- *Daily Safety Briefings*: Before the start of each workday, the model could analyze images from the construction site to identify any new or emerging hazards. These could be discussed during the daily safety briefing to ensure all workers are aware of the hazards and know how to avoid them.
- *Incident Reporting*: In the event of a safety incident, the model could analyze images from the incident to help determine the cause. This could provide valuable information for the incident report and for any subsequent investigations.
- *Safety Audits*: During safety audits, the model could be used to analyze a large number of images from the construction site in a short period of time. This could make the audits more efficient and comprehensive.

6.3.2 Legal Considerations

The implementation of a computer vision and natural language processing system for construction site safety management, while promising in its potential to enhance safety measures, must be carefully navigated within the confines of the legal framework. Several key legal considerations must be taken into account:

- *Privacy Laws*: Respect for workers' privacy rights is paramount. Informed consent should be obtained before recording, and the system should be used in a manner that respects privacy.
- *Data Protection Laws*: Our system processes personal data, so it must comply with data protection laws. This involves implementing robust data security measures and respecting individuals' rights over their data.

- *Employment Laws*: Laws about monitoring employees at work vary by jurisdiction. Employers may need to inform employees about the monitoring and restrict how the data can be used.
- *Safety Regulations*: The system must comply with existing safety regulations on construction sites. It should align with these regulations and contribute to a safer working environment.

To facilitate the practical implementation of the proposed methods, it is always recommended to understand the specific legal requirements in the jurisdictions where the system will be used. Furthermore, developing comprehensive policies for obtaining consent, protecting data, and complying with other legal requirements will be crucial steps towards implementation. A pilot study may also be beneficial to demonstrate the effectiveness and legality of the system in a real-world context.

6.3.3 Potential Unintended Consequences and Mitigation Strategies

While the integration of computer vision and natural language processing offers significant potential for enhancing construction site safety, it's important to acknowledge and address potential unintended consequences. Here are some key considerations:

- *Privacy Concerns*: Continuous monitoring could raise privacy concerns among workers. Clear communication about data collection, usage, and privacy protection is crucial. Implementing strict data access controls and anonymizing data where possible can help mitigate these concerns.
- *Over-reliance on Technology*: The risk of over-reliance on the technology, potentially leading to complacency, is a significant concern. Emphasizing that the technology is a tool to assist with safety, not a replacement for human judgement and vigilance, can help address this issue.
- *Misinterpretation of Outputs*: The risk of misinterpreting the system's outputs, especially if they are complex or technical, is another potential issue. Providing clear, user-friendly outputs and adequate training can help ensure correct interpretation and usage.
- *Technological Errors*: Like any technology, there's a risk of errors or malfunctions, which could lead to missed hazards or false alarms. Regular maintenance, testing, and updates can help ensure the technology functions correctly and reliably.

- *Workplace Stress*: Continuous monitoring could potentially increase stress among workers if they feel they are constantly being watched. Communicating that the purpose of the technology is to improve safety, not to monitor individual performance, can help alleviate this concern.

Acknowledging these potential unintended consequences and proactively implementing mitigation strategies can ensure that the benefits of this technology are realized while minimizing potential drawbacks. This approach aligns with the principle of responsible innovation, ensuring that technological advancements serve to enhance, rather than compromise, the well-being of individuals and communities.

6.4 Future Works

Each section of the modules in this research identifies the research limits in order to enhance the performance and feasibility of the proposed method. The key drawbacks, which are summarized in this thesis, are as follows, along with some recommended future research directions:

1) While the proposed method for semantic information extraction in construction safety management has shown promising results, the scale of the dataset used for training still needs to be expanded. To further improve the accuracy and effectiveness of the models, more images and types of objects and activities should be collected and labeled. This can enhance the capability of the models to identify safety hazards in construction sites, contributing to more accurate safety management practices.

2) The semantic similarity matching module currently used in this research is limited to simple sentences. More advanced technologies should be explored to improve its functionality in handling compound safety regulation sentences to transform these sentences into simpler ones. Furthermore, the matching algorithm could be optimized to handle more complex safety rule combinations. This can enhance the accuracy and effectiveness of the safety hazard identification process, contributing to a safer working environment on construction sites.

3) While the pre-trained models utilized in this research have shown promising performance, there is still room for improvement. Training these models on construction-

related data can further enhance their performance in identifying safety hazards on construction sites. This can improve the accuracy and effectiveness of the proposed vision-assisted BBS technique, contributing to more effective safety management practices in construction projects.

The proposed vision-assisted BBS technique can improve safety management practices on construction sites by automating site observation and safety hazard identification. However, to fully apply this method to BBS programs and safety management, more on-site case studies and applications are needed to integrate the proposed method into construction projects. This can provide more comprehensive and accurate safety hazard identification and prevention, contributing to a safer working environment for construction workers.

References

[1] Statista, U.S. construction industry share of GDP 2007-2020, Statista. (n.d.). https://www.statista.com/statistics/192049/value-added-by-us-construction-as-a-percentage-of-gdp-since-2007/ (accessed March 18, 2022).

[2] CPWR, Fatal and Nonfatal Injuries in Construction, CPWR |. (n.d.). https://www.cpwr.com/research/data-center/data-dashboards/fatal-and-nonfatal-injuries-in-construction/ (accessed January 3, 2023).

[3] S. Zhang, J. Teizer, J.-K. Lee, C.M. Eastman, M. Venugopal, Building information modeling (BIM) and safety: Automatic safety checking of construction models and schedules, Automation in Construction. 29 (2013) 183–195.

[4] S. Zhang, K. Sulankivi, M. Kiviniemi, I. Romo, C.M. Eastman, J. Teizer, BIM-based fall hazard identification and prevention in construction safety planning, Safety Science. 72 (2015) 31–45.

[5] B.H.W. Guo, Y. Zou, Y. Fang, Y.M. Goh, P.X.W. Zou, Computer vision technologies for safety science and management in construction: A critical review and future research directions, Safety Science. 135 (2021) 105130. https://doi.org/10.1016/j.ssci.2020.105130.

[6] B.H. Guo, Y.M. Goh, K.L.X. Wong, A system dynamics view of a behavior-based safety program in the construction industry, Safety Science. 104 (2018) 202–215.

[7] P. Martinez, M. Al-Hussein, R. Ahmad, A scientometric analysis and critical review of computer vision applications for construction, Automation in Construction. 107 (2019) 102947. https://doi.org/10.1016/j.autcon.2019.102947.

[8] B. Sherafat, C.R. Ahn, R. Akhavian, A.H. Behzadan, M. Golparvar-Fard, H. Kim, Y.-C. Lee, A. Rashidi, E.R. Azar, Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review, Journal of Construction Engineering and Management. 146 (2020) 03120002. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001843.

[9] S. Xu, J. Wang, W. Shou, T. Ngo, A.-M. Sadick, X. Wang, Computer Vision Techniques in Construction: A Critical Review, Arch Computat Methods Eng. 28 (2021) 3383–3397. https://doi.org/10.1007/s11831-020-09504-3.

[10] K.K. Han, M. Golparvar-Fard, Potential of big visual data and building information modeling for construction performance analytics: An exploratory study, Automation in Construction. 73 (2017) 184–198. https://doi.org/10.1016/j.autcon.2016.11.004.

[11] B. Akinci, F. Boukamp, C. Gordon, D. Huber, C. Lyons, K. Park, A formalism for utilization of sensor systems and integrated project models for active construction quality control, Automation in Construction. 15 (2006) 124–138.

[12] J. Yang, M.-W. Park, P.A. Vela, M. Golparvar-Fard, Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future, Advanced Engineering Informatics. 29 (2015) 211–224. https://doi.org/10.1016/j.aei.2015.01.011.

[13] D. Rebolj, N.Č. Babič, A. Magdič, P. Podbreznik, M. Pšunder, Automated construction activity monitoring system, Advanced Engineering Informatics. 22 (2008) 493–503. https://doi.org/10.1016/j.aei.2008.06.002.

[14] S.J. Prince, Computer vision: models, learning, and inference, Cambridge University Press, 2012.

[15] S. Raaijmakers, Deep Learning for Natural Language Processing, 1st edition, Manning, Shelter Island, NY, 2022.

[16] F. Ismail, A.E. Hashim, W. Zuriea, W. Ismail, H. Kamarudin, Z.A. Baharom, Behaviour Based Approach for Quality and Safety Environment Improvement: Malaysian Experience in the Oil and Gas Industry, Procedia - Social and Behavioral Sciences. 35 (2012) 586–594. https://doi.org/10.1016/j.sbspro.2012.02.125.

[17] W. Fang, P.E.D. Love, H. Luo, L. Ding, Computer vision for behaviour-based safety in construction: A review and future directions, Advanced Engineering Informatics. 43 (2020) 100980. https://doi.org/10.1016/j.aei.2019.100980.

[18] S. Xu, J. Wang, W. Shou, T. Ngo, A.-M. Sadick, X. Wang, Computer Vision Techniques in Construction: A Critical Review, Arch Computat Methods Eng. 28 (2021) 3383–3397. https://doi.org/10.1007/s11831-020-09504-3.

[19] W. Chu, S. Han, X. Luo, Z. Zhu, Monocular vision–based framework for biomechanical analysis or ergonomic posture assessment in modular construction, Journal of Computing in Civil Engineering. 34 (2020) 04020018.

[20] J. Kim, S. Chi, Multi-camera vision-based productivity monitoring of earthmoving operations, Automation in Construction. 112 (2020) 103121.

[21] B. Xiao, S.-C. Kang, Vision-based method integrating deep learning detection for tracking multiple construction machines, Journal of Computing in Civil Engineering. 35 (2021) 04020071.

[22] X. Yan, H. Zhang, H. Li, Estimating Worker-Centric 3D Spatial Crowdedness for Construction Safety Management Using a Single 2D Camera, Journal of Computing in Civil Engineering. 33 (2019) 04019030. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000844.

[23] X. Zhou, F. Kuang, J. Huang, X. Liu, K. Gryllias, Water-Lubricated Stern Bearing Rubber Layer Construction and Material Parameters: Effects on Frictional Vibration Based on Computer Vision, Tribology Transactions. 64 (2021) 65–81. https://doi.org/10.1080/10402004.2020.1803463.

[24] E. Rezazadeh Azar, B. McCabe, Automated Visual Recognition of Dump Trucks in Construction Videos, Journal of Computing in Civil Engineering. 26 (2012) 769–781. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000179.

[25] X. Luo, H. Li, X. Yang, Y. Yu, D. Cao, Capturing and Understanding Workers' Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian Nonparametric Learning, Computer-Aided Civil and Infrastructure Engineering. 34 (2019) 333–351. https://doi.org/10.1111/mice.12419.

[26] X. Luo, H. Li, D. Cao, Y. Yu, X. Yang, T. Huang, Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks, Automation in Construction. 94 (2018) 360–370. https://doi.org/10.1016/j.autcon.2018.07.011.

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

[28] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ArXiv:1409.1556 [Cs]. (2014). http://arxiv.org/abs/1409.1556 (accessed March 4, 2022).

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[31] W. Fang, L. Ma, P.E.D. Love, H. Luo, L. Ding, A. Zhou, Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology, Automation in Construction. 119 (2020) 103310. https://doi.org/10.1016/j.autcon.2020.103310.

[32] J. Rasmussen, Risk Management, Adaptation, and Design for Safety, in: B. Brehmer, N.-E. Sahlin (Eds.), Future Risks and Risk Management, Springer Netherlands, Dordrecht, 1994: pp. 1–36. https://doi.org/10.1007/978-94-015-8388-6_1.

[33] G.A. Howell, G. Ballard, T.S. Abdelhamid, P. Mitropoulos, Working near the edge: a new approach to construction safety, Proceedings IGLC-10, Garamado, Brazil. (2002).

[34] P.E.D. Love, P. Teo, J. Morrison, Unearthing the nature and interplay of quality and safety in construction projects: An empirical study, Safety Science. 103 (2018) 270–279. https://doi.org/10.1016/j.ssci.2017.11.026.

[35] T.R. Krause, K.J. Seymour, K.C.M. Sloat, Long-term evaluation of a behavior-based method for improving safety performance: a meta-analysis of 73 interrupted time-series replications, Safety Science. 32 (1999) 1–18. https://doi.org/10.1016/S0925-7535(99)00007-7.

[36] E.S. Geller, Behavior-Based Safety and Occupational Risk Management, Behav Modif. 29 (2005) 539–561. https://doi.org/10.1177/0145445504273287.

[37] D. Oswald, F. Sherratt, S. Smith, Problems with safety observation reporting: A construction industry case study, Safety Science. 107 (2018) 35–45. https://doi.org/10.1016/j.ssci.2018.04.004.

[38] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, 2016.

[39] W. Fang, L. Ding, P.E.D. Love, H. Luo, H. Li, F. Peña-Mora, B. Zhong, C. Zhou, Computer vision applications in construction safety assurance, Automation in Construction. 110 (2020) 103013. https://doi.org/10.1016/j.autcon.2019.103013.

[40] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE. 86 (1998) 2278–2324.

[41] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence. 39 (2017) 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

[42] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.

[43] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017: pp. 2961–2969.

[44] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: pp. 3431–3440.

[45] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual Relationship Detection with Language Priors, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016: pp. 852–869. https://doi.org/10.1007/978-3-319-46448-0_51.

[46] L. Cheng, Z. Yang, GRCNN: Graph Recognition Convolutional Neural Network for Synthesizing Programs from Flow Charts, ArXiv:2011.05980 [Cs]. (2020). http://arxiv.org/abs/2011.05980 (accessed November 25, 2021).

[47] Y. Yang, C.L. Teo, H. Daumé, Y. Aloimonos, Corpus-guided sentence generation of natural images, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, United Kingdom, 2011: pp. 444–454.

[48] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, BabyTalk: Understanding and Generating Simple Image Descriptions, IEEE Transactions on Pattern

Analysis and Machine Intelligence. 35 (2013) 2891–2903. https://doi.org/10.1109/TPAMI.2012.162.

[49] S. Li, G. Kulkarni, T.L. Berg, A.C. Berg, Y. Choi, Composing simple image descriptions using web-scale n-grams, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Portland, Oregon, 2011: pp. 220–228.

[50] J. Mao, W. Xu, Y. Yang, J. Wang, A.L. Yuille, Explain Images with Multimodal Recurrent Neural Networks, ArXiv:1410.1090 [Cs]. (2014). http://arxiv.org/abs/1410.1090 (accessed April 19, 2020).

[51] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and Tell: A Neural Image Caption Generator, ArXiv:1411.4555 [Cs]. (2015). http://arxiv.org/abs/1411.4555 (accessed April 19, 2020).

[52] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 652–663. https://doi.org/10.1109/TPAMI.2016.2587640.

[53] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ArXiv:1502.03044 [Cs]. (2016). http://arxiv.org/abs/1502.03044 (accessed April 19, 2020).

[54] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning, ArXiv:1612.01887 [Cs]. (2017). http://arxiv.org/abs/1612.01887 (accessed April 19, 2020).

[55] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: pp. 6077–6086.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, ArXiv:1706.03762 [Cs]. (2017). http://arxiv.org/abs/1706.03762 (accessed July 15, 2020).

[57] H. Kim, C.R. Ahn, D. Engelhaupt, S. Lee, Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement, Automation in Construction. 87 (2018) 225–234. https://doi.org/10.1016/j.autcon.2017.12.014.

[58] R. Akhavian, A.H. Behzadan, Simulation-based evaluation of fuel consumption in heavy construction projects by monitoring equipment idle times, in: 2013 Winter Simulations Conference (WSC), 2013: pp. 3098–3108. https://doi.org/10.1109/WSC.2013.6721677.

[59] K.M. Rashid, J. Louis, Automated Activity Identification for Construction Equipment Using Motion Data From Articulated Members, Front. Built Environ. 5 (2020). https://doi.org/10.3389/fbuil.2019.00144.

[60] T. Slaton, C. Hernandez, R. Akhavian, Construction activity recognition with convolutional recurrent networks, Automation in Construction. 113 (2020) 103138. https://doi.org/10.1016/j.autcon.2020.103138.

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

[62] H. Tajeen, Z. Zhu, Image dataset development for measuring construction equipment recognition performance, Automation in Construction. 48 (2014) 1–10. https://doi.org/10.1016/j.autcon.2014.07.006.

[63] J.C.P. Cheng, M. Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Automation in Construction. 95 (2018) 155–171. https://doi.org/10.1016/j.autcon.2018.08.006.

[64] W. Fang, L. Ding, B. Zhong, P.E.D. Love, H. Luo, Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach, Advanced Engineering Informatics. 37 (2018) 139–149. https://doi.org/10.1016/j.aei.2018.05.003.

[65] W. Fang, L. Ding, B. Zhong, P.E.D. Love, H. Luo, Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach, Advanced Engineering Informatics. 37 (2018) 139–149. https://doi.org/10.1016/j.aei.2018.05.003.

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

[67] H. Kim, S. Bang, H. Jeong, Y. Ham, H. Kim, Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation, Automation in Construction. 92 (2018) 188–198. https://doi.org/10.1016/j.autcon.2018.04.002.

[68] J. Kim, S. Chi, J. Seo, Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks, Automation in Construction. 87 (2018) 297–308. https://doi.org/10.1016/j.autcon.2017.12.016.

[69] M. Golparvar-Fard, A. Heydarian, J.C. Niebles, Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers, Advanced Engineering Informatics. 27 (2013) 652–663. https://doi.org/10.1016/j.aei.2013.09.001.

[70] E. Rezazadeh Azar, S. Dickinson, B. McCabe, Server-Customer Interaction Tracker: Computer Vision–Based System to Estimate Dirt-Loading Cycles, Journal of Construction Engineering and Management. 139 (2013) 785–794. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000652.

[71] C. Chen, Z. Zhu, A. Hammad, W. Ahmed, Vision-Based Excavator Activity Recognition and Productivity Analysis in Construction, (2019) 241–248. https://doi.org/10.1061/9780784482438.031.

[72] H. Luo, C. Xiong, W. Fang, P.E.D. Love, B. Zhang, X. Ouyang, Convolutional neural networks: Computer vision-based workforce activity assessment in construction, Automation in Construction. 94 (2018) 282–289. https://doi.org/10.1016/j.autcon.2018.06.007.

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

[74] H. Luo, M. Wang, P.K.-Y. Wong, J.C.P. Cheng, Full body pose estimation of construction equipment using computer vision and deep learning techniques, Automation in Construction. 110 (2020) 103016. https://doi.org/10.1016/j.autcon.2019.103016.

[75] J. Cai, Y. Zhang, H. Cai, Two-step long short-term memory method for identifying construction activities through positional and attentional cues, Automation in Construction. 106 (2019) 102886. https://doi.org/10.1016/j.autcon.2019.102886.

[76] J. Cai, Y. Zhang, L. Yang, H. Cai, S. Li, A context-augmented deep learning approach for worker trajectory prediction on unstructured and dynamic construction sites, Advanced Engineering Informatics. 46 (2020) 101173. https://doi.org/10.1016/j.aei.2020.101173.

[77] H. Kim, K. Kim, H. Kim, Data-driven scene parsing method for recognizing construction site objects in the whole image, Automation in Construction. 71 (2016) 271–282. https://doi.org/10.1016/j.autcon.2016.08.018.

[78] Y. Ham, M. Kamari, Automated content-based filtering for enhanced vision-based documentation in construction toward exploiting big visual data from drones, Automation in Construction. 105 (2019) 102831. https://doi.org/10.1016/j.autcon.2019.102831.

[79] S. Tang, D. Roberts, M. Golparvar-Fard, Human-object interaction recognition for automatic construction site safety inspection, Automation in Construction. 120 (2020) 103356. https://doi.org/10.1016/j.autcon.2020.103356.

[80] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Image retrieval using scene graphs, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: pp. 3668–3678. https://doi.org/10.1109/CVPR.2015.7298990.

[81] L. Zhang, J. Wang, Y. Wang, H. Sun, X. Zhao, Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge, Automation in Construction. 142 (2022) 104535. https://doi.org/10.1016/j.autcon.2022.104535.

[82] J. Johnson, A. Karpathy, L. Fei-Fei, DenseCap: Fully Convolutional Localization Networks for Dense Captioning, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 4565–4574. https://doi.org/10.1109/CVPR.2016.494.

[83] Y. Ding, J. Ma, X. Luo, Applications of natural language processing in construction, Automation in Construction. 136 (2022) 104169. https://doi.org/10.1016/j.autcon.2022.104169.

[84] C. Wu, X. Li, Y. Guo, J. Wang, Z. Ren, M. Wang, Z. Yang, Natural language processing for smart construction: Current status and future directions, Automation in Construction. 134 (2022) 104059. https://doi.org/10.1016/j.autcon.2021.104059.

[85] A.J.-P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining, Automation in Construction. 74 (2017) 39–54. https://doi.org/10.1016/j.autcon.2016.11.001.

[86] F. Zhang, A hybrid structured deep neural network with Word2Vec for construction accident causes classification, International Journal of Construction Management. 22 (2022) 1120–1140. https://doi.org/10.1080/15623599.2019.1683692.

[87] W. Fang, H. Luo, S. Xu, P.E.D. Love, Z. Lu, C. Ye, Automated text classification of near-misses from safety reports: An improved deep learning approach, Advanced Engineering Informatics. 44 (2020) 101060. https://doi.org/10.1016/j.aei.2020.101060.

[88] C.L. Yeung, C.F. Cheung, W.M. Wang, E. Tsui, A knowledge extraction and representation system for narrative analysis in the construction industry, Expert Systems with Applications. 41 (2014) 5710–5722. https://doi.org/10.1016/j.eswa.2014.03.044.

[89] M. Martínez-Rojas, R. Martín Antolín, F. Salguero-Caparrós, J.C. Rubio-Romero, Management of construction Safety and Health Plans based on automated content analysis, Automation in Construction. 120 (2020) 103362. https://doi.org/10.1016/j.autcon.2020.103362.

[90] X. Wang, N. El-Gohary, Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements, Automation in Construction. 147 (2023) 104696. https://doi.org/10.1016/j.autcon.2022.104696.

[91] J.-S. Kim, B.-S. Kim, Analysis of Fire-Accident Factors Using Big-Data Analysis Method for Construction Areas, KSCE J Civ Eng. 22 (2018) 1535–1543. https://doi.org/10.1007/s12205-017-0767-7.

[92] N. Xu, L. Ma, Q. Liu, L. Wang, Y. Deng, An improved text mining approach to extract safety risk factors from construction accident reports, Safety Science. 138 (2021) 105216. https://doi.org/10.1016/j.ssci.2021.105216.

[93] X. Pan, B. Zhong, Y. Wang, L. Shen, Identification of accident-injury type and bodypart factors from construction accident reports: A graph-based deep learning framework, Advanced Engineering Informatics. 54 (2022) 101752. https://doi.org/10.1016/j.aei.2022.101752.

[94] B. Zhong, H. Li, H. Luo, J. Zhou, W. Fang, X. Xing, Ontology-Based Semantic Modeling of Knowledge in Construction: Classification and Identification of Hazards Implied in Images, Journal of Construction Engineering and Management. 146 (2020) 04020013. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001767.

[95] M. Zhang, R. Shi, Z. Yang, A critical review of vision-based occupational health and safety monitoring of construction site workers, Safety Science. 126 (2020) 104658. https://doi.org/10.1016/j.ssci.2020.104658.

[96] S. Paneru, I. Jeelani, Computer vision applications in construction: Current state, opportunities & challenges, Automation in Construction. 132 (2021) 103940. https://doi.org/10.1016/j.autcon.2021.103940.

[97] W. Fang, P.E.D. Love, L. Ding, S. Xu, T. Kong, H. Li, Computer Vision and Deep Learning to Manage Safety in Construction: Matching Images of Unsafe Behavior and Semantic Rules,

IEEE Transactions on Engineering Management. (2021) 1–13. https://doi.org/10.1109/TEM.2021.3093166.

[98] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Advanced Engineering Informatics. 29 (2015) 239–251. https://doi.org/10.1016/j.aei.2015.02.001.

[99] B. Zhong, H. Wu, L. Ding, P.E.D. Love, H. Li, H. Luo, L. Jiao, Mapping computer vision research in construction: Developments, knowledge gaps and implications for research, Automation in Construction. 107 (2019) 102919. https://doi.org/10.1016/j.autcon.2019.102919.

[100] Y.-J. Cha, W. Choi, O. Büyüköztürk, Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks, Computer-Aided Civil and Infrastructure Engineering. 32 (2017) 361–378. https://doi.org/10.1111/mice.12263.

[101] H. Kim, H. Kim, Y.W. Hong, H. Byun, Detecting Construction Equipment Using a Region-Based Fully Convolutional Network and Transfer Learning, Journal of Computing in Civil Engineering. 32 (2018) 04017082. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.

[102] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, H. Omata, Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images, Computer-Aided Civil and Infrastructure Engineering. 33 (2018) 1127–1141. https://doi.org/10.1111/mice.12387.

[103] B.E. Mneymneh, M. Abbas, H. Khoury, Automated Hardhat Detection for Construction Safety Applications, Procedia Engineering. 196 (2017) 895–902. https://doi.org/10.1016/j.proeng.2017.08.022.

[104] H. Liu, G. Wang, T. Huang, P. He, M. Skitmore, X. Luo, Manifesting construction activity scenes via image captioning, Automation in Construction. 119 (2020) 103334. https://doi.org/10.1016/j.autcon.2020.103334.

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

[106] Y. Mo, D. Zhao, J. Du, M. Syal, A. Aziz, H. Li, Automated staff assignment for building maintenance using natural language processing, Automation in Construction. 113 (2020) 103150. https://doi.org/10.1016/j.autcon.2020.103150.

[107] H. Fan, F. Xue, H. Li, Project-Based As-Needed Information Retrieval from Unstructured AEC Documents, Journal of Management in Engineering. 31 (2015) A4014012. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000341.

[108] S. Li, H. Cai, V.R. Kamat, Integrating Natural Language Processing and Spatial Reasoning for Utility Compliance Checking, Journal of Construction Engineering and Management. 142 (2016) 04016074. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001199.

[109] J. Zhang, N.M. El-Gohary, Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking, Journal of Computing in Civil Engineering. 30 (2016) 04015014. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346.

[110] M.D. Martínez-Aires, M. López-Alonso, M. Martínez-Rojas, Building information modeling and safety management: A systematic review, Safety Science. 101 (2018) 11–18. https://doi.org/10.1016/j.ssci.2017.08.015.

[111] M.-W. Park, I. Brilakis, Continuous localization of construction workers via integration of detection and tracking, Automation in Construction. 72 (2016) 129–142. https://doi.org/10.1016/j.autcon.2016.08.039.

[112] B. Xiao, S.-C. Kang, Vision-Based Method Integrating Deep Learning Detection for Tracking Multiple Construction Machines, Journal of Computing in Civil Engineering. 35 (2021) 04020071. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000957.

[113] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Advanced Engineering Informatics. 29 (2015) 239–251. https://doi.org/10.1016/j.aei.2015.02.001.

[114] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009: pp. 248–255.

[115] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, L. Zhiwei, Dataset and benchmark for detecting moving objects in construction sites, Automation in Construction. 122 (2021) 103482. https://doi.org/10.1016/j.autcon.2020.103482.

[116] B. Xiao, S.-C. Kang, Development of an Image Data Set of Construction Machines for Deep Learning Object Detection, Journal of Computing in Civil Engineering. 35 (2021) 05020005. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000945.

[117] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, W.-Y. Ma, Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations, in: 2019: pp. 6609–6618. https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Unified_Visual-Semantic_Embeddings_Bridging_Vision_and_Language_With_Structured_Meaning_CVPR_2019_paper.html (accessed November 17, 2021).

[118] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: pp. 311–318.

[119] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004: pp. 74–81.

[120] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: pp. 4566–4575.

[121] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: European Conference on Computer Vision, Springer, 2016: pp. 382–398.

[122] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common Objects in Context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014: pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

[123] COCO - Common Objects in Context, (n.d.). https://cocodataset.org/#captions-leaderboard (accessed November 13, 2021).

[124] J. Choi, B.-J. Lee, B.-T. Zhang, Multi-focus attention network for efficient deep reinforcement learning, in: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[125] E. Sulem, O. Abend, A. Rappoport, BLEU is Not Suitable for the Evaluation of Text Simplification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018: pp. 738–744. https://doi.org/10.18653/v1/D18-1081.

[126] J. Novikova, O. Dušek, A. Cercas Curry, V. Rieser, Why We Need New Evaluation Metrics for NLG, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language

Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017: pp. 2241–2252. https://doi.org/10.18653/v1/D17-1238.

[127] S.-H. Han, M.-S. Kwon, H.-J. Choi, EXplainable AI (XAI) approach to image captioning, The Journal of Engineering. 2020 (2020) 589–594. https://doi.org/10.1049/joe.2019.1217.

[128] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Science Robotics. 4 (2019) eaay7120. https://doi.org/10.1126/scirobotics.aay7120.

[129] S. Du, M. Shehata, W. Badawy, Hard hat detection in video sequences based on face features, motion and color information, in: 2011 3rd International Conference on Computer Research and Development, ieee, 2011: pp. 25–29.

[130] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, ArXiv Preprint ArXiv:1301.3781. (2013).

[131] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, ArXiv Preprint ArXiv:1810.04805. (2018).

[132] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal Sentence Encoder for English, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018: pp. 169–174. https://doi.org/10.18653/v1/D18-2029.

[133] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019: pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

[134] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All you Need, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed January 10, 2023).

[135] S.R. Bowman, G. Angeli, Potts Christopher, C.D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical

Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2015.

[136] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2019. https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (accessed March 25, 2023).

[137] S. Marcel, Y. Rodriguez, Torchvision the machine-vision package of torch, in: Proceedings of the 18th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2010: pp. 1485–1488. https://doi.org/10.1145/1873951.1874254.

[138] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[139] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020: pp. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[140] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, Convolutional Neural Networks and Incremental Parsing. (2017).

[141] S. Banerjee, A. Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005: pp. 65–72. https://www.aclweb.org/anthology/W05-0909 (accessed August 29, 2020).

[142] R. Aharoni, Y. Goldberg, Split and Rephrase: Better Evaluation and Stronger Baselines, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics

(Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018: pp. 719–724. https://doi.org/10.18653/v1/P18-2114.

[143] S. Narayan, C. Gardent, S.B. Cohen, A. Shimorina, Split and Rephrase, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017: pp. 606–616. https://doi.org/10.18653/v1/D17-1064.

[144] J.A. Botha, M. Faruqui, J. Alex, J. Baldridge, D. Das, Learning To Split and Rephrase From Wikipedia Edit History, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018: pp. 732–737. https://doi.org/10.18653/v1/D18-1080.

[145] J. Kim, M. Maddela, R. Kriz, W. Xu, C. Callison-Burch, BiSECT: Learning to Split and Rephrase Sentences with Bitexts, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021: pp. 6193–6209. https://doi.org/10.18653/v1/2021.emnlp-main.500.

[146] N.D. Nath, A.H. Behzadan, Deep Learning Models for Content-Based Retrieval of Construction Visual Data, in: Computing in Civil Engineering 2019, American Society of Civil Engineers, Reston, VA, 2019: pp. 66–73. https://doi.org/10.1061/9780784482438.009.

[147] D. Gil, G. Lee, K. Jeon, Classification of images from construction sites using a deep-learning algorithm, in: ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, IAARC Publications, 2018: pp. 1–6.

[148] N.D. Nath, A.H. Behzadan, Deep Convolutional Networks for Construction Object Detection Under Different Visual Conditions, Frontiers in Built Environment. 6 (2020). https://www.frontiersin.org/article/10.3389/fbuil.2020.00097 (accessed March 18, 2022).

[149] D.K. Smith, M. Tardif, Building information modeling: a strategic implementation guide for architects, engineers, constructors, and real estate asset managers, John Wiley & Sons., 2009.

[150] J. Yang, B. Jiang, H. Song, A distributed image-retrieval method in multi-camera system of smart city based on cloud computing, Future Generation Computer Systems. 81 (2018) 244–251. https://doi.org/10.1016/j.future.2017.11.015.

[151] A. Latif, A. Rasheed, U. Sajid, J. Ahmed, N. Ali, N.I. Ratyal, B. Zafar, S.H. Dar, M. Sajid, T. Khalil, Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review, Mathematical Problems in Engineering. 2019 (2019) e9658350. https://doi.org/10.1155/2019/9658350.

[152] R. Raghavan, K. John Singh, A Review on Content Based Image Retrieval and Its Methods Towards Efficient Image Retrieval, in: M.N. Favorskaya, S.-L. Peng, M. Simic, B. Alhadidi, S. Pal (Eds.), Intelligent Computing Paradigm and Cutting-Edge Technologies, Springer International Publishing, Cham, 2021: pp. 363–371. https://doi.org/10.1007/978-3-030-65407-8_31.

[153] I.M. Hameed, S.H. Abdulhussain, B.M. Mahmmod, Content-based image retrieval: A review of recent trends, Cogent Engineering. 8 (2021) 1927469. https://doi.org/10.1080/23311916.2021.1927469.

[154] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, (2017). https://doi.org/10.48550/arXiv.1704.04861.

[155] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[156] F. Radenović, G. Tolias, O. Chum, Fine-tuning CNN image retrieval with no human annotation, IEEE Transactions on Pattern Analysis and Machine Intelligence. 41 (2018) 1655–1668.

[157] G. Tolias, R. Sicre, H. Jégou, Particular Object Retrieval With Integral Max-Pooling of CNN Activations, in: 2016: p. 1. https://hal.inria.fr/hal-01842218 (accessed January 27, 2022).

[158] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, USA, 2014: pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.

[159] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008. https://doi.org/10.1017/CBO9780511809071.

[160] F. Zhang, H. Fleyeh, X. Wang, M. Lu, Construction site accident analysis using text mining and natural language processing techniques, Automation in Construction. 99 (2019) 238–248. https://doi.org/10.1016/j.autcon.2018.12.016.

[161] L. Zhang, J. Wang, Y. Wang, H. Sun, X. Zhao, Automatic construction site hazard identification integrating construction scene graphs with BERT based domain knowledge, Automation in Construction. 142 (2022) 104535. https://doi.org/10.1016/j.autcon.2022.104535.

[162] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, (2019). https://github.com/facebookresearch/detectron2.

[163] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, (2017). https://doi.org/10.48550/arXiv.1412.6980.

[164] Y. Wang, B. Xiao, A. Bouferguene, M. Al-Hussein, H. Li, Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning, Advanced Engineering Informatics. 53 (2022) 101699. https://doi.org/10.1016/j.aei.2022.101699.

[165] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, Int J Comput Vis. 123 (2017) 32–73. https://doi.org/10.1007/s11263-016-0981-7.

[166] B. Hu, R.-J. Song, X.-S. Wei, Y. Yao, X.-S. Hua, Y. Liu, PyRetri: A PyTorch-based Library for Unsupervised Image Retrieval by Deep Convolutional Neural Networks, in: Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA, 2020: pp. 4461–4464. https://doi.org/10.1145/3394171.3414537 (accessed March 10, 2022).

[167] I. Ha, H. Kim, S. Park, H. Kim, Image retrieval using BIM and features from pretrained VGG network for indoor localization, Building and Environment. 140 (2018) 23–31. https://doi.org/10.1016/j.buildenv.2018.05.026.

[168] F. Cakir, K. He, X. Xia, B. Kulis, S. Sclaroff, Deep Metric Learning to Rank, in: 2019: pp. 1861–1870. https://openaccess.thecvf.com/content_CVPR_2019/html/Cakir_Deep_Metric_Learning_to_Rank_CVPR_2019_paper.html (accessed March 29, 2023).

[169] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations, in: 2016: pp. 1096–1104. https://openaccess.thecvf.com/content_cvpr_2016/html/Liu_DeepFashion_Powering_Robust_CVPR_2016_paper.html (accessed March 29, 2023).

[170] H.O. Song, Y. Xiang, S. Jegelka, S. Savarese, Deep Metric Learning via Lifted Structured Feature Embedding, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[171] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, (2011).

[172] H. Jun, B. Ko, Y. Kim, I. Kim, J. Kim, Combination of Multiple Global Descriptors for Image Retrieval, (2020). https://doi.org/10.48550/arXiv.1903.10663.

[173] E. Ramzi, N. Thome, C. Rambour, N. Audebert, X. Bitot, Robust and Decomposable Average Precision for Image Retrieval, (2021). https://doi.org/10.48550/arXiv.2110.01445.

[174] E.W. Teh, T. DeVries, G.W. Taylor, ProxyNCA++: Revisiting and Revitalizing Proxy Neighborhood Component Analysis, (2020). https://doi.org/10.48550/arXiv.2004.01113.