Deep Learning-based Framework of Summarizing Construction Videos for Vision-based Monitoring of Construction Sites

by

Bo Xiao

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

Doctor of Philosophy

in

Civil (Cross-disciplinary)

Department of Civil and Environmental Engineering

University of Alberta

ABSTRACT

In recent years, video monitoring of construction sites has become increasing popular worldwide, with the video footage captured containing important visual information concerning the progress of the given project. Video monitoring also improves the security at construction sites, serving as a deterrent against theft of materials and equipment. Furthermore, vision-based analysis of video footage is beneficial to construction management in terms of facilitating crew productivity, reducing safety risks, and optimizing site layouts. Despite offering a range of potential benefits, though, the efficient use of raw jobsite videos by construction professionals remains a challenge. In current practice, construction engineers have to manually browse the entire video to retrieve the desired information from a particular period of footage, and this manual inspection is a time-consuming and error-prone process. Meanwhile, storage of the video footage is challenging, especially considering the high resolution and long streaming time typical of construction site footage. Consequently, project managers have to recycle video footage every one or two weeks to free up digital storage space, discarding construction documentation that would have been invaluable as a long-term point of reference. To address these issues, this research proposes a deep learning-based framework to automatically distill raw video footage from construction sites into video highlights and text descriptions using a deep learning-based framework. To achieve this overarching goal, three specific objectives are pursued: (1) dataset development: developing an image dataset of construction machine images for deep learning object detection; (2) highlights detection: proposing a deep learning-based method for detecting video highlights from construction raw video footage; and (3) text generation: deploying deep learning image captioning methods to generate text descriptions from construction images. The outputs of the

proposed framework (i.e., video highlights and text descriptions) will help construction engineers to efficiently ascertain what is happening in construction site without the need to manually browse the original construction videos. Compared with the original raw footage, the video highlights and text descriptions require much less storage space, making it practical to retain them for a period of years rather than weeks. The proposed framework provides the foundation for several advanced applications that will benefit the construction management, including: (1) auto-generating reports from daily construction videos; (2) building a querying system that searches for clips of interest based on text descriptions; and (3) quantitatively analyzing construction productivity based on video highlights. The framework proposed in this research is focusing on summarizing videos of construction machines captured by stationary cameras, which can be expanded for processing other types of construction videos (e.g., workers and materials) in the future.

PREFACE

This thesis is the original work of Bo Xiao. This thesis is organized in a monograph format. Four journal papers related to this thesis have been published or submitted, which are listed below.

1. **Xiao, B**., and Kang, S. (2021). "Development of an Image Data Set of Construction Machines for Deep Learning Object Detection." *Journal of Computing in Civil Engineering*, 35(2), 05020005. Dr. Kang was the supervisory authority and was involved with concept formation and manuscript composition.
2. **Xiao, B**., and Kang, S. (2021). "Vision-Based Method Integrating Deep Learning Detection for Tracking Multiple Construction Machines." *Journal of Computing in Civil Engineering*, 35(2), 04020071. Dr. Kang was the supervisory authority and was involved with concept formation and manuscript composition.
3. **Xiao, B**., Yin, X., and Kang, S. (2021). "Vision-based Method of Automatically Detecting Construction Video Highlights by Integrating Machine Tracking and CNN Feature Extraction." *Automation in Construction*, 129, 103817. Dr. Kang was the supervisory authority and was involved with concept formation and manuscript composition. Dr. Yin was invloved in the methodology development and manuscript composition.
4. **Xiao, B**., Wang, Y., and Kang, S. (2021). "Deep Learning Image Captioning in Construction: A Feasibility Study", Under review for publication in *Automation in Construction.* Dr. Kang was the supervisory authority and was involved with concept formation and manuscript composition. Yingheng Wang was involved in the data collection and experiment conduction.

ACKNOWLEDGEMENTS

The journey of Ph.D. studies has been an amazing experience in my life. I am glad to have had the opportunity to study at the University of Alberta. Words cannot express the appreciation I feel towards my family, supervisor, exam committees, and friends. It was a pleasure carrying out my research with your support and collaboration.

First and foremost, I express my gratitude to my supervisor, Dr. Shih-Chung Kang, for his endless support, invaluable instruction, and wisdom. I wouldn't be here without his supervision. Meanwhile, I would like to thank my supervisory committee, Dr. Simaan AbouRizk, Dr. Hong Zhang, and Dr. Ming Lu, my exam committee chair, Dr. Ying-Hei Chui, my internal examiner Dr. Yasser Mohamed, and my external examiner Dr. Jack Chin Pang Cheng for their time, advice, and efforts to help me to strengthen my research.

I feel very fortunate to have worked with the following colleagues: Mr. Jason Yang, Mr. Zicong Huang, Dr. Yuan Chen, Dr. Chen Chen, Dr. Xianfei Yin, Ms. Jingwen Wang, Ms. Yuxuan Zhang, Mr. Yiheng Wang, Mr. Shuai Liu, Ms. Jieyu Cui, and Mr. Keith Lam, and I appreciate their kind help in my research.

Most especially, I offer my deepest gratitude to my parents, Mr. Shejun Xiao and Ms. Yan Liu, who have always encouraged me and provided guidance. Without their great support behind me, I would not have made it so far.

TABLE OF CONTENTS

ABSTRACT ii

PREFACE iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiv

Chapter 1: INTRODUCTION 1

1.1 Background 1

1.1.1 Vision-based monitoring in construction 2

1.1.2 Deep learning methods 4

1.2 Research Gap 5

1.3 Research Objectives and Scope 6

1.4 Thesis Organization 9

Chapter 2: LITERATURE REVIEW 10

2.1 Sensor-based Monitoring in Construction 10

2.2 Deep Learning Object Detection 11

2.3 Images Datasets for Object Detection 13

2.4 Vision-based Object Tracking 14

2.5 Video Highlight Detection 16

2.6 Deep Learning Image Captioning 18

2.7 Construction Applications of Vision-based Methods 19

Chapter 3: OVERVIEW OF THE PROPOSED PRAMEWORK 21

3.1 Introduction 21

3.2 Proposed Framework 21

Chapter 4: DEVELOPMENT OF AN IMAGE DATASET OF CONSTRUCTION MACHINES FOR DEEP LEARNING OBJECT DETECTION 24

4.1 Introduction 24

4.2 Methodology for Dataset Development 26

4.2.1 Machine category selection 26

4.2.2 Image collection 29

4.2.3. Image selection 30

4.2.4 Image annotation 34

4.3 Dataset Statistics 39

4.4 Algorithm Analysis 42

4.4.1 Algorithm selection 42

4.4.2 Analysis results 45

4.5 Discussion 49

4.6 Summary 52

Chapter 5: DEEP LEARNING-BASED METHOD OF AUTOMATICALLY DETECTING CONSTRUCTION VIDEO HIGHLIGHTS 54

5.1 Introduction 54

5.2 Methodology for Video Highlight Detection 56

5.2.1 Machine detection and tracking 58

5.2.2 Rule-based keyframe detection 64

5.2.3 CNN feature extraction and similarity evaluation 68

5.2.4 Video editing 71

5.3 Implementations and Evaluation Metrics 72

5.3.1 Implementations 72

5.3.2 Evaluation metrics 73

5.4 Case Study 1: Construction Gate 74

5.4.1 Experimental setup 74

5.4.2 Experimental results 78

5.4.3 Video highlights for construction gate control 79

5.5 Case Study 2: Earthmoving 80

5.5.1 Experimental setup 80

5.5.2 Experimental results 82

5.5.3 Video highlights for productivity analysis 83

5.6 Discussion 84

5.7 Summary 86

Chapter 6: GENERATING TEXT DESCRIPTIONS FROM CONSTRUCTION IMAGES BY ADOPTING DEEP LEARNING IMAGE CAPTIONING 88
6.1 Introduction 88
6.2 Methodology for Generating Text Descriptions 90
6.2.1 Linguistic schema and image annotation 91
6.2.2 Captioning dataset summary 94
6.2.3 Method selection 96
6.2.4 Evaluation metrics 99
6.3 Implementation and Evaluation Results 101
6.3.1 Implementation 101
6.3.2 Sentence level evaluation 102
6.3.3 Element level evaluation 104
6.4 Keyframe Captioning 106
6.5 Discussion 107
6.6 Summary 110
Chapter 7: CONCLUSIONS, CONTRIBUTIONS, AND FUTURE WORKS 112
7.1 Conclusions 112
7.2 Contributions 114
7.2.1 Academic contributions 115
7.2.2 Industrial contributions 116
7.3 Future Works 117

REFERENCES..... 120

LIST OF TABLES

Table 4-1. Descriptions of the construction machines selected in the proposed image dataset ... 28

Table 4-2. Description of selected deep learning algorithms ... 43

Table 4-3. Hyperparameter information for training selected deep learning algorithms ... 44

Table 4-4. Algorithm analysis results in terms of AP, mAP, and detection speed (the best performance is denoted in bold) ... 47

Table 5-1. Summary of predefined construction rules ... 64

Table 5-2. Specifications of test videos for construction gate case ... 76

Table 5-3. Construction rules applied to construction gate case ... 77

Table 5-4. Experimental results of proposed method in construction gate case ... 79

Table 5-5. Summary of machines accessing the gate ... 80

Table 5-6. Construction rules applied to earthmoving case ... 82

Table 5-7. Experimental results of proposed method in earthmoving case ... 83

Table 5-8. Duration and storage size of video highlights in construction gate case ... 85

Table 6-1. List of suggested activities for construction machines ... 93

Table 6-2. Statistics of Top 20 N-Gram of the captioning dataset ... 95

Table 6-3. Information on deep learning image captioning methods used for evaluation ... 99

Table 6-4. Sentence level evaluation results ... 103

Table 6-5. Element level evaluation results of tsfm-sc method ... 106

LIST OF FIGURES

Figure 1-1. Workflow of vision-based monitoring in construction management 4

Figure 1-2. Research objectives 7

Figure 3-1. Overview of the proposed framework for video summarization 22

Figure 4-1. Method for developing construction machine image dataset 26

Figure 4-2. Example images of construction machine objects in the proposed image dataset 27

Figure 4-3. Examples of applying the duplication criteria 31

Figure 4-4. Example images of machine size selection criteria 32

Figure 4-5. Example images selected in the proposed image dataset 33

Figure 4-6. Example images selected in the proposed dataset based on collection sources 33

Figure 4-7. Examples of annotation standards for occlusions and illuminations 36

Figure 4-8. Example of image annotation using LabelImg software 37

Figure 4-9. The developed user interface for the crowdsourcing platform mTurk 38

Figure 4-10. Number of objects and number of images for each type of construction machine in the ACID dataset 40

Figure 4-11. ACID dataset statistics 41

Figure 4-12. Learning curve graphs of training four deep learning algorithms 45

Figure 4-13. Confusion matrix on ACID validation set detected by Faster-R-CNN-ResNet101 48

Figure 4-14. Example images in ACID validation set under snowy, rainy, and night conditions detected by Faster-RCNN-ResNet101 49

Figure 5-1. Overview of proposed highlight detection method 57

Figure 5-2. Overall framework of CMT integrating YOLO-v3 58

Figure 5-3. Simulation of the association of two consecutive frames 60

Figure 5-4. Steps in calculating image hashing similarity 61
Figure 5-5. Example of keyframe selection applying the working zone rule 66
Figure 5-6. Example of keyframe selection using working interaction rule 68
Figure 5-7. Illustration of residual block 70
Figure 5-8. Illustration of the computation of TIoU 74
Figure 5-9. Example images from test videos in construction gate case 75
Figure 5-10. Overview of baseline method 78
Figure 6-1. Encoder–decoder architecture of deep learning image captioning 89
Figure 6-2. Methodology for generating text descriptions from construction images 91
Figure 6-3. Illustration of linguistic schema 92
Figure 6-4. Element distribution of captioning dataset 94
Figure 6-5. Architecture of baseline method 96
Figure 6-6. Architecture of attention method 97
Figure 6-7. Illustration of transformer decoder 98
Figure 6-8. Example captioning results produced by tsfm-sc 103
Figure 6-9. Example keyframe captioning results 107
Figure 6-10. Example captioning errors committed by tsfm-sc 109

LIST OF ABBREVIATIONS

RFID	Frequency Identiication
GPS	Global Positioning System
ID	Identification
3D	Three Dimensional
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
BIM	Building Information Modeling
LSTM	Long Short-Term Memory Neural Networks
UWB	Ultra-Wideband
R-CNN	Region-Based Convolutional Neural Networks
SVM	Support Vector Machine
YOLO	You Only Look Once
SSD	Single-Shot MultiBox Detector
R-FCN	Region-Based Fully Convolutional Neural Networks
VOC	Visual Object Classes
COCO	Common Objects in Context
SORT	Simple Online and Real-Time Tracking
IoU	Intersection over Union
mAP	Mean Average Recall
UI	User Interface
UAV	Flying Unmanned Vehicles
MTurk	Mechanical Turk
ACID	Alberta Construction Image Dataset
AP	Average Precision
fps	Frames per Second
TP	True Positive
FP	False Positive
FN	False Negative
SQL	Structured Query Language

CMT	Construction Machine Tracker
TIoU	Temporal Intersection over Union
LCY	Loose Cubic Yard
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
METEOR	Metric for Evaluation of Translation with Explicit ORdering
CIDEr	Consensus-Based Image Description Evaluation
SPICE	Semantic Propositional Image Caption Evaluation

Chapter 1: INTRODUCTION

1.1 Background

Construction is one of the largest indutrial sectors in Canada, contributing around 8.3% ($137.13 billion dollars) of Canada's gross domestic product in 2020 (Statistia 2021). However, the construction industry remains largely low-tech and labor-intensive, with relatively few construction companies investing in innovation (Yang et al. 2015). Manual obervation remains the primary method by which crew productivity and site safety are monitored, presenting an opportunity to improve the efficiency of this task through the introduction of automation. Sensor technologies, such as frequency identiication (RFID), global positioning system (GPS), and laser scanning, have been widely adopted in automated applications in construction for the purpose of expediting processes, improving productivity, and reducing safety risks. However, the deployment, management, and maintenance of sensor systems can be costly and time-consuming (Luo et al. 2018).

In recent years, video monitoring of construction sites has become increasingly popular in construction, as it allows project managers to monitor the status of their job sites remotely. Compared with other sensors (e.g., RFID, GPS, and laser scanners), the use of video cameras installed on site offers the advantages of lower cost, simple installation and maintenance, and larger monitoring range (Kim et al. 2019). In surveying 142 construction experts, Bohn and Teizer (2010) identified that construction cameras can reduce project budgets by boosting the efficiency of communication, resource management, and site security. In these respects, cameras are versatile tools in construction engineering for delivering high-quality and more economical projects.

Construction video footage captured by cameras contains important visual information about project progress and activities, while the analysis of this footage using computer vision methods is beneficial to construction management in terms of productivity analysis, safety control, and so forth (Xiao and Zhu 2018). For example, Chen et al. (2020) proposed a framework for recognizing excavator activities in video footage in order to automatically calculate the productivity of this equipment. To facilitate site safety, Chi and Caldas (2012) analyzed construction machines in video footage as a way of preventing potential collisions proactively in future operations. Besides the abovementioned benefits, construction videos are easily understood and interpreted by humans, and are widely adopted as a form of official project documentation (Zhou et al. 2012).

Currently, most existing researches in construction community are aiming to use automatic methods to help professionals to reduce engineering works (e.g., productivity analysis, site planning, and decision making). In fact, construction engineers have to spend a lot of time on non-engineering works including inspecting videos, organizing documentations, and writing daily reports. In the construction project lifecycle, systematic storage and organization of construction video footage is critical with respect to the retrieval, analysis, and documentation of construction activities. By adotping automatic methods to solve these non-engineering works, construction engineers can get rid of tedious works and put more effors on engineering works, which will eventually improve the productivity and safety of construction projects.

1.1.1 Vision-based monitoring in construction

The monitoring of construction sites using vision-based methods has emerged as a promising avenue of research within the construction automation field. Figure 1-1 provides the workflow of the adoption of computer vision in construction management applications. The data inputted in

the deployment of these vision-based methods is construction video footage. There are two main video capture methods in vision-based monitoring of construction sites: single camera and multi-camera. Single camera is adopted for monocular vision analysis (Chu et al. 2020), and multiple cameras are employed for stereo vision analysis (Kim and Chi 2020). The principal subjects of interest in construction video are machines (Kim et al. 2018b), workers (Park and Brilakis 2012), and materials (Song et al. 2006).

As concluded by Xu et al. (2020), vision-based methods can be categorized into low-level processing methods and high-level processing methods. Low-level processing refers to the retrieval of direct information (e.g., object pixel location, object category, and object ID) from construction videos, focusing on image-level processing. The low-level processing methods applied in construction management include feature extraction (Dalal and Triggs 2010), object detection (Xiao et al. 2021a), object tracking (Konstantinou et al. 2019), and image captioning (Liu et al. 2020), to name a few.

High-level processing refers to the task of gaining deeper understanding of the contents/scene in construction videos using the information retrieved from low-level processing, focusing on the holistic information in construction videos. The typical high-level processing methods in construction mangement include scene reconstruction (Yang et al. 2013), pose estimation (Chu et al. 2020), activity recognition (Chen et al. 2020), and 3D tracking (Lee and Park 2019). Various automated applications can then be built upon the high-level procesing methods in construction, such as productivity analysis (Kim et al. 2019), progress reporting (Park et al. 2018), safety control (Han and Lee 2013), and querying system (Ha et al. 2018).

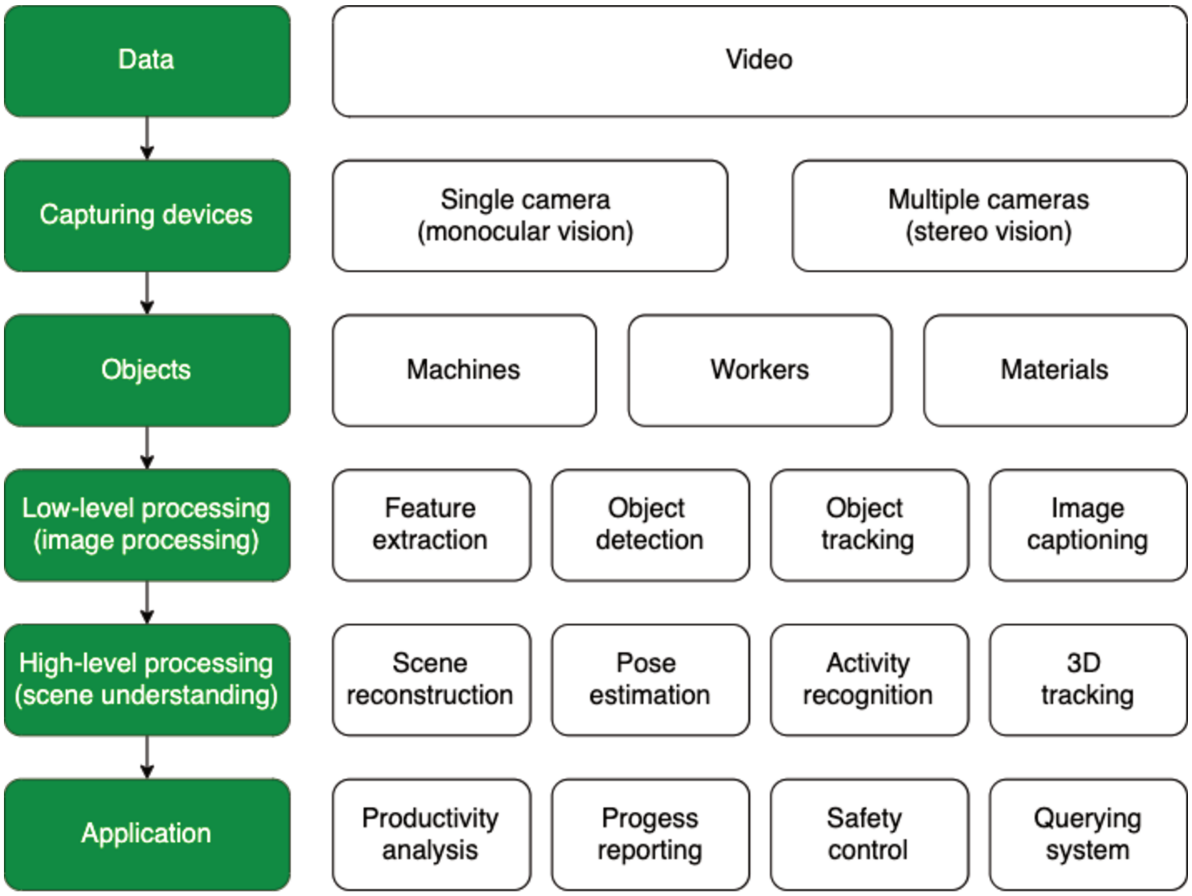


Figure 1-1. Workflow of vision-based monitoring in construction management

1.1.2 Deep learning methods

Deep learning is a subfield of machine learning based on artificial neural networks and representation learning. Compared with other machine learning methods, deep learning methods extract automated features rathe r than manually designing features, and as such they have been shown in recent studies to have achieved superior performance (Liu et al. 2016; Ren et al. 2017). Indeed, the computation power has grown exponentially in recent years. By performing parallel computations on graphics cards, deep learning methods can achieve high processing speeds in various tasks such as identifying objects from images, translating speech to texts, and recognizing activities from videos. Currently, deep learning methods are being widely employed

in many research fields, including human–computer interaction (Wang et al. 2018), medical imaging (Bouget et al. 2017), and surveillance (Kumar et al. 2017).

In construction, deep learning methods have been widely adopted for both low-level processing and high-level processing, as illustrated in Figure 1-1. The common deep learning architectures include convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term neural memory networks (LSTM), and 3D CNN, to name a few (LeCun et al. 2015). CNN has been widely employed for detecting construction objects (e.g., machines, workers, and materials) from videos (Xuehui et al. 2021), while RNN and LSTM have been integrated for automatically generating texts from construction images (Liu et al. 2020). Moreover, 3D CNN can be used for recognizing construction machine activities from videos (Chen et al. 2020). Most deep learning methods applied in construction management are supervised learning methods, which means the deep learning methods learn from human-created datasets. As such, the quality and quantity of the image datasets have a strong influence on the performance of deep learning applications in construction.

1.2 Research Gap

Despite offering a range of potential benefits, the use of raw construction video for construction site monitoring is challenged in three notable respects:

1. Retrieval of the desired information from construction video is time-consuming and labor-intensive. Construction professionals usually prefer to browse jobsite videos manually to retrieve the project information needed because videos provide visual information that can be understood by human eyes. However, manually browsing is time-consuming and causes fatigue. In this regard, some owners and engineers underestimate the value and utility of construction videos due to the difficulty of browsing these videos.

2. The sheer volume of video footage generated from continuously recording construction sites can become practically unmanageable. For instance, a one-hour video in 1080p resolution necessitates approximately 2 GB of digital storage space. Assuming one camera streams 3,000 hours in a one-year construction project, 6 TB of space is required for storing this footage. As such, many project managers prefer to delete videos at one- or two-week intervals, resulting in a loss of video records that may be of some value for future reference.
3. The practical efficiency of existing vision-based applications in construction is low. As illustrated in Figure 1-1, existing vision-based applications are built upon high-level processing methods, with these methods having a low processing speed and consuming considerable computational resources. Construction videos contain a significant number of redundant frames that can be removed without losing the relevant project management information. However, these redundant frames cannot be recognized and are integrally processed, wasting computational resources. As such, many vision-based applications have low processing speed and cannot be practically applied in construction management.

1.3 Research Objectives and Scope

To fill these gaps, this research proposes a deep learning-based framework to automatically summarize construction videos in the form of video highlights with corresponding text descriptions for vision-based monitoring of construction sites. The proposed framework is aiming reduce the efforts of construction engineers for inspection and management of construction videos. Through the adoption of the proposed framework, the generated video highlights and text descriptions can be used to replace the raw construction videos in project management, thereby reducing significantly the storage requirements and manual effort

associated with browsing video footage. As illustrated in Figure 1-2, the proposed framework is the holistically mid-level processing for vision-based monitoring of construction. It incorporates low-level processing (i.e., feature extraction, object detection, object tracking, and image captioning) for video summarization, with most of the irrelevant frames from the original video removed in an automated fashion. High-level vision-based processing methods can thereby be limited in scope to analyzing the video highlights and text descriptions rather than the entire raw video. As such, the practical efficiency of these vision-based applications can be improved. The key idea underlying the proposed method is to detect keyframes by analyzing moving trajectories of all construction objects appearing in the video, where the keyframes will be post-processed and captioned to generate video highlights and text descriptions, respectively.

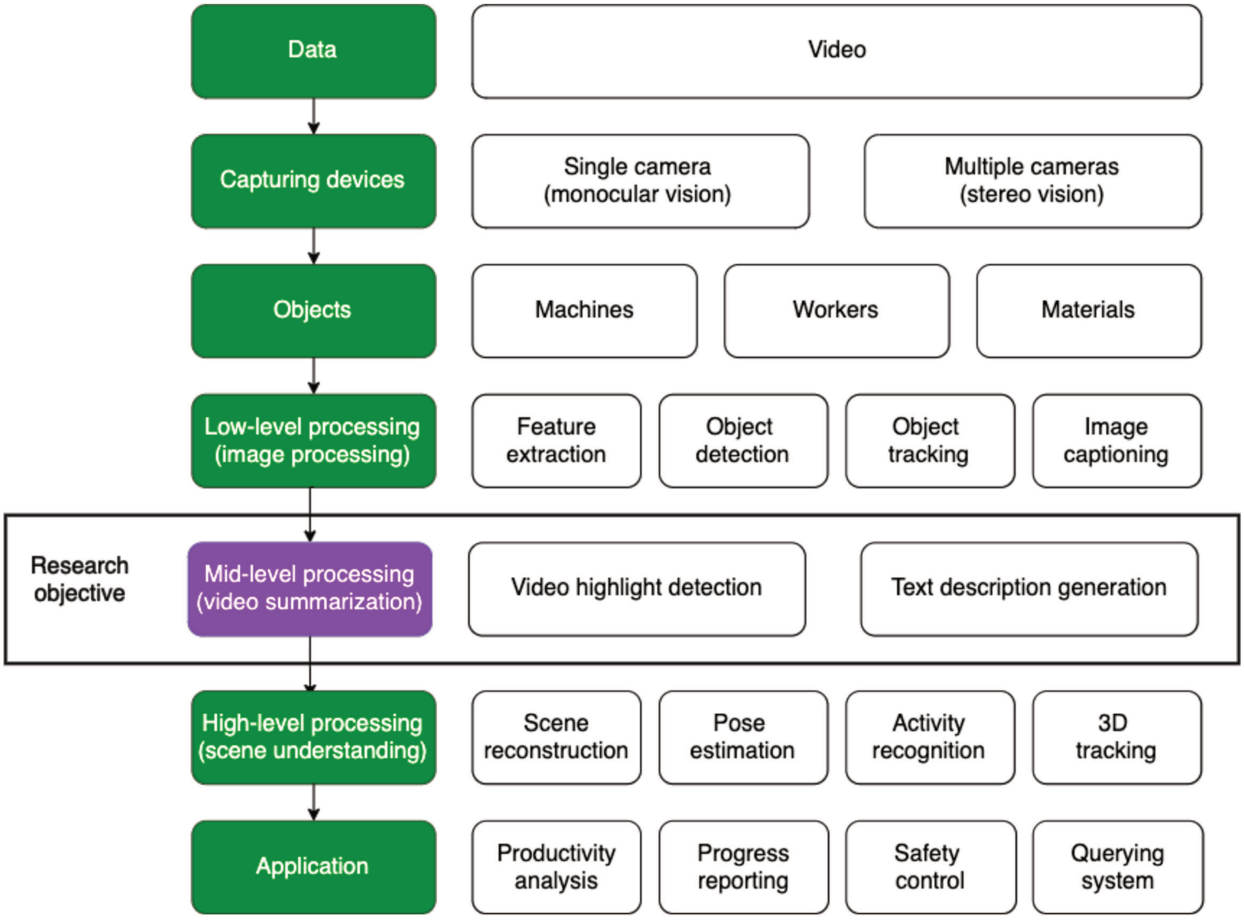


Figure 1-2. Research objectives

To achieve this goal, the following objectives are pursued in this research:

1. Dataset development: development of an image dataset for training deep learning object detection methods to recognize construction machines from videos. Since the proposed framework is based on detecting and tracking all construction objects, the development of such an image dataset is a fundamental step in this research.
2. Highlights detection: development of a deep learning-based method for automatically generating video highlights from construction videos. This objective addresses the question of how to define, detect, and produce video highlights that are useful in vision-based applications (e.g., productivity analysis and logistics management) in construction monitoring.
3. Text generation: adoption of deep learning image captioning as the basis for generating text descriptions from construction images. The focus here is on the use of image captioning methods to generate construction-related texts from keyframes of construction videos.

The scope of this research focuses on summarizing the construction videos captured by a stationary camera. It should also be noted that the proposed framework only tracks trajectories of construction machines (rather than workers or materials) for video summarization purposes, since, in many construction scenarios (e.g., earthmoving scenarios and gate scenarios), most keyframes are related primarily to equipment operations. Although tracking of workers and materials is not within the scope of this research, it can be achieved in future work by extending the proposed framework. The main scope of the present research, then, is the use of low-level image processing methods to summarize video footage of construction machines. Although high-level image processing and vision-based applications are outside the scope of this research,

Chapters 5 and 6 will describe how the proposed method can contribute to vision-based applications.

1.4 Thesis Organization

This thesis comprises seven chapters. Chapter 1 illustrates the research background, including brief introductions of the current practice of vision-based monitoring in construction and deep learning methods. The research gaps, objectives, and scope are also outlined in this chapter. In Chapter 2, a thorough review of the literature on sensor-based monitoring in construction, vision-based processing techniques, and construction applications of vision-based methods is provided. Chapter 3 provides an overview of the proposed framework, including a description of how the development of the proposed framework satisfies the research objectives.

Chapter 4 describes the development of an image dataset of construction machines for deep learning object detection. The methodology for collecting and annotating construction images is introduced, and an algorithm analysis is conducted to validate the feasibility of the developed dataset. Chapter 5 demonstrates the methodology for detecting video highlights from construction videos. Two case studies are conducted on the construction gate scenario and earthmoving scenario, respectively, to evaluate the proposed highlight detection methods. In Chapter 6, deep learning image captioning methods are incorporated in order to automatically translate keyframes into text descriptions. Moreover, a linguistic schema is proposed to annotate images of construction machines for training image captioning methods.

Finally, in Chapter 7, several conclusions are drawn, and the academic and industrial contributions of the research are summarized; the limitations of this research and potential areas of future works are also outlined in this chapter.

Chapter 2: LITERATURE REVIEW

This chapter outlines the relevant research in the following areas as presented in the literature: (1) sensor-based monitoring in construction; (2) deep learning object detection; (3) image datasets for object detection; (4) vision-based object tracking; (5) video highlight detection; (6) deep learning image captioning; and (7) construction applications of vision-based methods. A comprehensive literature review has been conducted to support understanding the state-of-the-art studies in the research community and research objectives of this thesis.

2.1 Sensor-based Monitoring in Construction

Sensor-based methods refer to monitor entities by various sensors such as GPS, RFID, UWB (ultra-wideband), and laser scanner. GPS tracks the location of an entity remotely through well-spaced satellites, and is a well-established monitoring system in construction scenarios (Li et al. 2005). Ergen et al. (2007) have integrated GPS in a precast storage yard to monitor the trajectories of construction components. Lu et al. (2007) have proposed a method that positions construction vehicles in building construction sites based on GPS. However, GPS-based methods are inconvenient to set up and lose precision in an indoor construction environment.

RFID is a wireless non-contact sensor using radio frequency waves to transfer data, and has been widely adopted for monitoring construction materials and workers. Song et al. (2006) have proposed a localization method of construction materials on jobsites based on RFID technology. Lee et al. (2012) have designed a RFID-based real-time locating system for construction safety monitoring; however, this tagging sensor-based technology does not perform adequately on modern construction sites because the deployment of sensors is costly.

UWB is another type of radio technology that transmits short-range pulses over a wide bandwidth range. The main advantage of UWB is this technology can provide precise 3D localization and requires lower power in harsh environments (Cheng et al. 2011). On construction sites, UWB can be used for real-time 3D tracking of workers to active tracking zones in order to improve workspace safety. Teizer et al. (2007) have employed UWB to monitor materials and ironworkers on construction sites. Shahi et al. (2012) have adopted UWB to track steel and timber in indoor construction projects. However, UWB systems require extra effort whereby professional engineers need to measure known positions by a total station in prior.

Laser scanners capture accurate 3D point clouds of an object's surface by combining two kinds of information: data from a laser being shone on the object and data from a moving camera. Zhang and Arditi (2013) have proposed a method to monitor the progress of a construction project using laser scanning technology. Bosché et al. (2015) have proposed a method to integrate laser scanning and building information modeling (BIM) to enhance construction site monitoring. Adán et al. (2018) have adopted laser scanning to reconstruct the BIM models of some components projected onto the wall. Laser scanners are expensive in terms of hardware and also require specialists to physically maintain and operate them.

2.2 Deep Learning Object Detection

Object detection is an important technique in machine learning that not only indicates the presence of a given class, but also indicates the position of instances (Kulchandani and Dangarwala 2015). Currently, state-of-the-art object detection methods are built up deep learning. A general type of deep learning object detection method produces a large number of candidate boxes, and then classifies these boxes using classification networks. This pattern is called two-stage detection, which has achieved top tier performance in terms of accuracy with respect to

several benchmarks. The representative approach is the region-based convolutional neural networks (R-CNN) (Girshick et al. 2014), which has adopted convolutional neural networks to extract image features and support vector machine (SVM) to classify candidate boxes. This two-stage detection method has been improved by other techniques, such as region-of-interest pooling (Girshick 2015), and anchor boxes (Ren et al. 2017), as these methods have achieved much faster speeds compared to their predecessors. Another promising deep learning object detection system is one-stage detection, which aims to further improve the processing speed. One-stage methods, such as YOLO (You Only Look Once) (Redmon et al. 2016) and SSD (Single-Shot MultiBox Detector) (Liu et al. 2016), store results prediction in the last convolutional layer, which enables the methods to achieve real-time speed.

In construction research community, a large number of researches have been conducted on advancing deep learning object detection methods specifically for detecting construction objects (e.g., construction machines and workers). Kim et al. (2018) has combined the region-based fully convolutional neural networks (R-FCN) and transfer learning techniques for robustly detecting construction equipment. Fang et al. (2018b) have proposed an improved faster region with convolutional neural networks for construction scenarios, which has achieved the average detection accuracy of 91% and 95% on worker and excavator, respectively. Arabi et al. (2019) has provided a comprehensive solution for construction equipment detection including the development and mobile-end deployment. Kolar et al. (2018) have proposed a deep learning detection method integrating transfer learning to detect safety guardrails in construction sites, which obtained the accuracy of 95.6% in testing.

2.3 Images Datasets for Object Detection

The image datasets for object detection contain a large number of images and each image has been manually annotated at the object level. The development of detection image datasets expands the potential of deep learning algorithms, which is an important reason why deep learning is successful (Russakovsky et al. 2015). Deep learning algorithms that have been trained on large datasets have the ability to detect objects from new scenarios to avoid the overfitting problem (Cogswell et al. 2016). There are several comprehensive datasets available to the public in the computer vision community; however, these datasets are mainly aimed at natural categories, such as vehicles, animals, and furniture. The PASCAL visual object classes (VOC) dataset provides 11,530 images of 20 object classes for training and testing object detection algorithms (Everingham et al. 2010). Furthermore, the Microsoft common objects in context (COCO) dataset contains 160,000 labelled images belonging to 91 categories (Lin et al. 2014). Recently, Google AI has constructed the open image dataset v4 containing 478,000 labelled images with 6,000+ categories (Kuznetsova et al. 2020).

There is limited research regarding the development of detection image datasets in construction scenarios. Tajeen and Zhu (2014) have developed an image dataset for evaluating construction equipment detection algorithms. However, that dataset was annotated to evaluate non-deep learning object detection algorithms in construction and the deep learning object detection algorithms were not considered. Also, that dataset includes only 5 machine types and contains only 2,000 annotated images, which need to be expanded for training deep learning object detection algorithms. Developing the annotated dataset is the fundamental step for all research that relied on deep learning object detection, while the quality and quantity of the dataset affects

the performance of vision-based applications. Therefore, more efforts are needed to develop the image datasets of construction machines for deep learning object detection.

2.4 Vision-based Object Tracking

Vision-based object tracking is conducted to address the tracking of single or multiple objects from videos. By integrating object detection, the tracking problem can be simplified to the association of detection results across frames (Milan et al. 2016). The vision-based tracking methods can be categorized into motion model methods and appearance model methods. The motion model methods associates the detection results based on the object's movements (e.g., trajectories and object bounding boxes). Rezatofighi et al. (2015) have proposed a joint probability data association method for tracking multiple objects by the hypothesis of their trajectories. Bo and Nevatia (2012) have proposed a conditional random field graph method to associate tracking trajectories. Bewley et al. (2016) have proposed a simple online sort and real-time tracking method (SORT) for multiple object tracking. In SORT, the detection results are tracked by the Kalman filter in each frame, where the detection and tracking results are associated by the intersection over union (IoU) matrix. Then, the Hungarian is employed to maximize the IoU matrix in order to assign tracking IDs to all objects. The association of motion model tracking can achieve high processing speed, while the tracking precision and robustness are impractical for complicated applications.

The appearance model tracking associates the detection results based on both the pixel region of detected objects and their pixel location information. Ross et al. (2008) have proposed a tracking method that integrates the particle filtering and the eigen images as the appearance model. In their work, the particle filtering predicts multiple locations of the object in the next frame based on possibility distributions and then the tracker selects one object as the tracking result which has

the most similar appearance with the object in the current frame. Yu et al. (2016) proposed a tracking method that extracts object features by GoogleNet (Szegedy et al. 2014) to build the association matrix by calculating the cosine distance between the extracted features. Choi (2015) has proposed a multiple object tracking method based on the aggregated local flow descriptor. Milan et al. (2017) have employed the recurrent neural networks to extract features from the detection regions for association and tracking. Although the association of appearance model tracking has better tracking performance, its processing speed reduces when the number of objects increases.

In construction, researchers have put much effort into developing vision-based object tracking methods for construction scenarios in terms of dealing with high-resolution images, frequent occlusions, and special features (e.g., vests for tracking workers). For tracking construction workers, Konstantinou et al. (2019) have proposed a vision-based object tracking method for workers in a complex environment based on both a filtering model and an appearance model. Angah and Chen (2020) have integrated the instance segmentation method into vision-based tracking of construction workers in outdoor environment. Park and Brilakis (2016) have developed a tracking method that integrates the SVM detection (Dalal and Triggs 2010) and particle filtering tracking (Ross et al. 2008) by bounding box location, size, and color histograms to track multiple construction workers under varied illumination scenarios.

For tracking construction machines, Xiao et al. (2021b) have integrated illumination enhancement methods into construction scenarios for tracking multiple machines at nighttime, which has achieved the robust performance in extreme lighting conditions. Zhu et al. (2016) adopted a particle filtering tracker for construction equipment to overcome the short-term occlusions typical on construction sites. Although object tracking methods are widely adopted in

vision-based construction monitoring, few studies have integrated object tracking into video summarization tasks.

2.5 Video Highlight Detection

Video highlight detection refers to producing short and representative clips from the full-length video, which has been used in sport highlights, the film industry, and egocentric videos (Liu et al. 2010). For example, Merler et al. (2019) developed multimodal excitement features to generate video highlights from a golf tournament and two international tennis tournaments, with the results having been closely aligned with the official video highlights. Wang et al. (2020) proposed a contrastive attention module as the feature representations to produce trailers from full-length movies. Yao et al. (2016) employed a pairwise deep ranking model to detect video highlights from first-person GoPro videos, achieving an accuracy of around 80% on over 100 hours of videos from YouTube. Moreover, video highlight detection has significantly reduced the manual editing and reviewing effort required in many applications.

A typical video highlight detection method extracts features from raw videos and then selects keyframes by analyzing changes in the feature space across frames. The video clips around keyframes, usually several seconds, are combined together to produce the video highlights for users (Lin et al. 2015). Feature extraction and keyframe selection are the main focuses in the computer vision community. A large number of features have been studied for the task of video highlight detection. For example, Laganière et al. (2008) integrated the spatio-temporal Hessian matrix to collect image features for video highlight detection. Liu et al. (2009) adopted the scale invariant feature transform to identify the boundary of video highlights. Deep neural network has also emerged as a promising method for extracting features from images by learning from human-created dataset. Mahasseni et al. (2017) employed the LSTM to summarize video

highlights. Xiong et al. (2019), meanwhile, adopted CNN technology to detect video highlights from Instagram videos.

Keyframes are a set of representative frames in videos that define the quality of the video highlights. One approach in this regard has been to calculate the Euclidean distance of every two continuous frames. The keyframes can then be identified as the points in the video footage where feature distance changes rapidly (Truong and Venkatesh 2007). Moreover, the clustering technique has been employed to extract keyframes. For instance, Mundur et al. (2006) developed a keyframe selection method based on Delaunay clustering. Other studies have employed a method of selecting keyframes by ranking all frames with a pre-defined importance score, such as entropy (Muhammad et al. 2020), context prediction score (Lin et al. 2015), or influence metric (Lu and Grauman 2013). However, for two reasons in particular, existing methods in computer vision are not able to efficiently detect construction video highlights: (1) keyframes in construction cannot be simply defined as the frames with image features change rapidly; and (2) video highlights are expected to be interpretable and flexible for construction management.

Researchers in the construction automation field have put efforts into developing video highlight detection methods to accommodate construction video characteristics. For instance, Chen and Wang (2017) developed construction-specific color, texture, and gradient features for extracting keyframes from videos. The developed methods were tested on four construction videos, and the experimental results suggested that color features generally outperform gradient and texture features. However, that study focused on exploring image features and did not utilize the content information. Ham and Kamari (2019) proposed a content-based keyframe selection method for construction videos captured by drones. However, their method was designed for drone videos, whereas it cannot be directly applied to videos captured by fixed-position cameras.

2.6 Deep Learning Image Captioning

Image captioning generates one or several sentences from an image to describe the scene information inside of this image, which is an interdisciplinary research topic of computer vision and natural language processing (Huang et al. 2019). Recently, deep learning methods have gained superior performance for image captioning, while the general framework of image captioning methods consists of an "encoder" to retrieve features from images and a "decoder" to generate texts from retrieved features (Hossain et al. 2019). Mao et al. (2014) have proposed a deep learning image captioning method that uses CNN as the "encoder" and RNN as the "decoder"; this work has been improved in the "show and tell" method by only inputting the visual features at the first time-step of RNN (Vinyals et al. 2017). Furthermore, attention mechanism has been widely applied in deep learning image captioning, which allows the neural networks to focus on its subset of inputs to select specific features (Gao et al. 2019; Xu et al. 2015). Recently, transformer attention has been implemented in image captioning studies(Vig 2019; Vig and Belinkov 2019; Zhang et al. 2019) and achieved reliable performance.

In construction management, image captioning can be used for scene analysis. By analyzing the sentences generated by image captioning methods, the major objects, activities, and interactions of objects can be retrieved. For example, Liu et al. (2020) have applied the CNN-LSTM captioning method for manifesting construction worker activity scenes. In that research, a linguistic schema used for annotating images of construction workers is proposed and three experiments are conducted to illustrate the feasibility of image captioning in construction. Bang and Kim (2020) have applied image captioning to drone images for vision-based monitoring of construction sites with achieving the mean average recall (mAP) of 45.52%. However, compared with other image processing methods, image captioning received limited attention in our

research community because this technique is relatively new. The performance of deep learning methods is still unclear and more investigations need to be conducted.

2.7 Construction Applications of Vision-based Methods

Construction video footage contains important visual information that can be used for productivity analysis, progress reporting, safety control, and querying system (Yang et al. 2015). For productivity analysis, Kim et al. (2018c) have conducted an interaction analysis of identifying the activities of earthmoving equipment based on vision-based tracking. Roberts and Golparvar-Fard (2019) have proposed an end-to-end solution of detection, tracking, and activity analysis of earthmoving equipment with high accuracy. Kim and Chi (2019) have developed a novel excavator action recognition method by integrating detection, tracking, and sequential pattern features. Kim and Chi (2020) have proposed a multi-camera vision-based productivity monitoring system of earthmoving operations based on detection and tracking excavators and dump trucks, which have achieved an accuracy of 97.6%.

Thanks to the availability of fixed-position cameras and smartphones, the number of images and videos have increased significant in construction sites on a daily basis. Naturally, these photography documentations can be used to report construction progress to project participates (Han and Golparvar-Fard 2014). Vision-based methods reconstructed the 3D construction scenarios from images or videos, which can be used to compare with BIM to obtain the deviation of progress (Ibrahim et al. 2009). For example, Brilakis et al. 2011a) have adopted the structure from motion technique to conduct the 3D reconstruction to report project progress. Chen et al. (2019) have integrated deep learning techniques for 3D reconstruction in construction sites, which has gained the 84.2% validation accuracy in experiments.

Safety is the most important concern in construction industry, whereas analyzing construction videos can enhance the site safety. Gualdi et al. (2011) have proposed a head hat detection method for monitoring the safety of construction workers using videos. Chi and Caldas (2012) have recognized construction machines in videos for the purpose of obtaining vehicle speeds in order to alert in the case of potential collisions. Nguyen and Brilakis (2018) have developed a vision-based system to detect the over-height vehicle in bridge and tunnels for improving site safety. Tang et al. (2019) have proposed a novel vision-based method to detect construction objects and forecast the potential collisions based on mixture density network and long short-term network, which can forecast the target location in the future 2 seconds. Zhong et al. (2020) have proposed a hybrid framework to extract construction procedural constraints from videos and compare with construction regulations in order to ensure site safety. Fang et al. (2020) have proposed a novel deep learning-based framework that combines data fusion and digital technologies to enhance construction sites safety.

The querying of construction images or videos helps to the project documentations, and vision-based methods can be used for efficiently querying in construction management. Brilakis and Soibelman (2005) have proposed a content-based search engine based on blind relevance feedback to retrieve construction images. Nath and Behzadan (2019) have investigated deep learning detection algorithms for retrieving construction visual data. Similarly, Ha et al. (2018) have proposed a BIM image retrieval method by implementing CNN networks. Li et al. (2020) have proposed a novel searching system named BIMSeek to retrieve BIM models by images or queries. Currently, the vision-based applications have to process the entire construction videos including the clips without any important information. By eliminating redundant frames in construction videos, the efficiency of vision-based applications can be remarkably facilitated.

Chapter 3: OVERVIEW OF THE PROPOSED PRAMEWORK

3.1 Introduction

As reviewed in the previous chapter, vision-based methods play an important role in automated monitoring of construction sites, and many research studies have been conducted in this domain. Management of the large volumes of video footage accumulating over the lifespan of a construction project, though, has proven challenging when using existing methods. In this chapter, a framework for automatically summarizing construction videos is outlined. The proposed framework can be divided into three main sections—dataset development, video highlight detection, and text description generation—each of which is described briefly in this chapter before being discussed in greater detail in subsequent chapters.

3.2 Proposed Framework

Figure 3-1 provides an overview of the proposed framework for construction video summarization. First, the construction engineer (user) inputs the raw construction video to the proposed framework through a web-based user interface (UI). Then, the object detection method detects all pre-defined classes of construction objects from each video frame. The object tracking method associates the detection results across frames in order to obtain the trajectories of construction machines appearing in the video, while the tracking results are stored in a database for further querying. Moreover, the keyframe detection recognizes the keyframes by analyzing the tracking trajectories and image features extracted from the raw construction video, whereas the keyframes can be used for producing video highlights.

Since this framework integrates deep learning object detection, an annotation image dataset for training is necessary. Therefore, this research also involves the development of an image dataset

for machine selection, image collection, image selection, and image annotation. Meanwhile, a linguistic schema is proposed to generate construction-related text annotations of the developed dataset for training deep learning image captioning methods. The trained image captioning model is adopted to produce text descriptions from keyframes of construction video. Finally, the video highlights and text descriptions are fed back through the web UI to construction engineers for future browsing, storage, and querying.

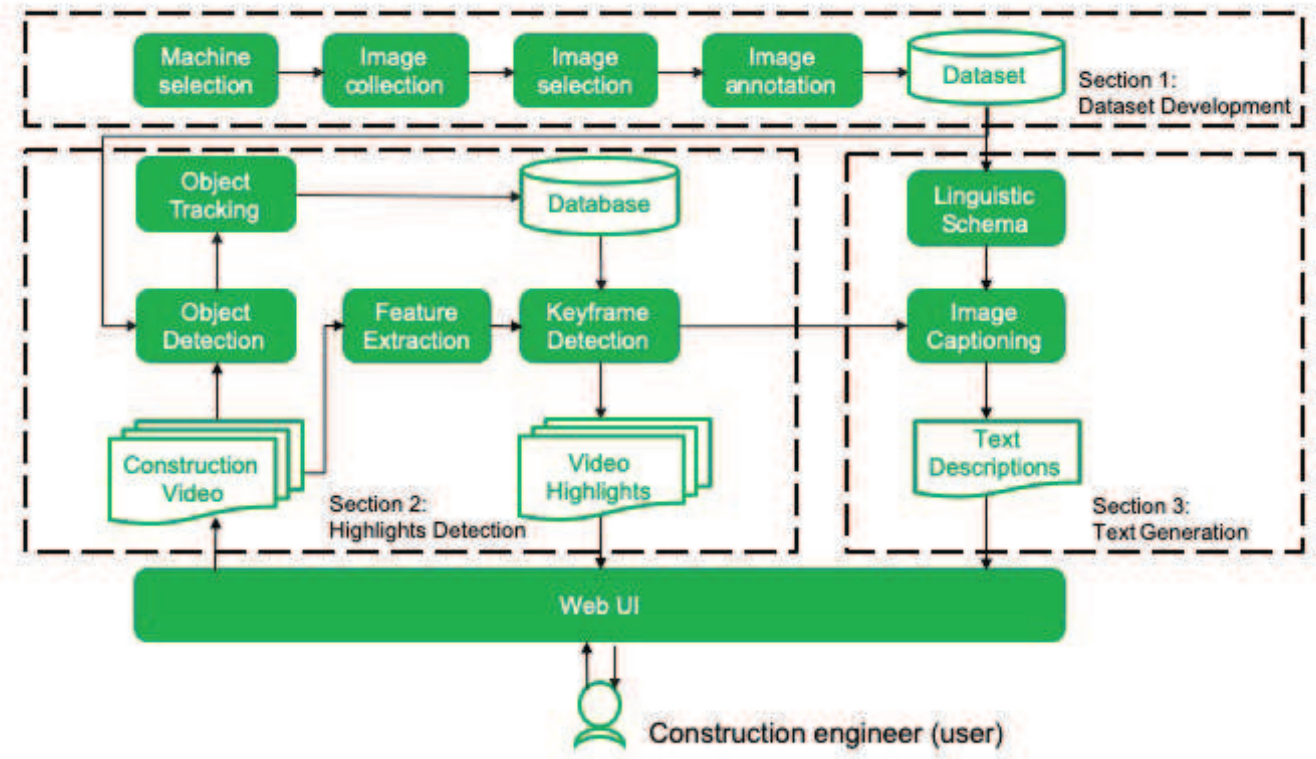


Figure 3-1. Overview of the proposed framework for video summarization

As demonstrated in Figure 3-1, the proposed framework can be divided into three sections as follows:

1. **Dataset development:** The dataset development serves research objective 1, where the main tasks are to develop a construction image dataset for deep learning object detection following the standard computer vision procedures. The performance of various existing

deep learning detection methods is also investigated. The dataset development is described in greater detail in Chapter 4.

2. **Highlight detection:** The development of a highlight detection method serves research objective 2, where the method detects construction video highlights by integrating object detection, object tracking, and feature extraction. Two case studies are conducted to evaluate the proposed highlight detection method, which is described in greater detail in Chapter 5.
3. **Text generation:** Development of an automated method to adopt deep learning image captioning for generating texts from construction images to fulfill the research objective 3. A novel linguistic schema is proposed in this section to bridge the gap between deep learning image captioning methods and construction management. Six deep learning methods are compared, and the best performing one is incorporated into the proposed framework for the purpose of producing text descriptions of keyframes. The details of the text generation are provided in Chapter 6.

Chapter 4: DEVELOPMENT OF AN IMAGE DATASET OF CONSTRUCTION MACHINES FOR DEEP LEARNING OBJECT DETECTION[1]

4.1 Introduction

Detecting construction resources (e.g., machines, workers, and materials) in images or videos is the first and fundamental step required in the development of automation to analyze construction videos. Once construction objects have been correctly recognized, a large number of construction monitoring tasks could be automated. For example, detecting excavators and dump trucks at the same time could automatically calculate the dirt-loading cycles in earthmoving projects (Chen et al. 2020). The continuous detection of machines and workers can prevent potential collisions and alert construction engineers in a timely manner (Zhu et al. 2017). Detection of construction materials identifies the material location in the supply chain, and enables effortless derivation of project performance indicators (Song et al. 2006).

Deep learning algorithms have achieved superior performance in terms of robustness and processing speed. Recent studies (Liu et al. 2016; Ren et al. 2017) indicate that deep learning algorithms can effectively detect objects in certain challenging scenarios, such as occlusions and illumination variations. This is because neural networks extract high-level features from images instead of manually designing features (e.g., edges and colors) (LeCun et al. 2015). Moreover, deep learning object detection algorithms are able to process in real-time or near real-time speed by integrating parallel computation and graphic cards (Zhao et al. 2019). Considering these

[1] A version of this chapter has been published in ASCE *Journal of Computing in Civil Engineering* as follows: Xiao, B., and Kang, S. (2021). "Development of an Image Data Set of Construction Machines for Deep Learning Object Detection." *Journal of Computing in Civil Engineering*, 35(2), 05020005. It has been reprinted with permision from the publisher.

advantages, deep learning object detection algorithms are widely employed in the field of construction automation to monitor productivity and safety (Fang et al. 2018; Kim et al. 2018a).

To apply deep learning object detection, a construction-specific image dataset, which includes machines, workers, and materials, is necessary to recognize the underlying relationships between construction objects and images. However, the construction research community is lacking such an image dataset for training deep learning object detection algorithms due to: (1) the accessibility of construction images is limited and the number of online resources offering construction images and videos is relatively limited. In industry, many construction engineers do not realize the value of construction videos, and they therefore dispose of all video footage once a given project is complete; (2) it is difficult to achieve a high degree of diversity (e.g., number of images, object categories, object size, and camera views) in construction image datasets to avoid the overfitting problem; and (3) annotating construction images in a manner that ensures high quality, low cost, and efficiency is challenging. In construction research, employing graduate students specialized in construction management is the typical method used to annotate datasets, and this is time-consuming and costly.

The primary objective of this chapter is to develop an image dataset specifically for construction machines. The development method focuses on how to collect construction images, how to select qualified images to ensure dataset diversity, and how to annotate construction images effectively. An image dataset has been developed in this research. A total of 10,000 images of ten types of machines (excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader, backhoe loader, tower crane, and mobile crane) have been collected and annotated manually. An algorithm analysis has been conducted on the developed dataset to validate its capacity for training deep learning object detection algorithms. The algorithm analysis demonstrates the

feasibility of the developed dataset in construction automation studies. Larger construction image datasets can subsequently be developed by following the same development method.

4.2 Methodology for Dataset Development

Figure 4-1 illustrates the proposed method of developing an image dataset specifically for construction machines. The development includes four main steps: machine category selection, image collection, image selection, and image annotation. Details will be provided in this section for each of the four steps. By following the proposed steps, larger image datasets can be developed and the dataset quality can be guaranteed.

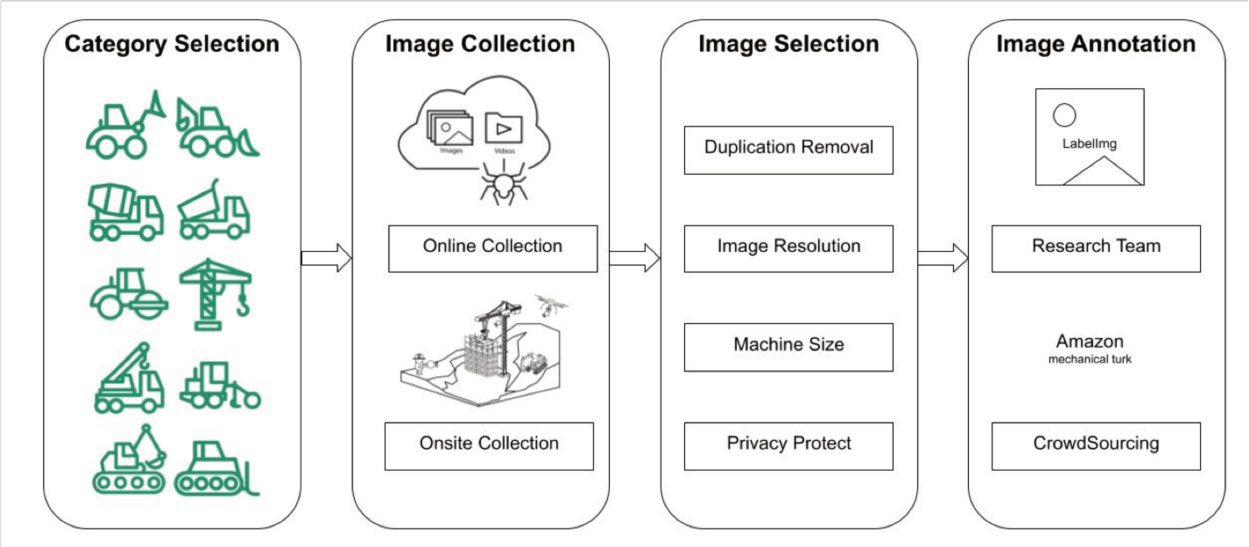


Figure 4-1. Method for developing construction machine image dataset

4.2.1 Machine category selection

Ten types of construction machines, namely excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader, backhoe loader, tower crane, and mobile crane, are chosen. The selected types of machines are considered to be common machine categories in construction scenarios, and these machines are involved in three main types of activities: excavating and lifting (excavator, backhoe loader, mobile crane, and tower crane), loading and hauling (dump

truck, concrete mixer truck, and wheel loader), and compacting and finishing (dozer, grader, and compactor). Figure 4-2 shows examples of each type of construction machine, and Table 4-1 presents descriptions of the selected construction machines.

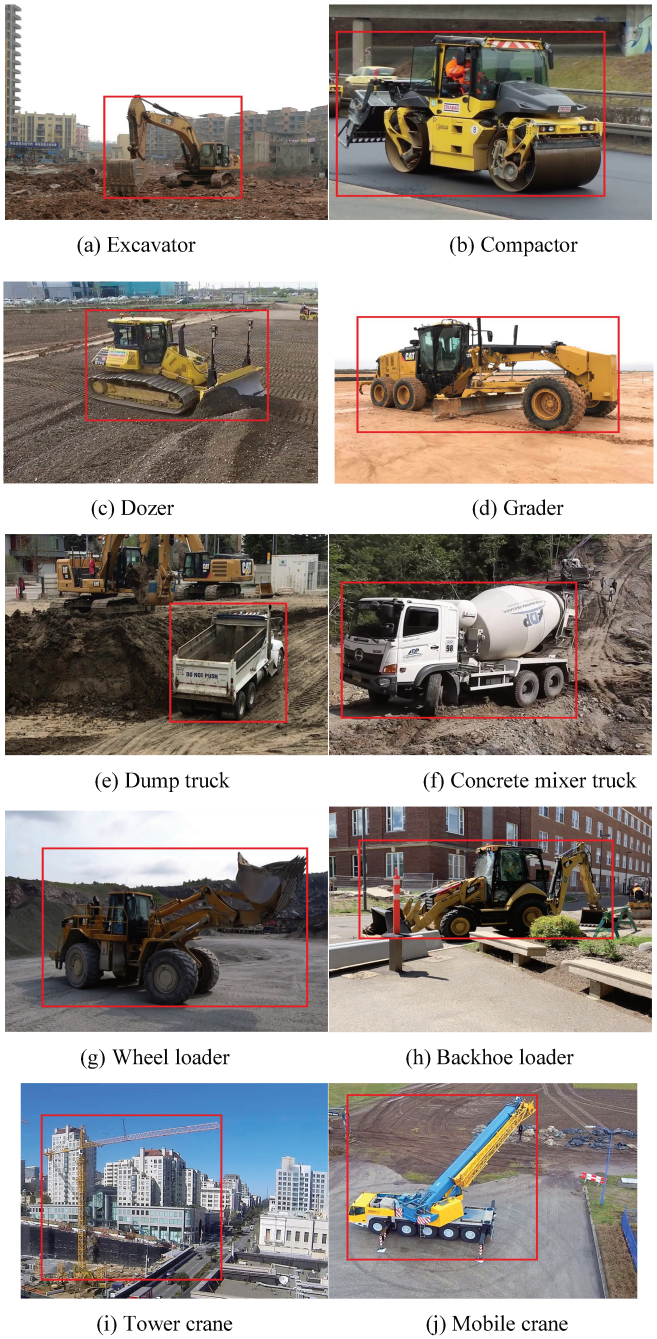


Figure 4-2. Example images of construction machine objects in the proposed image dataset

Table 4-1. Descriptions of the construction machines selected in the proposed image dataset

Construction Machine	**Description**	**Image Example**
excavator	Excavators are considered heavy construction equipment and consist of a boom, dipper, bucket and cab on a rotating platform known as the "house".	Figure 4-2(a)
compactor	A compactor is a type of machine used for compacting crushed rock as the base layer underneath concrete or stone foundations or slabs.	Figure 4-2(b)
dozer	A bulldozer is a crawler equipped with a metal plate and a claw-like device to loosen densely compacted materials	Figure 4-2(c)
grader	A grader is a construction machine with a long blade used to create a flat surface during the grading process.	Figure 4-2(d)
dump truck	A dump truck is used for transporting loads/dumps with a rear-hinged open-box equipped with hydraulic rams.	Figure 4-2(e)
concrete mixer truck	A concrete mixer truck is used to mix concrete and transport it to construction sites. It consists of a truck body and a mixer bucket.	Figure 4-2(f)
wheel loader	A wheel loader is heavy equipment machinery used in construction to load materials into or onto another type of machinery.	Figure 4-2(g)
backhoe loader	A backhoe loader is a type of heavy equipment that consists of a loader-style bucket on the front and a backhoe on the back.	Figure 4-2(h)
tower crane	A tower crane is a modern form of balance crane that consists of the same basic parts fixed to the ground on a concrete slab.	Figure 4-2(i)
mobile crane	A mobile crane is a machine with a truck body equipped with a crane.	Figure 4-2(j)

4.2.2 Image collection

To develop a construction machine image dataset, all images were collected in either one of two ways: online collection and onsite collection. For online collection, construction images and videos were downloaded from photo-sharing and video-sharing websites using a web crawler. For onsite collection, images and videos of real construction sites were collected three ways: flying unmanned vehicles (UAV), installing on-site cameras, and manually conducting site visits. All videos collected online and onsite have been converted to images in JPEG format.

Online images were downloaded from Google Images and Naver website using AutoCrawler software (YoongiKim 2018). AutoCrawler was asked to download 1,000 images from Google Images and 1,000 images from the Naver website for each class of construction machine by searching the machine names. It is difficult to download over 1,000 images per class from the photo-sharing websites since these websites have limited construction images. Meanwhile, it is observed that the dump truck, excavator, and wheel loader are relatively easier to find in the photo-sharing websites, while the mobile crane, tower crane, and concrete mixer truck are more difficult to find. YouTube was the main resource used for image collection. Totally 2,904 construction videos were collected from YouTube with an average duration of 6 minutes and each included at least one target machine. These videos include footage from multiple construction stages and various activities, and from different types of construction scenarios including day shifts, night shifts, raining, snowing, outdoor construction, and indoor construction. Videos collected from YouTube were downloaded and converted to images using a rate of one image per ten seconds considering that, in general, construction videos change slowly. To sum up, the total number of construction images collected online from Google Images, Naver, and YouTube is approximately 124,500.

For the onsite collection, the author and the research team members visited seven construction sites that are managed by the municipal government of Edmonton, Canada. During the site visits, 1,000 images were captured by cell phone cameras. Another 500 bird's-eye view images were captured by a UAV craft (DJI Mavic 2) that flies over these construction sites multiple times. Moreover, over 100 hours of construction videos captured from onsite cameras located in Xi'an City, China have been acquired. Similar to the procedure followed for the YouTube videos, these onsite videos have been converted to images at the rate of one image per ten seconds. In total, there are 37,500 images acquired from the onsite collection.

4.2.3. Image selection

After the image collection step, in total there are 162,000 images collected from online and onsite resources. If the image dataset is to avoid the overfitting problem in deep learning training, it needs to offer a high degree of variation in the images. In the image selection stage, qualified images are manually selected from all the 162,000 collected images. To ensure the dataset quality, four criteria are proposed in the framework for image selection, which are duplication removal, image resolution, machine size, and privacy protection.

(1) Duplication removal: Duplicate images were manually removed by researchers. For images retrieved from videos, although we only retrieve one image in ten seconds, there were many duplicate images because construction activities evolve very slowly, and the removal of the duplicate images was performed according to strict criteria. With respect to images extracted from the same video, to be considered for inclusion in the dataset each image must be significantly different from other images in terms of machine orientation, position, or illumination. Typically, 600 images are extracted from a 10-minute video (1 image per 10 seconds), and only 5 to 10 images of those images are chosen after duplication removal. Figure

4-3 shows an example of how the criteria of duplication are applied. To be noted, the duplication removal can also be processed by adopting automated duplicate removal algorithms based on measuring the similarity between images (Appalaraju and Chaoji 2017; Wang et al. 2014), which will be investigated in the future.

(a) not duplicated

(b) duplicated

Figure 4-3. Examples of applying the duplication criteria

(2) Image resolution: The requirement for image resolution is to only select images with a resolution larger than 608×608. Most of deep learning object detection algorithms resize the images into specific resolutions (e.g., 400×400, 1280×800) before convolution operation. In the proposed framework, any images with a resolution smaller than 608×608 are removed and this

process is automated. The resolution criteria are customized and should not be too small since the resolution of construction images is usually high.

(3) Machine size: Images are manually removed if machine objects are undersized or oversized with respect to the image size. The preferable machine size should between 1/42 to 3/4 of the whole image size. This is a soft regulation because the machine size is manually estimated, and if the construction machine is slightly oversized or undersized, this will not affect the dataset quality. An example of machine size selection is illustrated in Figure 4-4.

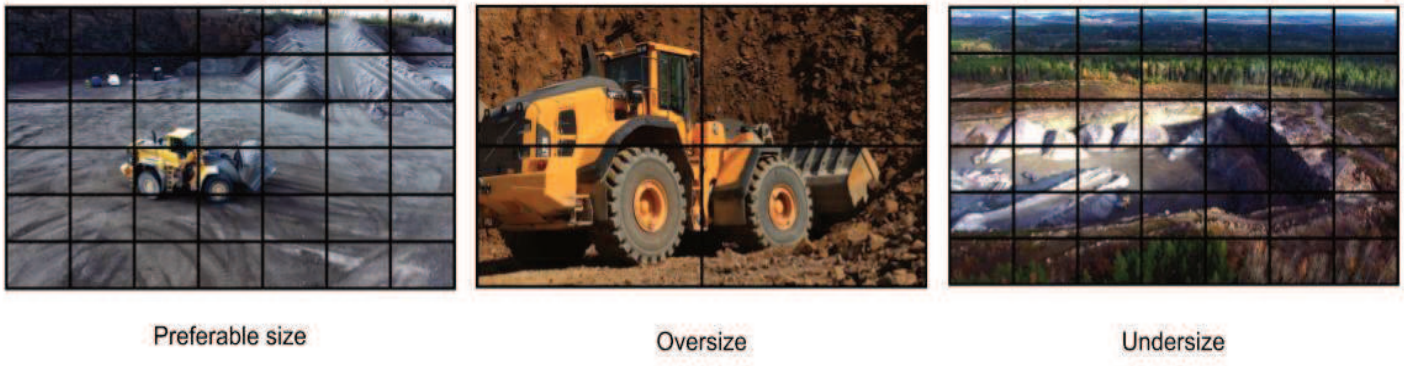


Figure 4-4. Example images of machine size selection criteria

(4) Privacy Protection: In the image collection step, some images were collected that contain workers and pedestrians. To avoid research ethical issues, all clear or partial clear human faces in the images are blurred.

After image collection and selection, 10,000 images were deemed to qualify for inclusion in the image dataset. Figure 4-5 shows some examples of the selected images. Figure 4-6 shows example images based on their collected resources (e.g., YouTube, Google Images, and UAV).

Figure 4-5. Example images selected in the proposed image dataset

Figure 4-6. Example images selected in the proposed dataset based on collection sources

For image selection, we first browsed all images to remove any images with duplication and oversize/undersize machines. The average browsing speed was 1,000 images per hour. There were 162,000 images collected from online and onsite sources. We spent around 162 man-hours in this step. After that, the computer program checked image resolutions to remove the images with low resolution (lower than 608×608). Then, a research team member then manually blurred all human faces appearing in the selected images. The blurring process took about 10 man-hours.

4.2.4 Image annotation

To be used as a dataset for object detection, the image annotation must include two components: class and position. Class refers to the object category, which is one of excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader, backhoe loader, tower crane, or mobile crane. The position is represented by an axis-aligned bounding box surrounding the construction machine object in the image. All annotation results were stored with XML format, which is a common format used in the computer vision community and is compatible with other formats. In the annotation step, half of the images are labelled by annotators at the University of Alberta, and half of the images are annotated through the Amazon crowdsourcing platform called Mechanical Turk (MTurk). To ensure the annotation quality, all annotations have been checked two times, first by one annotator and then by the author of the current study.

Annotation Standard

In order to ensure the annotation results are high quality, three standards adopted in the development of the VOC dataset have been employed and were strictly followed during the annotation process:

(1) Consistency: All annotations in the proposed image dataset are consistent in terms of the class definition, bounding box placement, and how to deal with occluded objects. The consistency also applies to how to deal with illuminations, how to control annotation quality, and how to annotate under snowing, raining, and night conditions.

(2) Correctness: All annotations in the proposed image dataset precisely describe the pixel-axis position of construction machine objects. The bounding boxes do not cut the objects, and the margins are within 5 pixels. The main objective is to annotate with as few errors as possible.

(3) Completion: All objects belonging to the ten classes are labelled. There may be more than one machine object in one image, and the annotations are exhaustive. All objects that can be identified were annotated, and occluded objects were annotated according to the part of the machine that can be seen.

Figure 4-7 shows two example images to illustrate the annotation standards in terms of dealing with occlusions and illuminations. The green color represents the objects that need to be annotated and the red color means the object should not be annotated. If the occluded area is less than 70% of the whole machine area and the machine can be easily identified upon sight, this machine needs to be annotated. For example, objects B, D, and E in Figure 4-7(a) need to be annotated. The objects I and J in Figure 4-7(b) should not be annotated because more than 70% of the objects are occluded and the contour of the dump trucks is incomplete.

For illumination, if the object's contours can be easily identified, this object needs to annotated. If the object is vague and incomplete because of illumination (too light or too dark), this object should not be annotated. For example, object E in Figure 4-7(a) needs to be labelled, and object F should not be labelled. Considering it may be difficult for annotators to discern the occluded

ratio, the 70% occlusion standard is not strict; however, annotators are required to annotate all images by following the same standards.

(a) example 1

(b) example 2

Figure 4-7. Examples of annotation standards for occlusions and illuminations

Manual Annotation

A software program named LabelImg (Tzutalin 2015) is used by the five members of the research team to annotate 5,000 images. The user interface of LabelImg is shown in Figure 4-8. The annotation process consists of three steps: (1) five annotators are given a tutorial that explains what this research is for and how to annotate images. Each annotator is required to annotate 100 images as practice and their annotations are checked and corrected, if necessary, by the author of the present study; (2) the five trained annotators are then required to annotate the construction images according to the annotation standards. A well-trained annotator can label 50 construction images per hour; and (3) to achieve consistency, all the completed annotations need to be checked by one other annotator and the author of the present study.

Figure 4-8. Example of image annotation using LabelImg software

Employing the research team to do the annotation is estimated to cost 170 man-hours for 5,000 images, where 50 man-hours are for training, 100 man-hours are for annotating, and 20 man-

hours are for checking. It should be noted that annotators cannot annotate images continuously for long durations because the annotation work quickly causes fatigue. The actual annotation work took a much longer time from start to finish. The pay rate of annotators is $15 per hour; therefore, the average cost for the research team to annotate one image is $0.51, and the average time cost is 0.034 man-hour per image.

Crowdsourcing Annotation

Crowdsourcing is a model that involves online participants in completion of a specific task (Standing and Standing 2018). Generally, tasks are posted to an online platform as an open call, and members of the crowd can self-select tasks to complete (Brawley and Pury 2016). In this research, half of the images have been annotated through the crowdsourcing platform mTurk. This platform allows requesters to self-design the user interface and functions using APIs. The developed user interface for this annotation task is shown in Figure 4-9.

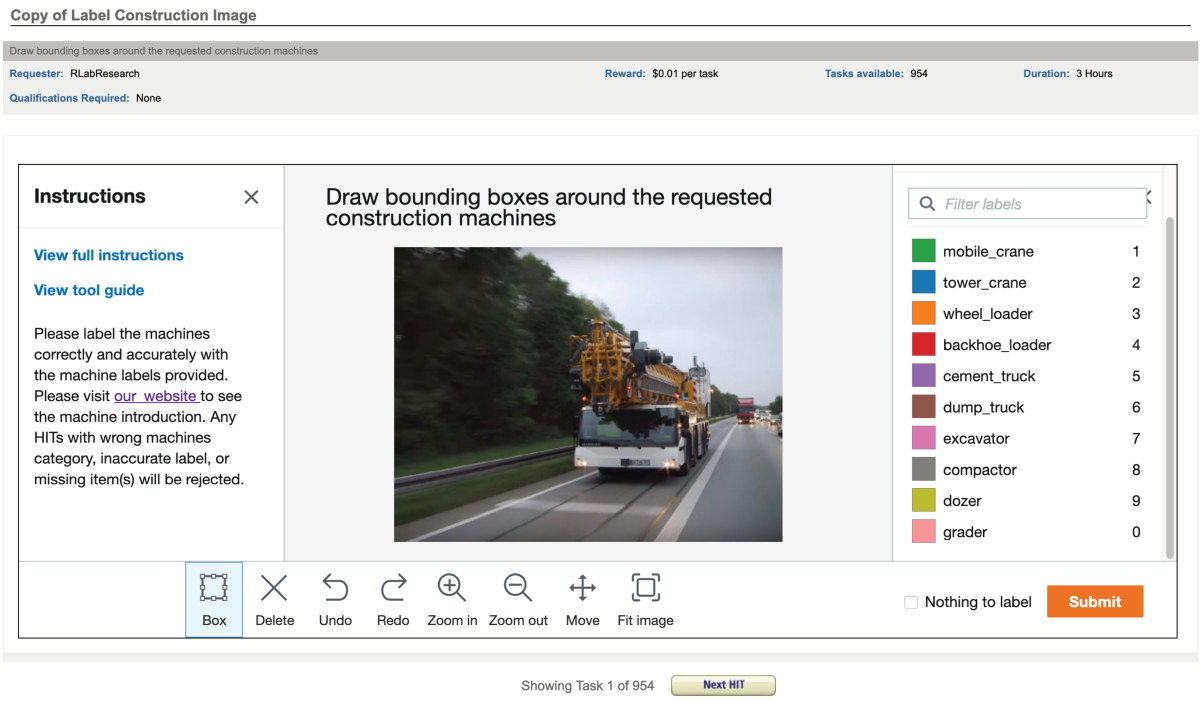


Figure 4-9. The developed user interface for the crowdsourcing platform mTurk

Crowdsourcing the annotation involves two steps: (1) the first step is to post images on the mTurk platform. In this research, 1,000 images were posted each time and the annotation time is about 1 hour. The incentive for each image is $ 0.10; and (2) the annotation results are then checked by one trained annotator and by the author of the present study. The annotator must spend a significant amount of time rejecting the images that do not meet the standards and re-labelling the images to meet the standards. This checking process costs 40 man-hours for 5,000 images. Using mTurk, the cost of annotating one image is $0.22 and the average time cost is 0.009 man-hours per image.

4.3 Dataset Statistics

In this chapter, the developed image dataset is named the Alberta Construction Image Dataset (ACID), also referred to herein as the ACID dataset. The dataset statistical analysis described in this section proves the diversity of ACID. ACID includes 10,000 images and 15,767 construction machine objects. Figure 4-10 depicts the number of objects and the number of images (containing at least one object in the respective category) for each type of construction machine. Dump trucks (3,713 objects), excavators (2,787 objects), and wheel loaders (1,823 objects) are recognized as the most frequent construction machines in ACID.

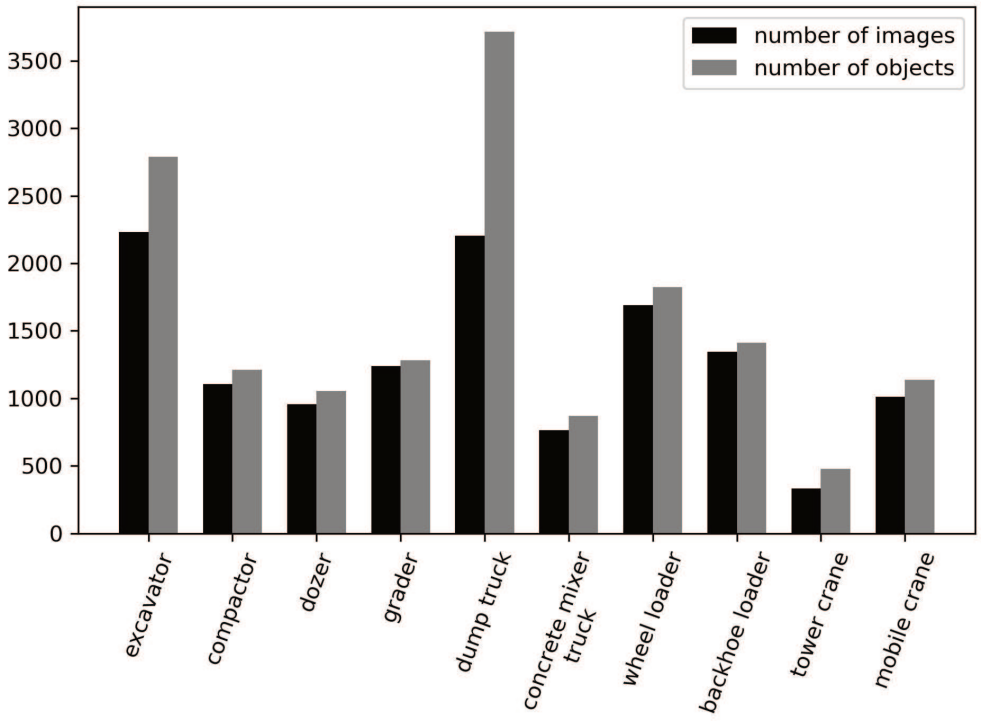
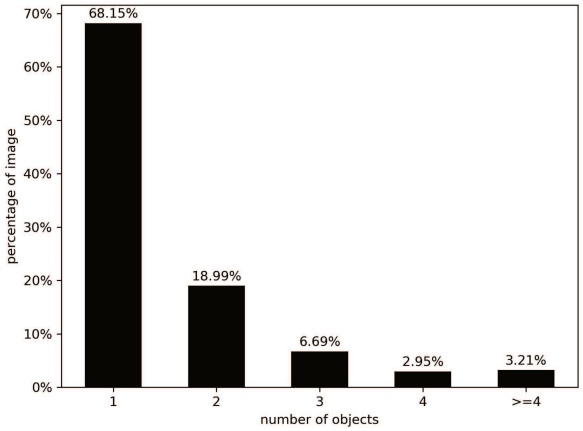


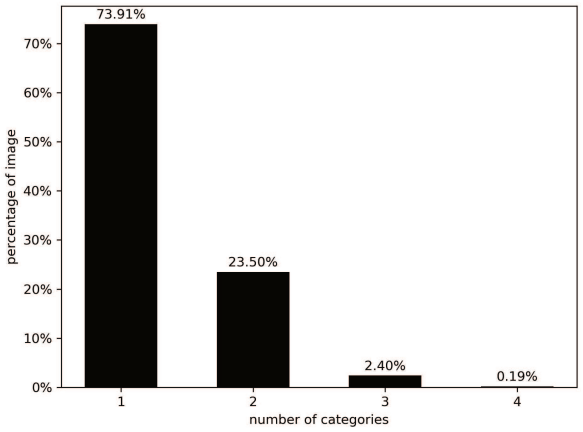
Figure 4-10. Number of objects and number of images for each type of construction machine in the ACID dataset

Figure 4-11(a) shows the number of annotated objects per image in ACID: 68.15% of the images only contain one object, and 18.99% of the images contain 2 objects per image. Figure 4-11(b) shows the number of machine categories per image: 73.91% of images contain one type of machine in the ACID. The distribution (number of objects and number of categories per image) of ACID is similar to that of the VOC dataset used in the computer vision community. In the VOC dataset, approximately 72% of the images contain only one type of object and approximately 54% of the images have only one object per image (Everingham et al. 2010). Figure 4-11(c) shows the distribution of the size of the bounding boxes that contain the object in ACID, wherein 39.94% of objects occupy a pixel area representing less than 5% of the entire image, and 35.68 % of objects occupy a pixel area between 20% and 40% of the entire image.

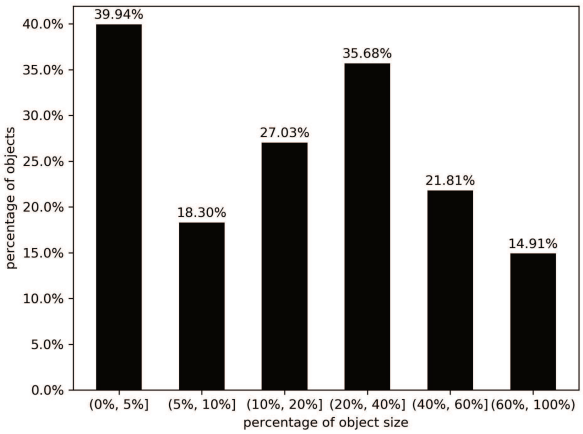
The object size distribution in ACID is more balanced than the VOC dataset, which in comparison has 95% of objects together representing 60% of the total area in VOC.



(a) Number of objects per image in the ACID dataset



(b) Number of categories per image in the ACID dataset



(c) Distribution of the bounding box size containing objects in the ACID dataset

Figure 4-11. ACID dataset statistics

4.4 Algorithm Analysis

Investigating the effectiveness of deep learning algorithms is a necessary section in this research, mainly because the performance of these algorithms in construction scenarios is unclear based on the review of the literature. Algorithm analysis is conducted on the ACID dataset, and this analysis is expected to help construction researchers select the proper algorithm for their construction applications. Meanwhile, other researchers are welcome to use ACID to evaluate other construction-specific object detection algorithms and to compare their results with the algorithm analysis results presented in this study.

4.4.1 Algorithm selection

Based on their performance in the context of computer vision datasets, four deep learning algorithms have been selected: YOLO-v3 (Redmon et al. 2016)**,** Inception-SSD (Liu et al. 2016)**,** Faster-RCNN-ResNet101 (Ren et al. 2017), and R-FCN-ResNet101 (Dai et al. 2016)**.** The detection mechanism of state-of-the-art deep learning object detection algorithms can be categorized as either one-stage detection or two-stage detection. In the computer vision community, one-stage detection frameworks (YOLO-v3 and Inception-SSD) are considered to perform better in terms of detection speed, while two-stage detection frameworks (R-FCN-ResNet101 and Faster-RCNN-ResNet101) perform better in terms of accuracy. Brief descriptions of the chosen algorithms have been summarized in Table 4-2.

Table 4-2. Description of selected deep learning algorithms

Detection Algorithm	Detection Mechanism		Brief Description
	One-stage	Two-stage	
YOLO-v3 (Redmon et al. 2016)	√		A uniform detection method with high speed, integrated multi-scale in training and priors anchor boxes for prediction.
Inception-SSD (Liu et al. 2016)	√		Single-shot multi-box detection method with classification that uses inception architecture. Of the one-stage detection algorithms, it is a high accuracy detection method.
Faster-RCNN-ResNet101 (Ren et al. 2017)		√	Detection with region proposal networks that uses ResNet101 to extract features from raw images and integrate ROI pooling before region proposal process.
R-FCN-ResNet101 (Dai et al. 2016)		√	A detection method via region-based fully convolutional network to reduce the proposal cost. Uses Resnet101 to extract features from images.

To conduct the algorithm analysis, the ACID dataset was divided into a training set (80%) and a validation set (20%). The four selected algorithms were trained on the training set with the default hyperparameter configuration, which has been summarized in Table 4-3 including input image size, training iterations, batch size, learning rate, and optimizer. Then, the trained models were tested on the validation set. The algorithm analysis was conducted on a computer with the following hardware configuration: a NVIDIA GTX 1080Ti graphic card, which has 11 GB of memory; an Intel Core i9-7920X@2.90 Hz CPU with 12 cores; and two 32 GB memory cards. YOLO-v3 is implemented using C language, and the rest of the three algorithms are implemented using the Tensorflow object detection API (Huang et al. 2017).

Table 4-3. Hyperparameter information for training selected deep learning algorithms

Algorithm	Input image size	Training iterations	Batch size	Learning rate	Optimizer	Momentum	Decay factor
YOLO-v3	608×608	200,000	32	0.0010	Momentum Decay	0.9	0.0005
Inception-SSD	300×300	200,000	32	0.0040	RMSprop	0.9	0.9
Faster-RCNN-ResNet101	1024×600	200,000	2	0.0030	Momentum	0.9	N/A
R-FCN-ResNet101	1024×600	200,000	2	0.0003	Momentum	0.9	N/A

Figure 4-12 shows the learning curve graphs (training vs. validation loss) of four deep learning algorithms training on ACID, where the red lines refer to training loss and green lines represent the validation loss. It is found the training loss decreases when the training step increases for four algorithms. To be noted, the validation loss of YOLO-v3 is missed because the validation function is not available in YOLO-v3 implementation. Also, the loss values for different algorithms are not necessarily comparable since the testing algorithms have implemented different loss functions. The learning curve graphs in Figure 4-12 have indicated all four deep learning algorithms are well-fitted on the ACID dataset. The training loss and validation loss have decreased in the training process, while the validation loss is higher than training loss. The learning curve graph shows limited degree of overfitting, which demonstrates the robustness and versatility of the ACID dataset. The learning curves of Faster-RCNN-ResNet101 and R-FCN-ResNet101 are not as smooth as YOLO-v3 and Inception-SSD due to the fact that they have smaller batch sizes.

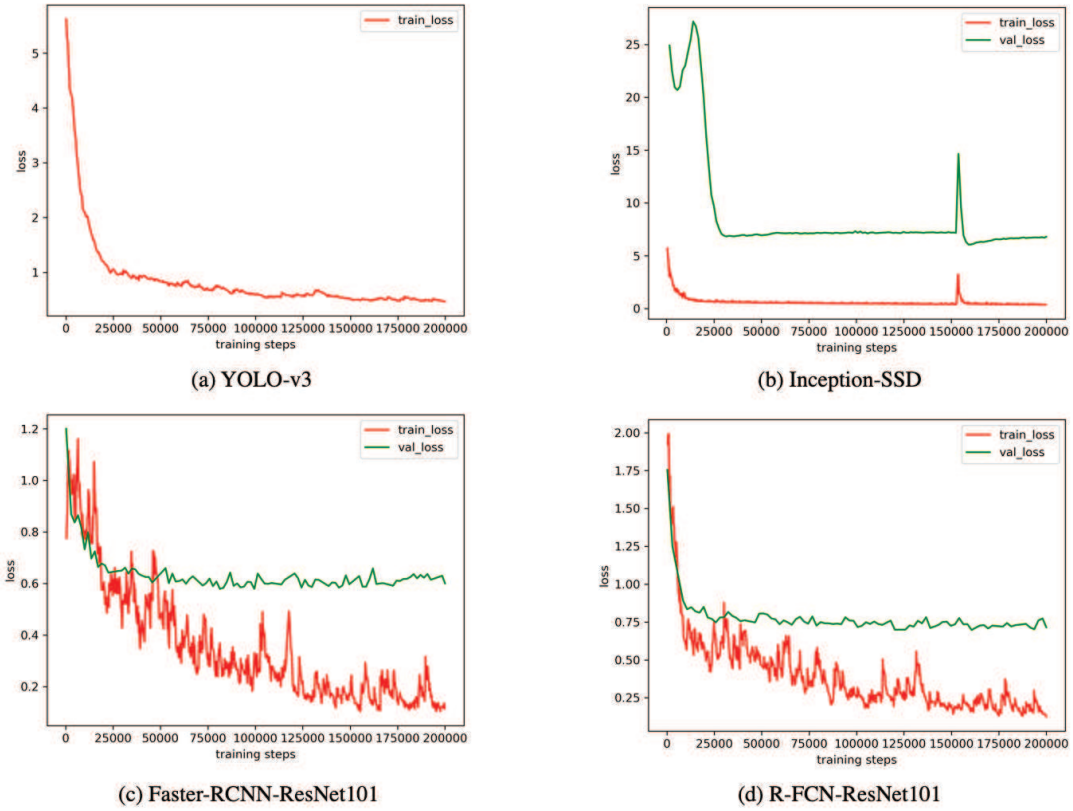


Figure 4-12. Learning curve graphs of training four deep learning algorithms

4.4.2 Analysis results

Mean average precision (mAP) is the evaluation metric used in algorithm analysis to describe the accuracy of object detection algorithms (Yilmaz and Aslam 2006), which is calculated by precision and recall (Davis and Goadrich 2006). Precision is a measurement of how accurate the object detection method is (Equation 4-1), while recall measures how well the object detection method can find all positives (Equation 4-2). Then, the average precision (AP) is the measurement of the average of precision at different recall levels r ($r \in \{0.1, 0.2, \ldots, 1\}$) for one class of object (Equation 4-3), and mAP is the mean of AP of all pre-defined classes (Equation 4-4). A higher mAP indicates better performance of an object detection algorithm in terms of both

accuracy and robustness. Frames per second (fps) is the criterion employed in algorithm analysis to evaluate the detection speed, which is calculated as the number of processed images per second.

$$Precision = \frac{TP}{TP+FP} \quad \text{(Equation 4-1)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{(Equation 4-2)}$$

$$AP = \frac{1}{11}\sum_{r\in\{0,0.1,0.2,\ldots,1\}} Precision@r \quad \text{(Equation 4-3)}$$

$$mAP = \frac{1}{N}\sum AP \quad \text{(Equation 4-4)}$$

where TP (true positive) is the number of correct detected bounding boxes. The correct detection box is determined by the IoU metric. If the IoU of a ground truth box and a detected bounding box is larger than 0.5, the detected bounding box is considered as a correct detection box; FP (false positive) is the number of negative instances that have been recognized as positive; FN (false negative) is the number of positive instances that have been recognized as negative, and; N is the number of pre-defined classes.

Table 4-4 shows a summary of the analysis results for the four selected deep learning object detection algorithms in terms of accuracy and speed. All four deep learning object detection algorithms have shown good performance in terms of detecting construction machines after having been trained on the ACID dataset. The Faster-RCNN-ResNet101 achieved the best performance in terms of accuracy with a mAP of 89.2%. The other two-stage algorithm, R-FCN-ResNet101, achieved the second-best performance in terms of accuracy with a mAP of 88.8%. For one-stage algorithms, the YOLO-v3 has achieved the mAP of 87.8%, and the Inception-SSD has gained the mAP of 83% in the algorithm analysis. All four deep learning object detection algorithms achieved a mAP of between 80% and 90%. The algorithm analysis results of ACID are similar to the results of the analysis conducted on the VOC dataset, which indicates the

complexities of ACID and VOC are at the same level. In terms of detection speed, YOLO-v3 achieved the best performance at 26.3 fps, and Inception-SSD ranked second with 20.8 fps.

Table 4-4. Algorithm analysis results in terms of AP, mAP, and detection speed (the best performance is denoted in bold)

	YOLO-v3	**Inception-SSD**	**Faster-RCNN-ResNet101**	**R-FCN-ResNet101**
AP (excavator)	**93.50%**	85.40%	92.50%	90.80%
AP (compactor)	**94.80%**	89.50%	92.30%	92.20%
AP (dozer)	89.90%	91.80%	**95.60%**	94.60%
AP (grader)	95.10%	96.00%	**98.70%**	98.30%
AP (dump truck)	**83.30%**	71.20%	81.50%	82.40%
AP (concrete mixer truck)	**94.90%**	90.80%	92.60%	94.30%
AP (wheel loader)	84.60%	83.00%	**90.60%**	88.60%
AP (backhoe loader)	95.60%	93.60%	**95.90%**	95.70%
AP (tower crane)	62.00%	54.40%	**64.40%**	64.20%
AP (mobile crane)	84.50%	84.30%	**88.40%**	84.30%
mAP	87.80%	83.00%	**89.20%**	88.80%
Detection speed (fps)	**26.3**	20.8	8.3	11.5

A confusion matrix (Luque et al. 2019) is a specific table for visualizing the errors of an machine learning algorithm. In this research, the confusion matrix has also been studied to visualize detection errors and present more details. Figure 4-13 shows the confusion matrix on the ACID validation set produced by the Faster-RCNN-ResNet101 algorithm. It is found the tower crane and mobile crane are prone to be mis-classified with each other from Figure 4-13. Around 14% of tower cranes have been mis-categorized into mobile cranes, while 20% of mobile cranes have been mis-classified as tower cranes.

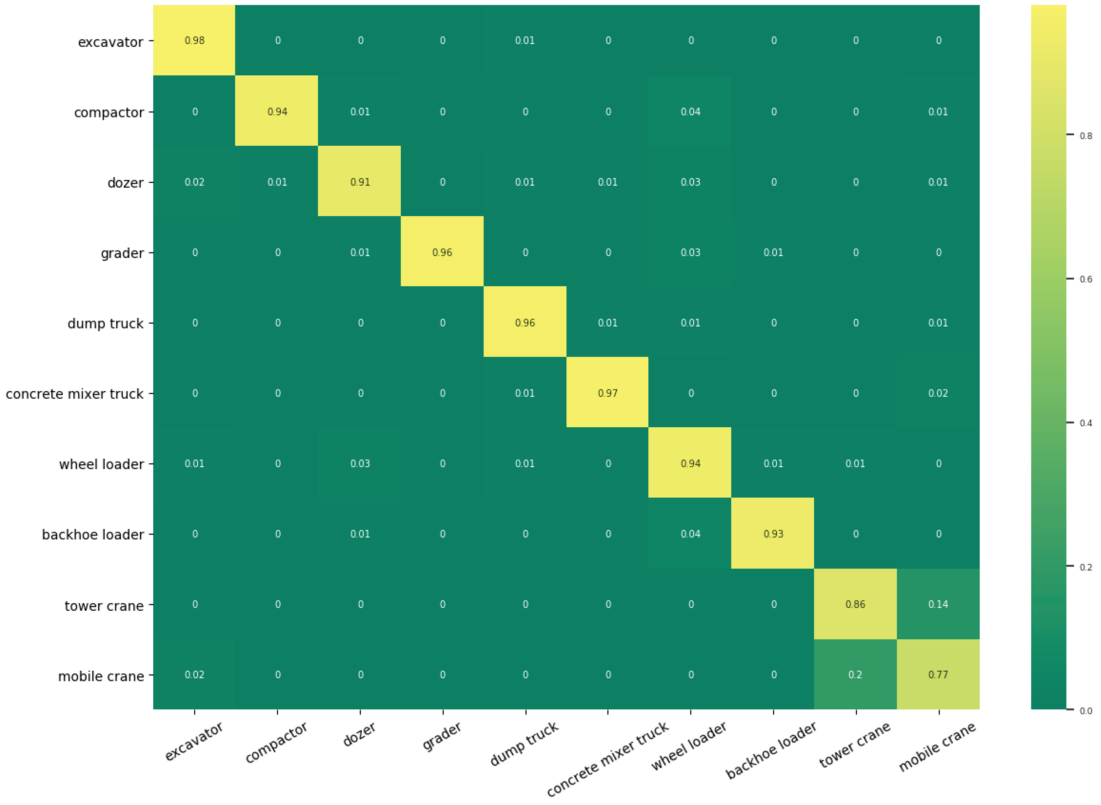


Figure 4-13. Confusion matrix on ACID validation set detected by Faster-R-CNN-ResNet101

The ACID dataset contains images captured from extreme conditions, such as snowy, rainy, and night. Detecting construction machines from the abovementioned conditions is important for specific construction scenarios (e.g., night construction). Figure 4-14 shows some example images from the ACID validation set under snowy, rainy, and night cases. In Figure 4-14, the construction images are detected by the Faster-RCNN-ResNet101 algorithm, which shows the ability of ACID for training models to detect construction machines in these extreme cases.

Figure 4-14. Example images in ACID validation set under snowy, rainy, and night conditions detected by Faster-RCNN-ResNet101

4.5 Discussion

The research in this chapter provides a standard image dataset for deep learning-based construction applications, which can be used as a platform for the evaluation and comparing of various deep learning object detection algorithms in construction scenarios. The research findings are discussed as follows:

- One-stage detection algorithms are considered more suitable for recognizing construction machines than two-stage detection algorithms. In the context of ACID, two-stage algorithms achieved higher accuracy compared to one stage algorithms; however, the difference was not significant and one-stage algorithms are much faster than two-stage algorithms. For example, YOLO-v3 achieved a mAP of 87.8%, which is only 1.4% lower than Faster-RCNN-ResNet101, while YOLO-v3 is three times faster than Faster-RCNN-

ResNet101 in terms of processing speed. It is noted that ACID dataset is easier than the PASCAL VOC dataset from two perspectives: (1) ACID only contains 10 classes of objects, which VOC contains 20 classes of objects; and (2) detecting construction machines is relatively easier than detecting natural objects (e.g., person, animal, and vehicles). When dealing with an easier detection dataset, it is reasonable that the difference between one-stage detection and two-stage detection is decreased. Therefore, one-stage detection algorithms with higher detection speeds and similar accuracies are more suitable for construction applications.

- The detection performance for tower crane and mobile crane is lower in comparison to the other construction machines based on the results of the algorithm analysis. For example, the Faster-RCNN-ResNet101 algorithm achieved an AP of 64.4% when detecting tower crane. It was found that the trained algorithms miscategorized mobile crane and tower crane, because some mobile cranes look very similar to tower cranes when in operation. As demonstrated in Figure 4-13, about 14% of tower cranes have been mis-categorized into mobile cranes, while 20% of mobile cranes have been mis-classified as tower cranes. This limitation may be solved in two ways: (1) categorize tower crane and mobile crane to crane; and (2) in specific construction applications, delete one type of crane and keep the other type of crane, depending on which is needed in a particular application.

- One of the main challenges with respect to developing ACID is the limited number of construction images that are available online. For example, all the images in the VOC dataset were downloaded from the Flickr website and the collection process was efficient in terms of time. In the case of the ACID dataset, the research team collected suitable

construction videos from YouTube by watching the video content, which was a time-consuming and labor-intensive process. Meanwhile, only 10,000 images of the 162,000 collected images qualified for the dataset according to the selection criteria, which is a qualifying ratio of approximately 6.2%. Although many construction sites have installed cameras for security and recording purposes, the captured videos are usually deleted after one week, which is a problem when trying to collect images from construction job sites.

- Online collection is recommended for collecting construction images, which is about 3 times more efficient than onsite collection. For online collection, we employed the research team to collect the videos from YouTube, which is estimated to cost 100 man-hours to collect 2,904 videos. Extracting images from YouTube videos was executed by the computer program, while the time and cost are minimal. For Google Images and Naver, the images were downloaded by the AutoCrawler software, while the time and cost can be ignored. It is estimated to cost 100 man-hours for collecting 124,500 images from online sources. The pay rate is $15 per man-hour; therefore, the average cost to collect one image from online sources is $0.012, and the average time cost is 0.0008 man-hours per image. For onsite collection, we visited seven construction sites located in Edmonton, Canada, once per month over a period of six months in order to capture the images from different construction stages. In each visit, we captured images by cell phones and the UAV for 2 man-hours (84 man-hours in total). Meanwhile, we have acquired 100 hours of videos from the construction sites directly and converted these videos to images by the computer program, while the time and cost can be ignored. It is estimated to cost 82 man-hours for collecting 37,500 images from onsite sources. The

average cost to collect one image from onsite sources is $0.0328, and the average time cost is 0.0022 man-hour per image.

- For annotating the images in the ACID dataset, the crowdsourcing performed more efficiently than did the research team using a manual approach. The crowdsourcing approach cost $0.22 per image and required 0.09 man-hour to annotate one image, which is 2.3 times cheaper and 3.8 times faster than annotation by the research team. Meanwhile, crowdsourced annotation reduces the amount of time that would have been required to train new annotators tasked with manual annotation. There are, however, two problems when using crowdsourcing: (1) the precision of crowdsourcing the annotation task is lower in comparison to the annotations conducted by the research team. Bounding boxes produced by crowdsourcing are often larger than the machine objects by more than the allowable 5-pixel tolerance; and (2) many crowdsourced workers did not correctly identify the construction machine types even though instructions were provided, and some machine objects were missed because the workers cannot identify the machine types.

4.6 Summary

In this chapter, a new image dataset, ACID, developed specifically for training deep learning object detection algorithms to recognize construction machines, was described. ACID contains 10,000 annotated images belonging to ten types of construction machines. Four state-of-the-art deep learning algorithms have been evaluated on ACID, and the results show ACID can be used to train deep learning object detection algorithms to detect construction machines with high accuracy and near real-time speed. ACID can be integrated with construction automation studies to recognize machines from images and videos.

The contributions of this research are three-fold: (1) an image dataset of construction machines is developed for deep learning object detection algorithms. Other construction datasets (e.g., workers and materials) can be built by following the same development method; (2) the efficiencies of two methods for annotating the construction dataset are compared. This research provides validation that crowdsourcing the annotation task is the more suitable option in the development of construction image datasets; and (3) an algorithm analysis is conducted on ACID. The results prove the feasibility of using deep learning object detection algorithms to recognize construction machines in images and videos.

Chapter 5: DEEP LEARNING-BASED METHOD OF AUTOMATICALLY DETECTING CONSTRUCTION VIDEO HIGHLIGHTS[2]

5.1 Introduction

Cameras have emerged as an important piece of equipment in construction management, widely used for remote monitoring of job sites. Indeed, construction videos contain important visual information that can serve multiple purposes in project management, such as crew productivity evaluation (Chen et al. 2020), material logistics management (Song et al. 2006), and safety control (Han and Lee 2013). As such, systematic storage of construction video footage is critical with respect to the retrieval, analysis, and documentation of construction activities throughout the project life cycle.

However, the management of construction video footage is difficult because construction videos have a long duration and only a few clips contain useful project information. For example, the footage captured during non-working hours has negligible value for management purposes. Even in working hours, most video clips are useless when project progress is slow. Meanwhile, a large number of video footage is generated in construction projects because cameras are streaming 24 hours per day in job sites. Storage of such amount of video data is challenged in project management because of the limited electronic disk spaces. By removing these unnecessary frames, processed video can replace the raw construction videos for productivity analysis, logistics management, and safety control. By attaching the time stamp and content information

[2] A version of this chapter has been published in *Automation in Construction* as follows: Xiao, B., Yin, X., and Kang, S. (2021). "Vision-based Method of Automatically Detecting Construction Video Highlights by Integrating Machine Tracking and CNN Feature Extraction." *Automation in Construction*, 129, 103817. It has been reprinted with permision from the publisher.

(e.g., objects and activities), the condensed video can be stored economically, more easily indexed, and efficiently retrieved for project management purposes.

Video highlight detection is a technology in computer vision that refers to the process of compactly depicting the original video and distilling its important contents into a short, watchable synopsis (Jiao et al. 2018). The video highlights allow users to obtain certain perspectives of a video without having to view the raw footage in its entirety. This technology has enjoyed success in the entertainment field (e.g., sports highlights and films). In construction, video highlight detection can be used to "distill" the raw construction videos and help project managers to quickly understand the salient developments at a given job site.

Generally, highlight detection methods select keyframes based on image feature changes and then combine clips around keyframes to produce video highlights. However, for three reasons in particular, these feature-based methods are not able to efficiently detect construction video highlights: (1) the performance of existing methods needs to be improved. Unexpected illumination changes in construction videos decrease the performance of feature-based methods; (2) keyframes in construction cannot be simply defined as the frames with image features change rapidly; and (3) the video highlights are expected to be interpretable and flexible for construction management.

To address these issues, this chapter proposes a vision-based method to detect video highlights from construction videos. The proposed method explores the context information from videos by tracking construction machines, and then selects object keyframes by analyzing the content information as prescribed by pre-defined construction rules. In parallel, CNN is employed to extract features from each frame, while the feature keyframes can be selected by calculating the feature changes. As such, the object keyframes and feature keyframes can be processed to

produce video highlights. The detected video highlights are expected to help project managers to efficiently retrieve and economically store their job site video footage.

5.2 Methodology for Video Highlight Detection

The overview of the proposed method of video highlight detection is illustrated in Figure 5-1. As shown in the figure, two types of keyframes are involved in generating video highlights: object keyframes and feature keyframes. Object keyframes are the frames that contain important construction management information related to continuous activities (e.g., machines accessing the working zone). Feature keyframes are the frames where the image feature changes significantly because of scene changes (e.g., camera zooming, edited changeover, task changes). In this research, the object keyframes are used to distill the important information from video clips in which construction machines appeared, while the feature keyframes are used to identify notable developments on the site by scanning the entire video.

First, the input video is processed by the machine detection and tracking module to produce the tracking results, including machine categories, machine ID, and the corresponding pixel locations of machines. The tracking results are stored in a database, and can be conveniently processed by structured query language (SQL). Then, a rule-based method is used to select object keyframes by applying pre-defined construction rules in analyzing the tracking results. These rules are deployed to explore the working zone, working status, and working interaction information of construction machines. For feature keyframe selection, the ResNet50 is employed to extract high-level features from all frames of the input video. The features across frames are evaluated using cosine similarity to select the keyframes that represent scene changes. Finally, object keyframes and feature keyframes are combined together in the video editing module to remove the duplicated keyframes and generate the video highlights.

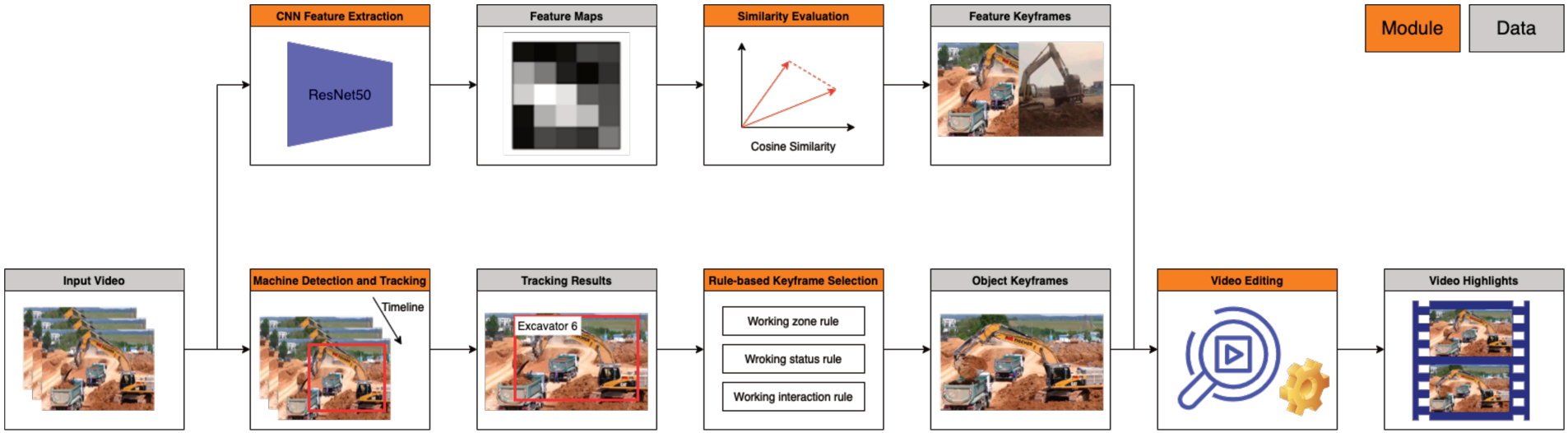


Figure 5-1. Overview of proposed highlight detection method

5.2.1 Machine detection and tracking[3]

The machine detection and tracking module tracks construction machines from the input video sequences by integrating object detection in order to generate information such as machine categories, IDs, and pixel locations. A robust tracking method that produces precise bounding boxes of construction machines is the foundation of the object keyframe selection. In the proposed video highlight detection method, a novel tracking method called construction machine tracker (CMT) is proposed by integrating the deep learning detection method YOLO-v3 for tracking multiple construction machines from videos. Figure 5-2 shows the overall framework of CMT, which consists of three major processes including detection, association, and linear assignment.

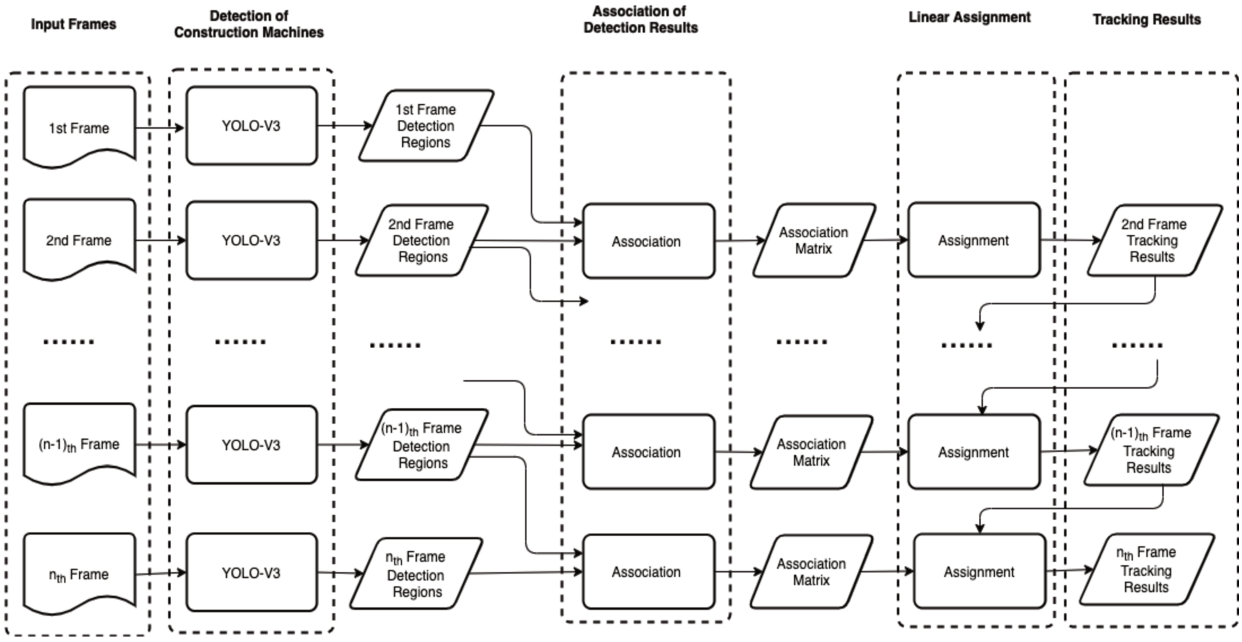


Figure 5-2. Overall framework of CMT integrating YOLO-v3

[3] A version of the machine detection and tracking module has been published in the ASCE *Journal of Computing in Civil Engineering* as follows: Xiao, B., and Kang, S. (2021). "Vision-Based Method Integrating Deep Learning Detection for Tracking Multiple Construction Machines." *Journal of Computing in Civil Engineering*, 35(2), 04020071. It has been reprinted with permision from the publisher.

Detection of construction machines

In the detection process, all images in the construction video are resized into 416×416, while the resized images are processed by YOLO-v3 to produce the detection results. In this process, the ACID dataset described in Chapter 4 is adopted for the purpose of training YOLO-v3. It is also important to track commuter cars in some construction scenarios (e.g., construction gate scenario); as such, we randomly select 2,000 car images containing 3,895 car objects from the COCO dataset to combine with ACID for training purposes. YOLO-v3 is selected because of its high processing speed and reliable perfromance, which has achieved the detection speed of 26.3 fps and the mAP of 87.8% in ACID dataset. Moreover, YOLO-v3 has adopted multiple-scale CNNs for detecting small objects, and this mechanism is helpful in construction scenarios since cameras are usually installed in high positions on construction sites and construction machines and therefore only cover small pixel areas.

Association of detection results

Once the detection windows, herein referred to as D_i (i.e., the set of all detection windows in frame i) has been produced, the association process matches D_i with the detection windows from the previous frame, D_{i-1}, based on machine categories and the similarity of detection. The similarity of detection is qualified by image hashing association and IoU association. The image hashing association determines the object relationship based on pixel features and the IoU association determines the pixel location relationship. Figure 5-3 shows a simulation of the association process where the blue windows represent D_i and the red windows represent D_{i-1}, which will together be used as an example to demonstrate the association process.

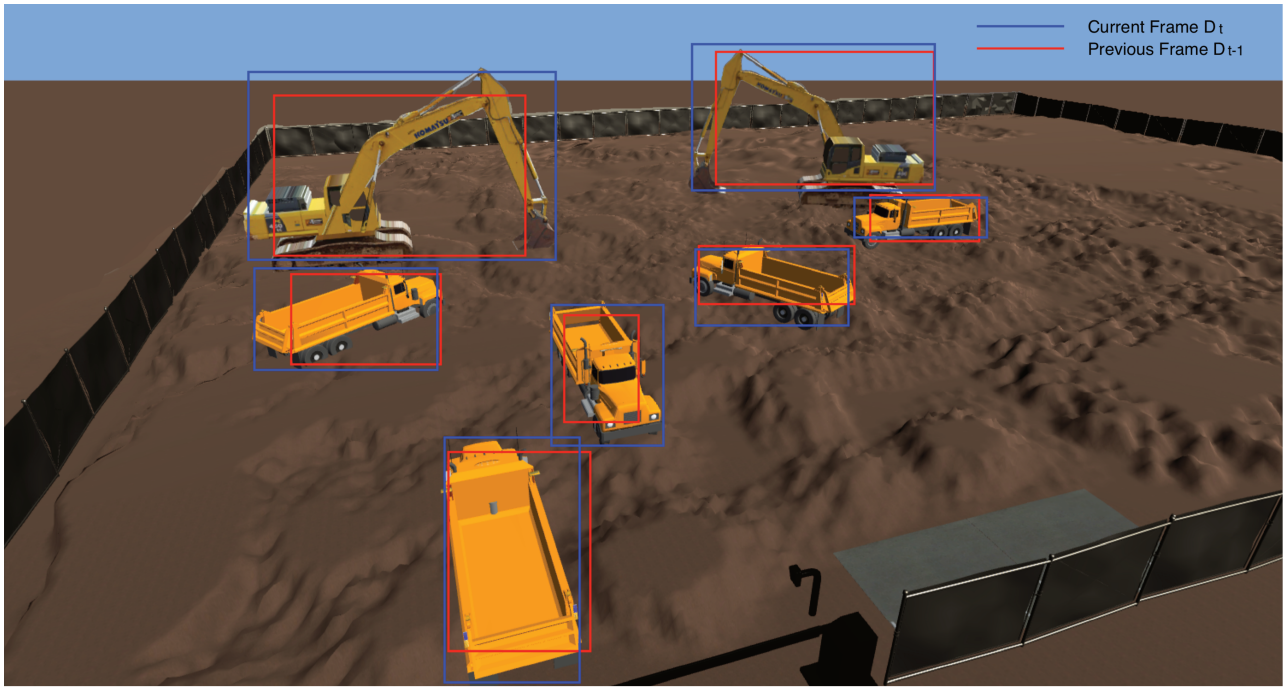


Figure 5-3. Simulation of the association of two consecutive frames

First, detection windows D_i and D_{i-1} are grouped by machine category before IoU association and hashing association, which means, for example, the excavator windows will only be associated with excavators and will never be associated with other machines. The machine category matching is aimed at preventing any mis-tracking and re-identification problems caused by two types of machines (e.g., excavator and dump truck) working in close proximity or causing occlusions.

IoU is commonly used to describe the pixel position relationship between two rectangular windows, which is defined in Equation 5-1.

$$IoU(a,b) = \frac{Area(a) \cap Area(b)}{Area(a) \cup Area(b)} \quad \text{(Equation 5-1)}$$

The variables a and b represent two individual windows. If IoU has a high value, it means there is more overlap between the two windows and they are more likely matched. For each type of

machine, the IoU for each pair of detection windows at two consecutive frames, D_i and D_{i-1}, is calculated to construct the IoU matrix $I_{machine}$. Two IoU matrixes are built based on the Figure 5-3 example for excavator and dump truck, respectively, and they are shown in Equations 5-2 and 5-3. In each matrix, the IoU_{xy} represents the IoU of detection window x at frame i and the detection window y at the previous frame $i-1$.

$$I_{dump\_truck} = \begin{bmatrix} IoU_{11} & IoU_{12} & IoU_{13} & IoU_{14} & IoU_{15} \\ IoU_{21} & IoU_{22} & IoU_{23} & IoU_{24} & IoU_{25} \\ IoU_{31} & IoU_{32} & IoU_{33} & IoU_{34} & IoU_{35} \\ IoU_{41} & IoU_{42} & IoU_{43} & IoU_{44} & IoU_{45} \\ IoU_{51} & IoU_{52} & IoU_{53} & IoU_{54} & IoU_{55} \end{bmatrix} \quad \text{(Equation 5-2)}$$

$$I_{excavator} = \begin{bmatrix} IoU_{11} & IoU_{12} \\ IoU_{21} & IoU_{22} \end{bmatrix} \quad \text{(Equation 5-3)}$$

The steps of calculating image hashing similarity (Coltuc 2000) are depicted in Figure 5-4. For each object, the original pixel region is converted to a greyscale image and resized to a scale of 8×8. Then, the average of all grey values of the 8×8 greyscale image is calculated and the pixels are then examined one by one. If the grey value is larger than the average, a 1 is added to the hash, otherwise, a 0 is added. Therefore, a 64-bit hash is generated for each object by performing the average hashing, and the similarity of two objects has been translated to the similarity between two hashes. The hashing similarity can be calculated by the number of bit positions in which the two bits are the same.

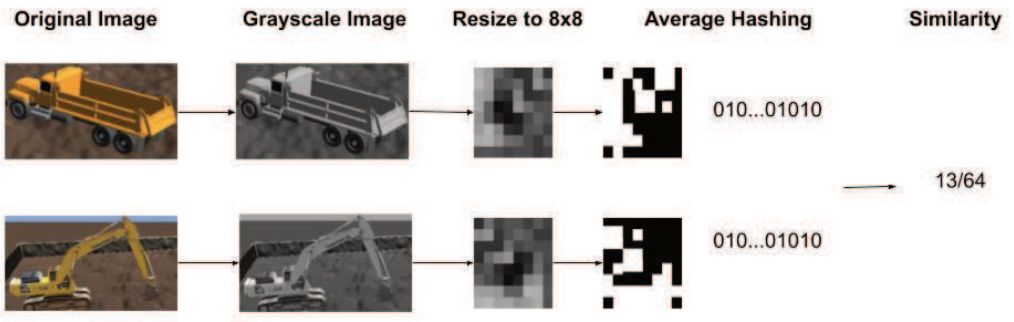


Figure 5-4. Steps in calculating image hashing similarity

Similar to IoU association, the hashing matrix $H_{machine}$ is constructed for each type of machine. Two hashing matrixes, as shown in Equations 5-4 and 5-5, are built from the Figure 5-3 example for excavator and dump truck, respectively. For each matrix, h_{xy} represents the pixel similarity of the detection window x at frame i and the detection window y at the previous frame $i-1$.

$$H_{dump\_truck} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} & h_{15} \\ h_{21} & h_{22} & h_{23} & h_{24} & h_{25} \\ h_{31} & h_{32} & h_{33} & h_{34} & h_{35} \\ h_{41} & h_{42} & h_{43} & h_{44} & h_{45} \\ h_{51} & h_{52} & h_{53} & h_{54} & h_{55} \end{bmatrix} \quad \text{(Equation 5-4)}$$

$$H_{excavator} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \quad \text{(Equation 5-5)}$$

Based on the IoU and image hashing association results, the final association matrix $A_{machine}$ can be calculated by linear combination using Equations 5-6 and 5-7.

$$A_{machine} = \alpha \times I_{machine} + (1-\alpha) \times H_{machine} \quad \text{(Equation 5-6)}$$

$$a_{xy} = A_{machine}[x, y] = \alpha \times IoU_{xy} + (1-\alpha) \times h_{xy} \quad \text{(Equation 5-7)}$$

where $0 < \alpha < 1$. The association matrix $A_{machine}$ describes the relationship between each pair of objects at the current frame and the previous frame for a specific machine. The larger value in the association matrix means these two detection windows have a higher likelihood of being matched. In this research, the α equals to 0.5.

Linear assignment

The assignment process aims to assign each detection window in D_i to one detection window in the previous frame D_{i-1}. For each association matrix $A_{machine}$, the assignment process can be formulated according to Equation 5-8.

$$maximum: \sum_{x=1}^{m} \sum_{y=1}^{n} a_{xy} * v_{xy} \quad \text{(Equation 5-8)}$$

$$v_{ij} = \begin{cases} 1, & if\ the\ detection\ window\ x\ is\ assigned\ to\ detection\ window\ y \\ 0, & if\ the\ detection\ window\ x\ is\ not\ assigned\ to\ detection\ window\ y \end{cases}$$

which is subject to:

$$\sum_{x=1}^{m} v_{xy} = 1,\ when\ y = 1, \dots, n \quad \text{(Equation 5-9)}$$

$$\sum_{y=1}^{n} v_{xy} = 1,\ when\ x = 1, \dots, m \quad \text{(Equation 5-10)}$$

where the variables m and n are the number of detection windows at the frame i and frame $i-1$, respectively, belonging to a particular machine category. In the present study, the Jonker-Volgenant algorithm is adopted to solve the linear assignment problem due to its high processing speed (Jonker and Volgenant 1987).

The assignment process returns one of three statements: matched windows, unmatched windows, and new entry windows. If a detection window d_x ($d_x \epsilon D_i$) is matched with a previous frame detection window d_y ($d_y \epsilon D_{i-1}$) and the corresponding $a_{xy} \geq 0.5$, d_i is the matched window and inherits the tracking ID. If a detection window d_x is matched with d_y, but the corresponding $a_{xy} < 0.5$, the d_x is considered a new entry window and d_y is the unmatched window while the threshold 0.5 is an empirical value. The detection windows that are not matched with a previous detection window are also categorized as new entry windows. If the detection window d_x in the previous frame cannot be matched with any window, d_y will be categorized as an unmatched window. The unmatched windows will be added to the detection results set D_i and will be associated with the detection results set D_{i+1} for the next frame. If unmatched windows cannot be updated to be associated with any windows in ten continuous frames, this window will be deleted for subsequent tracking.

Tracking results

A database is created to store the tracking results. The table contains nine attributes: frame number, time stamp, if_tracked, machine category, machine ID, cx, cy, w, and h. The frame number attribute indicates the sequencing of the current frame, and the time stamp attribute

shows the time of the current frame in the video, accurate to the second. Meanwhile, the if_tracked attribute is a Boolean value that indicates whether any machine has been identified in the frame. If there is a machine object in the current frame, the type of machine and its ID number will be stored in the machine category attribute and the machine ID attribute, respectively. The cx and cy attributes indicate the pixel coordinates of the centroid point of the bounding box, while the w and h attributes refer to the width and height of the machine bounding box, respectively. Using the database, the tracking results can be organized in a structured format within the database and conveniently analyzed by the rule-based keyframe detection module.

5.2.2 Rule-based keyframe detection

The purpose of this module is to select object keyframes by integrating predefined construction rules and tracking results. Three types of construction rules—working zone rule, working status rule, and working interaction rule— are proposed, where Table 5-1 summarizes the definition and target of each rule.

Table 5-1. Summary of predefined construction rules

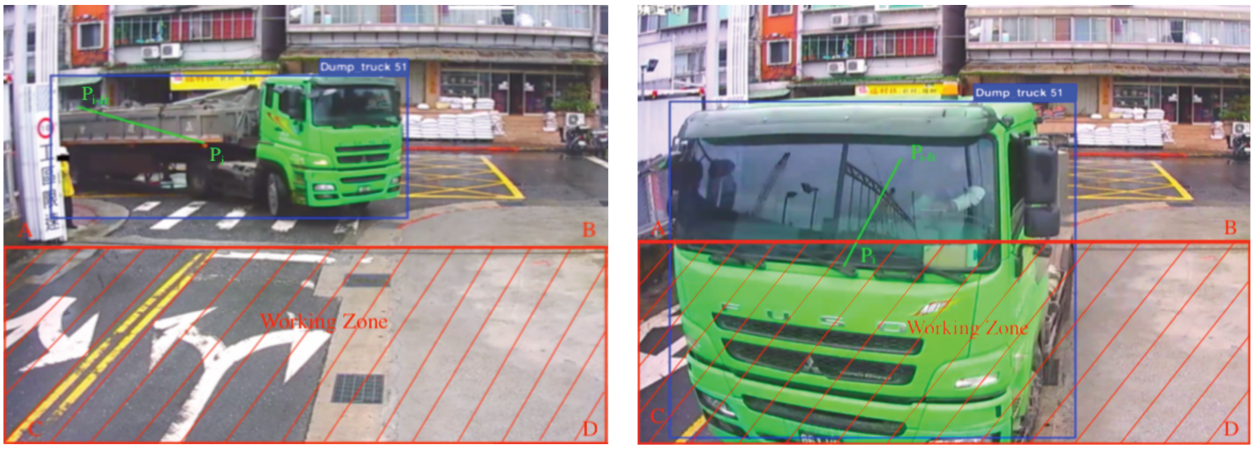
Rule name	Rule definition	Rule target
Working zone	Any frame that contains interested machines entering or leaving the working zone should be considered a keyframe.	Site safety and logistics management
Working status	Any frame that contains machine working status changes between working and idling in the working zone should be considered a keyframe.	Productivity analysis
Working interaction	Any frame that contains extensive overlap between cooperating machines in the working zone should be considered a keyframe.	Site safety and productivity analysis

Working zone rule

Working zone control is important for site safety and resource logistics in construction management. For instance, there is a risk of collisions between machines and pedestrians when machines access the working zone in some scenarios (e.g., road maintenance construction). The time stamp of machines accessing the working zone also indicates the actual scheduling information that can be compared with the planned schedules for logistics management purposes. Therefore, the frames that feature interested machines entering or leaving the working zone are selected as keyframes in the present study.

Equation 5-11 shows the judgement criterion underlying the working zone rule, where A_{ABCD} is the area of the working zone polygon ABCD, P_i is the pixel location of the machine object's central point in frame i, and fr is the frame rate of the video. Connecting the location of the central point at the current frame, i, and the location at frame $i - fr$ can generate a segment P_iP_{i-fr}. If the segment P_iP_{i-fr} has more than 0 intersections with the polygon area A_{ABCD}, frame i is selected as the keyframe. Figure 5-5 shows an example of application of the working zone rule. In Figure 5-5(a), segment P_iP_{i-fr} has no intersection with the working zone ABCD and should be ignored for keyframe selection. In Figure 5-5(b), the dump truck is entering the working zone, while segment P_iP_{i-fr} has one intersection with the working zone ABCD. Therefore, this frame should be selected as a keyframe based on the working zone rule.

$$\mathrm{Count}\left(P_iP_{i-fr} \cap A_{ABCD}\right) > 0 \qquad \text{(Equation 5-11)}$$



(a) Ignored frame (no intersection) (b) Keyframe (intersection with working zone)

Figure 5-5. Example of keyframe selection applying the working zone rule

Working status rule

Identification of the frames that contain working status changes of construction machines is an essential task for productivity analysis, as this information can be used to automatically calculate the machine idling time and efficiency factor. The working status rule selects keyframes in which the status of the interested machine changes from idling to working or from working to idling. This rule, it should be noted, is only interested in the machine status changes occurring in the working zone (i.e., the centroid point of the machine object must be in the working zone). (Even when a machine is idle, it should be noted, the pixel location of this object may change slightly because of the tracking bounding box precision.)

The judgement criterion underlying the working status rule at frame i is defined as per Equation 5-12, where cx_i and cy_i represent the x- and y-coordinates of the central point of the machine object, respectively, fr is the video frame rate, and $k \in \mathrm{N}$. Equation 5-12 calculates the average distance between the central points in the current frame, i, and its previous frame, $i-1$, in fr continuous frames. When the average distance is greater than d_1, the machine status is considered to be "working". The machine status is "idling", meanwhile, if the average distance is

less than d_2. When the average distance is between d_1 and d_2, the current frame indicates the machine is in transition between working and idling status, and as such it should be selected as a keyframe. The variables d_1 and d_2 are threshold values and need to be set for the given construction scenario.

$$d_1 > \frac{1}{fr}\sum_{k\in(i-fr,i]} \sqrt[2]{(cx_k - cx_{k-1})^2 + (cy_k - cy_{k-1})^2} > d_2 \qquad \text{(Equation 5-12)}$$

Working interaction rule

A certain level of interaction between two construction machines is often indicative of a meaningful moment with respect to crew productivity analysis and safety monitoring. For example, high overlap between the excavator and the dump truck in earthmoving represents a loading activity, which can be used for cyclic productivity calculation. High overlap between two dump trucks, meanwhile, may signify a potential collision and may be of interest for safety alerting purposes. In this research, the working interaction rule selects keyframes by analyzing the overlap between two interested construction machines in the working zone. To apply the working interaction rule, the IoU between two machine objects, m and n, at the current frame i is calculated by means of Equation 5-1, where $k \in \mathbb{N}$. If the average IoU in fr continuous frames is greater than threshold a (see Equation 5-13), the current frame is considered a keyframe. Figure 5-6 shows an example of an application of the working interaction rule. In Figure 5-6 (a), the dump truck and the wheel loader are overlapping. If these two machines are in the working zone and the average IoU is greater than a, this frame should be selected as the object keyframe. In Figure 5-6 (b), the excavator and dump truck have no interactions, so this frame will not be selected as a keyframe.

$$\frac{1}{fr}\sum_{k\in(i-fr,i]} IoU(m_k, n_k) > a \qquad \text{(Equation 5-13)}$$

(a) (b)

Figure 5-6. Example of keyframe selection using working interaction rule

To apply the abovementioned rules successfully in construction scenarios, two strategies need to be considered: (1) each type of construction rule should be considered a "blueprint", where several individual rules can be generated by changing the interested classes of machines (for example, two working interaction rules can be generated in the earthmoving scenario, where one focuses on the excavator and dump truck and another focuses on the wheel loader and dump truck); and (2) it is not necessary to apply all three types of construction rules to the same construction scenario. The procedure for generating individual rules consists of four steps: selecting the type of construction rule, defining the working zone, selecting the interested construction machines, and setting up threshold values (if needed). The keyframes detected by each individual rule are simply combined together and inputted to the video editing module.

5.2.3 CNN feature extraction and similarity evaluation

The CNN feature extraction and similarity evaluation are employed to detect feature keyframes. In the present study, feature keyframes are used for two purposes: (1) to represent video clips that have no machine objects; and (2) as an addition to object keyframes in video clips that do have machine objects, since feature keyframes are more effective than object keyframes for

describing scene changes (e.g., camera zooming, moving, and length transition). Compared with manually designed features, such as SIFT, CNN has been shown in previous studies to be more effective in representing construction images (Ha et al. 2018; Kolar et al. 2018).

In CNN feature extraction, all frames in the construction video are processed with the CNN neural networks to produce feature vectors for the purpose of representing original frames. In this research, the ResNet50 neural network (He et al. 2016) is employed for feature extraction due to its excellent performance in computer vision applications. The ResNet50 has 50 layers of neural networks for implementing the residual block, where the residual block is defined as per Equation 5-14.

$$y = \mathcal{F}(\mathrm{X}) + \mathrm{X} \qquad \text{(Equation 5-14)}$$

where X is the input feature map, $\mathcal{F}(\mathrm{X})$ is the feature map processed by the stacked layers, and y is the output feature map of the residual block.

As shown in Figure 5-7, the residual block is a "shortcut connection" that adds the outputs of the stacked layer $\mathcal{F}(\mathrm{X})$ to the input feature map X, where this residual learning solves the gradient vanishing problem in training the deep neural networks. In the CNN feature extraction module, all frames of the input video are first resized into 224×224 resolution. The resized frames are then inputted to the ResNet50, which has been pretrained on the ImageNet dataset (Russakovsky et al. 2015) for forward propagation. A vector with dimensions of 2,048×1 can then be extracted from the flatten layer as the output of this module.

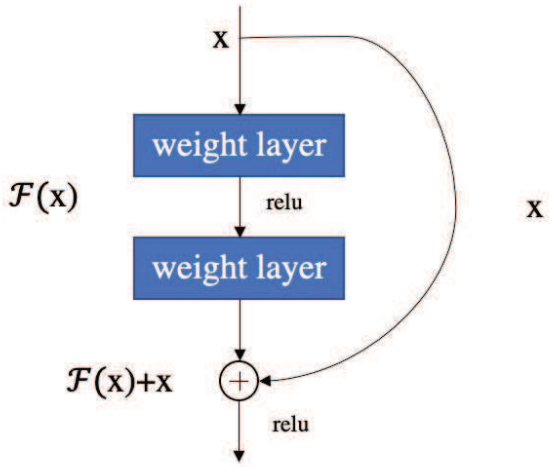


Figure 5-7. Illustration of residual block

The purpose of the similarity evaluation module is to select feature keyframes based on the average similarity AS at each frame. To calculate AS, we first define the similarity $S(m,n)$ of two frames (i.e., m and n) as the cosine similarity (Nguyen and Bai 2011) of their corresponding feature vectors (as shown in Equation 5-15).

$$S(m,n) = \frac{v(m)v(n)'}{\|v(m)\|\|v(n)\|} \quad \text{(Equation 5-15)}$$

where $v(m)$ and $v(n)$ are the feature vectors processed by ResNet50 for frame m and frame n, respectively, and $\|v(n)\|$ is the norm of vector $v(n)$.

Then, the average similarity $AS(i)$ at frame i (defined in Equation 5-16) is calculated as the average of similarity between the feature vectors of the frame i and the frame $(i-fr)$ in one continuous second where fr is the video framerate and $k \in \mathrm{N}$.

$$AS(i) = \frac{1}{fr}\sum_{k\in(i-fr,i]} S(k,k-fr) \quad \text{(Equation 5-16)}$$

If $AS(i)$ is smaller than threshold value, s, the current frame, i, is considered to be a feature keyframe. Here, the smaller the value of s that is adopted, the fewer feature keyframes will be

detected. In construction videos, continuous frames usually have high similarity because construction activities change in a relatively gradual manner. In the present case, the threshold s, at just 0.9, is relatively small. Because the role of the similarity evaluation module is to detect significant feature changes resulting from scene changes.

5.2.4 Video editing

The function of the video editing module is to produce video highlights based on detected object keyframes and video keyframes. This is carried out in two steps: redundancy removal and video concatenation. It should be noted that the detected object keyframes and feature keyframes are intervals of sets of frames rather than discrete frames. The object keyframes can be represented as $T_{object} = \{[s,e]_1, [s,e]_2, \dots, [s,e]_i\}$, where $[s,e]_i$ is a time interval of keyframes, i is the index, and s and e are the start- and end-frame number of the time interval, respectively. It is possible that the time interval may have only a few frames due to tracking errors. As such, any time intervals that have fewer than five frames ($e - s < 5$) are first removed. To generate useful and understandable video highlights, each video clip should be several seconds in length at a minimum in order for users to understand what is occurring in the highlight. In consideration of this, we expand the time interval $[s,e]_i$ to $[s',e']_i$ as per Equation 5-17. This equation calculates the median frame of the time interval $[s,e]_i$ and then finds the n seconds before and after the central frame as the basis for determining the new time interval, where the present research assigns n a value of 2. After this step, all time intervals have the same length of 4 seconds.

$$[s',e']_i = [floor(\tfrac{s+e}{2}) - n \times fr, floor(\tfrac{s+e}{2}) + n \times fr]_i \quad \text{(Equation 5-17)}$$

It is possible that the different construction rules will locate adjacent, overlapping, or identical keyframes. In other words, many time intervals in T_{object} are redundant and will need to be removed. For two continuous time intervals, we remove the first interval $[s',e']_i$ if $s'_{i+1} - s'_i \leq$

fr. If two continuous time intervals are close to one another ($2n \times fr > s'_{i+1} - s'_i > fr$), they are merged to a new interval $[s'_i - n \times fr, e'_{i+1} + n \times fr]$. The same process is conducted with respect to the feature keyframes $T_{feature}$.

The processed T_{object} and $T_{feature}$ can then be used to produce video highlights by extracting the corresponding frames from the original construction video and concatenated these frames together. It should be noted that the object keyframes and feature keyframes may be overlapping. In the present study, overlapping frames are not removed. Instead, all object keyframes and feature keyframes are merged in the final video highlights as the final keyframe set T. Users are thereby able to recognize whether a given video highlight frame belongs to object or feature keyframes.

5.3 Implementations and Evaluation Metrics

In this section, first the implementation of the proposed video highlight detection method is introduced. The evaluation metrics to validate the proposed method are also illustrated.

5.3.1 Implementations

The proposed method is programmed in Python 3.6, and the Opencv library is adopted for the video input/output. The YOLO-v3, originally programmed in C, is implemented via Python wrapper with an acceleration of CUDA 9.0 and Cudnn 7.0. In the CMT method, the Jonker-Volgenant algorithm was originally implemented in C++ language and integrated into the proposed method. Moreover, the rule-based keyframe selection module is built using the SQLite-Python library, whereas the construction rules are implemented using SQL queries. The ResNet50 is implemented using the Pytorch library, while the cosine similarity is built using the

scikit-kearn library. For video editing, the Moviepy library is employed to generate the final video highlights.The proposed method is tested in an Ubuntu 18.04 64-bit system environment.

For the hardware configuration, the proposed method is tested on a computer with a NVIDIA GTX 1080Ti graphics card, 11 GB memory, an Intel Core i9-7920X@2.90 Hz CPU with 12 cores, and two 32 GB memory cards. The processing speed when implementing the proposed method is approximately 7 frames per second. It should be noted that the graphics card specifications affect the speed of executing YOLO-v3 and ReNet50. As such, the processing speed can be increased by upgrading to an advanced graphics card or implementing parallel programming.

5.3.2 Evaluation metrics

Following the protocols set out in previous work (Zhang et al. 2016), precision, recall, and F1 score are employed as the evaluation metrics in the present study, where A denotes the video highlights generated by the proposed method, and B denotes the annotated ground truth video highlights. Precision, meanwhile, is the measurement of how accurate the highlight detection method is (see Equation 5-18), while recall measures how effective the highlight detection method is in identifying the correct highlight clips according to Equation 5-19. The F1 score is the harmonic mean of precision and recall as defined in Equation 5-20.

$$Precision = \frac{Number\ of\ correct\ highlight\ clips}{Number\ of\ highlight\ clips\ in\ A} \quad \text{(Equation 5-18)}$$

$$Recall = \frac{Number\ of\ correct\ highlight\ clips}{Number\ of\ highlight\ clips\ in\ B} \quad \text{(Equation 5-19)}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{(Equation 5-20)}$$

where the correct highlight clip is decided by the temporal intersection over union (TIoU) between the generated highlight clip a ($a \in A$) and the ground truth highlight clip b ($b \in B$), as

expressed in Equation 5-21. Figure 5-8 illustrates the process of computing TIoU. If the value of TIoU is greater than 0.5, the highlight clip a is considered to be a correct highlight clip.

$$TIoU = \frac{duration\ (a) \cap duration(b)}{duration\ (a) \cup duration(b)} \quad \text{(Equation 5-21)}$$

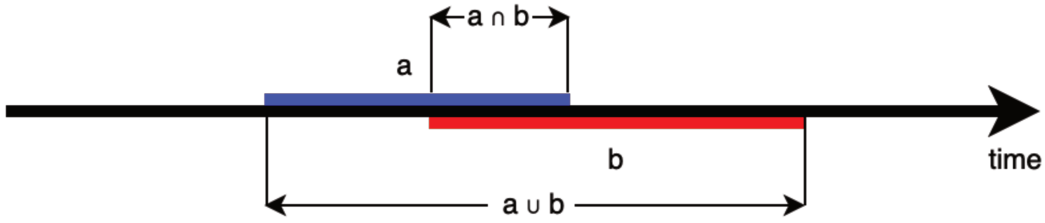


Figure 5-8. Illustration of the computation of TIoU

5.4 Case Study 1: Construction Gate

In case 1, the proposed video highlight detection method was tested on construction gate video footage captured from a gate specifically for machine traffic, with dump trucks, concrete mixer trucks, dozers, and cars all appearing in this footage. The construction gate scenario was adopted as a case study in this research for two reasons: (1) construction gate video footage contains important information about what equipment is present on the construction site at a given time (i.e., timestamped arrivals and departures of equipment), which is crucial for construction gate control; and (2) the need for video highlights is particularly pressing for gate scenarios since almost all construction gates feature video cameras capturing large volumes of raw video footage.

5.4.1 Experimental setup

Three one-hour gate videos captured from the same construction gate were used in the experiment. Figure 5-9 shows example images for each test video. The relevant information regarding the test videos (duration, resolution, frame rate, are number of highlights) is summarized in Table 5-2. Three test videos corresponding to different times of day, i.e., morning,

afternoon, and evening, were captured in order to investigate the feasibility of the proposed method under different illumination conditions.

(a) Gate-Video1 (captured in the morning)

(b) Gate-Video2 (captured in the afternoon)

(c) Gate-Video3 (captured in the evening)

Figure 5-9. Example images from test videos in construction gate case

Table 5-2. Specifications of test videos for construction gate case

	Duration (minutes)	Video Resolution	Frame rate (fps)	Number of highlights contained
Gate-Video1	60	1920×1080	12	20
Gate-Video2	60	1920×1080	12	19
Gate-Video3	60	1920×1080	12	14

To evaluate the performance of the proposed approach, the ground truth video highlights in each test video had to be manually annotated. Of course, the annotation of video highlights is an inherently subjective task since there is no absolute definition of what constitutes a highlight. However, construction engineers are likely to share similar points of view with regard to what constitutes a useful video highlight of construction site footage for construction management purposes based on their experience, knowledge, and intuition. In our research, five graduate students majoring in construction management were invited to manually identify highlights from construction video footage. Their annotations of these highlights consisted of a time stamp of the highlight and a short description (e.g., "From 27:01 to 27:05: A dump truck exits the gate and turns right"). A two-step strategy was implemented for video highlights annotation: (1) each participant was asked to find the video clips in which machines access the construction gate, the camera working state changes, or unusual activities occur, or other clips they think may be highlights; and (2) the author of this research manually browses the video highlights annotated by all participants to decide the final video highlights as ground truth. The annotated video highlights were then used to assess the proposed method.

In case 1, the working zone rule and working status rule were applied, with the detailed configurations of these two rules summarized in Table 5-3. The working zone rule was applied to

detect video highlights featuring machines accessing the gate. Machines suddenly stopping in the gate area, meanwhile, signified potential highlights that could be detected by the working status rule.

Table 5-3. Construction rules applied to construction gate case

Index	**Rule name**	**Machine/s of interest**	**Working zone**	d_1 **(pixel)**	d_2 **(pixel)**
1	Working zone	Dump trucks, concrete mixer trucks, dozers, and cars	[(0,600), (1920,600), (0,1080), (1920,1080)]	N/A	N/A
2	Working status	Dump trucks, concrete mixer trucks, dozers, and cars	[(0,600), (1920,600), (0,1080), (1920,1080)]	60	10

To validate the feasibility of the proposed highlight detection method, a baseline method has been proposed by removing the machine tracking module and the rule-based keyframe detection module from the proposed method. Figure 5-10 shows the overview of the baseline method. The baseline method retained the same CNN feature extraction module, similarity evaluation module, and video editing module as those in the proposed method. In the baseline method, however, the threshold s in the similarity evaluation module was set at 0.95, a higher value of s than that employed in the proposed method, in order to produce more video highlights.

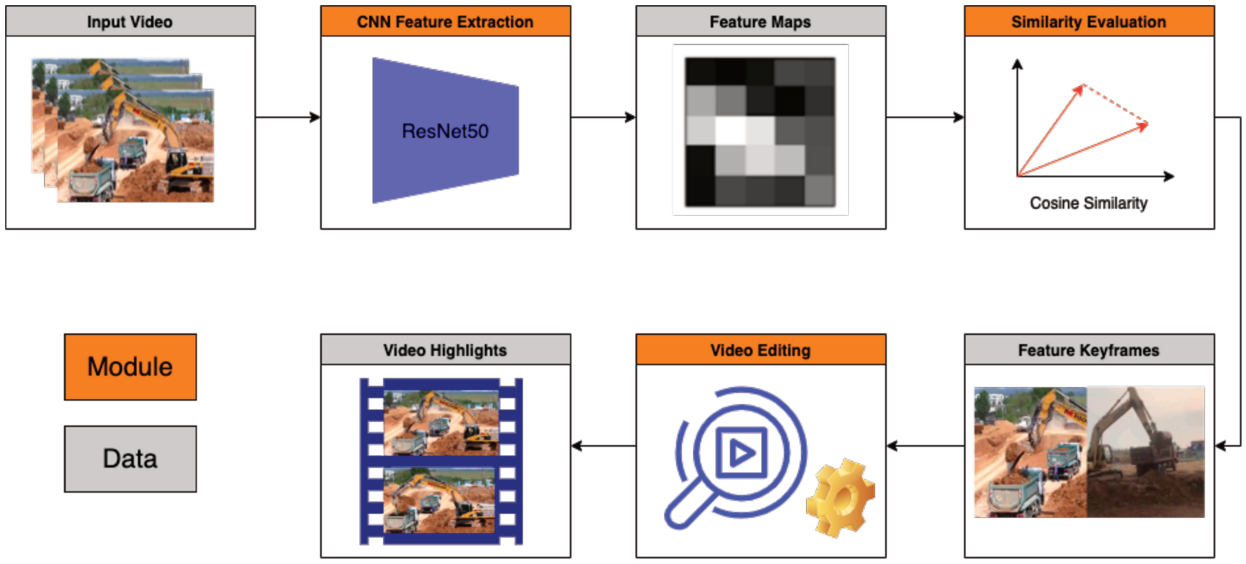


Figure 5-10. Overview of baseline method

5.4.2 Experimental results

Table 5-4 illustrates the experimental results in terms of precision, recall, F1 score, and the number of correct highlights. The proposed method achieved 87.7% on precision, 94.3% on recall, and 90.8% on F1 score on average, 13.8% higher on precision, 14% higher on recall, and 14.1% higher in terms of F1 score compared to the baseline method. Meanwhile, the proposed method detected 2.4 more correct video highlights from each video compared to the baseline method. The experimental results indicate that the proposed method is more robust and precise than the feature-based highlight detection method with respect to the construction gate scenario.

In the testing, the three videos represented different illumination conditions (i.e., morning, noon, and evening). It was found that the performance of the proposed method is stable (around 90% of F1 score) in dealing with different illumination conditions. In contrast, the baseline method achieved F1 scores of 82%, 79.1%, and 69%, respectively, for the three test videos. The baseline method was shown to be less effective in dealing with the night-time illumination condition (Gate-Video3) because the feature-based highlight detection method was sensitive to

illumination variations. The proposed method adopted object keyframes by tracking construction machines from videos, while the machine tracking module was built upon deep learning object detection. Therefore, the proposed method was found to be more robust than the feature-based method in detecting construction video highlights.

Table 5-4. Experimental results of proposed method in construction gate case

		Precision	Recall	F1 score	Correct highlights detected
Gate-Video1	Proposed method	90.0%	90.0%	90.0%	18
	Baseline method	84.2%	80.0%	82.0%	16
Gate-Video2	Proposed method	86.4%	100.0%	92.7%	19
	Baseline method	70.8%	89.5%	79.1%	17
Gate-Video3	Proposed method	86.7%	92.9%	89.7%	13
	Baseline method	66.7%	71.4%	69.0%	10
Average	Proposed method	87.7%	94.3%	90.8%	16.7
	Baseline method	73.9%	80.3%	76.7%	14.3

5.4.3 Video highlights for construction gate control

In construction management, gate control is a critical factor in achieving project success. Construction machines should access the gate at the scheduled time to complete their construction tasks, and the timestamp of machines accessing the gate should be recorded. Construction video highlights can serve the gate control purpose by providing video records and corresponding time stamp. Table 5-5 shows the actual number of instances of machines accessing the gate in the raw video, the number of instances of machines accessing the gate contained in the video highlights, and the accuracy of the three test videos. In case 1, the three

test videos showed 48 records of machines accessing the gate, while 45 access records were found to be contained in the detected video highlights, resulting in an accuracy of 93.8%. This result indicates that the video highlight detection method is reasonably effective for construction gate control, and that the generated video highlights can be useful as a form of project documentation for future reference.

Table 5-5. Summary of machines accessing the gate

	No. of machine accesses in original video	**No. of machine accesses in detected video highlights**	**Accuracy**
Gate-Video1	18	16	88.9%
Gate-Video2	18	18	100.0%
Gate-Video3	12	11	91.7%
Average	48	45	93.8%

5.5 Case Study 2: Earthmoving

Case study 2 focused on earthmoving, where the proposed method was tested on video footage of an excavator working with several dump trucks. Earthmoving, it should be noted, refers to a range of activities that involve excavating soil or rock and moving it to another part of the site, fundamental activities in all types of construction (e.g., residential building, roads, and bridges).

5.5.1 Experimental setup

In case 2, the proposed method was tested on a 40-minute earthmoving video with a resolution of 1280×720 and a frame rate of 30 fps. In the video, a Volvo EC210BLC excavator (bucket payload of 2.1 loose cubic yards (LCY)) works with several dump trucks in an outdoor construction environment and completes several earthmoving cycles. In each cycle, the

excavator digs soil and loads it into a dump truck. After the dump truck is fully loaded, it moves away and another dump truck approaches the excavator for the next cycle.

The earthmoving video footage was manually annotated to obtain the ground truth video highlights by following the same procedure described in reference to the construction gate case (i.e., participants were required to find the video clips of the excavator loading the dump truck, a change in status of the excavator, or any clips that may be of interest for construction management purposes). Through this process, 115 video clips were identified as video highlights in this case study. As with the other case, the feature-based highlight detection method was adopted as the baseline method to test the earthmoving video footage. The configuration of the baseline method was the same as in case 1.

In the earthmoving case, the working zone rule, working status rule, and working interaction rule were applied for detecting object keyframes. The details of these rules are summarized in Table 5-6. It should be noted that the working zone rule and working status rule only target the excavator, since the excavator is the major construction machine in this case and it governs the productivity of the whole crew. The working interaction rule, meanwhile, focuses on cases of overlap between the excavator and dump truck.

Table 5-6. Construction rules applied to earthmoving case

Index	Rule name	Machine/s of interest	Working zone	d_1 (pixel)	d_2 (pixel)	a
1	Working zone	Excavator	[(0,215), (1045,215), (0,720), (1045,720)]	N/A	N/A	N/A
2	Working status	Excavator	[(0,215), (1045,215), (0,720), (1045,720)]	20	10	N/A
3	Working interaction	Excavator and dump truck	[(0,215), (1045,215), (0,720), (1045,720)]	N/A	N/A	0.1

5.5.2 Experimental results

Table 5-7 shows the experimental results of the proposed method and the baseline method in terms of precision, recall, F1 score, and the number of correct highlights. The proposed method detected 111 video highlights, 104 of them being correct highlights, achieving a precision of 93.7%, recall of 90.4%, and F1 score of 92.0%. Meanwhile, the baseline method achieved a precision of 70.9%, recall of 63.5%, and F1 score of 67.0%. As can be seen, the proposed method outperformed the baseline feature-based method by a margin of 22.8% with respect to precision, 26.9% on recall, and 25.0% in terms of F1 score for the earthmoving case. It is also worth noting that, although the earthmoving case contains more extensive video highlights than the construction gate case, the proposed method achieved similar performance for both cases, underscoring the ability of the proposed method to deal with different construction scenarios.

Table 5-7. Experimental results of proposed method in earthmoving case

		Precision	Recall	F1 score	Correct highlights detected
Earthmoving-Video1	Proposed method	93.7%	90.4%	92.0%	104
	Baseline method	70.9%	63.5%	67.0%	73

5.5.3 Video highlights for productivity analysis

In the earthmoving case, the detected highlights were found to contain meaningful video clips of loading activities that would be useful for productivity analysis. As mentioned, in the earthmoving cycle, the excavator digs soil into the bucket and then loads it into a dump truck. Once the dump truck is fully loaded, it moves away and another dump truck approaches the excavator for the next cycle. As such, the number of cycles is equal to the number of loading activities, such that the excavator productivity can be calculated as per Equation 5-22, where the bucket payload per cycle is given by the excavator manufacturer (2.1 LCY, in this case).

$$Productivity = \frac{number\ of\ cycles}{time\ (hr)} \times \frac{bucket\ payload}{cycle}\ (LCY) \qquad \text{(Equation 5-22)}$$

In Earthmoving-Video1, the excavator has completed 99 work cycles in 40 minutes and the ground truth productivity is $311.85 LCY/hr$. By manually analysis the video highlights detected by the proposed method, the author of the present study found 93 video clips of the loading activity. In other words, if the video highlights are used for advanced vision-based method for productivity analysis, the ideally analyzed productivity of Earthmoving-Video1 can reach $292.95 LCY/hr$. The accuracy of the productivity analysis, then, is 93.9%, which means 93.9%

of the relevant productivity information can be retrieved from the detected video highlights without browsing the original construction videos.

5.6 Discussion

The experimental results indicate that the proposed method can successfully produce video highlights from construction videos for the purpose of reducing manual inspection efforts and digital storage requirements. The research findings identified in analyzing the test results are discussed below.

- The proposed video highlight detection approach exhibited better performance than the feature-based method for detection of construction video highlights. In experiments, the proposed method has achieved an average precision of 89.2%, recall of 93.3%, and F1 score of 91.1% for two case studies, respectively (4 videos in total), while the baseline method has achieved the average precision of 73.2%, recall of 76.1%, and F1 score of 74.3%. The proposed method outperforms the baseline method over 10.0% on three evaluation metrics. Technically, the proposed method achieved better performance than existing methods for two reasons: (1) adopting pre-defined construction rules (i.e., working zone, working status, and working interaction) to detect object keyframes by analyzing machine trajectories. As such, the proposed method explores the context information from construction videos and becomes more robust; and (2) employing ResNet50 to detect feature keyframes to describe scene changes in construction videos. The feature keyframes efficiently represent the video clips that have no construction machines, while improve the precision of the proposed method.
- Reducing the amount of construction video footage is a crucial benefit of applying video highlight detection in construction. In this regard, Table 5-8 provides a comparison of the

original raw video and the detected video highlights in terms of duration and storage size in reference to the two case studies. The average size of the original videos is 635.5 MB. After implementing with the proposed highlight detection method, the average size is reduced to 43.8 MB, a reduction in storage size requirement of approximately 93.1%. The average duration of video highlights is 2.77 minutes, while the original videos average 55 minutes in duration. The results indicate that the video highlights generated represent a more watchable synopsis of the raw video, meaning that the use of this method can reduce the amount of effort required in order to maintain construction video documentation.

Table 5-8. Duration and storage size of video highlights in construction gate case

	Original video		**Detected video highlights**	
	Duration (minutes)	**Storage size (MB)**	**Duration (minutes)**	**Storage size (MB)**
Gate-Video1	60	713.1	1.33	35.6
Gate-Video 2	60	713.5	1.30	39.3
Gate-Video3	60	713.3	1.15	27.8
Earthmoving-Video1	40	402.1	7.30	72.5
Average	55	635.5	2.77	43.8

- Compare to the baseline method, the proposed method is less sensitive to illumination changes, as demonstrated in the construction gate case. Most construction sites are outdoors, and as such illumination changes are frequent in construction video footage. Feature-based highlight detection methods are prone to errantly detect frames that contain significant illumination variations as keyframes, decreasing the accuracy of the highlight

detection. In contrast, the proposed method shows stable performance in dealing with illumination changes because it adopts the machine tracking module for object keyframe selection. The CMT tracking method, built upon YOLO-v3 object detection, shows excellent performance in tracking machine trajectories under illumination changes. In this respect, the proposed method exhibits reliable performance even in challenging construction scenarios.

- Compared with feature-based methods, the proposed method has better interpretability and flexibility because it integrates object keyframe selection with feature keyframe selection. The outputs of the proposed method include not only video highlights, but also the intuitive interpretation of selection rationale, such as a machine entering or leaving the frame. This information is beneficial for project management in terms of gate control and productivity analysis, as illustrated in the case studies. Furthermore, in the proposed method, the construction rules can be flexibly customized based on the particular needs of a given construction project. For example, the proposed method can generate video highlights that relate only to a specific construction machine (e.g., dump truck), or movement (e.g., machine leaving the site); this is not possible using feature-based highlight detection methods. The proposed method demonstrates the feasibility of rule-based highlight detection methods in construction scenarios.

5.7 Summary

An effective and efficient method for converting construction video footage into concise video data is in high demand in today's construction industry. In Chapter 5, a novel vision-based method has been proposed to generate video highlights from construction videos, and the objective 2 of this thesis has been achieved in this chapter. The proposed method consists the

following modules: machine detection and tracking, rule-based keyframe selection, CNN feature extraction, similarity evaluation, and video editing. Two case studies were conducted to validate the performance of the proposed method using construction gate and earthmoving video footage, respectively. The proposed method was found to achieve average precision of 89.2% and average recall of 93.3%, outperforming the feature-based highlight detection method. The proposed method can be integrated into several advanced applications that may potentially benefit construction management, including: (1) auto-generating reports from lengthy construction videos; (2) building a query system that searches for clips of interest in the video footage; and (3) quantitatively analyzing construction productivity based on video highlights.

The contributions of the research contained in Chapter 5 are three-fold. First, this research has proposed a novel method to detect video highlights from construction videos, while the proposed method outperforms the baseline method over 10% on robustness and precision. Second, three construction rules have been proposed for object keyframe detection: the working zone rule, the working status rule, and the working interaction rule. By integrating these rules, the detected video highlights are interpretable and flexible, meaning that the resultant construction videos are searchable, filterable, and manageable. Third, the proposed method is shown to be feasible in that it reduces the storage space requirement by over 90% while retaining most of the useful information for construction video documentation.

Chapter 6: GENERATING TEXT DESCRIPTIONS FROM CONSTRUCTION IMAGES BY ADOPTING DEEP LEARNING IMAGE CAPTIONING[4]

6.1 Introduction

Construction videos contain important visual information (e.g., working objects and their activities) that is of benefit for project management purposes. By analyzing construction videos using vision-based methods, many applications can be developed to automatically monitor crew productivity, identify safety risks, and so forth. For example, Chen et al. (2020) developed a vision-based system to calculate excavator productivity in earthmoving. Kolar et al. (2018), meanwhile, proposed a vision-based method for identifying safety guardrail at construction sites in order to prevent workers from accessing hazardous site areas.

Based on the target information, existing vision-based methods can be divided into object recognition, motion recognition, activity recognition, and scene analysis (Liu et al. 2020). The abovementioned four categories of methods target the retrieval of information pertaining to objects (Park and Brilakis 2012), movements (Zhu et al. 2016a), activities (Golparvar-Fard et al. 2013), and relationships between objects (Wang et al. 2019), respectively, from construction images or videos. However, these methods have two limitations: (1) the different scene information (e.g., objects, activities, and relationships between objects) in construction videos is retrieved separately, making it a time-consuming process; and (2) the retrieved scene information is usually combined based on pre-defined orders in order to generate a set of words or a sentence as the final output, but these results are prone to be incomplete. An automated method of generating a complete, concise, and correct sentence that contains the integral information

[4] A version of this chapter is under review for publication in *Automation in Construction* entitled as "Deep Learning Image Captioning in Construction: A Feasibility Study".

pertaining to objects, activities, and relationships is needed in order to improve vision-based monitoring of construction sites.

Image captioning refers to the formulation of one or several sentences to describe the contents of an image. Image captioning is a disciplinary technology rooted in computer vision and natural language processing (Hossain et al. 2019). By incorporating deep learning, image captioning methods can generate precise and concise text descriptions from images by training on annotated image datasets. As shown in Figure 6-1, a typical deep learning image captioning method is the encoder–decoder architecture, where a CNN encoder extracts embedding features from the images and a recurrent neural networks (RNN) decoder generates text based on the embedded features (Mao et al. 2014). By adopting deep learning image captioning, the scene information in construction images can be retrieved integrally in the form of a natural sentence.

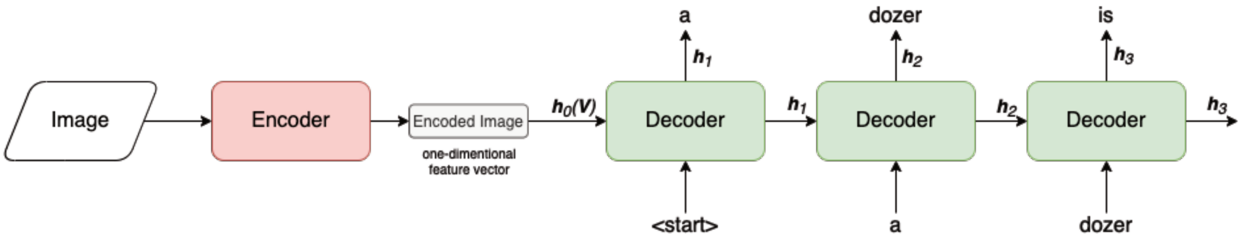


Figure 6-1. Encoder–decoder architecture of deep learning image captioning

Although deep learning image captioning has achieved considerable success within the computer vision community, the performance of deep learning image captioning in construction scenarios has yet to be substantiated. Most studies within this research domain only adopt the basic CNN-RNN method, while other advanced deep learning methods (e.g., attention) have not been tested in construction applications. Moreover, an annotated dataset of construction images is necessary in order to train deep learning image captioning algorithms to generate clear and professional

text descriptions for construction management purposes. A linguistic schema for annotating construction machine images is currently lacking within this research area.

The main objective of the research described in this chapter is to automatically generate text descriptions from construction images by adoption of deep learning image captioning. To achieve this goal, a linguistic schema for annotating construction machine images is proposed, and a captioning dataset is developed based on the ACID dataset described in Chapter 4. Moreover, six state-of-the-art deep learning methods from the computer vision community have been tested on the captioning dataset to investigate their performance in construction scenarios. Finally, the best performing image captioning method is integrated in the proposed video summarization framework to generate text descriptions of keyframes.

6.2 Methodology for Generating Text Descriptions

Figure 6-2 illustrates the methodology used in the present study for generating text descriptions from construction images. First, a linguistic schema for instructing the annotation of construction machine images is proposed. The images from the ACID dataset that is developed in Chapter 4 is annotated according to the linguistic schema, and a captioning dataset is developed in this process. Then, six deep learning image captioning methods are selected because of their reliable performance in computer vision community. The developed captioning dataset is divided into a training set and a validation set for testing six image captioning methods. In evaluation, six methods are compared at the sentence level using five evaluation metrics. After that, the best performing image captioning method is evaluated in the element level to indicate the feasibility of image captioning in construction management. The best performing method is ultimately integrated into the proposed video summarization framework to caption keyframes detected using the method described in Chapter 5.

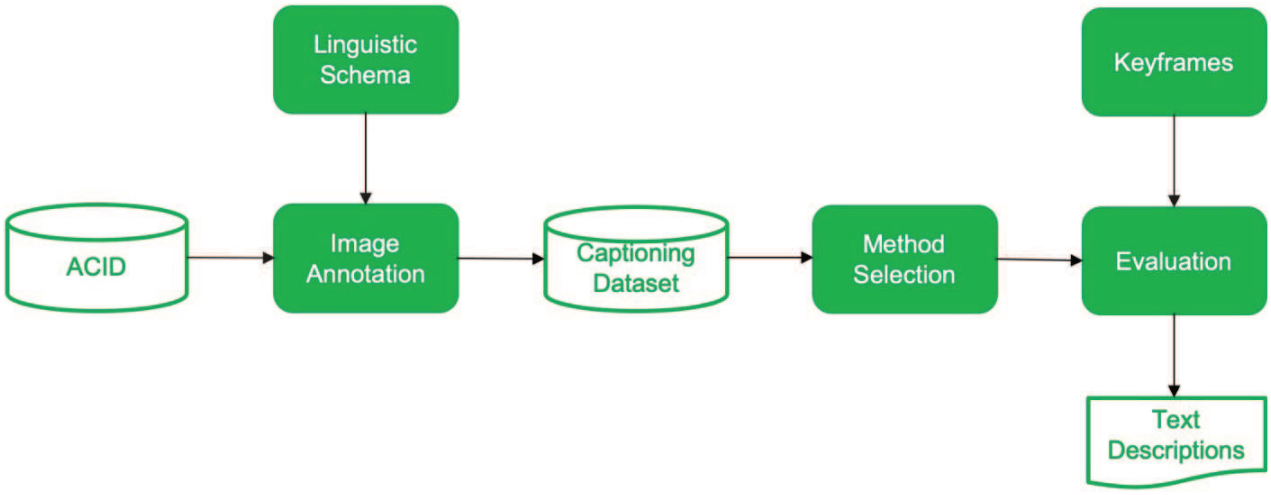


Figure 6-2. Methodology for generating text descriptions from construction images

6.2.1 Linguistic schema and image annotation

The linguistic schema informs the process of annotating construction images, which in turns plays an important role in the application deep learning image captioning in construction. For computer vision tasks, the annotators are required to describe images in their own words because the target images are captured from daily life. In construction, the text annotations of images should be professional and precise for construction management purposes. The annotators are required to use correct terms to describe construction objects, activities, and working contents using the linguistic schema rather than simply using their own words as in some other computer vision applications.

Figure 6-3 shows the linguistic schema used in this study for annotating construction machine images. First, the following elements should be deconstructed from the construction image according to the linguistic schema: (1) the primary machine object; (2) the machine object cooperating with the primary object; (3) the working contents (e.g., dirt, stone, and construction materials) of the primary object; (4) the activities of the primary machine; and (5) supplementary

information, such as color, count, and weather conditions. Then, the correct terms should be selected for matching each element deconstructed in the previous step. Finally, a logical and correct sentence should be formed using words to describe what is occurring in the construction image.

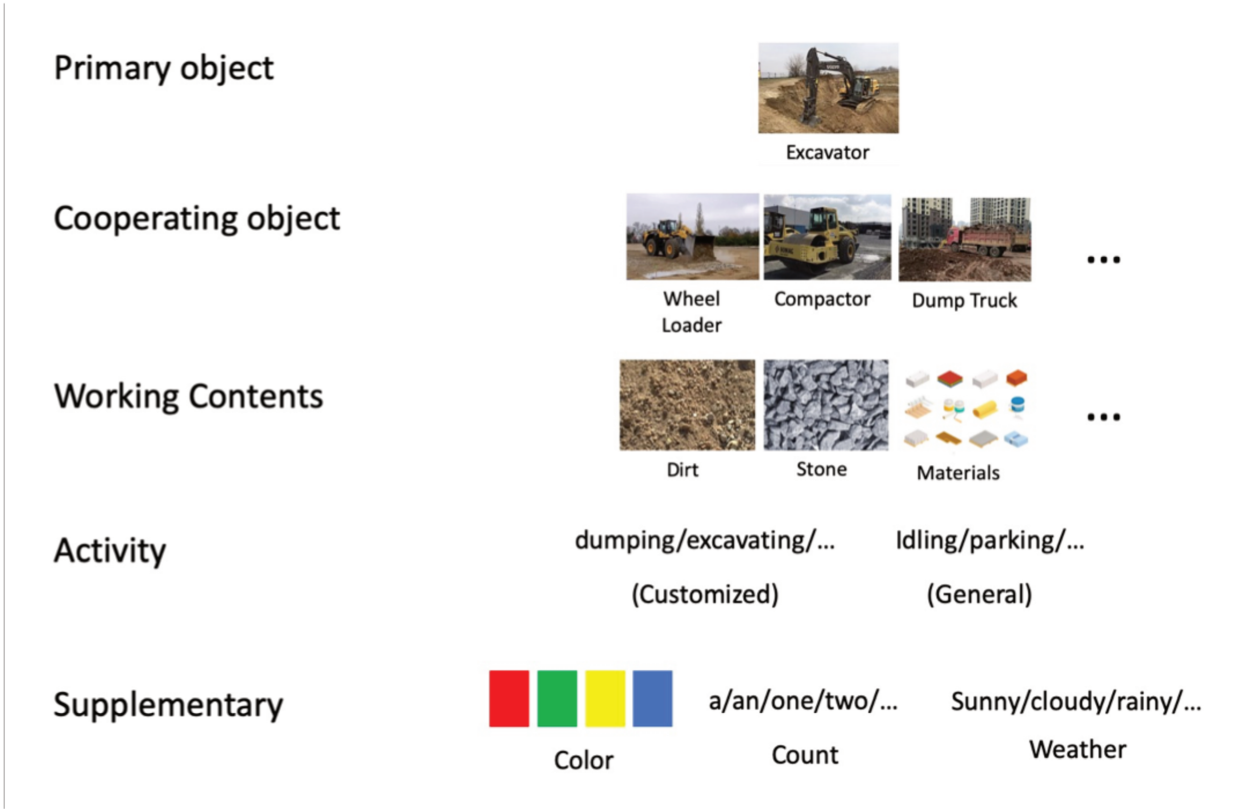


Figure 6-3. Illustration of linguistic schema

Since the construction images are drawn from the ACID dataset, the primary object and cooperating object terms must be selected from among the ten designated construction machine types: excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader, backhoe loader, tower crane, and mobile crane. Furthermore, the author of this study has provided a list of activities (as shown in Table 6-1) for each type of construction machines to serve the activity selection in the linguistic schema, where the general activities can be used for all types of machines. It should be noted that the annotators are encouraged to select the activity

of the primary object from the options listed in Table 6-1, although they are also permitted to use other activity terms based on their construction knowledge/background if needed.

Table 6-1. List of suggested activities for construction machines

Construction Machine	**Customized Activity**
Excavator	swinging/dumping/excavating/loading/etc.
Compactor	compacting/etc.
Dozer	grading/stripping/loosening/pushing/etc.
Grader	grading/stripping/loosening/pushing/etc.
Dump Truck	dumping/hauling/transferring/etc.
Concrete Mixer Truck	dumping/transferring/loading/etc.
Wheel Loader	dumping/excavating/loading/transferring/etc.
Backhoe Loader	dumping/excavating/loading/transferring/etc.
Tower Crane	lifting/transferring/swinging/etc.
Mobile Crane	lifting/transferring/swinging/etc.
General Activity	travelling/waiting/idling/driving/parking/etc.

In the image annotation process, 30 volunteered annotators from the University of Alberta with engineering background were participated in this annotation task. First, all annotators were given a half-hour presentation to introduce the research, including an overview of deep learning image captioning, annotation tasks, and the linguistic schema. Then, the volunteers were assigned a series of construction images and prompted to write one sentence describing the contents of each image. The research team then provide further instruction and feedback in reference to their annotations. Once the annotators had finished their tasks, the research team members manually checked the annotation results and resolved any errors identified.

6.2.2 Captioning dataset summary

Four thousand images from the ACID dataset were annotated to produce the captioning results. Meanwhile, a total of 8,226 captions were annotated, meaning that each construction image was annotated by two annotators on average. Figure 6-4 shows the element distribution in the captioning dataset, including the machine terms and activity terms. It can be observed that the excavator and dump truck are the two object terms appearing most frequently in the captioning dataset, while loading and dumping are the top two activity terms. Table 6-2 outlines the top 20 N-grams in the developed captioning dataset, indicating the most frequent 1-gram, 2-gram, and 3-gram terms and their quantities. The captioning dataset is divided into a training set (80%) and a validation set (20%) for evaluation purposes.

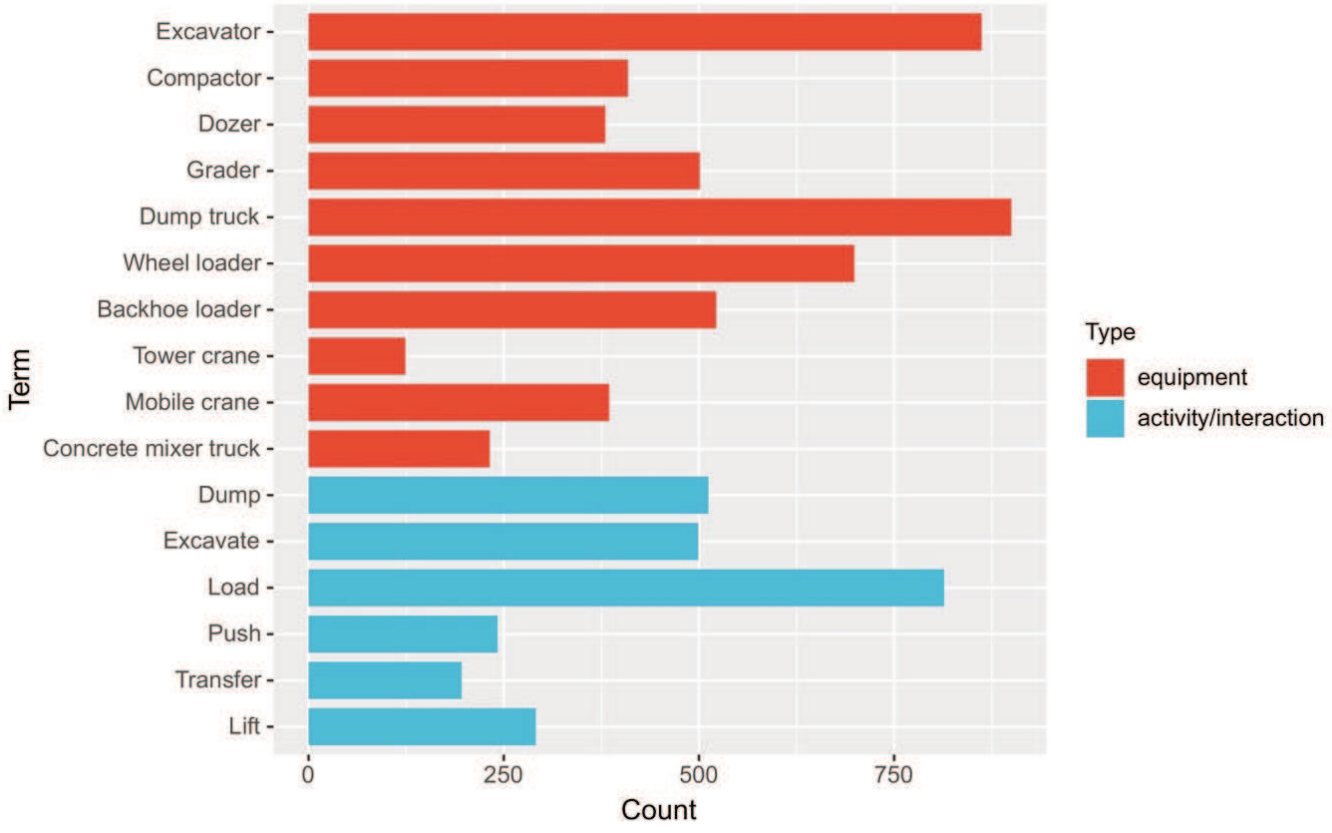


Figure 6-4. Element distribution of captioning dataset

Table 6-2. Statistics of Top 20 N-Gram of the captioning dataset

Index	1-Gram	Count	2-Gram	Count	3-Gram	Count
1	a	7795	loader is	2069	wheel loader is	1151
2	is	7568	dump truck	1575	a dump truck	979
3	the	5340	wheel loader	1392	a wheel loader	894
4	loader	2426	excavator is	1319	backhoe loader is	885
5	truck	2161	on the	1025	an excavator is	880
6	dump	1985	an excavator	1014	a backhoe loader	665
7	soil	1940	backhoe loader	988	a grader is	618
8	on	1926	a dump	983	mobile crane is	597
9	site	1819	truck is	916	a compactor is	508
10	excavator	1610	a wheel	898	a mobile crane	468
11	wheel	1431	grader is	896	dump truck is	445
12	road	1396	crane is	761	compactor is compacting	441
13	an	1217	compactor is	722	concrete mixer truck	401
14	backhoe	1053	mobile crane	715	a dozer is	387
15	to	1014	a backhoe	676	excavator is excavating	369
16	grader	959	a grader	645	mixer truck is	350
17	and	930	the soil	625	loader is loading	319
18	crane	928	dozer is	590	the dump truck	318
19	are	810	is loading	585	into a dump	318
20	compactor	802	the road	558	is travelling on	316

6.2.3 Method selection

The deep learning image captioning method selection is described in this section. As introduced in Section 6.1, most deep learning image captioning methods have an encoder–decoder architecture.

Baseline method (base)

The baseline method consisting of CNN and RNN networks is selected for evaluation. In the baseline method, the ResNet101 network (He et al. 2016) is employed as the encoder and the LSTM network (Vinyals et al. 2014) is adopted as the decoder. Figure 6-5 shows the architecture of the baseline method. It should be noted that most studies in construction (e.g., Bang and Kim 2020; Liu et al. 2020) have adopted baseline methods for captioning construction images.

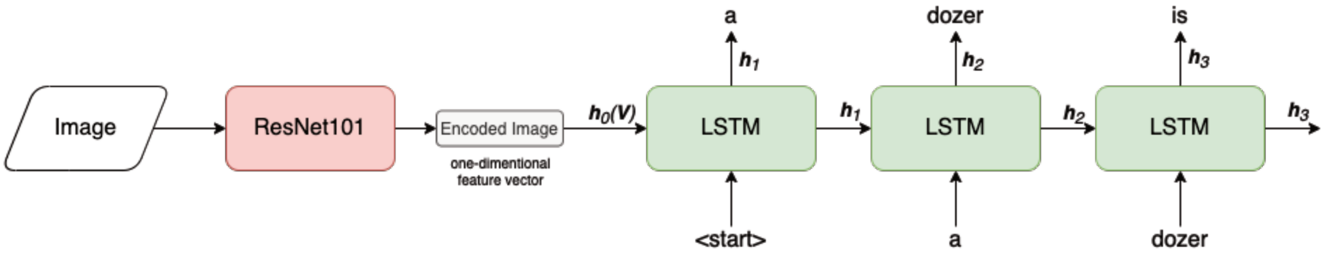


Figure 6-5. Architecture of baseline method

Attention method (att)

In this study, the attention method is selected as the decoder for testing the construction images (ResNet101 having been selected as the encoder). The attention decoder (Xu et al. 2015) allows neural networks to look at different parts of the image at different steps in the sequence, and this approach has enjoyed considerable success within the computer vision community. The architecture of the attention method is shown in Figure 6-6. Generally, attention decoder is a small neural network added to an LSTM neural network that takes the hidden state as input

(meaning it will look at the sequence generated thus far), and outputs a set of weights for the image feature that indicates what areas the LSTM should focus on (based on a high weight). The weights are applied to the image feature in order to obtain the feature content; the content will then be sent back to the LSTM to help generate the output. In contrast to the baseline method, the attention method changes LSTM to an attention network, while the encoder networks remain the same.

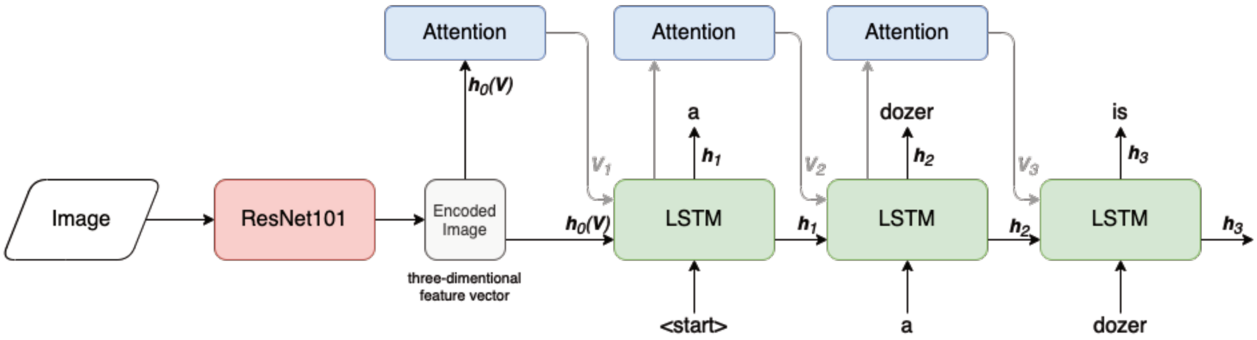


Figure 6-6. Architecture of attention method

Transformer method (tsfm)

The transformer decoder (Vaswani et al. 2017) is a multi-head attention mechanism that has achieved better performance than the attention decoder in computer vision. The author of this study has implemented the transformer method by integrating ResNet101 encoder and transformer decoder. As shown in Figure 6-7, the transformer decoder consists of multi-head attention layers, normalization layers, and feed-forward layers. The multi-head attention layers are a set of parallel attention networks that calculate the attention weights, while the feed-forward layers, such as LSTM, are responsible for conducting the bulk of the decoding work.

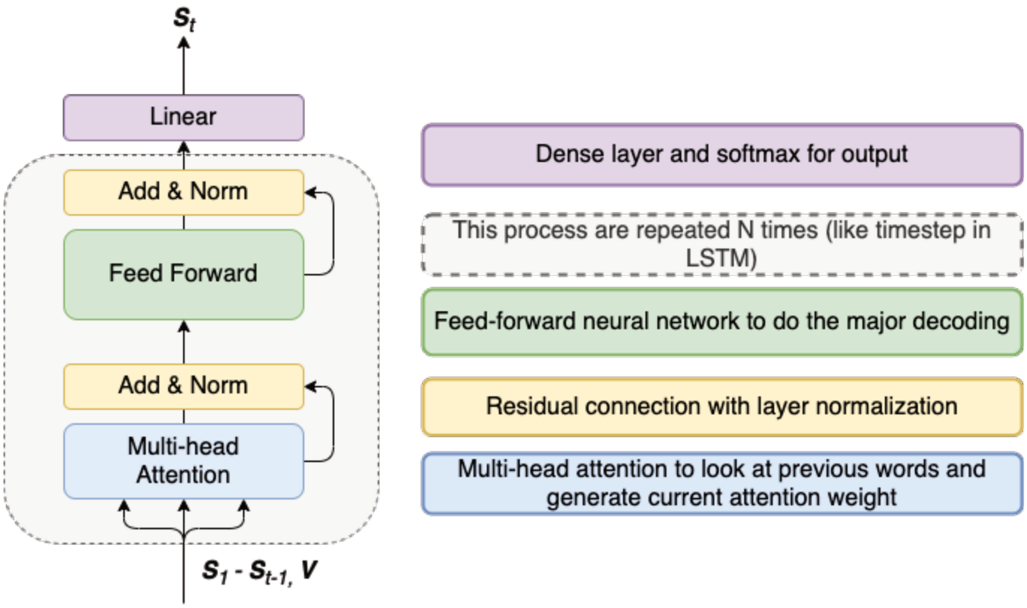


Figure 6-7. Illustration of transformer decoder

Self-critical training strategy (sc)

It should be noted that applying specific training strategies will improve the performance of deep learning image captioning methods. In this regard, the self-critical training strategy (Rennie et al. 2017) is integrated in the present study. This strategy adopts reinforcement learning for the purpose of training deep learning image captioning methods, and it uses a non-differentiable task metric to optimize the entire training process. In the present study, the self-critical strategy is applied to all three of the methods described above (i.e., base, att, and tsfm).

To summarize, six deep learning image captioning methods are selected for testing on the developed captioning dataset: base, base-sc, att, att-sc, tsfm, tsfm-sc. Table 6-3 summarizes the salient information concerning these methods in terms of encoder, decoder, and whether the self-critical strategy has been applied.

Table 6-3. Information on deep learning image captioning methods used for evaluation

Method	Encoder	Decoder	Self-critical
base	ResNet101	LSTM	
base-sc	ResNet101	LSTM	√
att	ResNet101	Attention	
att-sc	ResNet101	Attention	√
tsfm	ResNet101	Transformer	
tsfm-sc	ResNet101	Transformer	√

6.2.4 Evaluation metrics

At present, no single general metric for evaluation of image captioning methods in computer vision is proposed. For the purpose of this study, five automatic evaluation metrics are adopted in order to assess the performance of deep learning image captioning methods by comparing the ground truth sentences and the generated sentences: Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Consensus-based Image Description Evaluation (CIDEr), and Semantic Propositional Image Caption Evaluation (SPICE). For these metrics, a higher value indicates better captioning performance. It should be noted that the scale of values for CIDEr is 0 to 10, while the scale for the other four metrics is 0 to 1.

BLEU

BLEU (Papineni et al. 2001) measures the overlap between the predicted single word or n-gram (sequence of n adjacent words) and a set of reference sentences. BLEU only measures the word

match and sentence length match and does not take the semantic meaning of the words into account. Variations of BLEU include BLEU-1, BLEU-2, BLEU-3, and BLEU-4, where the number appearing after the hyphen signifies the number of words used up to n-grams (the variations listed here are the ones adopted in the present study).

ROUGE-L

ROUGE (Chin-Yew 2004) utilizes n-grams to measure the recall score of the generated sentences relative to the reference sentences. The most widely used version of ROUGE, ROUGE-L, is adopted in the present study. ROUGE-L computes the recall and precision of the longest common subsequences between the candidate and reference sentences.

METEOR

METEOR (Lavie and Agarwal 2007) introduces semantic matching for automatic evaluation. It includes lexical match, stemmed words match, synonym match, and paraphrase match. The METEOR score is calculated by mapping the unigrams of the candidate and reference sentences and measuring their alignment.

CIDEr

CIDEr (Vedantam et al. 2014) first converts the words in both the candidate and reference sentences into their root forms and then measures the co-existence frequency of the n-grams in both sentences. During the measurement, the term frequency inverse document frequency is applied. The most commonly used version today is CIDEr-D, as it is capable of preventing outlier scores resulting from poor human judgment. The present study adopts CIDEr-D version.

SPICE

SPICE (Anderson et al. 2016) calculates the score by measuring the similarity between the scene graph tuples of the candidate and reference sentences. The scene graph includes objects, their attributes, and relationships extracted from the sentence.

6.3 Implementation and Evaluation Results

In this section, the implementation of six deep learning image captioning methods is described., The evaluation results at the sentence level and element level are then presented.

6.3.1 Implementation

All six deep learning image captioning methods are implemented in the Python language. The encoder (i.e., ResNet101) and decoders (i.e., LSTM, attention, and transformer) adopted in this study are implemented by the Pytorch library. The ResNet101 is pretrained on the ImageNet dataset, while the Opencv library is employed for image input/output. In terms of hardware, the evaluation is conducted on a computer that features two NVIDIA GTX 1080 Ti GPUs (11 GB each), an Intel Core i9-7920X@ 2.90 Hz CPU with 12 cores, and two 32 GB memory cards. The testing environment uses the Ubuntu 16.04 system.

In the training process, all images are resized to 256×256 and normalized based on a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225]. Moreover, the sentence annotations are tokenized in order to divide them into lists of single words without punctuation. As part of the tokenization process, the <start> and <end> label are added to the beginning and ending of each token list to indicate the start and end of each annotation. The number of training epochs is 30 for the non-self-critical training methods (i.e., base, att, and tsfm). For base-sc and att-sc, the models are first trained for 30 epochs by optimizing the cross-entropy loss and then

training for 20 epochs using the self-critical training strategy. For tsfm-sc, the model is trained for 15 epochs for the traditional strategy and for 5 epochs for the self-critical training strategy.

6.3.2 Sentence level evaluation

The deep learning image captioning methods are trained on the training set and then validated on the validation set. Table 6-4 summarizes the validation results for the deep learning image captioning methods. Among the six methods, the tsfm-sc achieves the best performance on the task of captioning construction images, attaining a BLEU-1 score of 0.606, BLUE-2 of 0.506, BLEU-3 of 0.427, BLEU-4 of 0.349, METEOR of 0.287, ROUGE-L of 0.585, CIDEr of 1.715, and SPICE of 0.422, underscoring the feasibility of the transformer decoder and self-critical training strategy in deep learning image captioning. In computer vision, it should be noted, the up to date leading scores in the COCO captioning challenge are a BLEU-1 of 0.795, BLEU-2 of 0.635, BLEU-3 of 0.485, BLEU-4 of 0.363, ROUGE of 0.573, METEOR of 0.277, CIDEr of 1.196 and SPICE of 0.213, and these metrics are close to the performance observed in the present study. This indicates that the deep learning image captioning methods under consideration attain comparable results in construction applications to those observed in the computer vision community. Figure 6-8 shows example captioning results in the validation set produced by the tsfm-sc method, which can describe the contents of construction images with correct descriptions in most cases.

Table 6-4. Sentence level evaluation results

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
base	0.587	0.477	0.398	0.320	0.274	0.560	1.499	0.394
base-sc	0.546	0.460	0.390	0.318	0.253	0.567	1.549	0.358
att	0.570	0.463	0.385	0.308	0.267	0.549	1.408	0.382
att-sc	0.576	0.484	0.412	0.340	0.274	0.582	1.702	0.418
tsfm	0.586	0.474	0.392	0.311	0.273	0.556	1.427	0.388
tsfm-sc	**0.606**	**0.506**	**0.427**	**0.349**	**0.287**	**0.585**	**1.715**	**0.422**

Note: The best performance is denoted in bold.

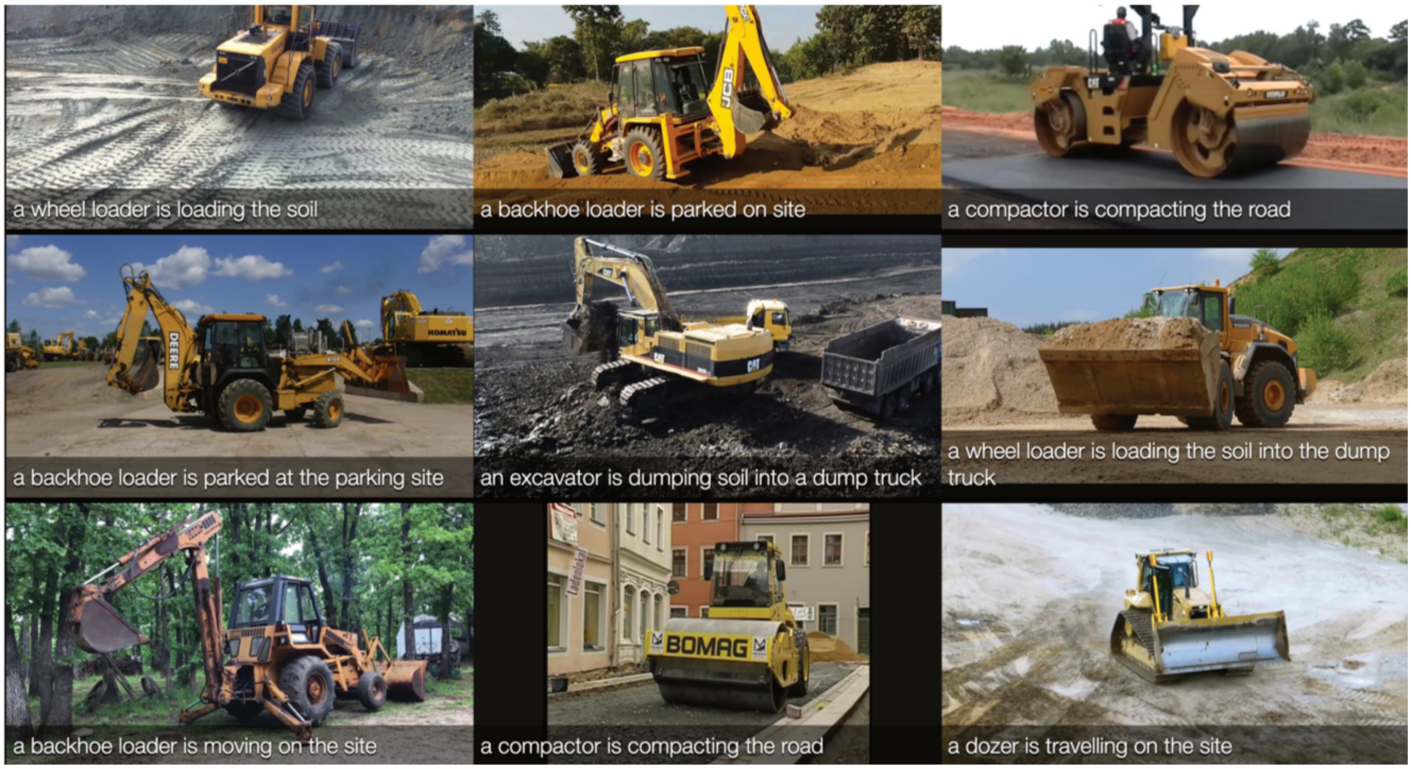


Figure 6-8. Example captioning results produced by tsfm-sc

The base method is found in this experiment to rank second in performance, achieving a BLEU-1 score of 0.587, BLEU-2 of 0.477, BLEU-3 of 0.398, BLEU-4 of 0.320, METEOR of 0.274,

ROUGE-L of 0.560, CIDEr of 1.499, and SPICE of 0.394, and outperforming the attention-based method (att) and the transformer-based method (tsfm) in construction scenarios. In computer vision, in contrast, the att and tsfm achieve better performance than the base. This result demonstrates the different characteristics of image captioning in construction versus in computer vision. In fact, the degree of difficulty of image captioning in construction is lower than that in computer vision applications. As such, a simpler method (base) can achieve comparable performance to more advanced methods (i.e., att and tsfm).

6.3.3 Element level evaluation

Image captioning provides a holistic solution for understanding the scene information (i.e., objects, activities, and relationships between objects). By analyzing the generated sentences, this scene information can be retrieved for various uses related to construction object and activity recognition. For example, construction machine objects can be extracted from sentences to replace the object detection methods. To validate the feasibility of deep learning image captioning methods for construction scene analysis, the tsfm-sc method is evaluated at the element level in terms of its ability to recognize machine objects in images.

As with the sentence level evaluation, in the element level evaluation the tsfm-sc method is trained on the training set and validated on the validation set. In the evaluation process, the machine objects are extracted from the generated sentences and ground truth sentences for the purpose of comparison. Following the evaluation metrics for object detection, the precision, recall, and F1 score are used for element level evaluation. The calculation of these metrics is illustrated in Equations 6-1 to 6-3.

$$Precision = \frac{TP}{TP+FP} \quad \text{(Equation 6-1)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{(Equation 6-2)}$$

$$F1 = 2 \times \frac{Precison \times Recall}{Precision + Recall} \quad \text{(Equation 6-3)}$$

where TP (true positive) is the number of machine objects appearing in both generated sentences and ground truth sentences. FP (false positive) is the number machine objects appearing in the generated sentences but not appearing in the ground truth sentences, and FN (false negative) is the number of machine objects appearing in the ground truth sentences but not in the generated sentences.

Table 6-5 shows the element level evaluation results. As can be seen in the table, the tsfm-sc is found to achieve, on average, a precision of 91.1%, recall of 83.3%, and F1 score of 86.6% for the validation set, meaning that it has comparable performance with state-of-the-art object detection methods in construction scenarios. The tsfm-sc method achieves the highest precision (94.6%) recognizing dozer objects and the highest recall (92.9%) recognizing grader objects. It should be noted that, in current practice, image captioning methods cannot serve as a replacement for object detection methods when it comes to recognizing construction objects. Image captioning focuses on the primary and cooperating machine objects, while the small machine objects in the background are ignored. Object detection methods, on the other hand, are capable of recognizing all machine objects appearing in the given image or video. In real cases, however, recognizing the major construction objects is sufficient for the purpose of construction management, and this can be achieved using deep learning image captioning methods.

Table 6-5. Element level evaluation results of tsfm-sc method

Scene Element	Precision	Recall	F1 score
Excavator	87.6%	89.9%	88.7%
Compactor	92.9%	90.8%	91.9%
Dozer	94.6%	76.8%	84.8%
Grader	92.9%	92.9%	92.9%
Dump truck	79.6%	89.9%	84.4%
Concrete mixer truck	91.9%	61.8%	73.9%
Wheel loader	93.1%	77.7%	84.7%
Backhoe loader	90.9%	85.7%	88.2%
Tower crane	100.0%	82.1%	90.2%
Mobile crane	87.0%	85.7%	86.3%
Average	91.1%	83.3%	86.6%

6.4 Keyframe Captioning

Captioning the keyframes detected using the highlight detection described in Chapter 5 is an important objective of the research described in this chapter. In this regard, the proposed video summarization framework is able to generate text descriptions from raw construction videos. As introduced in section 5.2.4, the keyframes T detected from construction videos are a set of image sets, while $T = \{[s, e]_1, [s, e]_2, \dots, [s, e]_i, \}$, where i is the index, and s and e represent the start and end frames of each highlight clip. In this study, only the median frame, m_i, of a given clip, $[s, e]_i$, is selected for captioning. As such, the set of median keyframes $M = [m_1, m_2, \dots, m_i]$ is captioned by the deep learning image captioning method tsfm-sc.

Figure 6-9 shows example results of keyframe captioning in the construction gate case produced by the tsfm-sc method. In most keyframes, the image captioning method has generated precise text descriptions. For example, in keyframe#4, the generated sentence is “a dump truck is coming to the construction site”. It is also noted that the keyframe captioning results have errors in a few cases. For example, in keyframe#2, the captioning result is “a concrete mixer truck is travelling on the road”, whereas it is in fact a dump truck that appears in the keyframe. Notwithstanding these errors, the results show that deep learning image captioning methods can be used to generate useful and professional text descriptions from construction images to assist in project management. Potential uses include image searching engine, daily report generation, and construction video documentation. The generated texts make construction video highlights searchable, filterable, and manageable.

Figure 6-9. Example keyframe captioning results

6.5 Discussion

In this chapter, a linguistic schema has been proposed for annotation of construction images with professional sentences. The performance of six deep learning image captioning methods has been investigated in construction scenarios, and the best-performing method has been

incorporated into the video summarization framework for the purpose of captioning keyframes. The evaluation results indicate that the objective underlying the research described in this chapter has been successfully achieved. The specific research findings of note are outlined as follows:

- The self-critical training strategy has been shown to improve the performance of image captioning methods in construction. In the evaluations, the three methods in the present study adopting self-critical training obtained an average BLEU-1 of 0.576, BLEU-2 of 0.483, BLEU-3 of 0.410, BLEU-4 of 0.336, METEOR of 0.271, ROUGE-L is 0.578, CIDEr of 1.655, and SPICE of 0.399. For the methods not adopting self-critical training, the average performance was 0.581 for BLEU-1, 0.471 for BLEU-2, 0.392 for BLEU-3, 0.313 for BLEU-4, 0.271 for METEOR, 0.555 for ROUGE-L, 1.445 for CIDEr, and 0.388 o for n SPICE. With the exception of BLEU-1 and METEOR, the methods adopting self-critical training outperformed the methods without self-critical training. The results indicate that applying specific strategies in training can improve the performance of image captioning methods.
- The transformer encoder performs better than the attention decoder in construction scenarios. In this study, two methods—tsfm and tsfm-sc—employed the transformer decoder. These two methods achieved an average BLEU-1 of 0.596, BLEU-2 of 0.490, BLEU-3 of 0.410, BLEU-4 of 0.330, METEOR of 0.280, ROUGE-L is 0.571, CIDEr of 1.571, and SPICE of 0.405 in the evaluation. For the attention-based methods (att and att-sc), the average results were 0.573 for BLEU-1, 0.474 for BLEU-2, 0.399 for BLEU-3, 0.324 for BLEU-4, 0.271 for METEOR, 0.566 for ROUGE-L, 1.555 for CIDEr, and 0.400 for SPICE. The transformer-based method achieved higher performance than

attention-based method, consistent with the results reported within the computer vision domain.

- Figure 6-10 shows some example captioning errors produced by the tsfm-sc method. The captioning errors frequently encountered were mainly: (1) mis-recognition of primary machines (e.g., in failure #3, the primary machine wheel loader is misidentified as a grader); (2) mis-recognition of activities (e.g., in failure #4, the wheel loader is driving on the road, whereas it is misidentified as being engaged in dumping soil; and (3) mis-recognition of working contents (e.g., in failure #7, the tsfm-sc fails to recognize the working contents and yields the unreasonable sentence "a mobile crane is lifting the construction site"). To reduce captioning errors, the captioning dataset needs to be expanded by adding more annotated construction images.

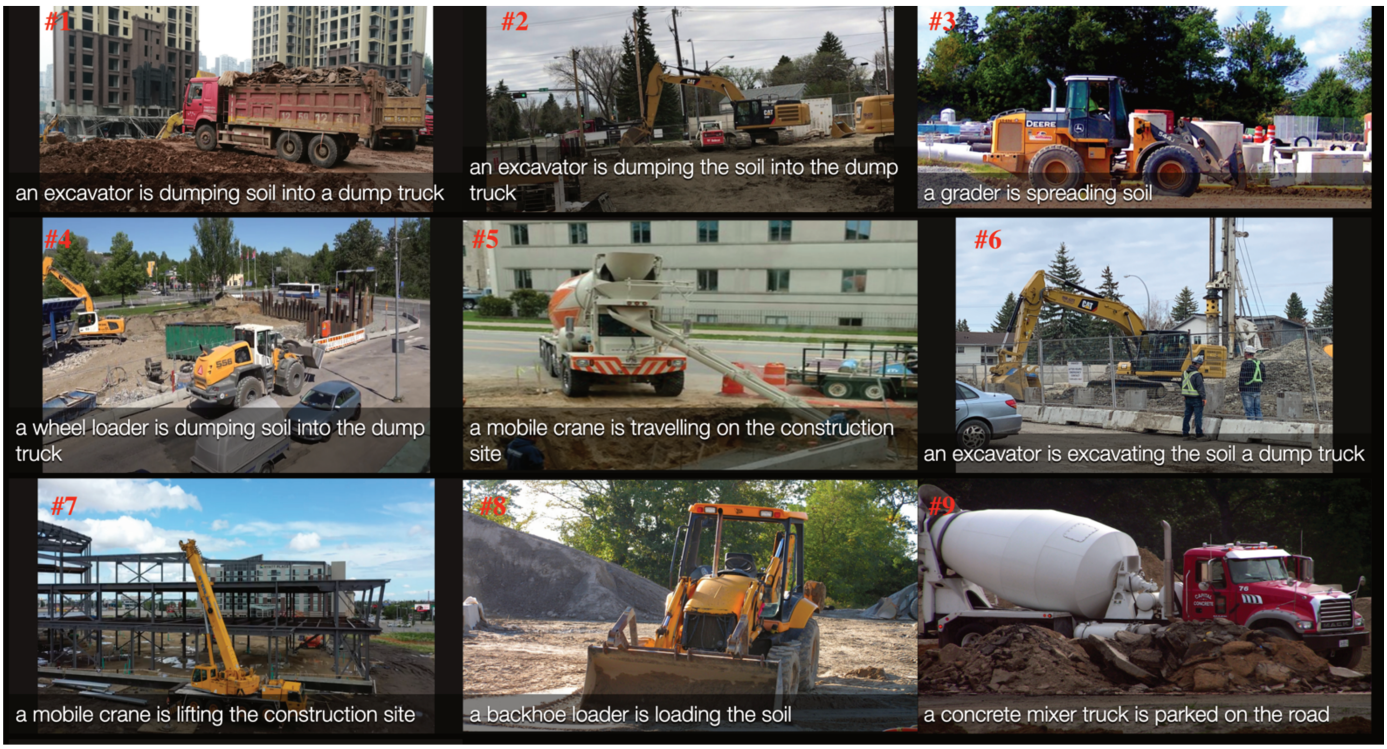


Figure 6-10. Example captioning errors committed by tsfm-sc

- Image captioning methods have practical implications for construction management. The sentences generated by image captioning methods contain rich information about construction objects, activities, and relationships that constitutes an integral data package for vision-based construction applications. Moreover, generating text descriptions from construction images/videos helps with documentation of construction site footage by fulfilling the role of "text-index". Construction engineers can obtain information of interest by simply querying the generated text descriptions. By reviewing video highlights and text descriptions together, the project manager can readily gauge the daily project progress to inform decision-making and resource allocation in construction projects.

6.6 Summary

This chapter proposed a linguistic schema for deconstructing construction machine images into primary objects, cooperating objects, activities, working contents, and supplementary information. Using the linguistic schema, professional descriptions can be annotated for the purpose of training deep learning image captioning methods. A captioning dataset was developed containing 4,000 images and 8,226 sentences. In turn, six deep learning image captioning methods were tested on the captioning dataset, with the tsfm-sc method achieving the best performance. The tsfm-sc was then employed for scene element analysis and incorporated into the proposed video summarization framework.

The contributions of the research described in this chapter are three-fold. First, an annotated image dataset has been developed for training deep learning methods for captioning of images featuring construction machines. The developed dataset can also be used for other studies in the construction automation field. Second, six deep learning image captioning methods have been

compared, and their performance in construction applications investigated. Third, an analysis of the efficacy of various image captioning methods in identifying construction objects has been conducted using the tsfm-sc method, demonstrating the potential of applying image captioning in scene element analysis. In conclusion, the use of deep learning image captioning methods has been shown to provide a tremendous opportunity as an advanced vision-based application that makes construction videos searchable, filterable, and manageable.

Chapter 7: CONCLUSIONS, CONTRIBUTIONS, AND FUTURE WORKS

7.1 Conclusions

Construction videos contain important visual information about projects that allows project managers to monitor jobsites remotely. Automatically analyzing construction videos using vision-based methods is beneficial to construction management in terms of expediting processes, improving productivity, and reducing safety risks. However, the current practice of using raw construction video for vision-based sites monitoring is challenged in three notable respects: (1) retrieval of the information of interest from construction videos is time-consuming and labor-intensive because construction videos are un-structured data; (2) digital storage of construction videos requires large a large amount of disk space, considering the long streaming time and high resolution associated with construction site footage; and (3) existing vision-based applications are inefficient because, although many or most frames contain little relevant project information, significant computational resources are consumed in processing them.

To fill the above gaps, this research proposed a deep learning-based framework to summarize construction videos into video highlights and text descriptions for vision-based monitoring of construction sites. The main idea here is to convert un-structured video data into structured video highlights and text descriptions, thereby significantly reducing the inspection time and storage requirements. In this way, vision-based applications can then be limited in scope to processing only the detected video highlights and text descriptions rather than having to process entire raw construction videos, thereby increasing markedly the efficiency of the analysis. Meanwhile, this research is aiming to reduce the efforts on non-engineering works (video inspection, documentation, and management) in projects instead of replacing engineers on decision making

tasks. To accomplish the goal of this research, three main objectives have been pursued as summarized below:

1. *Development of an image dataset of construction machines for deep learning object detection.* A useful image dataset of construction machines for training deep learning object detection is not currently available due to the limited accessibility of construction images, the time- and-labor-intensiveness of manual annotations, and the knowledge base required in terms of both construction and deep learning. The present study developed a comprehensive image dataset, called ACID, specifically for construction machines. In ACID, 10,000 images belonging to ten types of construction machines are compiled and annotated with machine types and their corresponding positions on the images. To validate the applicability of this image dataset, four existing deep learning detection algorithms were trained on ACID: YOLO-v3, Inception-SSD, R-FCN-ResNet101, and Faster-RCNN-ResNet-101. The mAP was found to be 83.0% for Inception-SSD, 87.8% for YOLO-v3, 88.8% for R-FCN-ResNet101, and 89.2% for Faster-RCNN-ResNet-101. The average detection speed of the four algorithms was found to be 16.7 fps, a speed that satisfies the needs of most studies in the field of automation in construction.
2. *Development of a deep learning-based method for detecting video highlights from construction videos by exploring context and feature information.* To obtain and store useful video footage systematically and concisely, the present study proposed a vision-based method to automatically generate video highlights from construction videos. The proposed method categorizes construction keyframes into feature keyframes and object keyframes. The proposed approach was validated through two case studies: a gate scenario and an earthmoving scenario. In the experiments, the proposed method achieved

89.2% precision and 93.3% recall, outperforming the feature-based method by 16.1% and 17.2% on precision and recall, respectively. Meanwhile, the proposed method was shown to reduce the digital storage requirement by 93.1%. The proposed approach offers potential benefits to construction management in terms of significantly reducing video storage space and efficiently indexing construction video footage.

3. *Adoption of deep learning-based image captioning as the basis for generating text descriptions from construction images.* Generating text descriptions from construction images can help engineers to expeditiously ascertain an image's contents, and provides a "text index" for construction image/video documentation. In this study, a linguistic schema for annotating construction machine images was proposed. A captioning dataset based on the developed ACID dataset was then developed by following the proposed linguistic schema. Six deep learning image captioning methods built upon the encoder–decoder architecture were trained and validated on the captioning dataset. The best performing method, tsfm-sc, was then applied for scene element analysis, achieving an F1-score of 86.6% on the validation set. The tsfm-sc method was then integrated with the proposed video summarization method for captioning keyframes, while the contents of the keyframes were successfully converted into text descriptions.

7.2 Contributions

This research makes several notable contributions to the body of knowledge on vision-based monitoring of construction sites. The academic and industrial contributions are outlined in this section.

7.2.1 Academic contributions

1. A novel deep learning-based framework of construction video summarization has been proposed. The concept of video summarization builds on existing research on vision-based monitoring in construction by introducing a mid-level image processing method. The proposed framework increases the processing efficiency compared to current vision-based applications in construction.
2. A method for developing construction image dataset for deep learning object detection is proposed that consists of four main steps: category selection, image collection, image selection, and image annotation. By following the same development method, other construction datasets (e.g., workers and materials) can be built in future work.
3. A novel method for construction video highlight detection has been proposed in this research that outperforms the feature-based method over 10% in terms of recall and precision. In the video highlight detection method, three construction rules have been proposed for object keyframe detection: the working zone rule, the working status rule, and the working interaction rule. As results, the detected video highlights in this research are interpretable and flexible.
4. This study demonstrated how to generate professional text descriptions from construction images by adopting deep learning image captioning methods. A linguistic schema for annotating construction machine images was proposed, and a captioning dataset was developed based on the linguistic schema. Meanwhile, the feasibility of state-of-the-art deep learning captioning methods was investigated in this research. The investigation results can be used to instruct researchers in the construction field on selecting the proper image captioning method for a given application.

7.2.2 Industrial contributions

1. Providing a standard image dataset of construction machines for deep learning object detection and image captioning. The dataset is one of the most important resources for applying deep learning methods in the construction industry. A standardized dataset can improve the performance of vision-based applications adopted deep learning in construction projects. The developed ACID dataset presented in this thesis has been made available to the broader construction community on the following website (www.acidb.ca). To date, the ACID dataset has been used by over 130 research groups from universities/institutions, demonstrating the impacts of the present research.
2. The proposed framework reduces the manual inspection requirement in the analysis of construction video footage. In the current practice of construction management, engineers typically prefer to manually browse videos to retrieve the desired information because video data is easily understood, but this manual process is highly time consuming due to the sheer volume of video data to analyze. In this context, through two case studies, the proposed framework has been shown to reduce the average construction video duration from 55 minutes length to 2.77 minutes. What this demonstrates is that the proposed framework is able to remove redundant video clips from raw construction videos and thereby reduce manual inspection efforts.
3. The proposed framework reduces the digital storage requirement for construction videos. In the experiments carried out as part of this research, the proposed framework reduced the video storage requirement by 93.1% while retaining most of the relevant content for construction management purpose. The video documentation in construction projects can thereby be improved. Assuming the available digital storage space is sufficient for about

one month of raw video, the same storage capacity could accommodate up to 2 years of video footage when applying the proposed framework.

4. The text descriptions generated in this research facilitate construction management tasks in terms of indexing and documenting construction videos. For indexing, the text descriptions can be used for querying the video highlights by incorporating a search function. (Text-based searching is more efficient and precise than content-based searching and tag-based searching.) Meanwhile, the text descriptions contain information about construction objects, activities, and relationships that is integral for generating daily reports and documenting project progress.

7.3 Future Works

To improve the performance and feasibility of the proposed framework, the research limitations are identified, while avenues of research that can be pursued in future work are recommended:

1. The scale of the developed ACID dataset needs to be enhanced in terms of the number of images and the number of classes of construction machines. Compared with other comprehensive image datasets in the computer vision community, the number of images in ACID is relatively low. For example, the COCO2014 dataset has around 160,000 images from 91 categories. To maximize the capacity of deep learning algorithms in construction applications, more images and classes of construction machines will be added to the ACID dataset in the future.
2. The annotations of the ACID dataset need to be extended. At present, ACID can only be used for training deep learning object detection and image captioning methods. By annotating the datasets at the pixel level, we can train object-segmentation algorithms and

produce masks of objects, which are more precise than bounding boxes. Therefore, annotating the ACID dataset at the pixel level is another important area of future work.

3. The detection of keyframes in the video highlight detection method presented herein needs to be enhanced. At present, the parameters of pre-defined construction rules are manually set up, and the parameters in one scenario are not typically generically applicable to other scenarios. Future work will focus on developing an automated process to set up the parameters of construction rules using machine learning techniques. Meanwhile, the ResNet50 was adopted in the proposed framework for feature keyframe detection, while other feature extraction networks may perform better than ResNet50 and need to be investigated in the future.
4. More encoder neural networks need to be investigated in construction scenarios. In the present study, the author implemented only the same CNN networks (i.e. ResNet101) for all deep learning image captioning methods, with more emphasis placed on the decoder networks (i.e., LSTM, attention, and transformer). Therefore, in future work, the ResNet101 will be replaced with more advanced encoder networks. Additionally, the present research only investigates "encoder–decoder" image captioning methods, whereas there are some deep learning image captioning methods available built upon other mechanisms (e.g., captioning by detection) whose application to construction warrants investigation in future work.
5. This research explored the feasibility of deep learning image captioning in construction management, while more successive researches need to be conducted in the future. Currently, the text descriptions only contain limited project information in terms of objects, activities, weather, and so forth. However, the image captioning methods adopted

in this research cannot provide more granular information such as project stage of this construction image. In the future, the proposed linguistic schema will be improved in order to annotation construction images from a more comprehensive way. As such, the deep learning image captioning methods will be able to retrieve more in-depth project information from construction images.

6. The proposed framework introduced a new concept named mid-level processing in the vision-based monitoring of construction sites, while the development of novel vision-based applications is not in the scope of this research. In the future, more applications based on the proposed video summarization framework need to be developed. By integration of the proposed framework, advanced vision-based applications, such as daily log generation, earthmoving productivity calculation, and gate logistics reminding, can be developed in an efficient manner.

REFERENCES

Adán, A., Quintana, B., Prieto, S. A., and Bosché, F. (2018). "Scan-to-BIM for 'secondary' building components." *Advanced Engineering Informatics*, 37, 119–138.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). "SPICE: Semantic propositional image caption evaluation." *2016 European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, 382–398.

Angah, O., and Chen, A. Y. (2020). "Tracking multiple construction workers through deep learning and the gradient based method with re-matching based on multi-object tracking accuracy." *Automation in Construction*, Elsevier, 119, 103308.

Appalaraju, S., and Chaoji, V. (2017). "Image similarity using Deep CNN and curriculum learning." *ArXiv*, ID: 1709.08761.

Arabi, S., Haghighat, A., and Sharma, A. (2019). "A deep learning based solution for construction equipment detection: from development to deployment." *ArXiv*, ID: 1904.09021.

Bang, S., and Kim, H. (2020). "Context-based information generation for managing UAV-acquired data using image captioning." *Automation in Construction*, 112, 103116.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). "Simple online and realtime tracking." *2016 IEEE International Conference on Image Processing*, IEEE, Phoenix, USA, 3464–3468.

Bo Yang, and Nevatia, R. (2012). "An online learned CRF model for multi-target tracking." *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2034–2041.

Bohn, J. S., and Teizer, J. (2010). "Benefits and barriers of construction project monitoring using

high-resolution automated cameras." *Journal of Construction Engineering and Management*, 136(6), 632–640.

Bosché, F., Ahmed, M., Turkan, Y., Haas, C. T., and Haas, R. (2015). "The value of integrating Scan-to-BIM and Scan-vs-BIM techniques for construction monitoring using laser scanning and BIM: The case of cylindrical MEP components." *Automation in Construction*, 49, 201–213.

Bouget, D., Allan, M., Stoyanov, D., and Jannin, P. (2017). "Vision-based and marker-less surgical tool detection and tracking: a review of the literature." *Medical Image Analysis*, 35, 633–654.

Brawley, A. M., and Pury, C. L. S. (2016). "Work experiences on MTurk: Job satisfaction, turnover, and information sharing." *Computers in Human Behavior*, 54, 531–546.

Brilakis, I., Fathi, H., and Rashidi, A. (2011). "Progressive 3D reconstruction of infrastructure with videogrammetry." *Automation in Construction*, 20(7), 884–895.

Brilakis, I., and Soibelman, L. (2005). "Content-based search engines for construction image databases." *Automation in Construction*, 14(4), 537–550.

Chen, C., Zhu, Z., and Hammad, A. (2020). "Automated excavators activity recognition and productivity analysis from construction site surveillance videos." *Automation in Construction*, 110, 103045.

Chen, J., Kira, Z., and Cho, Y. K. (2019). "Deep learning approach to point cloud scene understanding for automated scan to 3D reconstruction." *Journal of Computing in Civil Engineering*, 33(4), 04019027.

Chen, L., and Wang, Y. (2017). "Automatic key frame extraction in continuous videos from construction monitoring by using color, texture, and gradient features." *Automation in*

Construction, 81, 355–368.

Cheng, T., Venugopal, M., Teizer, J., and Vela, P. A. (2011). "Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments." *Automation in Construction*, 20(8), 1173–1184.

Chi, S., and Caldas, C. H. (2012). "Image-based safety assessment: Automated spatial safety risk identification of earthmoving and surface mining activities." *Journal of Construction Engineering and Management*, 138(3), 341–351.

Chin-Yew, L. (2004). "ROUGE: A package for automatic evauluation of summaries." *Text Summarization Branches Out*, ACL, 74–81.

Choi, W. (2015). "Near-online multi-target tracking with aggregated local flow descriptor." *ArXiv*, ID: 1504.02340.

Chu, W., Han, S., Luo, X., and Zhu, Z. (2020). "Monocular vision–based framework for biomechanical analysis or ergonomic posture assessment in modular construction." *Journal of Computing in Civil Engineering*, 34(4), 04020018.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. (2015). "Reducing overfitting in deep networks by decorrelating representations." *ArXiv*, ID: 1511.06068.

Coltuc, D. (2000). "Robust image hashing." *Encyclopedia of Information Science and Technology, Third Edition*, IGI Global, 5998–6008.

Dai, J., Li, Y., He, K., and Sun, J. (2016). "R-FCN: Object detection via region-based fully convolutional networks." *2016 Conference on Neural Information Processing Systems*, Barcelona, Spain, 379–387.

Dalal, N., and Triggs, B. (2010). "Histograms of rriented gradients for human detection." *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE,

886–893.

Davis, J., and Goadrich, M. (2006). "The relationship between Precision-Recall and ROC curves." *23rd international conference on Machine learning - ICML '06*, ACM Press, New York, USA, 233–240.

Ergen, E., Akinci, B., and Sacks, R. (2007). "Tracking and locating components in a precast storage yard utilizing radio frequency identification technology and GPS." *Automation in Construction*, 16(3), 354–367.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). "The pascal visual object classes (VOC) challenge." *International Journal of Computer Vision*, 88(2), 303–338.

Fang, W., Ding, L., Love, P. E. D., Luo, H., Li, H., Peña-Mora, F., Zhong, B., and Zhou, C. (2020). "Computer vision applications in construction safety assurance." *Automation in Construction*, 110, 103013.

Fang, W., Ding, L., Zhong, B., Love, P. E. D., and Luo, H. (2018). "Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach." *Advanced Engineering Informatics*, 37, 139–149.

Gao, L., Li, X., Song, J., and Shen, H. T. (2019). "Hierarchical LSTMs with adaptive attention for visual captioning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1112–1131.

Girshick, R. (2015). "Fast R-CNN." *2015 IEEE International Conference on Computer Vision*, IEEE, Las Condes, Chile, 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation." *2014 IEEE Conference on Computer Vision*

and Pattern Recognition, IEEE, Columbus, USA, 580–587.

Golparvar-Fard, M., Heydarian, A., and Niebles, J. C. (2013). "Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers." *Advanced Engineering Informatics*, 27(4), 652–663.

Gualdi, G., Prati, A., and Cucchiara, R. (2011). "Contextual information and covariance descriptors for people surveillance: An application for safety of construction workers." *EURASIP Journal on Image and Video Processing*, 1–16.

Ha, I., Kim, H., Park, S., and Kim, H. (2018). "Image retrieval using BIM and features from pretrained VGG network for indoor localization." *Building and Environment*, 140, 23–31.

Hackley, C. 2016. "Construction 2016 album." Accessed June 10, 2016. https://www.flickr.com/photos/hackleypubliclibrary/albums/with /72157667687353571.

Ham, Y., and Kamari, M. (2019). "Automated content-based filtering for enhanced vision-based documentation in construction toward exploiting big visual data from drones." *Automation in Construction*, 105, 102831.

Han, K. K., and Golparvar-Fard, M. (2014). "Automated monitoring of operation-level construction progress using 4D BIM and daily site photologs." *Construction Research Congress 2014*, ASCE, Atlanta, Georgia, 1033–1042.

Han, S., and Lee, S. (2013). "A vision-based motion capture and recognition framework for behavior-based safety management." *Automation in Construction*, 35, 131–141.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, 770–778.

Heavy Equipment Channel. 2020. "Heavy equipment fails 2020 dangerous idiots excavator

operator construction truck fail win." Accessed April 22, 2020. https://www.youtube.com/watch?v=Z5mNKFrOt_w.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). "A comprehensive survey of deep learning for image captioning." *ACM Computing Surveys*, 51(6), 1–36.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K. (2017). "Speed/accuracy trade-offs for modern convolutional object detectors." *2017 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, 3296–3297.

Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). "Attention on attention for image captioning." *2019 IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea, 4633–4642.

Ibrahim, Y. M., Lukins, T. C., Zhang, X., Trucco, E., and Kaka, A. P. (2009). "Towards automated progress assessment of workpackage components in construction projects using computer vision." *Advanced Engineering Informatics*, 23(1), 93–103.

Jiao, Y., Li, Z., Huang, S., Yang, X., Liu, B., and Zhang, T. (2018). "Three-dimensional attention-based deep ranking model for video highlight detection." *IEEE Transactions on Multimedia*, 20(10), 2693–2705.

Jonker, R., and Volgenant, A. (1987). "A shortest augmenting path algorithm for dense and sparse linear assignment problems." *Computing*, 38(4), 325–340.

Kim, H., Ham, Y., Kim, W., Park, S., and Kim, H. (2019). "Vision-based nonintrusive context documentation for earthmoving productivity simulation." *Automation in Construction*, 102, 135–147.

Kim, H., Kim, H., Hong, Y. W., and Byun, H. (2018a). "Detecting construction equipment using

a region-based fully convolutional network and transfer learning." *Journal of Computing in Civil Engineering*, 32(2), 04017082.

Kim, J., and Chi, S. (2019). "Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles." *Automation in Construction*, 104, 255–264.

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

Kim, J., Chi, S., and Seo, J. (2018b). "Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks." *Automation in Construction*, 87, 297–308.

Kolar, Z., Chen, H., and Luo, X. (2018). "Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images." *Automation in Construction*, 89, 58–70.

Konstantinou, E., Lasenby, J., and Brilakis, I. (2019). "Adaptive computer vision-based 2D tracking of workers in complex environments." *Automation in Construction*, 103, 168–184.

Kulchandani, J. S., and Dangarwala, K. J. (2015). "Moving object detection: Review of recent research trends." *2015 International Conference on Pervasive Computing*, IEEE, Louis, USA, 1–5.

Kumar, K., Shrimankar, D. D., and Singh, N. (2017). "Event BAGGING: A novel event summarization approach in multiview surveillance videos." *2017 International Conference on Innovations in Electronics, Signal Processing and Communication*, IEEE, Shillong,India, 106–111.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., and Ferrari, V. (2018). "The open images dataset V4: Unified

image classification, object detection, and visual relationship detection at scale." *arXiv*, ID: 1811.00982.

Laganière, R., Bacco, R., Hocevar, A., Lambert, P., Païs, G., and Ionescu, B. E. (2008). "Video summarization from spatio-temporal features." *2nd ACM workshop on Video summarization - TVS '08*, ACM Press, New York, USA, 144–148.

Lavie, A., and Agarwal, A. (2007). "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments." *Second Workshop on Statistical Machine Translation*, ACL, Prague, Czech Republic, 228–231.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning." *Nature*, 521(7553), 436–444.

Lee, H.-S., Lee, K.-P., Park, M., Baek, Y., and Lee, S. (2012). "RFID-based real-time locating system for construction safety management." *Journal of Computing in Civil Engineering*, 26(3), 366–377.

Lee, Y. J., and Park, M. W. (2019). "3D tracking of multiple onsite workers based on stereo vision." *Automation in Construction*, 98, 146–159.

Li, H., Chen, Z., Yong, L., and Kong, S. C. W. (2005). "Application of integrated GPS and GIS technology for reducing construction waste and improving construction efficiency." *Automation in Construction*, 14(3), 323–331.

Li, N., Li, Q., Liu, Y. S., Lu, W., and Wang, W. (2020). "BIMSeek++: Retrieving BIM components using similarity measurement of attributes." *Computers in Industry*, 116, 103186.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). "Microsoft COCO: Common objects in context." *2014 European Conference on Computer Vision*, Springer, Zurich, Switzerland, 740–755.

Lin, Y.-L., Morariu, V. I., and Hsu, W. (2015). "Summarizing while recording: Context-based highlight detection for egocentric videos." *2015 IEEE International Conference on Computer Vision Workshop*, IEEE, Santiago, Chile, 443–451.

Liu, D., Gang Hua, and Tsuhan Chen. (2010). "A hierarchical visual model for video object summarization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12), 2178–2190.

Liu, G., Wen, X., Zheng, W., and He, P. (2009). "Shot boundary detection and keyframe extraction based on scale invariant feature transform." *2009 Eighth IEEE/ACIS International Conference on Computer and Information Science*, IEEE, Shanghai, China, 1126–1130.

Liu, H., Wang, G., Huang, T., He, P., Skitmore, M., and Luo, X. (2020). "Manifesting construction activity scenes via image captioning." *Automation in Construction*, 119, 103334.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., and Berg, A. C. (2016). "SSD: Single shot multibox detector." *arXiv*, ID: 1512.02325.

Lu, M., Chen, W., Shen, X., Lam, H.-C., and Liu, J. (2007). "Positioning and tracking construction vehicles in highly dense urban areas and building construction sites." *Automation in Construction*, 16(5), 647–656.

Lu, Z., and Grauman, K. (2013). "Story-driven summarization for rgocentric v2ideo." *2013 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Portland, USA, 2714–2721.

Luo, X., Li, H., Cao, D., Dai, F., Seo, J., and Lee, S. (2018). "Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by

convolutional neural networks." *Journal of Computing in Civil Engineering*, 32(3), 04018012.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). "The impact of class imbalance in classification performance metrics based on the binary confusion matrix." *Pattern Recognition*, 91, 216–231.

Mahasseni, B., Lam, M., and Todorovic, S. (2017). "Unsupervised video summarization with adversarial LSTM networks." *2017 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, 2982–2991.

Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. L. (2014). "Explain images with multimodal recurrent neural networks." *arXiv*, ID: 1410.1090.

Merler, M., Mac, K.-N. C., Joshi, D., Nguyen, Q.-B., Hammer, S., Kent, J., Xiong, J., Do, M. N., Smith, J. R., and Feris, R. S. (2019). "Automatic curation of sports highlights using multimodal excitement features." *IEEE Transactions on Multimedia*, 21(5), 1147–1160.

Milan, A., Leal-Taixe, L., Reid, I., Roth, S., and Schindler, K. (2016). "MOT16: A Benchmark for Multi-Object Tracking." *ArXiv*, ID: 1603.00831.

Milan, A., Rezatofighi, S. H., Dick, A., Reid, I., and Schindler, K. (2017). "Online multi-target tracking using recurrent neural networks." *31st AAAI Conference on Artificial Intelligence*, San Francisco, USA, 4225–4232.

Muhammad, K., Hussain, T., and Baik, S. W. (2020). "Efficient CNN based summarization of surveillance videos for resource-constrained devices." *Pattern Recognition Letters*, 130, 370–375.

Mundur, P., Rao, Y., and Yesha, Y. (2006). "Keyframe-based video summarization using Delaunay clustering." *International Journal on Digital Libraries*, 6(2), 219–232.

Nath, N. D., and Behzadan, A. H. (2019). “Deep learning models for content-based retrieval of construction visual data.” *2019 Internaltional Conference of Computing in Civil Engineering*, ASCE, Atlanta, USA, 66–73.

Nguyen, B., and Brilakis, I. (2018). “Real-time validation of vision-based over-height vehicle detection system.” *Advanced Engineering Informatics*, 38, 67–80.

Nguyen, H. V., and Bai, L. (2011). “Cosine Similarity Metric Learning for Face Verification.” *2011 Asian Conference on Computer Vision*, Springer, Pondicherr, India, 709–720.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). “BLEU: A method for automatic evaluation of machine translation.” *40th Annual Meeting on Association for Computational Linguistics*, ACL, Morristown, USA, 311.

Park, J., Cai, H., and Perissin, D. (2018). “Bringing information to the field: Automated photo registration and 4D BIM.” *Journal of Computing in Civil Engineering*, 32(2).

Park, M.-W., and Brilakis, I. (2012). “Construction worker detection in video frames for initializing vision trackers.” *Automation in Construction*, 28, 15–25.

Park, M.-W., and Brilakis, I. (2016). “Continuous localization of construction workers via integration of detection and tracking.” *Automation in Construction*, 72, 129–142.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). “You only look once : Unified , real-time object detection.” *2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, 779–788.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). “Faster R-CNN: Towards real-time object detection with region proposal networks.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). “Self-critical sequence

training for image captioning." *2017 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, 1179–1195.

Rezatofighi, S. H., Milan, A., Zhang, Z., Shi, Q., Dick, A., and Reid, I. (2015). "Joint probabilistic data association revisited." *2015 IEEE International Conference on Computer Vision*, IEEE, Las Condes, Chile, 3047–3055.

Roberts, D., and Golparvar-Fard, M. (2019). "End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level." *Automation in Construction*, 105, 102811.

Ross, D. A., Lim, J., Lin, R.-S., and Yang, M.-H. (2008). "Incremental learning for robust visual tracking." *International Journal of Computer Vision*, 77, 125–141.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). "ImageNet large scale visual recognition challenge." *International Journal of Computer Vision*, 115(3), 211–252.

Shahi, A., Aryan, A., West, J. S., Haas, C. T., and Haas, R. C. G. (2012). "Deterioration of UWB positioning during construction." *Automation in Construction*, 24, 72–80.

Song, J., Haas, C. T., and Caldas, C. H. (2006). "Tracking the location of materials on construction job sites." *Journal of Construction Engineering and Management*, 132(9), 911–918.

Standing, S., and Standing, C. (2018). "The ethical use of crowdsourcing." *Business Ethics: A European Review*, 27(1), 72–80.

Statistia. (2021). "GDP at basic prices of the construction industry in Canada 1997-2020." <https://www.statista.com/statistics/519742/gdp-for-construction-sector-in-canada/>.

Smith, B. 2014. "Along the road album." Accessed July 23, 2016. https://

www.flickr.com/photos/byzantiumbooks/albums/72157645494430658.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). "Going deeper with convolutions." *Journal of Chemical Technology & Biotechnology*, 91(8), 2322–2330.

Tajeen, H., and Zhu, Z. (2014). "Image dataset development for measuring construction equipment recognition performance." *Automation in Construction*, 48, 1–10.

Tang, S., Golparvar-fard, M., Naphade, M., and Gopalakrishna, M. M. (2019). "Video-based activity forecasting for construction safety monitoring use cases." *2019 Internaltional Conference of Computing in Civil Engineering*, ASCE, Atlanta, USA, 204–210.

Teizer, J., Lao, D., and Sofer, M. (2007). "Rapid automated monitoring of construction site activities using ultra-wide band." *24th International Symposium on Automation and Robotics in Construction*, Madras, India, 23–28.

Truong, B. T., and Venkatesh, S. (2007). "Video abstraction." *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1), 3.

Tzutalin. (2015). "LabelImg." *GitCode*, <https://github.com/tzutalin/labelImg>.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is all you need." *arXiv*, ID: 1706.03762.

Vedantam, R., Zitnick, C. L., and Parikh, D. (2014). "CIDEr: Consensus-based image description evaluation." *arXiv*, ID: 1411.5726.

Vig, J. (2019). "A multiscale visualization of attention in the transformer model." *arXiv*, ID: 1906.05714.

Vig, J., and Belinkov, Y. (2019). "Analyzing the structure of attention in a transformer language model." *arXiv*, ID: 1906.04248.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). "Show and tell: A neural image caption generator." *arXiv*, ID: 1411.4555.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2017). "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 652–663.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y. (2014). "Learning fine-grained image similarity with deep ranking." *arXiv*, ID: 1404.4661.

Wang, K., Yan, X., Zhang, D., Zhang, L., and Lin, L. (2018). "Towards human-machine cooperation: self-supervised sample mining for object detection." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, 1605–1613.

Wang, L., Liu, D., Puri, R., and Metaxas, D. N. (2020). "Learning trailer moments in full-length movies." *arXiv*, ID: 2008.08502.

Wang, Y., Liao, P.-C., Zhang, C., Ren, Y., Sun, X., and Tang, P. (2019). "Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety." *Advanced Engineering Informatics*, 42, 101001.

Xiao, B., Zhang, Y., Chen, Y., and Yin, X. (2021a). "A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation." *Advanced Engineering Informatics*, 50, 101372.

Xiao, B., Lin, Q., and Chen, Y. (2021b). "A vision-based method for automatic tracking of construction machines at nighttime based on deep learning illumination enhancement." *Automation in Construction*, 127, 103721.

Xiao, B., and Zhu, Z. (2018). "Two-dimensional visual tracking in construction scenarios: A

comparative study." *Journal of Computing in Civil Engineering*, 32(3), 04018006.

Xiong, B., Kalantidis, Y., Ghadiyaram, D., and Grauman, K. (2019). "Less is more: Learning highlight detection from video duration." *arXiv*, ID: 1903.00859.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). "Show, attend and tell: Neural image caption generation with visual attention." *arXiv*, ID: 1502.03004.

Xu, S., Wang, J., Shou, W., Ngo, T., Sadick, A.-M., and Wang, X. (2020). "Computer vision techniques in construction: A critical review." *Archives of Computational Methods in Engineering*.

Xuehui, A., Li, Z., Zuguang, L., Chengzhi, W., Pengfei, L., and Zhiwei, L. (2021). "Dataset and benchmark for detecting moving objects in construction sites." *Automation in Construction*, 122, 103482.

Yang, J., Park, M.-W., Vela, P. A., and Golparvar-Fard, M. (2015). "Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future." *Advanced Engineering Informatics*, 29(2), 211–224.

Yang, M.-D., Chao, C.-F., Huang, K.-S., Lu, L.-Y., and Chen, Y.-P. (2013). "Image-based 3D scene reconstruction and exploration in augmented reality." *Automation in Construction*, 33, 48–60.

Yao, T., Mei, T., and Rui, Y. (2016). "Highlight detection with pairwise deep ranking for first-person video summarization." *2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, 982–990.

Yilmaz, E., and Aslam, J. A. (2006). "Estimating average precision with incomplete and imperfect judgments." *15th ACM international conference on Information and knowledge*

management, ACM, New York, USA, 102.

YoongiKim. (2018). "AutoCrawler." *Git code*, <https://github.com/YoongiKim/AutoCrawler>.

Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., and Yan, J. (2016). "POI: Multiple object tracking with high performance detection and appearance feature." *2016 European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, 36–42.

Zhang, B., Titov, I., and Sennrich, R. (2019). "Improving deep transformer with depth-scaled initialization and merged attention." *arXiv*, ID: 1908.11365.

Zhang, C., and Arditi, D. (2013). "Automated progress control using laser scanning technology." *Automation in Construction*, 36, 108–116.

Zhang, K., Chao, W.-L., Sha, F., and Grauman, K. (2016). "Video summarization with long short-term memory." *2016 European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, 766–782.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X. (2019). "Object detection with deep learning: A review." *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212–3232.

Zhong, B., Xing, X., Luo, H., Zhou, Q., Li, H., Rose, T., and Fang, W. (2020). "Deep learning-based extraction of construction procedural constraints from construction regulations." *Advanced Engineering Informatics*, 43, 101003.

Zhou, W., Whyte, J., and Sacks, R. (2012). "Construction safety and digital design: A review." *Automation in Construction*, 22, 102–111.

Zhu, Z., Park, M.-W., Koch, C., Soltani, M., Hammad, A., and Davari, K. (2016a). "Predicting movements of onsite workers and mobile equipment for enhancing construction site safety." *Automation in Construction*, 68, 95–101.

Zhu, Z., Ren, X., and Chen, Z. (2016b). "Visual tracking of construction jobsite workforce and

equipment with particle filtering." *Journal of Computing in Civil Engineering*, 30(6), 04016023.

Zhu, Z., Ren, X., and Chen, Z. (2017). "Integrated detection and tracking of workforce and equipment from construction jobsite videos." *Automation in Construction*, 81, 161–171.