SEMANTIC AND SPATIO–TEMPORAL UNDERSTANDING FOR COMPUTER VISION DRIVEN WORKER SAFETY INSPECTION AND RISK ANALYSIS

BY

SHUAI TANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Civil Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Associate Professor Mani Golparvar-Fard, Chair
Professor Khaled El-Rayes
Associate Professor Liang Liu
Associate Professor Nora El-Gohary
Associate Professor Derek Hoiem

ABSTRACT

Despite decades of efforts, we are still far from eliminating construction safety risks. Recently, computer vision techniques have been applied for construction safety management on real-world residential and commercial projects; they have shown the potential to fundamentally change safety management practices and safety performance measurement. The most significant breakthroughs of this field have been achieved in the areas of safety practice observations, incident and safety performance forecasting, and vision-based construction risk assessment. However, fundamental theoretical and technical challenges have yet to be addressed in order to achieve the full potential of construction site images and videos for construction safety.

This dissertation explores methods for automated semantic and spatio–temporal visual understanding of worker and equipment and how to use them to improve automatic safety inspections and risk analysis: (1) a new method is developed to improve the breadth and depth of vision-based safety compliance checking by explicitly classifying worker–tool interactions. A detection model is trained on a newly constructed image dataset for construction sites, achieving 52.9% mean average precision for 10 object categories and 89.4% average precision for detecting workers. Using this detector and new dataset, the proposed human-object interaction recognition model achieved 79.78% precision and 77.64% recall for hard hat checking; 79.11% precision and 75.29% recall for safety vest checking. The new model also verifies hand protection for workers when tools are being used with 66.2% precision and 64.86% recall. The proposed model is superior to methods relying on hand-made rules to recognize interactions or that reason directly on the outputs of object detectors. (2) to support systems that proactively prevent these accidents, this thesis presents a path prediction model for workers and equipment. The model leverages the extracted video frames to predict upcoming worker and equipment motion trajectories on construction sites. Specifically, the model takes 2D tracks of workers and equipment from visual data –based on computer vision methods for detection and tracking– and uses a Long Short-Term Memory (LSTM) encoder-decoder followed by a Mixture Density Network (MDN) to predict their locations. A multi-head prediction module is introduced to predict locations at different future times. The method is validated on an existing dataset TrajNet and a new dataset of 105 high-definition videos recorded over 30 days from a real-world construction site. On TrajNet dataset the proposed model significantly outperforms Social LSTM. On the new dataset, the presented model outperforms conventional time-series models and achieves average localization errors of 7.30, 12.71 and 24.22 pixels for 10, 20, and 40 future steps, respectively. (3) A new construction worker safety analysis method is introduced that evaluates worker level risk from site photos and videos. This method evaluates worker state, which is based on workers' body pose, their protective equipment use, their interactions with tools and materials, the construction activity being performed, and hazards in the workplace. To estimate worker state, a visual–based Object–Activity–Keypoint (OAK) recognition model is proposed that take 36.6% less time and 40.1% less memory while keeping comparably performances compared to a system running individual models for each sub-task. Worker activity recognition is further improved with a spatio-temporal graph model using recognized per-frame worker activity, detected bounding boxes of tools and materials, and estimated worker poses. Finally, severity levels are predicted by a trained classifier on a dataset of images of construction workers accompanied with ground truth severity level annotations. In the test dataset, the severity level prediction model achieves 85.7% cross-validation accuracy in a bricklaying task and 86.6% cross–validation accuracy for a plastering task.

ACKNOWLEDGMENTS

I would like to first thank my advisor, Professor Mani Golparvar-Fard for his support, guidance, and care during my Ph.D. at the University of Illinois. Mani, you took me under your wing when I was uncertain and helped me all the way to strive as a creative and independent researcher. Your teaching will keep inspiring me for my future career. I'm grateful for this wonderful journey and the most remarkable time in my life.

I would also like to express gratitude to other members of my doctoral committee, Professor Khaled El-Rayes, Professor Liang Liu, Professor Nora El-Gohary, and Professor Derek Hoiem. Thank you for your insights, feedback, and supports. This thesis would not have been possible if it were not for your patience, knowledge, and help.

My life in Champaign is surrounded by many talented people who I'm privileged to know, including my RAAMAC family Jacob Lin, Dominic Roberts, Wilfredo Torres Calderon, Amir Ayman Elsayed Mahfouz Hassan Ibrahim, Juan Diego Nunez Morales, Fouad Amer, Kevin Han, Yumo Chi, Yeritza Perez, Gustavo Garcia; my broader University of Illinois family Zhizhong Li, Jiajun Lu, Youjie Li, Rajbir Kataria, Tanmay Gupta, Chuhang Zou, Mike Yeh, Kevin Shih, Ernest-John Ignacio, and many people whom I haven't had a chance to say goodbye before they graduated. I have spent several hard winters in Champaign and your company always warm my heart. Graduate school life couldn't have been this enjoyable without all of you by my side.

I thank my industry partners for their time and effort to support my research. I want to thank the National Science Foundation (NSF) for funding my research through grants CMMI-1544999 and CMMI-1446765. I have done three fascinating internships with wonderful people during my Ph.D. training, I would like to thank my mentors Dr. Maryam Rezvani at Autodesk, Dr. Liang Yu, and Peng Chen at Alibaba Cloud, Dr. Milind Naphade and Murali Gopalakrishna at Nvidia, for their support and help. Thanks for introducing me to the amazing technical world.

To my parents, Xianbin Tang and Ni Hu, words can not express how much I love you. You are always my role models and my backbone. Thank you for your unconditional love and selflessness to support me in overcoming many challenges. To my girlfriend, Jiaxin, you are my inspiration to be better. Thank you for your love and patience.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Despite decades of efforts to achieve the “zero-accident” safety vision, construction safety remains to be a challenging problem today. In the past decade, efforts for improving safety management practices have not yet met a significant nation-wise safety enhancement in the United States. According to a report of the Occupational Safety and Health Administration (OSHA), 1008 fatalities were reported from construction sites in 2018 [1], taking 21.1% of the total annual workplace fatalities and reaching a new highest point since 2011 (Fig 1.1). Unsafe activities, dangerous site conditions, and ergonomic risks result in a large number of worker fatalities and injuries. The Bureau of Labor Statistics (BLS) [2] reports that, in 2018 alone, fall through the surface and existing opening causes 60 deaths, exposure to electricity causes 86 deaths, struck-by or caught-in equipment and objects causes 132 deaths, and pedestrian struck-by vehicle causes 56 deaths. BLS also reports a growing number of non-fatal injuries since 2011. Categorized by the source of injuries, in 2018, 16,960 injuries were caused by parts and materials, 8,240 injuries were caused by powered/non-powered hand tools, 5,300 injuries were caused by vehicles. When categorized by the part of the body affected, 11,110 back injuries, 6,640 head injuries, 4,850 foot injuries, and 13,930 hand injuries were reported. The number of construction fatalities and injuries has brought severe financial impacts to the industry. An early study [3] estimated, on average, one fatal accident results in a 4 million dollars loss, and per non-fatal accident causes 42,000 dollar loss. Another recent article from Midwest Economic Policy Institute [4] reports, from 2011 to 2016, the average 867.8 annual construction worker fatalities cost the United States nearly 5 billion dollars in lost production, loss of family income, pain and suffering costs, and reduced quality of life every year. The tremendous losses in construction companies’ financial well–being and workers’ quality of life calls for future research to identify, assess, and manage unsafe practices on construction sites. This thesis points out three critical gaps

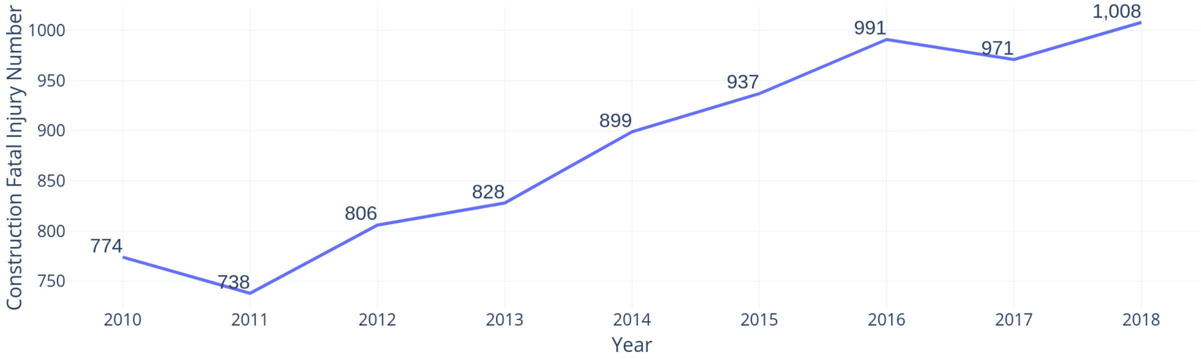


Figure 1.1: Construction Workers Total Fatalities from financial year 2010 to 2018

in the practical and theoretical knowledge of current construction safety:

1. **Unrecognized hazards and underestimated risks associated with behaviors**. Behavior–based safety programs are the most commonly implemented safety control on construction sites, whose scopes are protective equipment noncompliance, workers' exposure to hazardous areas, and failure to follow the safety procedure [5]. Historical behavior data are used as training materials for toolbox talk and safety courses [6, 5]. However, such data only represent a small subset of potential scenarios that unfortunately resulted in injuries [7]. The generalization from this knowledge may not faithfully reflect all hazard scenarios and their degrees of severity in the future. Partially because of this, a large portion of hazards is not recognized. Albert *et al.* [8] discovers on average only 46% of hazards are recognized by workers in diverse projects. A study conducted in Australia [9] shows workers failed to recognize 57% of hazards in the work environment. Haslam *et al.* [10] analyze 100 individual accidents and find out 42% of accidents are associated with inadequate hazard recognition skill.

2. **Insufficiency of paper–based safety inspections**. Safety inspections, in the form of paper-based checklists, document unsafe conditions, and hazards on construction sites. Manual safety inspection for a construction site is typically scheduled on a weekly to monthly basis by a small group of safety engineers in the U.S. Researchers [11, 12, 13, 14, 15, 16, 17] have considered many bottle–necks faced by the manual safety inspections: (1) the prolonged checklist-filling time diverts attention paid to

safety practices; (2) many safety–critical changes are left unrecognized and unreported between two scheduled safety inspections; (3) the manual inspection processes are subjective, costly and not comprehensive to avoid biases; (4) an unwelcome focus on the numbers instead of the content [17]. Besides these practical drawbacks, in a safety analysis cycle [18] the slow, costly and manual safety inspections also prevents obtaining timely feedback after corrective actions are made. It ultimately undermines the efforts to validate factors for systematically site safety improvement [12, 14, 19, 20]. As a consequence, the current safety management often stays as a tactical response to fatalities and injuries instead of strategical planning to prevent accidents [13].

3. **Safety quantification and measurement**. Commonly used safety indicators are based on historical injury data, such as the Incident Rate (IR) and Days-away-from-Work. While providing long–term safety performance measures and serving as benchmarks across projects and companies, they are reactive and can only be updated after accidents happen. For a particular project, they often provide too few data points to analyze and to proactively predict future accidents [21]. Also, they do not support continuous learning to further improve safety conditions when no injury occurs in a project for a long period [22]. Proactive safety measurement techniques, such as construction risk assessment and job hazard analysis, are also costly and labor–intensive [7].

The key concept to address these gaps is the automation in the perception, the prediction, and the risk analysis in safety practices [23, 24, 25, 5]. The improved perception technologies for safety observations allow safety practices, both good or bad, to be timely reported to the management and cover every safety–critical location in a construction site. Previous studies have articulated the importance of advancement in the safety sensing technologies [13, 24, 26, 18, 27] that far more safety non-compliance cases and hazards can be recognized than the manual approach, and that resolving the recognized cases ultimately renders a safer work environment. Early forecasting on imminent accidents and near-miss incidents prevents workers from being harmed and helps construction companies decide on initiatives that can proactively improve safety and lower their insurance premiums. Risk analyses based on manual and automatically recorded data are crucial to building more objective standards for risk assessment. A reduction in the number of observed safety non-compliance levels and worker ergonomic risks indicates an improvement of safety [26, 28, 29]. When implemented at the enterprise–level, automatic risk assessment can also serve as a common benchmark to compare one project against one another and prioritize safety training on projects [30].

Improving the perception of safety practices is the first and fundamental step towards this grand vision. To achieve this, it is crucial to make good use of massive and inexpensive construction site data that are already collected. Recently, a growing volume and types of visual data have been collected on construction sites. Today, construction sites generate a significantly great number of visual data [31], about 325,000 images are taken by professional photographers, 95,400 images by webcams, and 2000 images weekly by the construction project team at a typical commercial building project about 750,000 squared-feet. The ground robot, Unmanned Aerial Vehicle (UAV), and laser scanner have been deployed to automatically collect indoor and outdoor visual data (see Fig. 1.2). These data contain information regarding construction resources' appearances, geometries, distances, locations, activities, and motions history. This information is essential to make safety inspections for safety rule checking, hazard identification, accident prevention, etc. Some images collected by the safety team have already been used as visual evidence along with written safety reports (see example in Fig. 1.3). The recent growth in the number and types of construction site visual data provides an opportunity to incorporate them to develop new means to understand site safety conditions. A recent study [18] shows incorporating manually image review is helpful in a continuous safety management feedback loop.

1.1 Point of Departure

Several commercially available solutions (e.g. AutoDesk BIM 360, SafeSite, Safety-Report) provide visual data collection features for safety inspections using mobile devices. These solutions allow site images and videos to be categorized, tagged, and uploaded by workers and safety inspectors. Any safety observation or comment made by onsite personnel can be attached along with the uploaded visual data. This adaptation to digitized safety reports

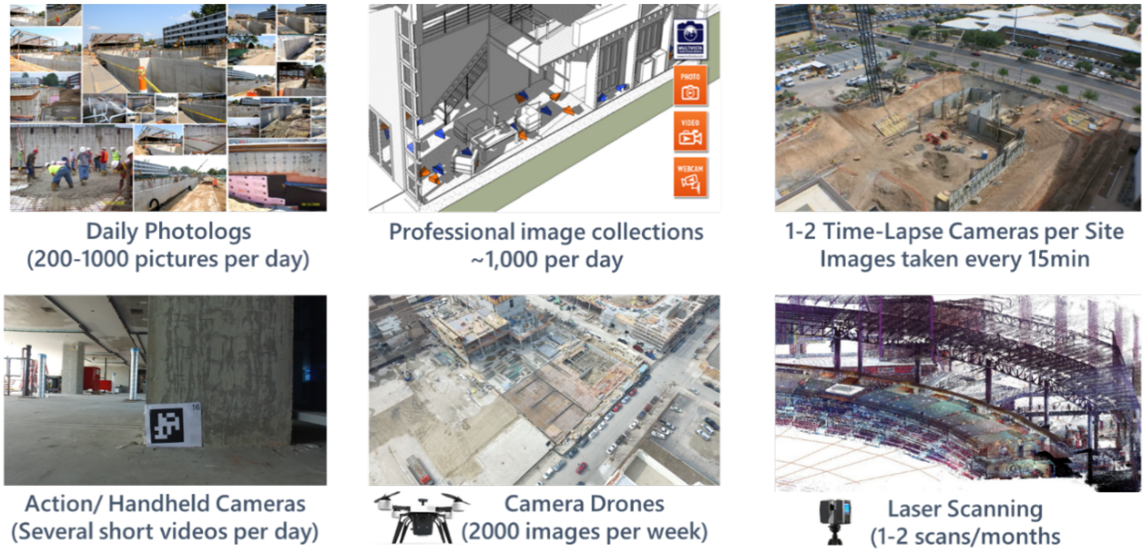
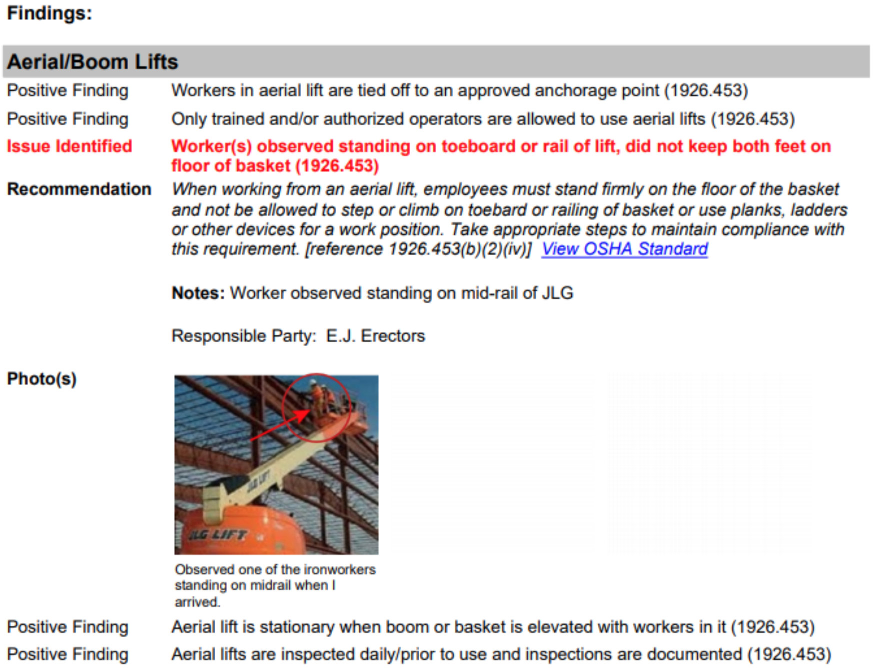


Figure 1.2: Various forms of visual data and their frequency of capture



Findings:

Aerial/Boom Lifts

Positive Finding	Workers in aerial lift are tied off to an approved anchorage point (1926.453)
Positive Finding	Only trained and/or authorized operators are allowed to use aerial lifts (1926.453)
Issue Identified	**Worker(s) observed standing on toeboard or rail of lift, did not keep both feet on floor of basket (1926.453)**
Recommendation	*When working from an aerial lift, employees must stand firmly on the floor of the basket and not be allowed to step or climb on toebard or railing of basket or use planks, ladders or other devices for a work position. Take appropriate steps to maintain compliance with this requirement. [reference 1926.453(b)(2)(iv)]* *View OSHA Standard* **Notes:** Worker observed standing on mid-rail of JLG Responsible Party: E.J. Erectors
Photo(s)	Observed one of the ironworkers standing on midrail when I arrived.
Positive Finding	Aerial lift is stationary when boom or basket is elevated with workers in it (1926.453)
Positive Finding	Aerial lifts are inspected daily/prior to use and inspections are documented (1926.453)

Figure 1.3: An example of safety observation report with image evidence. Image credit to Safety-Reports.

brings down the granularity of safety update frequency from a monthly to a daily basis. From a company's perspective, these solutions provide well–structured safety reports and a constant influx of safety information that can be stratified at the sub-contractor, project, and regional levels. Such archive of safety information opens new opportunities for more efficiently selecting sub-contractors and more accurately conducting accident investigations. However, this feature still depends on manual tagging and descriptions and does not economically solve the fundamental problem of low frequency in safety inspections.

Methods to provide automatic safety inspections rely on computer vision and machine learning models to process the collected visual data. Integrated solutions have been offered from site data collections to hazard and non-compliance identification. For instance, SmartVid.io uses object detection and voice recognition to automatically tag images portraying over 50 safety–related entities, such as workers, PPE, equipment, and material. The resulting tags are used for conducting basic PPE compliance checkings, such as housekeeping, fall protection, and PPE compliance. Skycatch, DroneDeploy, and SiteTrax.io provide image collections from UAV and a broad range of downstream computer vision applications from surface mapping, object detection, and asset tracking. In the risk assessment field, worker ergonomic risk assessment is carried out by feeding visually recognized body joint angles to standard ergonomic severity assessment systems [32, 29, 33].

However, the current computer vision and machine learning methods applied to construction safety in the existing literature haven't addressed some fundamental theoretical and technical problems in automating safety inspections and risk analysis. For example, existing automatic safety observation tools, powered by deep learning object detection models, of Smartvid and SiteTrax do not consider workers' activities and their surrounding work environment when conducting PPE compliance checking. Such as knowing the fall arrest systems are only required if workers operate on elevated surfaces. The hazardous proximity indicator for collision prediction [26, 13] still passively assess risks by creating buffer zones, it does not consider the dynamic of moving workers and equipment. Existing visual–based ergonomic assessment methods don't consider workers' activity, their PPE use, their interactions with tools and materials, and the workplace context as variables in the safety analysis. In addition to these theoretical limitations, early studies [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44] on visual–based productivity and safety applications often validate their methods on datasets with limited size and variance, making the generalization of their methods across projects challenging.

1.2 Computer Vision Methods Applied and Contributions Made

Applied computer vision and machine learning research for construction safety is a rapidly evolving topic and researchers closely follow the latest research in the general computer vision community. However, problem formulation, data collection, and validation procedures remain to be the main challenges in this domain.

As previously mentioned, localization of construction resources is the fundamental input required by the majority of construction safety applications, such as for PPE compliance checking, worker and equipment tracking, and worker pose estimation. Currently, many localization tasks in the computer vision domain are formulated as object detection that detects tight bounding boxes around object instances. This dissertation follows this task design of localizing for its efficiency and simplicity. Based on the classic object detection task, this dissertation expands the scope of current computer vision-based safety applications by employing and developing advanced computer vision models that propose for a better semantic and spatio-temporal understanding of construction resources from site photos and videos. The developed advanced computer vision models can be directly integrated into object detection models, or directly taking the output bounding boxes of object detection models as the input, or reuses intermediate object detection model results. This is achieved by employing the classic Faster RCNN [45] object detection framework, which naturally allows multiple types of outputs in addition to bounding boxes to be generated from single images.

In particular, this dissertation explores the following three major directions that are based on the classic object detection task, their approaches and contributions are:

i **Semantic**: the human-object interactions between worker and their PPE and tools use are modeled and recognized directly from construction site static images. An end-to-end computer vision model is proposed that first detects worker, equipment, and tools, and recognizes workers' interactions with detected tools and equipment.

ii **Spatio-temporal**: the motion trajectories of workers and equipment are captured from site videos. Then a new forecasting model is proposed based on recurrent neural network formulation and a sequence generation task. The forecasted locations are checked against pre-defined safety regions to predict fall and collision hazards and sends notifications to the management.

iii **Semantic+Spatio-temporal**: a risk analysis method is proposed to predict worker-level severity levels based on an estimated worker state feature vector consisting of worker's activity, worker interactions with tools, equipment, and material, workers' body pose, and surrounding hazard condition. All components of the worker state feature vector are generated from a stack of consecutive frames. To generate the per-frame visual recognition results, a variant of Mask RCNN [46] is used which generates objects, keypoints, and activity labels at 26 frames per second. To refine per-frame activity recognition, a spatio-temporal graph model is proposed that takes individual frames' object, activity, and keypoint features. A pre-trained model is used to lift 2D worker pose to 3D.

Despite the recent massive growth of construction site photos and videos, obtaining large-scale, diverse, and balanced construction datasets is still challenging. As a consequence, few public released models can be widely deployed on construction sites. The two main data collection challenges are: first, collections of diverse construction site photos; second, effective and economical annotation procedures. This is due to the fact that public available construction site images are scarce, and annotating construction resources requires expert knowledge, sometimes years of training. This dissertation employs and extends the previous crowd-sourcing annotation method [47] by breaking down complex annotation tasks into well-defined steps and using the combination of crowd-sourced annotators and in-house expert annotators. This dissertation also demonstrates obtaining online construction site photos using web–crawling techniques. The datasets presented in this dissertation are the largest and the most diverse in their own tasks compared to the existing publicly available datasets.

The constructed image datasets are used to verify and validate the proposed models: (i) the human-object interactions recognition model is first verified in a new construction site image dataset with abundant object and interaction annotations. The best model is validated on retrieving PPE non-compliant workers from the existing image repository and outperforms alternative strategies using detected bounding boxes; (ii) the proposed motion trajectory forecasting model is verified in a newly constructed construction worker and equipment long-term trajectory dataset. The proposed forecasting model also outperforms previous generic trajectory forecasting models on a short-term trajectory dataset collected on a university campus and streets. The proximity hazard prediction application is validated in a real–world construction site and demonstrated good potential in helping safety managers to identify workers crossing excavation area events; (iii) The proposed computer vision models to generate worker state vectors are verified in a newly constructed single worker video dataset recording plastering and bricklaying operations. The proposed risk analysis method is validated on a set of construction image dataset and ground-truth human annotations on single worker severity levels. The full worker state is shown to be more informative to predicting frame-wise worker severity level, compared to an alternative approach predicting severity levels using an image classifier.

The next section describes the research objectives and discusses their significance. The rest of the chapters are organized as in the following order: Chapter 3 presents and validates a novel method to recognize worker-equipment interactions for PPE compliance checking; chapter 4 presents and validates a novel method to forecast the worker and equipment motion trajectory; chapter 5 present a novel machine learning–based worker level risk analysis method that leverages worker activity, worker pose, their PPE use, their interactions with tools and materials, and the workplace context. Each chapter provides the background, the introduction of the problem, the methodology, and the evaluation process.

1.3 Research Objectives

The overarching goal of this dissertation is to improve PPE compliance checking, forecast unsafe conditions, and behaviors, and enhance worker–level risk analysis using existing image and video data collected from construction sites. This is done by researching semantic and spatio-temporal visual understanding of construction site images and videos. This overarching goal is achieved with the following research objectives:

Objective 1:

Create and validate a crow-sourced method to build construction image datasets based on construction safety knowledge and existing construction visual data.

Research Questions:

(1)What are the similar language structures of safety rules? (2) What kind of construction resources and their relations can be visually detected by computer vision models? (3) can we use train non-experts in construction and teach them to apply safety rules and annotate on construction images? (4) How do we validate their annotations?

Significance:

It is challenging to build large–scale construction dataset. Our research is one of the first attempts to analyze semantic components of construction safety rules and annotate their visual correspondences using a crowd-sourced approach. The safety rule parsing and crowd-sourced method are the key components to identify safety checking items that can be automated by computer vision and machine learning models.

Objective 2:

Create and validate a computer vision method to improve visual-based PPE compliance checking concerning workers' interactions with the tools and equipment.

Research Questions:

(1) How are visual-based PPE compliance checking results used to improve safety inspection? (2) How can we assess the severity of PPE non-compliance under various working scenarios? (3) Can we make visual PPE compliance checking and severity assessment robust under cluttered scenes and computationally efficient? (4) can we discover the interactions of construction resources from safety regulations and automatically recognize their visual correspondence.

Significance:

Since visual-based PPE compliance checking is a well-recognized computer vision application for construction safety. Making it more robust to real-world data and being able to assess the severity of PPE incompliance further cuts down the time for reviewing and validating safety reports. It also makes it possible to establish a common benchmark at the enterprise level to compare PPE compliance performance across projects.

Objective 3:

Create and validate a machine learning method to forecast worker and equipment motion trajectories for hazardous proximity indicators.

Research Questions:

(1) What is missing in the location-based proximity hazard indicator? (2) Can we make it more proactive by predicting the motion trajectories of moving workers and equipment? (3) Can we make the predictions with respect to the context information?

Significance:

Proactive proximity hazard identification and accident prevention is an appealing idea to save lives and cut down the cost. Struck-by prevention and too-close near-miss reporting are not yet fully proactive without considering the dynamic of moving workers and equipment.

Objective 4:

Create and validate computer vision models for jointly recognizing workers' activity, their body pose, their PPE use, and their interactions with tools and materials.

Research Questions:

(1) The worker activity, their tools use, and worker body pose can be closely related, can we perform these three different tasks simultaneously through optimizing model architecture? (2) Can we design the model in a way that tasks can provide contextual information to each other? (3) How can we use model outputs to describe safety–critical information of a worker?

Significance:

Visual–based worker activity recognition, PPE and tools detection, and worker pose estimation are conventionally treated as different tasks performed by very different models. their unification under a general model framework has not been explored. Despite that they are closely related, their mutual information is not used for improving the performances of individual tasks.

Objective 5:

Create and validate a machine learning–based approach for worker–level risk analysis using visually recognized work information as input.

Research Questions:

(1)How to formulate risk scores for a worker at a given time step in terms of probability and severity of the worker's status? (2) Can a worker's severity level displayed in a single image being evaluated consistently among safety practitioners? (3) Can we predict human–rated worker level severity levels with the visually recognized worker information?

Significance:

Worker risk analysis is often explored in the context of proximity and ergonomic health. Seldom effort is made on investigating comprehensive information of workers' status, including their activity, their body pose, their PPE use, tools, and material use, as well as the workplace context can be used to predict worker severity level. On the other hand, visual–based PPE compliance tools haven't considered risk severity level. This research automatically collects workers' status and exam how they can be informative in predicting human assessment on severity level.

CHAPTER 2: HUMAN-OBJECT INTERACTION RECOGNITION FOR AUTOMATIC CONSTRUCTION SITE SAFETY INSPECTION[1]

2.1 Introduction

Safety inspections on construction sites are a vital part of any company's injury prevention efforts. Despite the improvement in safety education and practices, the ever-growing desire for higher productivity is negatively impacting safety on construction sites [8]. For example, in the United States, around 20% of fatal injuries occur on construction sites [1], while construction workers make up less than 10% of the total workforce [2]. Noncompliance with proper protection and incorrect use of tools often result in environmental harm, object contact, and body part injuries. In 2017 alone, 145 fatalities were due to exposure to harmful substances or environments, and another 133 fatalities were caused by contact with objects and equipment [1]. In addition, in 2017, the U.S. Bureau of Labor Statistics (BLS) reported 8,280 injuries caused by powered or non-powered hand tools, 6,560 head injuries, 4,850 foot injuries, and 13,530 hand injuries [48]. Ensuring that workers wear safety gear and use tools correctly makes a difference. The financial losses due to accidents on construction sites are substantial, often around billions of dollars every year. According to a previous study [3], the average total cost of a fatal accident is $4 million and that of non-fatal accident is above $42,000 (values in 2002 dollars).

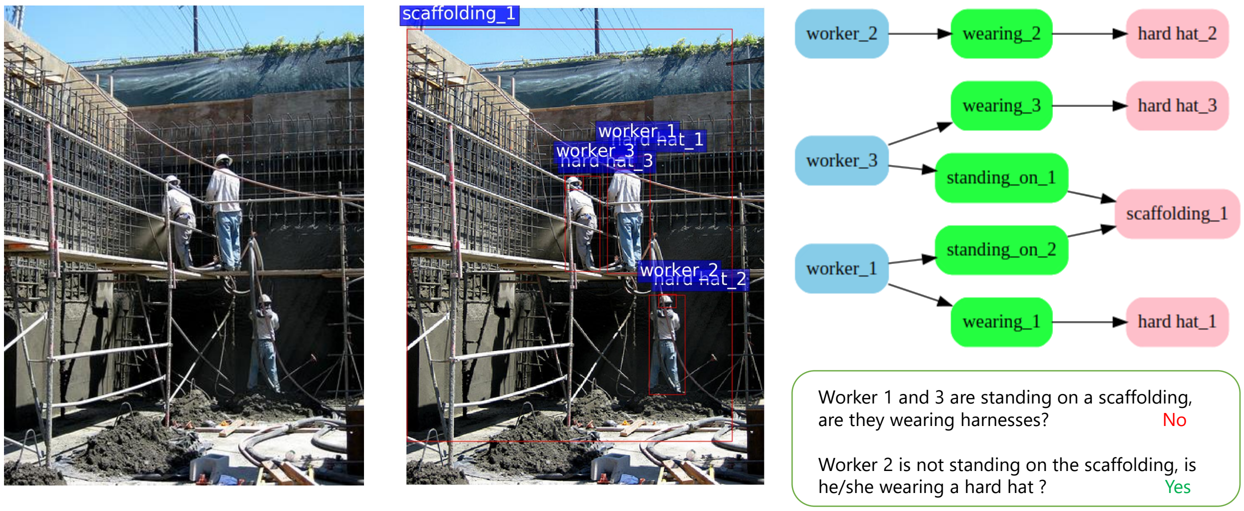
Timely, effective, and accurate safety inspections are essential for evaluating and improving construction safety. Safety inspections are typically performed by manual observers who produce biweekly or monthly written reports. The frequency at which this process is carried out does not allow hazards to be recognized and eliminated swiftly. Furthermore, text-based written safety reports often do not describe safety hazards at a sufficient level of detail [23]. To bridge these two gaps in practice, efforts need to be made on automatically conducting

[1] This chapter in whole or in part is published in the Automation in Construction journal.

safety inspections on massive amounts of data that is rich in details and easily comprehended by humans. Visual data from construction sites is a viable candidate for this purpose. Today, construction sites generate hundreds to thousands of photos and videos on a daily basis [31]. However, the majority of visual data is underutilized or used for progress tracking and as-built documentation purposes [23, 25]. Researchers have been investigating use of computer vision methods on construction visual data. As generic computer vision methods [49, 50, 46] become more potent and accessible, a new opportunity arises to incorporate the massive and unmanageable amount of construction site visual data. One example is to complement today's safety inspection and documentation practices [25]. Photos are already taken on job sites on a daily basis, particularly when they are pooled from document management systems. Hence, automatically checking safety compliance from site photos increases frequency of site safety inspections. This procedure is particularly valuable if project teams are understaffed. The effectiveness of such a procedure has already been demonstrated by recent commercial solutions. For example, SmartVid's artificial intelligence (AI) engine, which has been successfully applied to more than 1000 projects, detects hard hat, gloves, and safety vests for worker safety. Identified potential safety compliance issues are tagged and sent to safety managers for review and further comment. Corrections or manual annotations also help improve completeness and accuracy of data for future machine learning training and development purposes. When used at the enterprise level, such automated solutions help companies benchmark their projects against one another and prioritize safety training on projects. Comparing incident rates recorded via these systems against average industry numbers also helps construction companies decide on initiatives that can proactively improve safety and lower their insurance premiums.

Research on applying computer vision for safety inspection is still at an early stage [34, 35, 36, 37, 38, 39, 51, 32]. Many methods are not tested for robustness to occlusion, variation in object size and appearance, and differences in construction scenes [39, 51, 32]. Vision-based worker and equipment safety proximity checking methods often convert objects' 2D locations to 3D by camera calibration or monocular depth estimation [52]. Activity recognition from temporal data has also been investigated for safety applications. Previous researches on recognizing single worker and equipment's activity assume a few types of actions are performed during limited temporal intervals [40, 41, 42, 43, 44, 53]. Group and multi-agent activity recognition for construction often leverage the spatial relations between workers and equipment. Cai *et al.* [54] apply hand-made spatial cues, such as head pose and body orientation, to recognize groups of workers and equipment and then classify group activity. Kim *et al.* [55] improve individual's activity recognition by designing rule-based post-processing that leverages interacting excavators and dump trucks' object type, reconstructed 3D locations, and individual actions. Similarly, complex safety inspection tasks, such as fall protection, can not be handled easily only using detector outputs [56]. Existing safety gear compliance checking methods, including SmartVid's AI engine, often rely on rule-based post-processing. For example, checking hardhat compliance by determining whether the hard hat box overlaps with the uppermost part of the worker box [39] will fail if the worker is not in an up-right posture. In this paper, the authors expand on this further by learning to recognize workers' interactions in static 2D images, as opposed to recognizing them using hand-made rules on detected construction objects. To learn the interactions, one needs a uniform and scalable interaction representation. Tang and Golparvar-Fard [57] present a framework to correlate construction objects with linguistic constraints in site images. This framework is extracted from textual safety rules and is associated with their visual correspondences in site images. However, Tang and Golparvar-Fard [57] only provide early examples of these correlations, and different interactions were not formalized in a consistent manner. The absence of representations that are both structured and formal impedes data annotation and learning tasks. In this article, the authors build on Tang and Golparvar-Fard [57] and formalize interactions in a uniform structure as presented in human-object interactions (HOIs).

Many safety checking tasks can be formulated as HOI recognition. In a job hazard analysis, individual steps of a job are often described by action verbs or action phrases, such as "holding tools", "using grinders", and "climbing ladders". Potential hazards are associated with each step, and control measures are suggested. Similarly, the severity of hazards recognized from images can be more accurately evaluated by recognizing workers' interactions with the tools and the equipment. For instance, not wearing a face shield is not necessarily a noncompliance when the worker is not using a tool that produces sparks, heat, or strong light; not wearing hand protection while using tools is more critical than not wearing hand protection in the office trailer. Based on these observations, the authors propose a learned HOI model which improves the performance of existing vision-based safety checking methods and prevents false alarms. The authors argue that such a model not only improves existing safety gear compliance checking tasks, but also highlights critical noncompliance incidents. The authors present a number of experiments to validate their claims. First, the proposed HOI recognition model is compared and validated with a previous HOI method and a rule-based method. The method is more effective at retrieving actual interaction instances. Second, to demonstrate safety gear compliance checking, the proposed HOI model is compared with alternative checking strategies, such as using a rule-based HOI method and using object detection alone. For checking hard hat and safety coloring compliance, the model achieves better precision and recall compared to the rule-based method. For checking hand protection compliance while using tools, the HOI formulation outputs significantly fewer false positives than when object detection is applied alone. Also for this example, the proposed HOI model achieves better precision and recall compared with the rule-based alternative. These experiment results suggest a learned HOI model has practical value and improves existing vision-based safety checking methods. The authors will discuss these in greater detail in the Experiment Results and Discussions section. Figure 2.1 shows an example of the model output. The authors also introduce an approach for collecting and annotating construction images to build a new dataset upon which the proposed HOI models are trained and validated.



Worker 1 and 3 are standing on a scaffolding, are they wearing harnesses? No

Worker 2 is not standing on the scaffolding, is he/she wearing a hard hat ? Yes

Figure 2.1: Example of HOI model output. (**Left**) The input image. (**Middle**) The object detector branch recognizes all targeted construction resource instances. (**Right**) The HOI branch recognizes interactions between detected object instances. Once objects and HOIs are recognized, fall protection safety questions such as "Are worker_1 and worker_3 having personal fall arrest systems?" and "Is worker_2 wearing a hard hat?" can be answered directly using model outputs. Best viewed in color and high definition.

2.2 Related Work

Researchers on vision-based construction safety checking have been intensively investigating how to recognize various construction resources from construction visual data. The majority of previous work utilizes vision-based object detection methods to recognize and localize construction resources. In this section, the authors first review previous work on improvements on object detection and on the application of object detection to construction. Next, as the authors propose learning an HOI recognition model on object detection results for vision-based compliance checking, a review of existing vision-based HOI recognition methods is provided. The authors also introduce the closely related concept of "Scene Graphs", which can be used to integrate interaction instances into a semantically rich description of an image.

2.2.1 Recognizing Construction Resources for Construction Safety

Computer vision has been applied to construction safety in the context of recognizing and localizing construction resources such as workers, trucks, and excavators [34, 58, 36]. Detection of construction materials [59, 60], however, is often used to help progress monitoring. Researchers have proposed various representations to capture object locations in construction site images. Motion [58, 34], geometry [61], and appearance cues have been used to generate "object proposals" in the detection pipeline. These object proposals are described by various types of features, such as Histograms of Oriented Gradients and Hue-Saturation-Value color histograms [36]. Discriminative classifiers such as Support Vector Machines are used to categorize these object proposals as belonging to one of several target object categories.

Since 2015, major progress on the traditional two-stage "propose-then-classify" approach has greatly improved object detection models' accuracy and speed (e.g. FastRCNN [62], FasterRCNN [45], and MaskRCNN [46]). FasterRCNN, for example, takes an image as the input and detects objects by generating and then classifying object proposals in a single forward pass of the model. The object proposals can be reused for other tasks, such as generating the object instance segmentation masks in the detected bounding box or estimating person body keypoints from a detected person box. These tasks (object classification, instance segmentation, and keypoint estimation) can be regarded as running in parallel. MaskRCNN [46] is an example of this parallelism design. In MaskRCNN, both object classification and instance segmentation are built on top of proposed object bounding box regions. Region Proposal Network (RPN) [45] generates a set of proposal bounding boxes using the convolutional image features. Regional features of bounding boxes are extracted by Region of Interest Pooling (RoIPool), also from the convolutional image features. RoIPool evenly divides each box-shape region into cells such that regional features of all boxes will have equal dimensions. For each cell, RoIPool quantizes the image features covered by that cell's coordinates, then selects the max value as the representative value for this cell. MaskRCNN introduced an enhanced version of RoIPool called RoIAlign. For each cell, RoIAlign bilinearly interpolates the image features covered by that cell and obtains feature values at the center of that cell. This improves alignment between the box regions and regional feature values. For more details on RoIAlign, readers are encouraged to look into He *et al.* [46]. In this paper, the authors adhere to this parallelism design and build a sub-module network on top of FasterRCNN to recognize HOIs. This design enables a fast and integrated model to detect construction resources and recognize HOI between workers and tools/equipment.

The deep learning-based object detectors [63, 49, 45] often perform well with abundant training data. However, the construction-specific datasets used in recent work [64, 65, 66, 57, 56] are much smaller than generic object detection datasets [67, 68]. Despite this, deep learning-based object detectors still achieve comparable results to traditional object detectors. This is because the previous construction object detection datasets often consist of a few construction sites, small in the number of objects in the image, and lack intra-class variance of the objects within the same class. The authors present a new dataset that spans a larger number of construction sites. The authors also present construction resources bounding box annotations with higher intra-class variations in order to examine modern object detectors for more diverse and realistic construction situations.

Action recognition using ordinary images in the construction focuses on workers and excavators. Researchers investigate part-based keypoint features of a single object [61, 40] and visual appearance of a single object [47] to recognize its actions. The recognized actions are primarily used for productivity monitoring. On the other hand, automatic safety inspection from ordinary images often does not take actions into account. Modeling and recognizing interaction of workers with their environment and resources from site images, which is vital from a safety perspective, is a heavily underdeveloped area of research. Work that leverages actions for safety monitoring has thus far been conducted with rule-based approaches [35, 39, 42]. Luo *et al.* [65] apply rule-based method to recognize construction activities. They presented a predefined score matrix called Relevancy Network which is applied on the detected instances. By leveraging predefined semantic relations and spatial distances of detected objects, the most likely activity is determined. While the Relevancy Network performs well on the tested job site, it can not be directly applied to a different site, as the spatial layouts of detected objects might be completely changed.

2.2.2 HOI Recognition and Scene Graph

An HOI instance is often expressed in triplet form $< Subject - Predicate - Object >$. To recognize an HOI instance, the model needs to simultaneously localize both subject and object and predicting the interaction between them. Prior work on HOI recognition relies on hand-crafted features. Gupta *et al.* [69] classify interactions using spatial and functional constraints such as typical locations of human pose and manipulative objects; Yao *et al.* [70] define graph connections of objects and body parts and then apply a conditional random field on object detection and human pose estimation outputs; Prest *et al.* localize and track objects over time with respect to person locations [71]. In contrast, recent research on HOI recognition [72, 73, 74] explores the rich bounding box regional features from RoIPool and combines object detection and HOI recognition in a single model that is trained in an end-to-end fashion. The proposed HOI recognition model is inspired by these works. However, generic HOI studies focus on classifying the right interaction class of an object pair from multiple likely answers, e.g., holding, riding, and sitting between a person and a bike. In this paper, the proposed HOI recognition model is designed to search and determine whether certain interactions are present. This is a reasonable simplification because often the key interaction between worker and a type of construction object is required for identifying safety compliance issues. Xiong *et al.* [75] present a recent effort on learning to recognize workers' interactions for hazard identification. They model relations between two objects on construction sites with a conditional random field. Xiong *et al.* also treat visual relations as triplet forms and hazards as certain relations between a pair of objects. They provide early explorations of the concept. There are few key differences between Xiong *et al.* and this paper. First, in this paper, extensive experiments on different models are conducted. Multiple safety inspection tasks are thoroughly examined to validate the proposed HOI models against alternative vision-based strategies. These verification significantly extend the exploration by Xiong et al. Second, as proven in [72, 73, 74], exploring all potential object pairs can be effectively conducted. Third, a newly constructed large dataset is presented in this paper, as well as the detailed description of dataset construction.

Because triplet form has simple formulation and carries rich semantic information(e.g., time, the location, and logical order of events), triplet form offers an advantage and convenience in formulating many scene understanding tasks. For instance, HOI recognition is treated as an intermediate step to scene graph generation [76, 77], which captures a global snapshot of all objects' relations to each other in the scene using a directed graph. Scene graphs can also be constructed from visual relations as shown in [75]. In the last section, the authors will explain how scene graphs can be used for safety education.

2.3 Approach for Building an HOI Dataset for Construction Safety

While image datasets exist for workers, and many safety gear detection, an HOI dataset for construction sites is not presented in previous literature. In addition, existing object detection datasets for construction are often collected from a few construction sites and are accompanied by a small number of ground truth annotations. Models trained on these datasets often do not generalize well on other sites. To validate the proposed HOI model, this section describes the authors' approach for constructing a more diverse and safety-relevant object detection dataset and subsequently the first HOI dataset for construction. Inspired by a similar dataset building procedures [74], the authors conduct the following steps. First, a selection of safety rules is used to identify the set of construction objects and HOI classes. These keywords are used to find relevant images from public resources. Second, a crowd-sourced approach is used to annotate objects of the targeted classes. This approach consists of three parts: annotator qualification, annotation quality control, and annotation post-assessment. Third, a similar approach is used for HOI annotation. HOI annotators are asked to link previously annotated boxes given the interaction specified. All HOI data are post-assessed and validated. In the following section, the authors will introduce the HOI dataset building procedure in greater detail.

2.3.1 Defining the Scope of Safety Inspection Tasks

The same safety rule collection presented by Tang and Golparvar-Fard [57] is used here. It reviews over 1,000 generic safety rules covering various safety aspects, such as the general work environment, safety gear, elevated surfaces, and contamination control, etc. Tang and Golparvar-Fard [57] select safety rules that can be checked from visual appearance. For each selected safety rule, linguistic entities such as "worker", "wearing", and "hardhat" are manually parsed to linguistic semantic role labels: a single "Agent", a single "Verb", and zero to multiple "Theme"s. However, these semantic role labels are not necessarily triplets. In this work, they are further selected and parsed. When the semantic role "Agent" is realized by workers, the "Agent" role is treated as the Subject. Each tool and equipment realizing semantic role "Theme" is treated as an Object. A "Verb" or phrase derived from "Verb" (e.g., "wearing" and "standing_on") is treated as the Predicate between each pair of Subject and Object. In terms of the Cognitive Reliability and Error Analysis Method (CREAM) [78], this triplet formulation captures human errors related to error mode Object. It recognizes human errors by recognizing workers not interacting with the right objects when conducting activities; for instance, a worker standing on scaffolding but not wearing a fall arrest system, or a worker using a power tool but not wearing eye protection. Table 2.1 shows examples of safety rules and their primary HOIs.

The refined safety rule collection identifies 22 types of construction resources to be annotated; the authors list the definitions of these construction resources in Sec. 2.3.4. Ideally, all 22 classes should be included in the HOI experiments. Nevertheless, the authors discover visual correspondences for some of the HOI classes that are extremely rare in the currently available public databases. Examples include "workers-standing_on-guardrail" or "worker-riding_on-vehicle". Therefore, the authors report HOI recognition results only for three interactions and 10 construction resources (see Table 2.2 and Table 2.6 accordingly), as by themselves these categories provide an abundant number of HOI instances for evaluations.

Table 2.1: Examples of Selected Safety Rules and Their Primary HOIs

Safety Rules	**Primary HOIs**
Is appropriate foot protection required where there is the risk of foot injuries from hot, corrosive, poisonous substances, falling objects, crushing or penetrating actions?	worker-using-hand tool worker-using-power tool worker-wearing-foot protection
Are appropriate safety glasses, face shields, and similar equipment used while using hand tools or equipment that might produce flying materials or be subject to breakage?	worker-using-hand tool worker-wearing-eye protection worker-wearing-face protection
Are goggles or face shields always worn when grinding?	worker-wearing-eye protection worker-using-power tool
Is approved respiratory equipment provided and used when appropriate during spraying operations?	worker-wearing-respiratory equipment
Are welders and other workers nearby provided with flash shields during welding operations?	worker-wearing-face protection worker-using-power tool

Table 2.2: HOI Predicate Definitions

Predicates	**Definition**
standing_on	Workers supported by elevated surfaces against gravity. Include worker standing, sitting, and climbing on scaffolding planks, bracing, pipe.
using	Instances that show workers using hand and power tools in any intermediate stage of a construction activity. Carrying the tools on the waist belt and holding the tools during transportation or being idle are not considered as using. This criteria is checked by examining workers' body poses and surrounding context.
wearing	Workers wearing protective equipment (e.g., hard hat, hand protection, ear protection) as a means to protect body parts from sources of injury, e.g., sound, light, physical impact, heat, chemical, dust.

2.3.2 Image Collection

The authors design a guideline for image data collection that is inspired by Lin *et al.* [67]. Open image resources are explored, and construction site images are crawled, then manually selected. Such selection aims to collect images of high data variety in order to reflect real-world scenarios. The authors use a web-crawling tool with the construction safety HOI class and object categories as keywords to retrieve images from Flickr and Google Images, and also they extensively searched on ImageNet.

Flickr. The authors use the *group_id* and *tags_mode* Flickr APIs to retrieve images from construction-related user image groups. As of September 2017, the largest group had collected over 38,900 construction site images. Flickr also allows searches through image tags. By setting *tag_mode* to "ANY" and "ALL", images whose user-defined tags match any or all search keywords will be returned. The authors use object names and their synonyms in "ANY" mode for each search and use "ALL" mode to search tuple $< subject, object >$ pattern tags (e.g. "worker, hard hat"). The max number of the image returned is set as 10000 for each search.

Google Images. HOI triplets (e.g., "worker climbing ladder") are used to crawl images on Google Image. Also, the prefix "construction" is added on each search to resolve ambiguous terms. The maximum number of images returned is set to 10,000 for each search.

ImageNet. ImageNet [79] is a large dataset of images depicting a broad variety of objects, of which only a portion are construction-specific. The authors manually selected images from the following ImageNet entries: "repairman", "structure", "scaffolding", "hard hat", "scaling ladder". In total, 1,170 images were selected from the ImageNet dataset.

Unsuitable Images. A few rounds of reviews are conducted to select final candidate images. The goal for review is to select realistic images as taken from real-world construction sites. The most desired images depict construction workers in non-frontal/non-iconic views [67] and with a substantial amount of construction site environment context. Five common unsuitable properties are identified; one unsuitable image often has one or more of these properties. (see Fig. 2.2). Unsuitable images are removed because they are unrealistically rendered or with limited visibility.

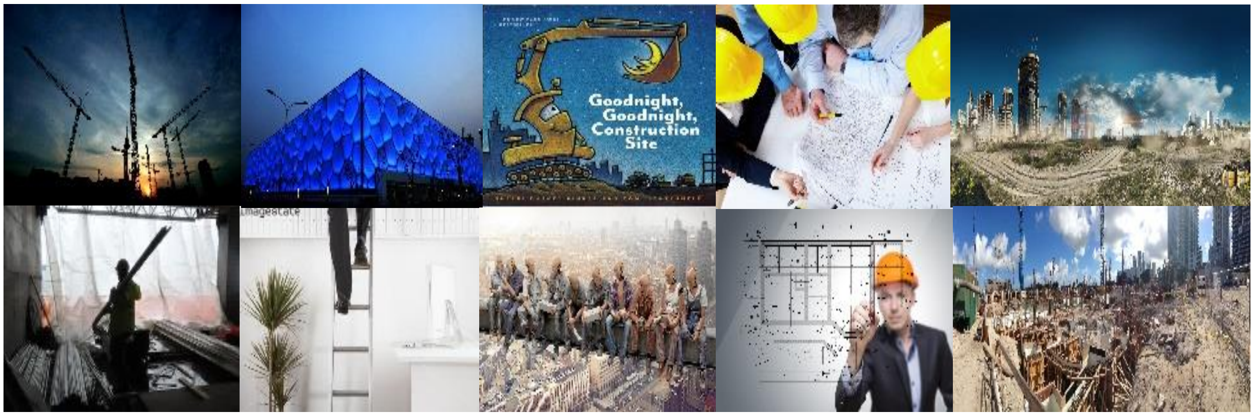


Figure 2.2: Examples of unsuitable images returned by a web crawler. **First column**: incorrect exposure; **second column**: non-construction content; **third column**: non-realistic content; **fourth column**: lack construction background; **fifth column**: panorama.

a. **Incorrect Exposure**. Images that are either overexposed or underexposed provides poor information of construction sites and objects. Silhouettes, common in Flickr images, also fall into this category.

b. **Irrelevant Content**. Despite search among specific user groups or adding prefixes to search terms, a lot of irrelevant images are returned from the crawler, especially in Google Image search.

c. **Nonrealistic Content.** This differs from Irrelevant Content, as images depict construction scenes or objects that are not reflective of real-world scenarios, such as cartoons or computer-rendered images.

d. **Lack of Background.** Crawling on Google Image returns many construction images of objects backdropped against a white background. These images are not included, as they do not show objects in their natural context. Artificial background introduces unnecessary data biases that do not appear in the real world.

e. **Panorama/Fisheye Images** Objects are often heavily distorted in these images. They are unsuitable because the appearance and geometry of objects in these images often do not agree with those of the majority of construction daily photos.

2.3.3 Image Annotation

Annotating a large and diverse image collection is challenging. An hourly paid small group of annotators with construction knowledge can provide reliable annotations, though the annotation process is likely to be slow. Crowd-sourcing annotations distributes tasks across many more annotators, and each annotator gets paid by the number of tasks completed. This approach allows for rapid and cost-effective completion of tasks but requires a significant amount of quality control. In this paper, the authors use the crowd-sourcing platforms and design verification methods to obtain highly confident annotations for both object instances and HOI instances.

Object Annotations. A crowd-sourced data annotation team is hired to annotate each of 22 construction objects in the final candidate image set. To train the annotators, first the authors provide generic but also distinctive definitions for each object class (Table2.3). The edge cases for each object class are articulated with three to five positive and negative instances (e.g., sunglasses are not an eye protection). A subset of 207 fully annotated images is provided to the annotation team as examples of annotations of sufficient quality. Another 200 images are annotated by the authors as the golden set; these images, rather than their annotations, are also sent to the annotation team.

The authors here describe the object annotation quality control procedure similar to Liu and Golparvar-Fard [47]. It involves three major steps: (1) qualification of annotators based on their performance on a small portion of the pre-annotated data; (2) a quality control process based on comparing submitted and ground truth annotations on the golden set images, which are randomly inserted in the annotation tasks; (3) a post-assessment phase, which collects and leverages multiple annotators' answers using majority voting, this generates the final annotations for each image. The Annotations for all images are accepted only if the performance on the golden set satisfied the desired threshold. The readers are encouraged to look into Liu and Golparvar-Fard [47] for detailed explanations. This stage took less than a week to complete; examples of annotated object instances is shown in Fig. 2.3. Table 2.3 and 2.4 shows the object classes and their statistics.

HOI Annotations. The HOI annotations are conducted on Amazon Mechanical Turk

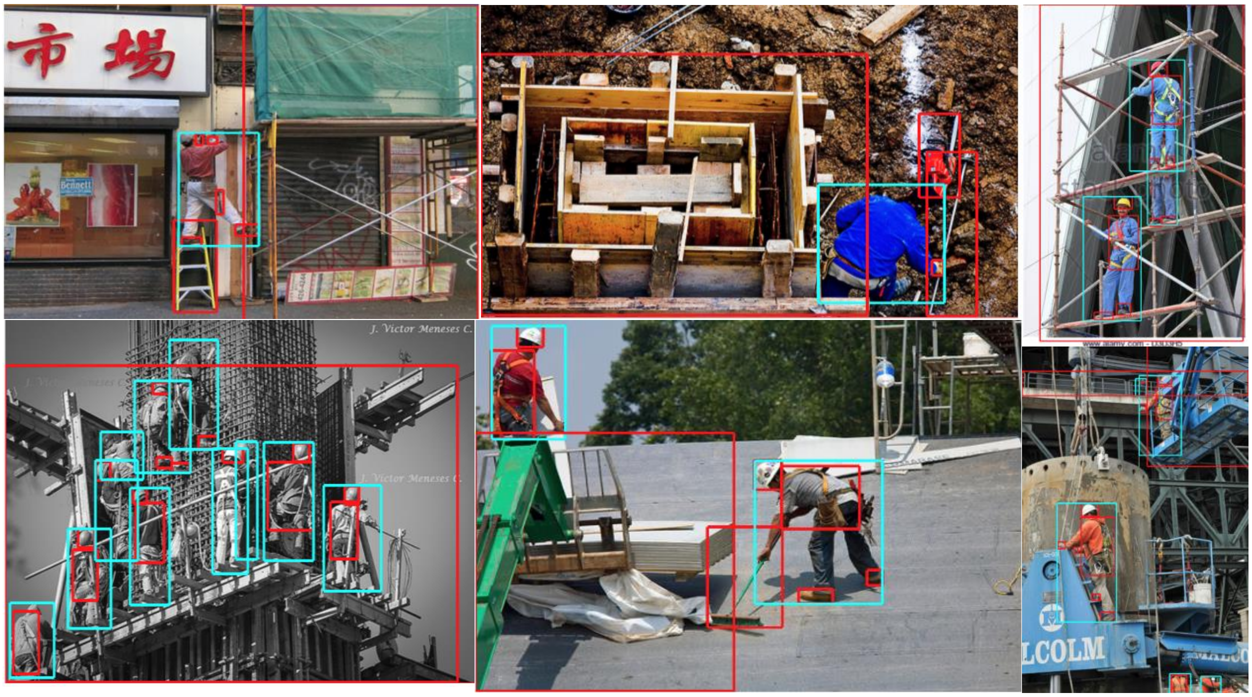


Figure 2.3: Bounding box examples of construction resource instances. Each blue box contains a single worker; red boxes contain instances of other object categories that a worker might interact with.

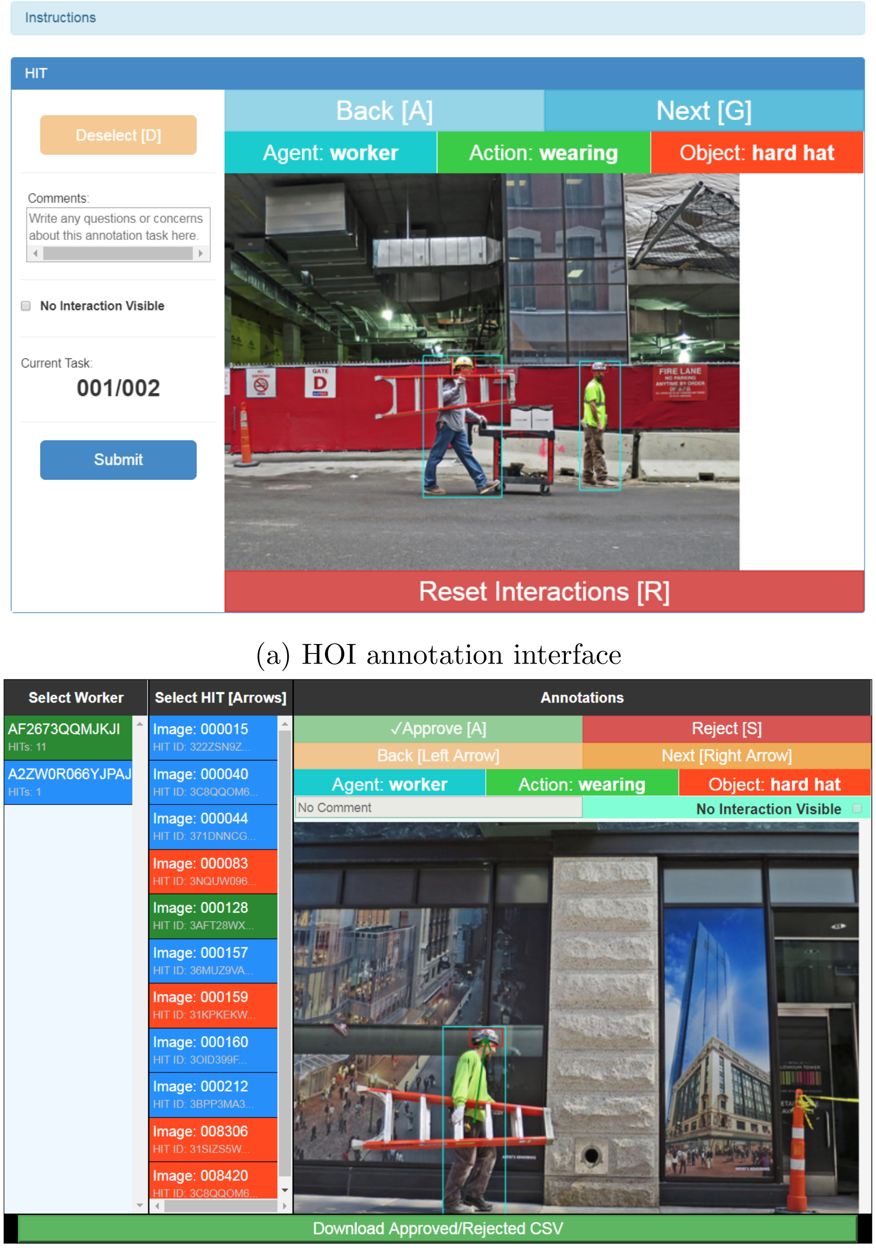
(AMT) because this task uses only a subset of annotated objects, i.e., workers, safety gears, tools, and scaffolding. Linking bounding boxes is also easier compared to the previous object annotation task. The annotation interfaces that were used to create the HICO-DET [74] are adapted and modified. Our HOI annotation pipeline includes an annotation interface and a review interface (see Figure 2.4). In the annotation interface, every AMT worker is assigned to a Human Intelligence Task (HIT) to finish one to six HOI annotation jobs given an image and drawn bounding boxes. For each job, only one HOI class and all boxes of relevant objects are displayed, e.g., if construction workers and hand protections are displayed, AMT workers are asked to annotate all "wearing" interaction instances. The annotation procedure is shown in Figure 2.5. AMT workers' annotations are reviewed offline, and a HIT is approved only when all its constituent tasks are approved.

The authors do not use a golden set of HOI instances here, as this task is simpler compared to object annotation in the previous round. Missing and incorrectly annotated HOI instances are rectified after annotator-reviewer communication via AMT. Annotators submitting random results are quickly identified and rejected by human reviewers. In total, 78 AMT workers participated in the HOI labeling. On average, each AMT worker completed 50.02 HITs, though only 18 AMT workers completed over 50 HITs. The HOI labeling took around 60 man-hours in total with an average per-image completion time of 52.11 seconds. The entire review process took one proficient reviewer around 20 man-hours with an average approval rate of 86.05%. These statistics imply that labeling HOIs for the construction site can be conducted by non-expert annotators with satisfactory approval rate, the major bottleneck of this process being the number of proficient reviewers. This HOI annotation stage was finished within two days.

2.3.4 Dataset Statistics

Twenty-two construction object categories from 32 safety rules (examples can be found in Sec. 2.3.1) were annotated. The web-crawling from three online sources returns 14,253 images. 4,565 images were selected as the final candidates and then annotated. In total, 37,735 object instances were returned. Box size distribution (see Table2.4) in the dataset is close to that of Microsoft Common Objects in Context (MS-COCO), in the proposed dataset the proportion of small, medium, large boxes take 42.63%, 32.88%, and 24.49% , while in MS-COCO they are 41%, 34%, and 24%, respectively. Small boxes have areas less than 32-by-32 pixels, large boxes have areas greater than 96-by-96 pixels. Object detectors are expected to have similar performance on these two datasets under MS-COCO standard object detection evaluation metrics. The proposed dataset contains more annotated workers than the dataset introduced by Luo *et al.* (2018) [65]. On average, there are 2.29 workers per image in the presented dataset, and only 0.38 worker in that of Luo *et al.* Over 22,000 instances of safety gear are annotated in the presented dataset. The ratio between annotated worker and safety gear is nearly 1:2.2, this gives a rich context to detect interactions between workers and safety gear. Table 2.5 is a comparison of object statistics in the proposed dataset with construction datasets from previous work. Using 3,311 images, the authors obtained 13,479 highly confident HOI annotations; their distribution is shown in Table 2.6.

The dataset for object detection is split into training and testing sets by the 80/20 conven-



(a) HOI annotation interface

(b) HOI review interface

Figure 2.4: HOI AMT interfaces. (a) **HOI annotation interface**. The AMT worker links box pairs that satisfy the displayed HOI class. If no such HOI instance can be identified, the AMT worker can check the No interaction visible box. After finishing all HOI class, the worker is requested to submit the HIT; (b) **Review interface**. Reviewers can browse all HITs submitted by AMT workers. In the left two columns, blue means not reviewed yet, green means HIT or worker accepted, and red means HIT rejected.

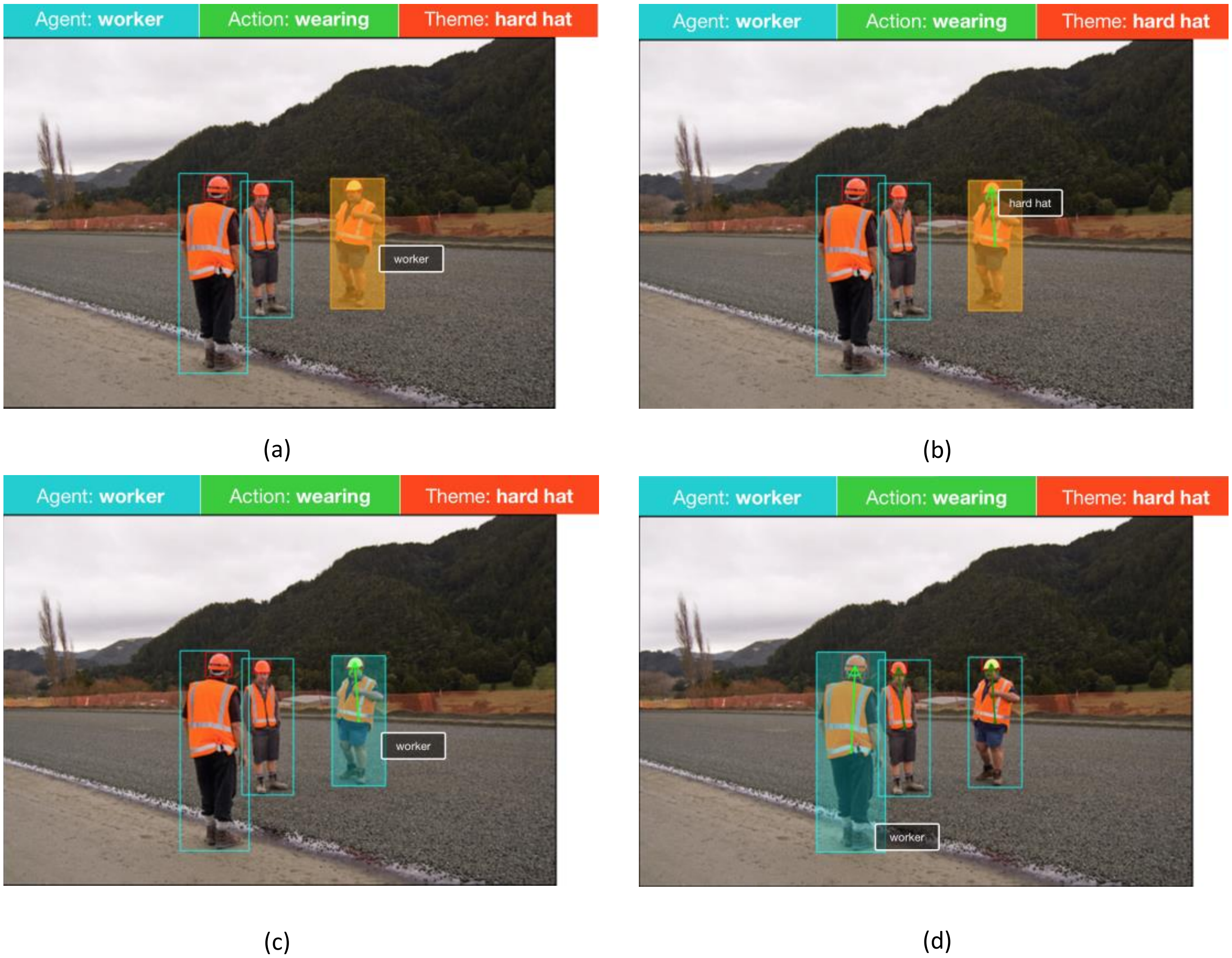


Figure 2.5: An example of "worker-wearing-hard hat" instance annotated using the presented AMT HOI annotation interface. (a) find/click on the worker box to select a worker. (b) drag the mouse to the top of the corresponding hardhat detection box. (c) release mouse when the "arrow" links the two detection boxes and when the annotation of "worker-wearing-har dhat" instance is complete. (d) repeating steps a to c until all workers are associated with the hard hats they are wearing.

Table 2.3: Object Class Definition and Note for Annotating HOI Dataset

Cat.	Name	Definition	If in HOI dataset
Worker	**Worker**	All labors.	Included
Safety Gear	**Hard hat**	Type I and type II hard hats of classes G, E and C	Included
	Foot Protection	Leather boots and rubber boots	Included
	Safety Coloring	High visualization color safety vests or uniforms	Included
	Hand Protection	Cotton, leather and insulated gloves	Included
	Eye Protection	Safety glasses, goggles, and weld goggles	Included
	Fall Arrest System	Body harness and connector	Included
	Ear Protection	Ear muffs	Few wearing instances
	Face Protection	Face shields and flash shields	Included
	Respiratory Equip.	Respirators and dust masks	Included
Tools	**Hand Tool**	Hammers, concrete floats, shovels and hack saws	Included
	Power Tool	Grinders, welders, drills, round saws and breakers	Included
	Ladder	Metal and wooden scale ladders	No three points of contact violations
	Compressed Air	Nail guns, jackhammers and sandblasters	No interactions
Vehicle	**Lifting Equip.**	Scissor lift and suspended scaffold	No standing on rail interactions
	Other Vehicle	Mobile cranes, trucks, pickup trucks and cars	Few driving and no riding instances
	Excavator	Crawler excavators	Few driving and no riding instances
	Backhoe Loader	Side shift and center pivot backhoe loaders	Few driving and near identical
	Fueling Tank	Fuel tanks, fuel trailers and fuel trucks	No interaction
Others	**Scaffolding**	Steel, patented, single and kwik-stage scaffold	Included
	Guardrail	Fixed barriers with top rails and midrails	No standing on interaction
	Concrete Paste	Concrete paste for flatwork	No interactions

Table 2.4: Full Object Instance Distribution

Name	All	Small	Medium	Large
Worker	10492	1397	4205	4890
Hardhat	6751	4706	1861	184
Foot Protection	6240	4757	1441	1088
Safety Coloring	3942	1079	1775	1088
Hand Protection	3542	2620	878	44
Hand Tool	1253	461	570	222
Other Vehicle	1008	36	204	768
Eye Protection	681	523	150	8
Ladder	621	62	263	296
Fall Arrest System	608	111	336	161
Power Tool	492	106	254	132
Scaffolding	419	0	12	407
Excavator	384	16	65	303
Lifting Equipment	258	6	48	204
Guardrail	255	19	73	163
Concrete Paste	200	5	38	157
Ear Protection	150	84	59	7
Face Protection	121	40	64	17
Backhoe Loader	96	4	17	75
Compressed Air Tool	96	17	46	33
Respiratory Equipment	87	35	39	13
Fueling Tank	39	4	12	23

Table 2.5: Construction Site Object Detection Dataset Comparison

Dataset	**Images**	**Classes**	**Instances**
Chi *et al.* [34]	750	3	1282
Memarzadeh *et al.* [36]	7508	3	7508
Azar *et al.* [37]	6070	1	6070
Park *et al.* [39]	3320	2	6402
Kim *et al.* [64]	<3000	5	2920
Luo *et al.* [65]	7790	22	13984
This paper	4565	22	37735

tion. The HOI recognition train and test splits are subsets of the object detection train and test splits. A total of 3,652 images (30,277 object instances) were used for object detection training, and 913 images (7458 object instances) were used for testing. For HOI recognition, the training set contains 2,657 (80.2%) images and 10,815 HOI instances, the testing set contains 654 images and 2,664 HOI instances (see Table 2.6). The HOI training and testing sets have similar class distributions.

2.4 Detecting and Recognizing HOI

The design of the HOI recognition model is presented in this section. As mentioned in the related work, the authors take advantage of the "parallel" model design in FasterRCNN. A sub-network to recognize HOI instances is built on top of a FasterRCNN object detection model. The overall idea of the HOI recognition subnetwork is to leverage detected object boxes to perform pairwise classifications of all worker and nonworker box pairs.

2.4.1 Network Architecture

The proposed model is shown in Figure 2.6. The model consists of three branches: region proposal, object detection, and HOI recognition. RPN [45] and RoIAlign combined are

Table 2.6: HOI Annotation Statistics

Predicates	**Objects**	**Instances**
standing_on	Scaffolding	302
using	Hand Tool	573
	Power Tool	278
wearing	Eye Protection	461
	Face Protection	88
	Foot Protection	3496
	Hand Protection	2090
	Hard hat	3638
	Fall Arrest System	400
	Respiratory Equipment	58
	Safety Coloring	2095

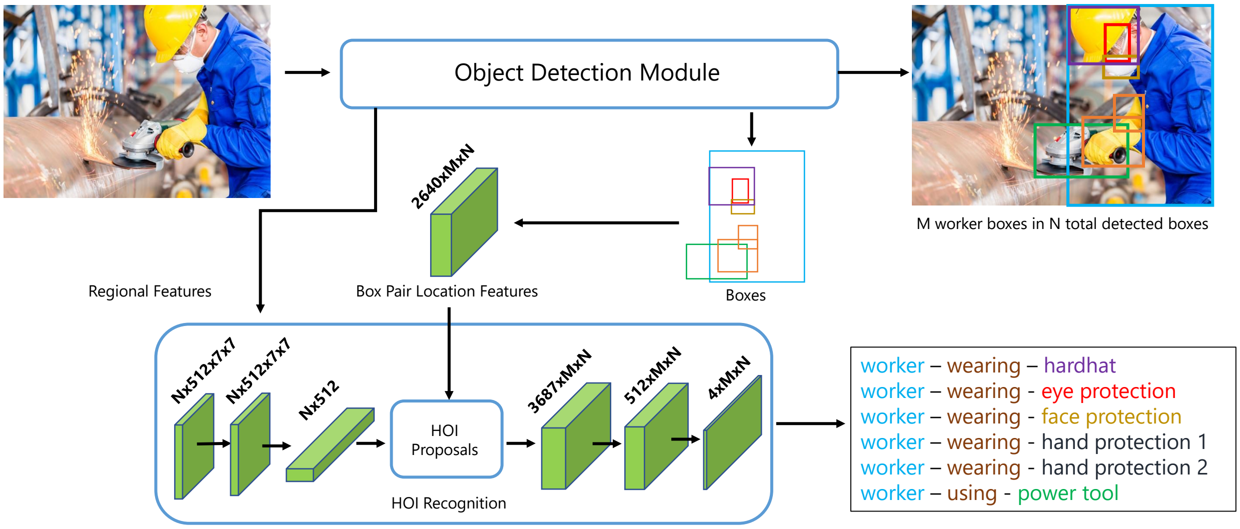

Object Detection Module
M worker boxes in N total detected boxes
2640xMxN
Regional Features
Box Pair Location Features
Boxes
Nx512x7x7
Nx512x7x7
Nx512
HOI Proposals
HOI Recognition
3687xMxN
512xMxN
4xMxN
worker – wearing – hardhat
worker – wearing - eye protection
worker – wearing - face protection
worker – wearing - hand protection 1
worker – wearing - hand protection 2
worker – using - power tool

Figure 2.6: Network Architecture. Staring from the last convolutional layer (layer4) from the ResNet50+FPN model, the network adds an HOI Recognition branch on the Faster RCNN framework. In this branch object detector results, layer4 appearance features, and HO pair's Spatial Features are used to construct HOI proposals. For M worker boxes detected out of N detected boxes, HOI proposal generates at most $M \times N$ human-object pairs to be classified.

treated as the region proposal to provide boxes and their regional visual features to be shared by both the object detection branch and HOI recognition branch. In addition to visual features from RPN, a new spatial feature is introduced to encode the spatial relation of two bounding boxes (see Sec. 2.4.2); this feature is generated from bounding box predictions and directly used for HOI recognition.

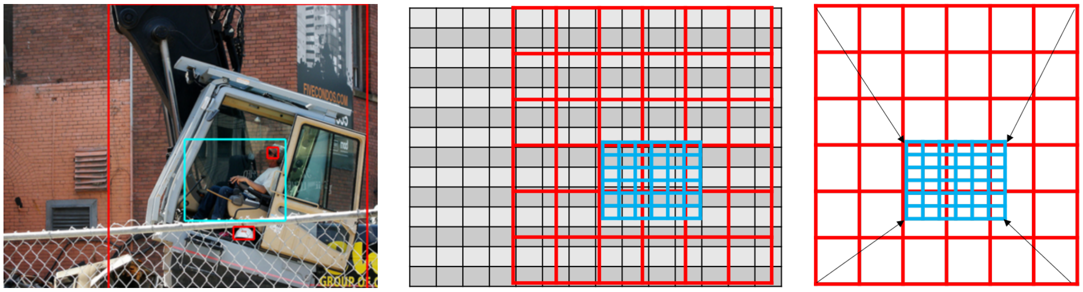
Object and HOI Proposals. The object proposals are generated with RPN, each object proposal is parameterized by the coordinates of the top-left corner and the bottom-right corner. The HOI proposals are generated by enumerating all possible pairings between worker and tool or equipment. Let M be the number of detected worker instances and N be the total number of detected instances in an image. There are thus $M \times (N - M)$ valid HOI proposals.

Object Detection. Object detection first applies RPN on the last convolutional feature maps (layer4) of a Resnet50 + FPN model [80]. For each image, the *layer4* feature is a 3D tensor whose depth is 256 and height and width are proportional to image height and width. Region of Interest (RoI) features [62], which are extracted by RoIAlign with box proposals from *layer4* features, are forwarded to two fully connected layers to predict object categories and regress bounding box coordinates as in FasterRCNN [45].

HOI Recognition. The HOI recognition branch concatenates three features for the HOI proposals: (1) HOI spatial feature, which is generated from detected objects' box coordinates, while the detail is introduced later; (2) linguistic cue, which is the class probability for the object box from the object detection branch; (3) HOI appearance feature, by concatenating RoI features of detected human box and object box. The RoI feature is handled by a specialized mininet: a 1×1 convolutional layer followed by a ReLU activation layer, then an average pooling layer whose kernel width and height are the same as the RoI feature's width and height (see Figure 2.6, HOI recognition branch). This mininet is meant for feature transfer with the minimum number of parameters. All three features are reformulated into a 4D tensor whose dimension is $Batch\_size \times nChannel \times M \times N$, where $nChannel$ is the length of concatenating all three HOI features and $Batch\_size$ is always 1. This tensor is referred to as the *HOI feature map.* HOI classification on an HOI feature map is similar to generating segmentation masks in Fully Convolutional Network [81] for semantic segmentation. The final output of the HOI recognition model is a 2D mask of size M × N, where each element in the mask contains an HOI class index or a background index indicating the absence of target interaction.

2.4.2 Modeling Spatial Correlation in HO Boxpairs

The aforementioned spatial feature is explained here. This spatial feature discretely encodes a Human-Object box pair's spatial location. The motivations of this discret box pair feature are: (1) each box's position should be represented relative to the entire image; (2) the HO box pair's relative position is invariant to image scale (see Fig. 2.7); (3) spatial features should be represented discriminatively by one-hot vectors. Firstly, the entire image is divided to a $S \times S$ grid (Fig. 2.7b). Two one-hot vectors are used to represent the location of individual cells in the grid; for example, cell (2,1) is represented as $([0, 1, 0, ..., 0], [1, 0, ..., 0])$. For each box, P points are evenly sampled on every edge, resulting in $4(P - 1)$ points representing a single box's location. Each point lies in one cell of the image grid, whose location is represented by a $2S$ length vector, so the entire box is encoded by an $8 \times S(P - 1)$ length vector. To capture the relative location feature, a distance metric is proposed in terms of the number of cells between the $i'th$ point in the human box and the $i'th$ point in the object box (Fig. 2.7c). The distance on one axis between two points is an integer value between $[-S + 1, S - 1]$, so another one-hot vector of size $2 \times (2S - 1)$ is used to encode this distance. Thus, a Human-Object box pairs relative distance is encoded in a vector of length $8 \times (P - 1) \times (2S - 1)$. To construct the final HOI spatial feature of a Human-Object pair, the human box's encoded coordinates, the object box's encoded coordinates, and the box pair's relative distance are concatenated, resulting in a feature of length $8 \times (P - 1) \times (4S - 1)$. In the experiment $S = 14$ and $P = 7$; the author simply follows conventions of ROI features to choose these parameters.



(a) RGB image and boxes

(b) Sparse box locations

(c) Encoding box pair relations

Figure 2.7: Procedure to encode box pair relation. (a) Image with blue box drawn around worker and red boxes drawn around other objects. (b) Encoding box coordinates by determining box sample point locations in image grid; only the boxes border coordinates are sampled. (c) Encoding HO box pair's relative position; 4 arrows are shown as 4 out of 24 links to corresponding points between human box and object box; distance between corresponding sample points on box border is computed in terms of number of cells.

2.5 Experiment

2.5.1 Implementation Details

Loss Functions. Object classification and smooth L1 bbox regression losses are used to train the object detection branch [45, 46]. For the HOI branch, since the output tensor is a $4 \times M \times N$ tensor (3 interactions and the background class), the ground truth HOI labels are organized as an $M \times N$ matrix; then cross-entropy is used to compute the HOI prediction loss.

Batch Preparation. Only one image is used in every batch as in Faster RCNN. Batch size for object classification is 256 object proposals from RPN. The positive:negative example ratio is kept at approximately 1:3 to avoid extreme data imbalance. In the HOI recognition branch, as the number of objects detected varies for every image, the HOI proposal number changes drastically. Positive HOI examples are often rare in an image, so the authors apply a hard negative examples sampling strategy. For every image, scores for every HOI proposal are computed; this is done during the forward pass. Then HOI batch size is set as 64. Positive examples are filled first, then the rest of the batch is filled by highly confident false positive prediction and sorted by their scores in descending order.

Inference. The model detects object bounding boxes (e.g., workers, hard hats, power tools), and then recognizes HOI instances (e.g., worker wearing hardhats), in an image by a single forward pass, similar to the cascaded inference in Gkioxari *et al.* [73]. From the object detection branch, a large amount of detected boxes are generated, and then a Non-Maximun Suppression (NMS) operation is applied to remove the overlapping boxes. At most the top 100 boxes, which have high detection scores and are deemed to be non-overlapping, are preserved. Boxes whose scores are higher than 0.5 are used to generate the HOI proposals.

Evaluation Metrics. Correctly detected bounding boxes for an object class are called true positives (TP), incorrectly detected boxes are called false positives (FP), missed ground truth boxes are called false negatives (FN). The precision of detecting an object class is defined as the ratio between TP and TP+FP. Recall of an object class is defined as the ratio between TP and TP+FN. A detected box is a true positive when: (1) the Intersection of Union (IoU) between this box and a ground truth box is above a threshold and the highest among all unmatched ground truth boxes; (2) the detected box's object class is the same as the ground truth box it matches to; (3) the detection score is above a threshold. For a given IoU threshold, a precision-recall curve is generated by calculating precision and recall for each detection score threshold. The Average Precision (AP) of a class is the area under the precision-recall curve. Object detection models are evaluated by mean Average Precision (mAP) from all object classes. The PAscal Visual Object Classes (VOC) [68] mAP is evaluated by a single IoU threshold of 0.5; it is used as the main evaluation metric for object detection in this paper. An HOI recognition result is considered correct when: (1) both detected boxes are correct and have at least 0.5 IoU with their corresponding ground truth boxes; (2) the interaction is predicted correctly. HOI recognition is evaluated by Top-K recall or Recall@K [82, 77]. Top-K recall is defined as, given the K most confident HOI prediction instances sorted in descending order, the ratio of ground truth HOI examples being included in these K predictions. Top-K recall is used because persons in the image center region are more likely to be annotated than persons in the background or edges of the image [74]. Using mAP to evaluate HOI recognition can falsely penalize positive but not annotated HOI instances. The average number of ground truth HOI examples per image in the HOI test set is 4.11; the authors report K in the experiments as 5, 10, and 15 by inspections of experiment results. When K is greater than 15, Recall@K scores do not change much and become saturated. Because the number of recognized HOI instances in an image is also determined by the number of detected objects, it can be lower than K. The average number of ground truth HOI examples per image is also used as a metric. In general, the closer to this number, the better the HOI recognition model is.

Training Specifications. A Pytorch implementation of FasterRCNN is used. Ten object classes relevant to the HOI experiments (i.e., worker, eye protection, face protection, foot protection, hand protection, hard hat, safety coloring, scaffolding, hand tool, and power tool) are used to evaluate the object detection model. The Faster RCNN model is fine-tuned from an MS-COCO object detection model, the final output layer is re-initialized with random weights, and output classes are changed from 91 to 11, including the background class. The object detection model is trained with the stochastic gradient descent (SGD) optimizer with momentum 0.9. The object detection model is firstly trained with a learning rate $1E-3$ for the first 70,000 steps and $1E-4$ for the following 25,000 steps. To detect small objects better, a small 4×4 region proposal anchor is added to the original three larger anchors. Otherwise, the training setup is the same as the training for the VOC07 [68] dataset. The HOI model shares convolutional layers with the trained object detection model up to the last convolutional layer (layer4) of ResNet50+FPN. When training the HOI model, weights up to the layer4 are fixed, and the layers after this are trained from scratch by the SGD optimizer with 0.9 momentum and by 13,000 steps. The learning rate for training HOI recognition model is set as $1E-3$. The weight decay is applied for training HOI recognition model and is set to $1E-4$ for each step as a means of regularization. All weights converge when the loss is stable and the test set performance no longer increases. All models are trained with one NVIDIA Tesla K40 GPU.

2.5.2 HOI Recognition Experiments

The authors design a rule-based HOI recognition baseline (*"Rule-based"*) to compare with the proposed learning-based HOI recognition model. This baseline consists of the following rules: (1) All non-worker objects detected will be assigned to their closest detected worker instance, the distance between two bounding boxes is measured by the L2 distance of each box's center. (2) The interaction of an HO pair is always assigned to the most frequent non-background interaction from the training set (e.g., whenever a hard hat box and a worker box are paired, their interaction is always assigned as "wearing"). This baseline heavily exploits class and location biases to emulate previous work that uses a rule-based method for activity recognition [65]. As a comparison to the previous HOI recognition models, the authors reimplement HO-RCNN, proposed by [74]. The HO-RCNN has two main differences from the authors'. First, the human-object pair spatial features are different. Second, HO-RCNN is a multi-stream model and sums up different streams' outputs and applies a softmax layer for the final prediction.

Second, a sanity check is conducted by predicting HOI results from an untrained HOI recognition model (model *Random*) to validate that the bounding boxes and interactions are indeed correlated in the presented HOI dataset, such that a random guess cannot predict them. Also, an ablation study is conducted to examine and validate each feature's contribution to the proposed HOI recognition model. Model *RoI* uses only concatenated HO pair RoI features; *SP.* uses only the proposed HO spatial feature; other models use various combinations of *RoI*, *SP.*, and linguistic features.

The authors present an experiment on applying a single HOI recognition result for safety gear compliance checking. Consider the task of checking workers wearing hard hats. A false negative occurs when: a worker does not wear a hard hat but the system says the worker does, or the system missed this worker. A false positive occurs when: the system says a worker does not wear a hard hat but the worker does or the detected worker is a false detection. The same definitions apply to safety coloring checking. The proposed method checks hard hat compliance as in the following steps: (1) An image is taken in and workers and hard hats are detected; (2) the HOI model checks whether there is "wearing" interaction between each detected worker-hard hat pair; (3) for a detected worker, if there is no hard hat detected or a detected worker is not recognized "wearing" any hardhat, that worker is marked as an incidence; (4) evaluation is made by matching bounding box overlaps between the marked incidents, and the actual incidents, then precision and recall are calculated. A match should have an IoU greater than 0.5. All HOI recognition results are used in an image.

The same procedure applies to safety coloring checking. The authors apply three different checking strategies. Strategy "Multi-class detection" reports workers only if there is not a single hard hat/safety coloring detected. Strategy"Rule-based HOI" performs per-worker checking based on the *Rule-based* HOI method. Strategy "Learned HOI" uses our final HOI model (*RoI.+SP.+Ling.*) for per-worker checking.

Checking hand protection is not required when, for instance, workers are inside an office trailer or near rotating machinery. The authors apply the proposed HOI recognition model for finding workers using hand tools or power tools but not wearing hand protection; this criterion is covered in: OSHA 1910.138(a) General requirements. "Employers shall select and require employees to use appropriate hand protection when employees' hands are exposed to hazards such as those from skin absorption of harmful substances; severe cuts or lacerations; severe abrasions; punctures; chemical burns; thermal burns; and harmful temperature extremes." Object detection models, such as in Smartvid AI engine, cannot directly check this criterion because they do not leverage the context in which the workers are operating. The authors examine different strategies to check this task. Strategy "Worker detection only" flags all detected workers as potential incidents. Strategy "Multi-class detection" flags all worker instances where at least one worker and one tool are detected but hand protections are not. Strategy "Rule-based HOI" applies the *Rule-based* method to identify incidents. Strategy "Learned HOI" applies the final HOI recognition model (*RoI.+SP.+Ling.*) by first classifying whether a worker is using hand tools or power tools, then classifying whether this worker is wearing hand protection.

2.6 Experiment Results and Discussions

2.6.1 Object Detection Error Analysis Tool

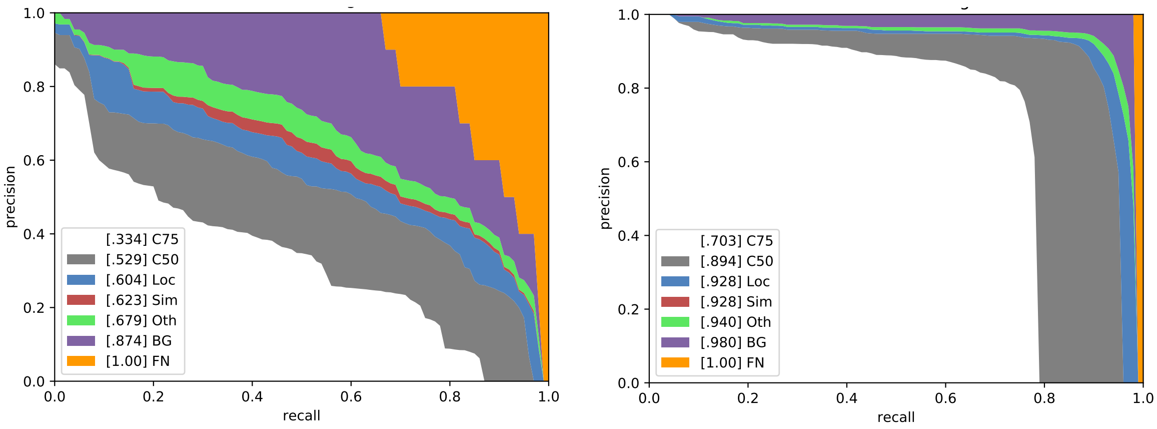
To understand the root cause of detection error, an object detector error analysis inspired by Hoiem *et al.* [83] is used for analyzing different false positive errors in the trained object detector. Such analysis shows six types of errors, and in such an order their corresponding precision-recall (PR) curves are generated. The investigated types of errors are: (1) PR curve at $IoU = 0.75$ (**C75**); (2) PR curve at $IoU = 0.5$ (**C50**), the same as the VOC AP metric; (3) PR curve at $IoU = 0.1$ (**Loc**), which error by mis-localization is ignored to show the impact of incorrect object localization; (4) PR curve ignoring false positives from detecting similar objects (**Sim**), which in the proposed dataset affect only safety gear classes; (5) PR curve by considering a detection correct when it matches to any class label (**Oth**); (6) PR curve after removing all background and class confusion false positives (**BG**); the last PR trivially representing $AP = 1$ when all other false negatives are corrected (**FN**). Each curve represents the object detection performances if all previous errors, including itself, are corrected. The color-coded areas show the relative importance of each type of error analyzed. Figure 2.8 shows examples of the object detection error analysis.

2.6.2 Object Detector Results and Analysis

The best-performing Faster RCNN detector with ResNet50 + FPN features reaches overall 52.9% mAP. In particular, this model achieves 89.4% AP for worker detection. Full detector results are shown in Table 2.7. Based on the error analysis of all object detection results (Fig. 2.8), resolving the confusions with the background (purple area in Fig. 2.8a) can significantly improve overall detection performance by 19.5% mAP. This detection error analysis result (Fig. 2.8) confirms many previous observations [57, 65, 31] that high intra- and inter-category appearance variabilities are serious hindrances to object detection methods for construction resources. Future research is suggested to disambiguate similar patches between background structure and relevant objects. The overall worker detection AP in the trained detection model reaches 89.4% (Fig. 2.8b). This number is 29.4% higher than the number for worker detection reported in Luo *et al.* [65], despite the experimental results being evaluated on a dataset with more inter-class variance (Table2.5). The overall high performance of the trained object detection shows it is sufficient to use detected object instances for training the HOI models and evaluate the proposed method from raw image inputs.

Table 2.7: Faster RCNN Object Detector Results

Class	**AP (%)**	**Class**	**AP (%)**
Worker	89.4	Hard hat	93.6
Eye protection	28.0	Safety coloring	82.8
Face protection	25.1	Scaffolding	42.0
Foot protection	67.8	Hand tool	20.6
Hand protection	66.1	Power tool	14.2



(a) All object instances.

(b) All worker instances.

Figure 2.8: Trained Faster RCNN detector error analysis. (a) shows the error analysis for all subjected classes; (b) shows the error analysis for worker.

Table 2.8: HOI Recognition Experiment Recall Results. Note two bounding box sources are used: one from detected objects (Det.), the other from ground truth object annotations (GT).

Boxes	Method	Recall@5(%)	Recall@10(%)	Recall@15(%)
Det.	Random	0.450	0.713	0.938
	Rule-based	51.40	57.32	63.25
	HO-RCNN	51.00	58.50	60.13
	RoI	27.36	41.19	48.49
	SP	50.17	58.58	59.72
	RoI+Ling.	37.82	50.22	55.94
	RoI+SP.	50.38	58.39	59.87
	RoI+SP.+Ling.	55.59	62.58	63.25
GT	Rule-based	74.66	93.50	95.49
	RoI+SP.+Ling.	71.05	85.51	87.57

2.6.3 HOI Recognition Results and Analysis

HOI recognition models' Recall@K results are shown in Table 2.8. HOI recognition models trained with detected and ground truth bounding boxes are presented in Table 2.8 Box Sources "Det." and "GT", respectively. The best-performing HOI recognition model and *Rule-based* model are retrained with the ground truth bounding boxes in the train set. This setting eliminates the box localization and object classification errors to evaluate HOI recognition assuming a perfect object detection is achieved.

When trained on the detected boxes, model *RoI+SP.+Ling.* performs better than the rest of the models. Compared with the *Rule-based* model, the best model outperforms significantly on Recall@5 and Recall@10 but achieves the same result on Recall@15 metric, because all highly confident detected object instances are used. The proposed model notably outperforms the HO-RCNN by large margins on all metrics. This validates the authors' incentive to incorporate multiple types of features into the proposed HOI recognition model. Spatial features are shown to be strong cues; using these alone produces better results than when they are not used, as can be shown by comparing the SP. and ROI+Ling. rows in Table 2.8. Linguistic features also help with HOI recognition, as one of the main differences between the proposed model and HO-RCNN is that linguistic cues are also used. More qualitative results from the best-performing HOI recognition model are displayed in Fig. 2.9. Discussion on safety compliance checking will show how the improvement in recognizing HOI results in a better approach to perform automatic safety checking.

A comparison between the *Rule-based* model and the *RoI+SP.+Ling.* model using detected boxes (Box Source "Det." in Table 2.8) and ground truth boxes (Box Source "GT" in Table 2.8) reveals an interesting finding on the impact of object detection for HOI recognition. When HOI recognition model is trained with the detected boxes, the *RoI+SP.+Ling.* model outperforms the *Rule-based* model at all evaluation metrics. When the HOI recognition model is trained with the ground truth boxes, *RoI+SP.+Ling.* model receives a significant boost in performance, e.g. from 55.59% to 71.05% for Recall@5. However, the *Rule-based* model outperforms the *RoI+SP.+Ling.* model significantly on all metrics. The perfect object detection setting almost doubles the HOI recognition performance. However, it is unrealistic to expect object detection outputs to be completely reliable in real-world scenarios. Thus, the proposed HOI model remains the more viable solution in practice.

Table 2.9 shows the HOI recognition results for hard hat and safety coloring. Both show high performance and thus are considered reliable for compliance checking. Table 2.10 shows results from hard hat and safety coloring compliance checking. For both safety gear checking tasks, the learned HOI model achieves both higher precision and recall than the rule-based model. Since both methods use the same object detection results, the learned HOI recognition model correctly recognizes more workers wearing hard hats/safety coloring, and therefore produces fewer false positives. And the model correctly classifies those who do not wear hardhats/safety coloring, hence producing fewer false negatives. Compared with using object detection results alone, applying HOI clearly achieves much better recall of the actual noncompliance with only slightly losing in precision in the safety coloring example.

Table 2.11 reports quantitative results of different strategies for checking hand protection conditioned on the use of tools. Strategy "Worker detection only" recalls most of the actual instances, but includes a very large number of false positives. In practice, this creates a

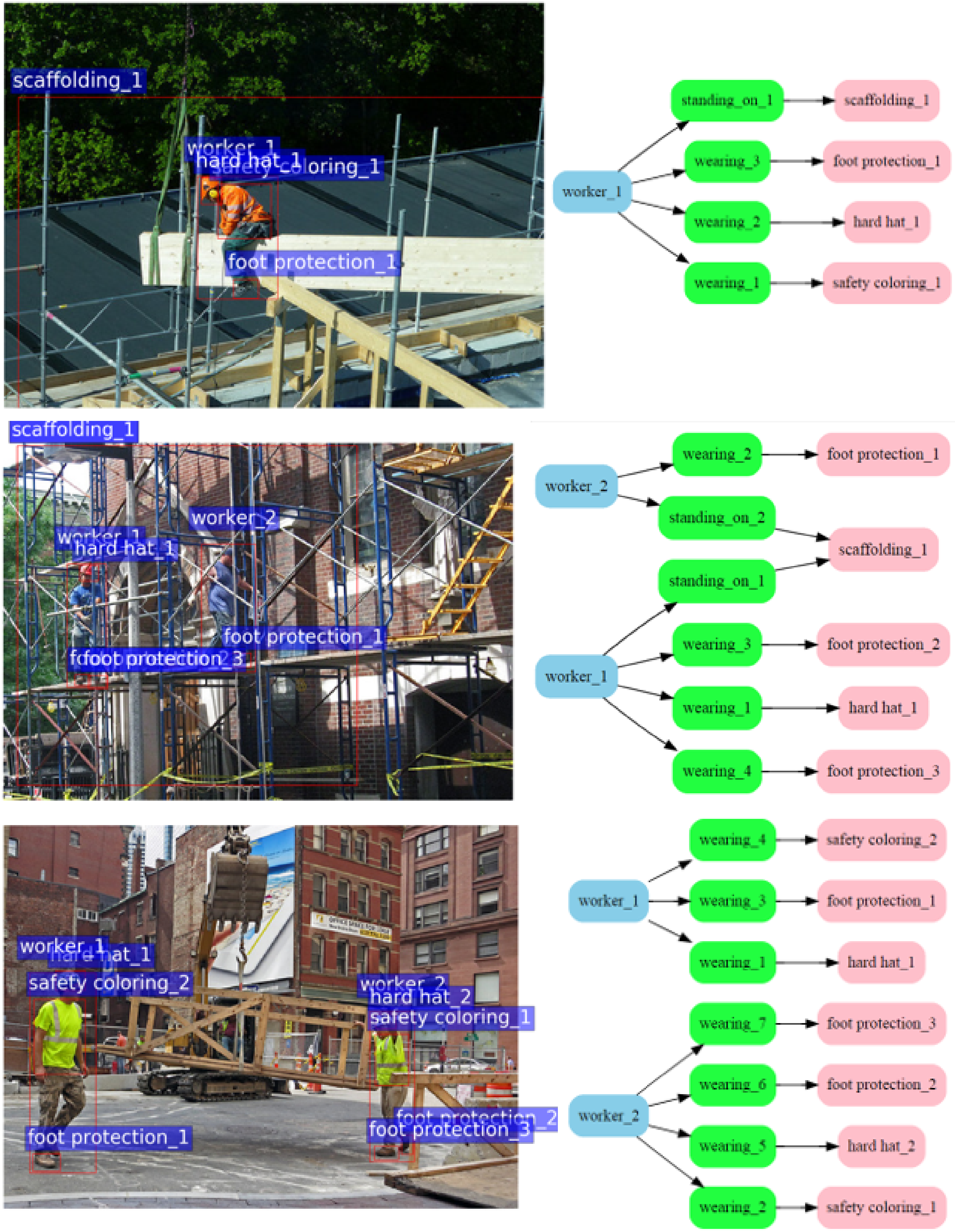


Figure 2.9: More HOI recognition results from the best model.

Table 2.9: HOI Recall for Hard Hat and Safety Coloring. Model trained with the detected bounding boxes.

Safety Gear	Method	Recall@5(%)	Recall@10(%)	Recall@15(%)
Hard Hat	Rule-based	76.02	78.50	78.50
	RoI+SP.+Ling.	84.67	88.97	88.97
Safety Coloring	Rule-based	72.04	79.45	79.79
	RoI+SP.+Ling.	76.02	80.82	81.84

Table 2.10: Precision and Recall for Hard Hat and Safety Coloring Compliance Checking

Criteria	Strategy	Precision (%)	Recall (%)
Not wearing hard hat	Multi-class detection	72.04	18.84
	Rule-based HOI	78.32	74.68
	Learned HOI	79.78	77.64
Not wearing safety coloring	Multi-class detection	80.83	59.21
	Rule-based HOI	78.43	71.78
	Learned HOI	79.11	75.29

Table 2.11: Precision and Recall for Hand Protection Compliance Checking

Criteria	Method	Precision (%)	Recall (%)
Not wearing hand protection while using hand tools or power tools	Worker detection only	7.50	87.80
	Multi-class detection	24.14	51.21
	Rule-based HOI	65.21	60.00
	Learned HOI	66.20	64.86

burden of manually removing false alarms. Compared with “Worker detection”, the strategy “Multi-class detection” which also includes hand protection and tools detection, reduces the false positives but suffers from a large increase in false negatives. In practice, this leaves too many incidents undetected and may need a second round of manual inspection of all images. HOI methods significantly improve precision over just using detection results, which means more relevant incidents will be reported to human reviewers. Compared to the “Multi-class detection” strategy, both HOI methods cover more actual incidents. The proposed method “learned HOI” improves recall by 4.86% compared to “Rule-based HOI”. These results show that applying HOI results is potentially a better strategy than relying on object detection results alone, and a learned HOI model is practically more feasible than a rule-based HOI method.

Over 1,000 construction sites in the United States have successfully applied object detection models for tagging safety gear, equipment, and workers on images taken from their sites. In practice, tags are reviewed, added, and modified by safety personnel. The automatically generated and human-verified tags provide a means of safety observation at larger scale and more frequently. These improvements potentially enable executives to understand safety trends within a project and benchmark across different projects. The proposed HOI model highlights relevant instances for human reviewers and potentially further reduces their workload. It also allows checking results with context to be reported to executives such that a better informed decision can be made. Nevertheless, the current HOI recognition model is limited in two ways. First, HOI models depend on the object detection performance. As suggested in the object detection error analysis, confusion with background patches is the main obstacle that requires more attention. An improved object detection model in the proposed method reduces false negatives in safety gear compliance checking. Second, many safety critical interactions between workers and equipment/tools are rarely recorded in visual data; more research on HOI should be done to transfer the knowledge of worker interactions from one type of equipment to another. For example, an HOI recognition model that recognizes a worker climbing a ladder should also recognize a worker climbing the cross-bracing of a scaffolding.

2.7 Future Applications of HOI Recognition in Construction Safety

Many potential applications can benefit from the rich semantic scene information that HOI provides. In this paper, three construction safety applications are listed. Note that for each application the following three factors are considered: (1) **Model Maturity** (model robustness and efficiency to be widely deployed); (2) **Confidence from Management** (how much trust do construction managers have in the machine learning models); (3) **Level of complexity to detect objects in construction sites.** (the level of material and equipment clutter, light conditions and weather, ans so forth. determines the level of difficulty to train a production-grade object detection and HOI recognition model). The authors list potential applications for safety inspections, monitoring, and education.

2.7.1 Assistive Automatic Safety Report Generation

Mobile devices are now commonly used to capture and share construction job sites' "as-is" conditions. Reporting job site safety condition by digitized reports, 2D photos, and 3D scans has largely reduced the latency in safety communication. Despite this, safety reporting is primarily performed manually. The simplest form of automation in safety reporting is to customize the safety report template according to the job site. For instance, when a worker is captured in the image, a set of personal protective equipment safety checks should be suggested from the application, and the inspector's confirmation is expected to complete corresponding checking entries.

2.7.2 Proactive Safety Performance Measurement

HOI information also helps with computing a proactive safety index. In safety management, proactive safety indexes are calculated from safety observations [14]. Several proactive index metrics had been proposed by [20, 84], indicating a strong correlation between daily construction activity statistics to the overall safety condition of the entire job site. Retrieving recognized interactions and discovering their correlations with safety incidences can be helpful to establish new proactive indicators. The triplet form of HOI recognition has the advantage of expressing queries as natural language rather than just as a list of keywords.

2.7.3 Safety Education

Recent progress on safety education uses interactive virtual reality-based training platforms [85] for new construction workers to increase their understanding and awareness of common safety hazards on the job site. The mechanism to create game content can be tailored to fit construction scenarios. Recent research in computer vision propose that scene graphs, a data structure used in for representing logical events in games, can be modeled by a set of HOI instances and directly generated from images. As a vision task, scene graph generation [77, 76] can be regarded as the inverse process of computer graphics rendering. Once scene graphs are generated from an image or a video, the same information can be rendered in a safety VR training environment. The advantages is that the rendered scene can play back what has been observed in the real world. Figure 2.10 presents the general workflow of this concept: (1) input images are first read by a multi-level scene understanding model; (2) detection results are combined into a scene graph representation; scene graphs depict how the machine comprehends image content, and it resembles human interpretation of the scene; (3) a generated scene graph can be rendered with a predefined 3D model library. Note that in the scene graph example present in Figure 2.10, the rendering is created with the Google SketchUp 3D model warehouse.

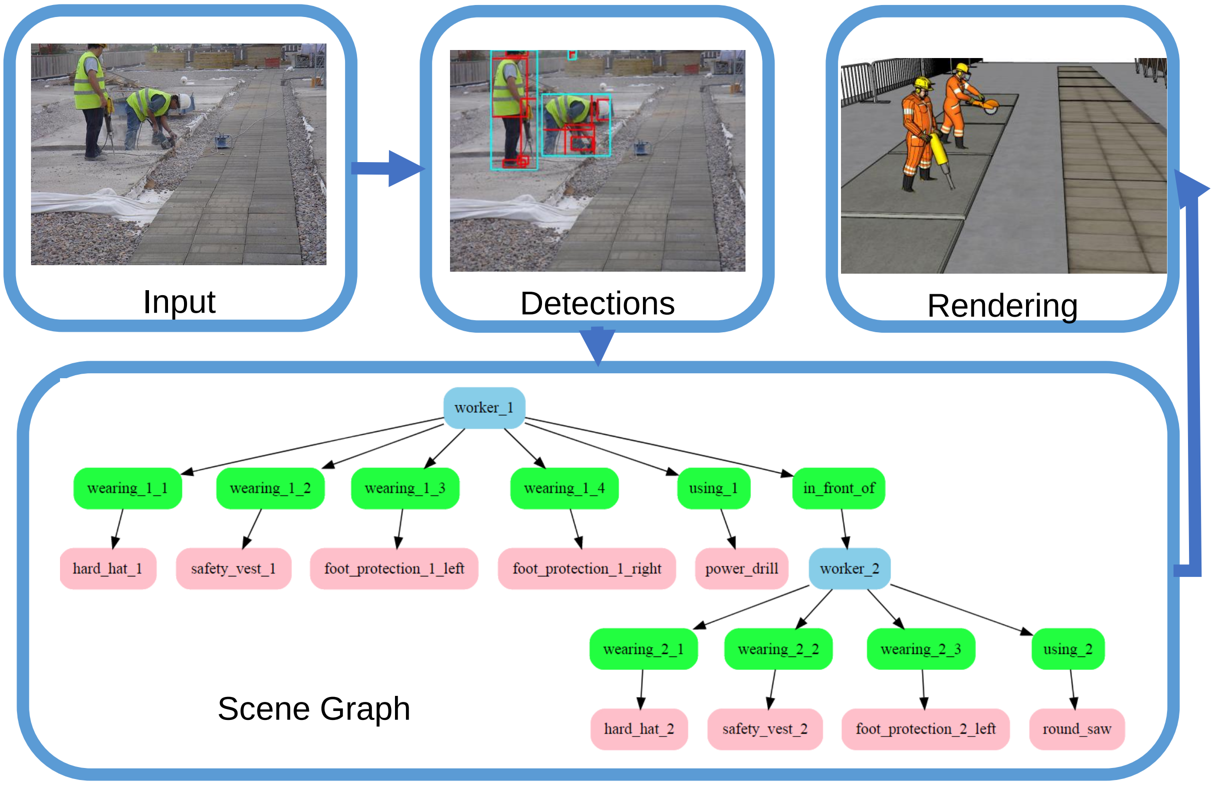


Figure 2.10: Illustrating a potential HOI use case for safety training. Realistic safety training scenes can be rendered in virtual reality environment using scene graph representations captured from real world construction images.

2.8 Conclusion

This paper improves vision-based safety checking methods. While the majority of existing automatic safety checking methods detect only safety-relevant resources, the authors propose an additional HOI recognition model to directly recognize the worker interactions (i.e., wearing protective equipment, using tools, and standing on equipment). The proposed HOI recognition model achieves Top-5 recall of 55.59%, outperforming a baseline of hand-made rules on detected objects and another previous HOI recognition model. There are two main practical benefits from the proposed approach. First, safety gear compliance checking is performed directly using HOI recognition results. Using examples of hard hat and safety coloring checking, the authors validate the HOI model and show that it is better than the hand-made rule-based method both in precision and recall. Second, the proposed method considers workers' interactions and highlights potential compliance issues. For the example of checking workers not wearing hand protection while using hand/power tools. Using HOI recognition results in significant improvement in precision compared with using object detection alone. For this example, a learned HOI recognition model is better in terms of precision and recall than the rule-based method. Experimental results validate the claim that applying HOI improves vision-based safety checking. The validations point out that visual understanding of workers' activity and tools they used are crucial for automated safety checking systems, and such information cannot be easily extrapolated by object detection results alone. Other potential use cases, such as safety report generation and virtual reality based safety training, are also discussed in the paper. All experiments are supported by a newly constructed safety-related object detection and HOI recognition dataset for construction sites.

CHAPTER 3: VIDEO-BASED MOTION TRAJECTORY FORECASTING METHOD FOR PROACTIVE CONSTRUCTION SAFETY MONITORING SYSTEMS[1]

3.1 Introduction

Construction is one of the least safe industrial sectors of the U.S. economy. Annually accounting for one in five worker deaths in the U.S., the Bureau of Labor Statistics (BLS) and the Occupational Health and Safety Agency (OSHA) have both reported that between 2011 and 2017, construction fatality number has grown from 781 to 971 incidents [1]. These statistics show workers-on-foot (referred to as workers) frequently operate in hazardous conditions due to nearby construction equipment and the surrounding working environment. In 2017 alone, 118 workers died from vehicle accidents and collapse hazards [2], including (1) struck-by-vehicle in work zone or objects falling from vehicle or machinery, (2) caught in or compressed by equipment or objects, (3) struck, caught, or crushed in collapsing structure, equipment, or material. Although OSHA regulations, company-wise policies, and best practices have been successfully established and followed for many years, it is still challenging to empower individuals to remain fully alert at all times. Improving safety monitoring helps workers and equipment operators make better-informed decisions regarding their safety without impeding their tasks at hand. Over the past decade, safety alerting systems have been developed that proactively inform workers and equipment operators about potential safety incidents. Examples of these systems include audio alerting systems in the equipment cabins or tags and safety vests that inform workers. To work reliably, these systems require real-time identification and communication of worker and equipment current and future locations.

Identifying, analyzing and recording proximity of workers and equipment to hazards is a critical part of any proactive construction safety management system seeking to avoid exces-

[1] This chapter in whole or in part is published in the Journal of Computing in Civil Engineering.

sive proximity and the potential struck-by accidents/near-misses this can incur. While safety altering systems have significantly advanced over the past decade, methods for identification and prevention of safety hazards are still in their infancy. Recent reports by the Construction Industry Institute [86] and the BLS data [2] shows while safety programs are widely implemented, the need for faster and more productive delivery of projects, is still negatively impacting safety and as such, construction fatality and non-fatal injury rates have been steadily growing. While lagging indicators used in today's practices help executives understand these safety trends and determine which projects need more attention, at the project-level these metrics may not be effective in offering superintendents and project managers with actionable insight to proactively identify and prevent safety hazards. Many researchers advocate proactive safety management by adapting leading indicators [22, 87, 88, 89, 90, 91, 92, 93]. In contrast to the lagging indicators, the leading indicators do not rely on past injury data and promote positive feedback from safety-related practices. Another body of work leverages safety-related observation data to measure leading indicators [13, 94, 95, 96, 97, 98, 99] or enable real-time safety alert systems [23, 100, 101, 102]. Extended reviews on real-time location tracking systems are provided by Teizer *et al.* (2015) [103] and Soltanmohammadlou *et al.* (2019) [104].

A critical component of these methods and systems is real-time visibility to current states of workers and equipment on site using remote sensing technology [23, 105, 106]. For example, location monitoring and quantification of worker's proximity to hazard zones provides an objective measurement of the hazard severity [13, 94]. In addition to locations, Wang and Razavi (2018) [106] suggest assessing proximity safety using other factors such as worker velocity, blind spots and resource orientations. However, estimating the future locations of tracked objects is rarely discussed in the literature. Without considering the future locations, estimation of proximity events and using that information in systems that alert stakeholders on safety risk may be limited. In the absence of such systems and relevant information, workers, equipment operators, and safety managers may not have enough response time to prevent accidents. A system that proactively alerts project stakeholders on upcoming safety issues should be able to track and forecast future locations of workers and construction equipment.

Nevertheless, forecasting future locations of workers and equipment based on previous observations is challenging. Most related literature such as [107, 108, 104] explores the application of real-Time Localization Systems (RTLS) via wireless network, Global Positioning System (GPS), Ultra-wideband (UWB), and Radio Frequency Identification (RFID) that recognizes unauthorized action, entrance of workers inside a predefined risk area around equipment, indoor hazard assessment, and real-time collision prevention. Despite their benefits, systems that tag workers are still faced by privacy concerns. More research on non-intrusive technologies and data handling protocols is needed to increase workers' acceptance.

The vision-based technologies provide an alternative solution to track workers and equipment. The combination of visual data and computer vision algorithms provide an easy, inexpensive, and rapid mechanism for generating a large body of operational knowledge, and naturally in a non-intrusive way [103, 24, 109, 31, 25]. As most computer vision algorithms are developed to resemble human vision, the direct outputs, such as class labels, bounding boxes, segments, are more intuitive for human comprehension. Also, visual data is rich in semantic information such that the same set of data can also be reused for different safety applications [25]. However, vision-based approaches face critical shortcomings. The current visual perception algorithms for construction often do not generalize well due to intra-class variance in real-world data. Occlusion is a critical limitation that hinders the accuracy of tracks [110, 104]. Other challenges in vision-based tracking for construction, such as camera locations and field of view, data collection and converting image-based trajectory to world coordinates, have been discussed in other works [111, 112, 113, 114]. Considering their benefits and limitations, several companies have already implemented these systems on their sites; e.g., Skanska's use of RTLS system [115], for checking worker and equipment arrival and leaving time. Oxblue [116] uses site visual data for tracking and estimating activity levels. While the authors focus on vision-based methods in this manuscript as an example, the underlying algorithms offered for forecasting locations can be applied to non-visual based tracks as well. The authors take advantage of the advent of modern computer vision techniques in detection and tracking [46, 50, 117], as well as GPU technology. Compared to vision-based tracking methods previously applied to construction applications, the presented tracking pipeline is superior in terms of speed, accessibility and tracks' quality and details are introduced in the Dataset and Implementation Details section.

To provide insight into how safety issues can be identified proactively, forecasting arbitrary length trajectories in real-world data is not trivial. The uncertainty often originates from the length of historical data, semantic scene layout, surrounding objects' motion, and actors' intention [118, 119, 120, 121, 122]. Existing methods are often applied to simple scenes; for example, a person walking in a parking lot [118]. Template-based tracking methods, such as Karman filter [110], bear similarity to path forecasting as they also estimate the state of the next step from past observations. However, in practice, they are used as a robust estimation of instrument measures, not for iterative future sequence generation [123]. This paper takes advantage of readily available tracked worker and equipment data from single camera construction site videos and offers a new method for actively reporting future events of workers and equipment that has the potential for real-time safety alerting systems. In addition to exhaustive experiments on forecasting models, the applicability of the forecasting model based on visual feedback is also demonstrated by putting the model into a real-world safety monitoring application prototype. Note, single-camera videos are always affected by illuminations, occlusions, and change in the camera field of view. In the following sections, the authors first offer an overview of the relevant methods and gaps in knowledge. Next, the method, experimental setup and verification steps are discussed in detail.

3.2 Related Work

Over the past two decades, a large body of research has focused on data-driven approaches to discover safety leading indicators. These methods have been explored with both project-level and operation-level data to correlate observations with safety measurements. Computer vision methods such as object detection and object tracking have been frequently employed to recognize construction resources. A complete review of these methods is out of the scope of these papers and readers are encouraged to look into [103, 124, 109, 47]. Instead, the most relevant works are highlighted in this section:

Vision-based detection and tracking in construction. Previous work has introduced vision-based tracking methods to construction sites and a number of most recent publications achieve promising results. The majority of these works formulate tracking resources in 2D or 3D as a data association problem and as such, they focus on tracking a template-based detector in consecutive frames [125, 113, 111]. Because these methods rely on the most recent observations for tracking, they do not perform well under occlusions. Robust estimation methods such as [110, 112] have been introduced that address the occlusion problem for single object tracking on construction sites. Kim *et al.* (2019) [99] is the most recent example where it is shown that state-of-the-art real-time object detection improves the localization of workers and equipment. In the past decade, the advent of deep neural network features has revolutionized object detection methods in the computer vision domain. MASK-RCNN [46] and YOLO [50] detectors are two outstanding examples that are built for generic object detection. Their designs favor performance and inference time respectively but are both generally reliable. However, a set of their experimental results on MS-COCO [67] suggests that modern object detectors still struggle in drawing bounding boxes around objects that look cluttered or small in the image. Object tracking has been a challenging task [126]. However, recent works such as [117] show that it suffices to address tracking problems by applying rudimentary optimization techniques to object detector outputs. These methods enable fast prototyping on construction site safety applications from off-the-shelf models. Nevertheless, it is worth noticing while workers appear very similar to generic person instances, construction equipment such as mobile cranes and backhoe loaders are not included in common datasets and can not be easily detected. For more details on such datasets and methods, readers are encouraged to look into [127].

Motion trajectory and activity forecasting using visual data. Motion trajectory and activity forecasting are similar topics and have been explored extensively in the computer vision community [128]. The former, also referred to as *path prediction* in some work, predicts one or more agents' future 2D/3D locations, the latter predicts future low-level human activities in videos. Specifically, forecasting requires enough temporal information to differentiate ambiguous predictions. For example, the actions of getting-on and getting-off the car may appear the same in some transition states. On the other hand, over-exploiting the demonstration may lead to the model constraining itself and overlooking other faithful alternative predictions, this is a part of a phenomenon known as mode collapse in generative models[129].

Path prediction. Early work employs clustering, Kalman filter, linear regression, auto-regression and non-linear Gaussian process [130, 131]. These methods often only work under laboratory settings and simple scenarios. The *inverse reinforcement learning* [118, 120, 121] framework is recently introduced to this topic, it models the actor's strategies under certain states, and also take into account the effect of static semantic environment on pedestrians. A typical constraint of this type of method is that it needs at least some estimation of the endpoints to forecast the path connecting them. Others perform path prediction in real-world data by regressing sequential locations using Long Short-term Network (LSTM) model[119, 132, 122, 133], this is also known as an imitation learning problem. Many LSTM based methods are adapted from the sequence prediction model proposed by Graves (2013) [123], which combines LSTM and Mixture Density Network(MDN). Pedestrians and road users interactions figure prominently in prior work. Socially acceptable behavior in path prediction has been extensively studied in [119, 122, 121, 132, 134]. For example, Social LSTM[119] introduced social-pooling to combine LSTM states of nearby moving pedestrians. The impact from the static world had been investigated by Kitani *et al.* (2012) [118], where semantic segmentation masks are used to estimate the rewards of different states. Nevertheless, these methods are often developed for predicting short sequences and their application in long-term forecasting is not investigated.

Activity prediction stems from activity classification and is also a closely related topic to path prediction. The majority of research conducted on activity classification benchmarks also provides tasks for activity prediction; readers are referred to previous work [135, 136]. The objective of activity prediction is to predict immediate follow-up activities from an observed sequence of activities. A detailed survey of relevant work is provided by Kong and Fu (2018) [128]. Activity prediction is connected with path prediction because the forecasting formulation can often be applied to both tasks. In the Method section, the authors discuss how the presented path prediction model can be directly adapted to predict activities from construction site videos. Validating this model for activity prediction in construction is beyond the scope of this manuscript and is considered as part of the future work.

Proactive safety management for proximity hazards. Several works in the literature have focused on leading indicators and methods for proactive safety management at the project-level. For example, Hallowell and Gambatese (2009) [11] identified highly effective components of a safety program and quantified each component's ability to mitigate safety risks. Hinze *et al.* (2013) [87] characterized leading indicators as passive and active measures that are predictive over an extended period of time or are able to initiate corrective actions in a short time, respectively. Hallowell *et al.* (2013) [14] reviewed proactive metrics from safety related practice and suggested near-miss reporting to be considered among the top priorities. Sheenhan *et al.* (2016) [88] provided evidence of the link between leading and lagging indicators and suggested the middle management plays a moderating role. Guo and Yiu (2016) [89] described a pragmatic method to identify leading indicators through conceptualization, indicator generation, validation and revision. These works suggest that leading indicators can be identified from non-injury safety data, such as safety observations and near-miss reports. Useful leading indicators should explain the status to workers and executives and support their decision making.

At the object-level, proactive management using workers and equipment data has been extensively explored. Workers and equipment's location data has been used to identify proximity events and assess risks [13, 94, 95, 98]. Proactive construction management systems have been explored using observations of unsafe actions and positions [137, 138, 97, 102, 56]. Real-time alert systems have also been investigated based on proximity checking [23, 100]. These system all benefit from employing low-latency remote sensing technologies. Triax (2020) [139] is a recent example of these systems that has been successfully implemented on 100s of jobsites. Despite their benefits, their implementation still faces resistance from the union workers. For example, on a project in the state of Illinois, union ironworkers refused to wear these safety clips due to their intrusiveness. In contrast, construction site visual data is non-intrusive and has been traditionally used for many construction applications [24, 31], including predictive progress analytic [140]. Recently, visual data has been used for proactive safety measurement. Fang *et al.* 2018 [56] proposes a personal fall protection inspection model for steeplejack workers and provide a fall accident analysis based on the configuration of recognized resources. Another recent attempt is made by SmartVid [141]. Similar to earlier works of Khosrowpour *et al.* (2013) [142], their solution conducts personal protective equipment recognition outputs from images collected from job sites. Together with other features such as project types and weather, their method also predicts the chance of accidents in the future. Nevertheless, solely relying on visual data for safety inspections has several limitations: cameras can not monitor objects out of view or occluded; cameras need to be calibrated to recover objects' 3D world coordinates in most cases. In this paper, the authors do not explicitly address these limitations and conduct all experiments under normal conditions. This is because the primary goal here is to present and validate a method that can leverage existing visual feeds from jobsites to provide actionable insight to project management. Since site cameras are almost always available on every construction site, the application of this method may offer an opportunity to prevent accidents that may otherwise go unnoticed.

The method presented in this paper is developed following this philosophy: construction safety management systems should not only generalize well to different input data, tasks, and occasions but also faithfully explain why the predictions are made. The outputs of these systems should inform when, where and how to prevent accidents.

The most closely related work to ours is Kim *et al.* (2019) [114] where a proof of concept is offered. Specifically, Kim *et al.* (2019) [114] retrain Social GAN [122] models using the data provided in the same work and test model performances on forecasting paths of a worker, an excavator and a wheel loader in 916 sequence frames. While Social GAN is a strong baseline for short-term (in terms of time steps versus the actual time) dense trajectories, our work focus on long-term predictions in long sequences. Our models are trained and tested on a larger dataset which has over 3000 tracks and contains more than one million steps. The longest track in our presented dataset contains 1996 steps. The maximum prediction length in Kim *et al.* (2019) [114] is 16 steps ahead, whereas in experiments conducted in this article, predictions on 10, 20, and 40 steps ahead are reported. Nevertheless, the authors compare Kim *et al.* (2019) [114] with the proposed method, the results show that our motion trajectory forecasting model significantly outperforms Kim *et al.* (2019) [114](see section Experiments).

3.3 Method

In this section, the authors formulate the problem of using construction site videos to forecast workers and equipment trajectories. In brief, for each tracked object, the model takes the 2D image plane coordinates of an object along with its contextual features at the current step as the input. These features are used to generate 2D Gaussian mixture distribution parameters. The mode of the Gaussian mixture is then treated as the predicted location for one future time step of each tracked object. The model outputs a sequence of locations for future time steps; i.e., 10, 20, and 40 steps ahead.

Prediction formulation. This paper makes an analogy between the path prediction problem and the sequence generation problem presented in Graves (2013) [123], where LSTM is used to capture the long term dependency in multidimensional sequential data, and is successfully validated in hand-writing stroke generation. Similarly, path prediction can be considered as prediction future locations from previously observed location sequences. Note, this formulation is different from that of a tracking problem because no observation in the future is used in forecasting. In the following, a high-level overview of the forecasting problem formulation is offered and discussions on how this model has been adapted for our task is described later. Let vector $\mathbf{x} = (x_1, x_2, ..., x_T)$ is the input to past trajectories of construction resources (i.e., tracking trajectories of workers and equipment) from videos obtained through deep learning algorithms and its prediction target vector $\mathbf{o} = (o_1, o_2, ..., o_T)$. O representing all possible values for o_t. The goal is to find the most likely o_t, i.e. which maximizes $Pr(o_t|x_{1:t})$. A latent variable y_t and functions $\mathcal{Y}$,$\mathcal{H}$ are introduced, then target can be written as $\hat{o_t} = argmax_{\hat{o_t}} Pr(O|y_t)$ and $y_t = \mathcal{Y}(\mathcal{H}(x_{1;t}))$. When $\mathcal{H}$ is a single-layer LSTM and $\mathcal{Y}$ is a fully connected layer, y_t can be obtained from encoding equation:

$$h_t = \mathcal{H}(W_x x_t + W_h h_{t-1} + b_h) \tag{3.1}$$

and output equation

$$y_t = \mathcal{Y}(W_y h_t + b_y) \tag{3.2}$$

where $\mathcal{Y}$ denotes the output layer function parameterized by weight matrix W_y and bias unit b_y, W_x is the weight matrix to embed input data, W_h is the weight matrix to process hidden state from last step, b_h is the bias units of the LSTM. So the likelihood of target vector $Pr(\mathbf{o})$ is

$$Pr(\mathbf{o}) = \prod_{t=1}^{T} Pr(\hat{o}_t|y_t) \tag{3.3}$$

and the loss function $\mathcal{L}(\mathbf{o})$ to train LSTM model is

$$\mathcal{L}(\mathbf{o}) = -\log Pr(\mathbf{o}) = -\sum_{t=1}^{T} \log Pr(\hat{o}_t|y_t) \tag{3.4}$$

This model can be trained with standard backpropagation through time and gradient descent optimizers.

Applicability for construction tasks. Although the aforementioned prediction formulation is designed for path prediction, it can be easily extended to other forecasting tasks and potentially used to automate workforce assessment applications such as the one described in Liu and Golparvar-Fard (2015) [47]. In path prediction, both x_t, model input at time t, and o_t, model output at time t, are continuous image plane coordinate (w, h). For activity prediction, a closely related task to path prediction, both x_t and o_t can be one-hot vectors of length C, where C is the number of target activity categories. Because the output space is no longer continuous, MDN can be simply replaced by linear layers followed by softmax. While such applications are very plausible, such work is outside the scope of the current manuscript and is considered as future work.

Using contextual cues for activity forecasting. Inspired by previous works that investigate the social and environment impact on path prediction [118, 119, 122, 121, 132, 134],the authors make similar analogies to construction sites.

- **Social feature**. Many previous works consider the interactions between tracked objects[119, 122]. However, both static and moving objects can affect workers and equipment behavior on construction sites. For instance, a worker may try to avoid an excavator in motion that is backing up. The Occupancy Map feature (denoted as $Feat_{Social}$) in Alahi *et al.* (2016) [119] provides a simple way to capture social features from both static and dynamic objects and can be precomputed to speed up the training.

- **Attribute feature**. Interaction patterns may differ based on the object's attributes[120], this paper uses one-hot object class (i.e., worker and vehicle) vectors as a feature (denoted as $Feat_{Attribute}$) to differentiate tracked objects.

Output layer and multi-head predictions. In multidimensional sequence generation, MDN is often used to address the uncertainty derived from one-to-many mapping between the input sequence and the output sequence. Its advantage is demonstrated in Graves (2013)[123]. Hence MDN is used here as the output layer $\mathcal{Y}$. Contrarily to most prior work that define the target variable as the offset between two consecutive steps $\hat{o}_t = x_{t+1} - x_t$, in this work target variable $\hat{o}_t$ is the actual location x_{t+s} in s steps. Given h_t. MDN outputs parameters of a mixture Gaussian distribution at each step, such that output equation is rewritten as

$$y_t = \{\pi_t^j, \mu_t^j, \sigma_t^j, \rho_t^j\}_{j=1}^{M} = \mathcal{Y}(W_y h_t + b_y) \tag{3.5}$$

where M is the number of mixture components, processing of Gaussian parameters is the same as Eq.19-22 in Graves (2013) [123]. Using multiple MDN output layers is, in spirit, similar to conducting sequence to sequence prediction which alleviates error propagation in long sequences. A set of prediction steps is noted as $\mathbf{s} = (s_1, s_2, ...s_K)$. Therefore, the likelihood function and loss function can be expressed as:

$$Pr(\hat{o_{kt}}|y_t) = \sum_{j=1}^{M} \pi_t^j \mathcal{N}(\hat{o_k t}|\mu_k t^j, \sigma_k t^j, \rho_k t^j) \tag{3.6}$$

$$\mathcal{L}(\mathbf{o}) = -\sum_{k=1}^{K} \sum_{t=1}^{T} \log Pr(\hat{o_{kt}}|y_t) \tag{3.7}$$

where $\mathcal{N}$ is the 2D Gaussian function parameterized by $(\mu_t^j, \sigma_t^j, \rho_t^j)$. At test time, $\hat{o}_t = \mu_t^{argmax_j \pi_t^j}$.

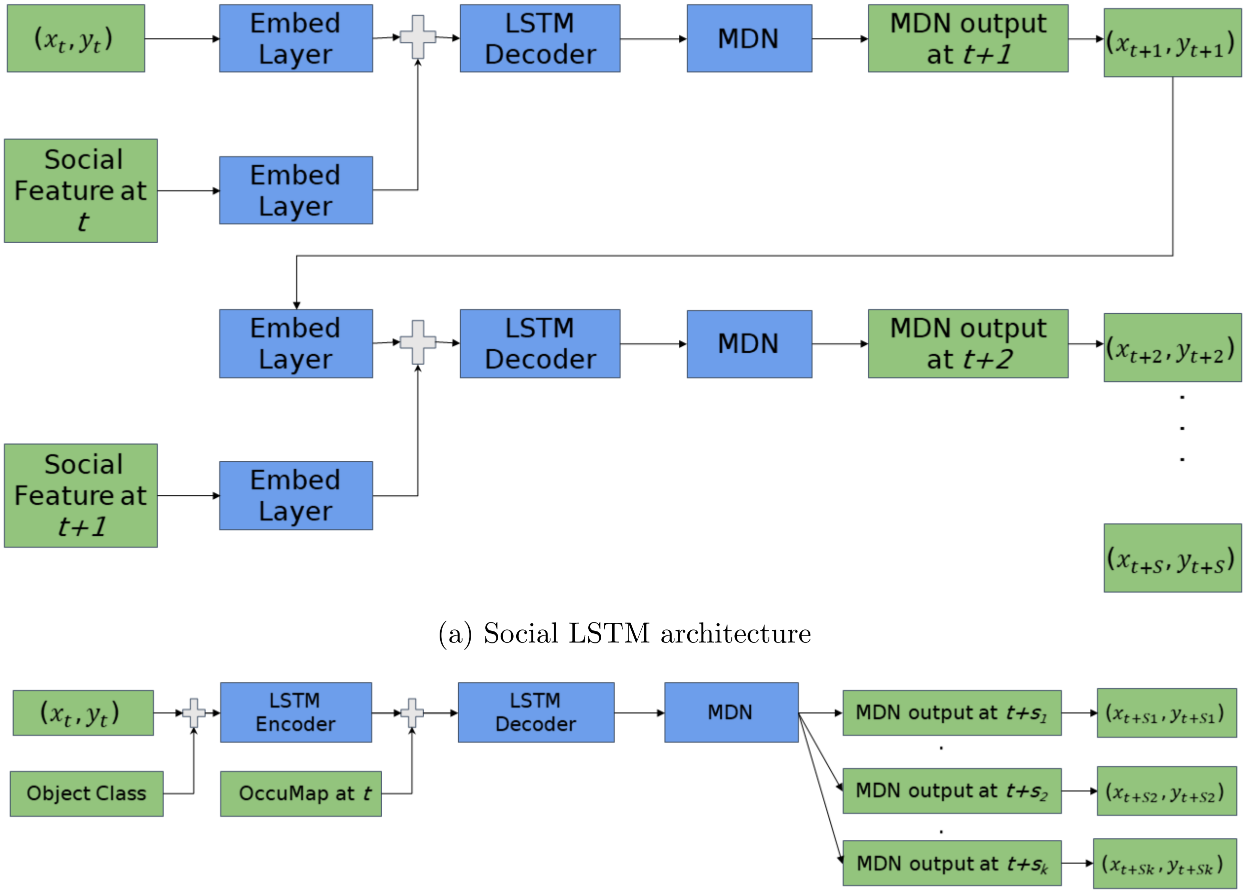
Embedding contextual data. The models illustrated in Fig. 3.1. Our model architecture is presented in Fig. 3.1b. Different from Social LSTM (Fig. 3.1a), who uses two separate multi-layer perceptions to embed current locations and social tensor then decode with a single LSTM layer. The proposed method employs rigorous embedding of contextual features. Specifically, the input image coordinates x_t is first concatenated with $Feat_{Attribute}$ before fed into the LSTM encoder, whose embedding latent feature is concatenated with $Feat_{Social}$. Then the concatenated LSTM states are passed through an LSTM decoder, the decoder outputs directly to MDN, which generates mixed Gaussian parameters for $\mathbf{s}$. The intuition is that different types of contextual features may work in a hierarchical way of contributing to decision-making instead of being crammed together. Variations to this design have been tested and whose results are outperformed. The detail numbers are not reported in this paper.

3.4 Dataset and implementation details

To validate the proposed method, two real-world datasets are used for experiments. The first, Voyager dataset, contains 105 videos from a fixed camera recorded for 30 days on one construction site. The recorded area roughly covers a 270-meter by 35-meter region where many underground utilities are installed during the recording. The site-camera configuration is shown in Fig. 3.2; the second dataset, the TrajNet dataset [143], is built from overhead camera footage. The TrajNet dataset is derived from commonly used path prediction benchmarks in previous works and is currently the largest benchmark for evaluating generic pedestrian trajectory forecasting models.

Voyager data collection and processing. Over 1,000 minutes of 1080p high-definition videos are collected from a single construction site (see examples frames in Fig. 3.3). Trajectory data is automatically generated with a detect-match schema using a state-of-the-art real-time tracker [117], that is:

- All videos are extracted at 20 frames per second and all extracted frames are detected by an off-the-shelf Mask RCNN model [46] trained on MS-COCO [67] dataset.
- With the confidence thresholds of 0.5, detection results of three MS-COCO categories, *person*, *truck*, *car* are selected from every frame.



(a) Social LSTM architecture

(b) Proposed model architecture

Figure 3.1: Proposed model architecture compared against Social LSTM [119] architecture. Social LSTM model (3.1a) predicts future path in an iterative manner, while the presented model (3.1b) generates and updates samples of the future paths within fixed future time intervals.

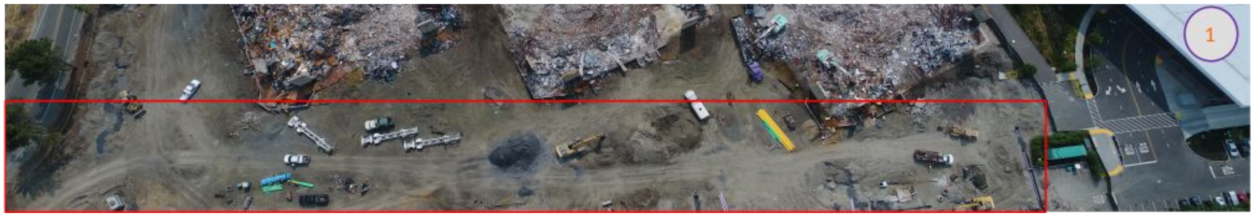


Figure 3.2: Site-camera configuration for the voyager dataset. The red bounding box marks the area monitored, the circle on the top right corner marks the position of the camera mounted on a tripod. The monitored area is 270 meters in width and 35 meters in height.

- Detected instances for each video are then processed by SORT [117] tracker to obtain raw trajectories, the life-span for each track is set as 20 frames to allow re-identifying occluded instances;
- Tracking results are interpolated and smoothed to fill in missing values.

Trajectories of *construction equipment* and *passenger vehicles* are merged into *vehicle* category, as the off-the-shelf object detection model is not trained on equipment such as excavators and mobile cranes. The Mask-RCNN model often confuses them with trucks and cars, but these instances are very robust to be differentiated from the rest of COCO categories so that highly confident vehicle instances can still be obtained. For other construction site video collections, worker and vehicle tracks can be easily obtained following the presented tracking pipeline. In total, 17397 person tracks and 33037 vehicle tracks are recognized for the Voyager dataset. Note that not all tracks are used for experiments, a few criteria are applied to find admissible tracks. In general, admissible tracks are selected where they last at least 30 frames but not longer than 2000 frames, and whose endpoints L2 distance is longer than 50 pixels. Admissible tracks from the 76 videos are used as training and validation sets (that we refer to collectively as *trainval*). 20% of tracks in *trainval* set are used for model selection, admissible tracks from the last 29 videos are used as the *test* set. In the *trainval* set there are 1630 person and 1752 vehicle tracks, in the *test* set, there are 143 person tracks and 161 vehicle tracks. Fig. 3.4 shows admissible tracks' length and duration distributions in both the *trainval* set and the *test* set, these two sets are very similar. Fig. 3.3 shows examples of extracted frames, the top row is from the *trainval* and the bottom row is from the *test*, one can tell the site layout has drastically changed during the recording. Although the aerial view of the site is obtained as in Fig.3.2 and track conversion from the image plane coordinates in pixels to the world coordinates in meters is possible, all experiments results are reported in pixels. This is because creating a 3D scene from visual data and projecting the forecasted motion trajectories within that scene or establishing homography transformation between two scenes are mutually exclusive research tasks to path prediction. However, for reporting purposes, the mapping conducted between reference object's image size and physical size shows that for experiments conducted, each pixel in the near-camera

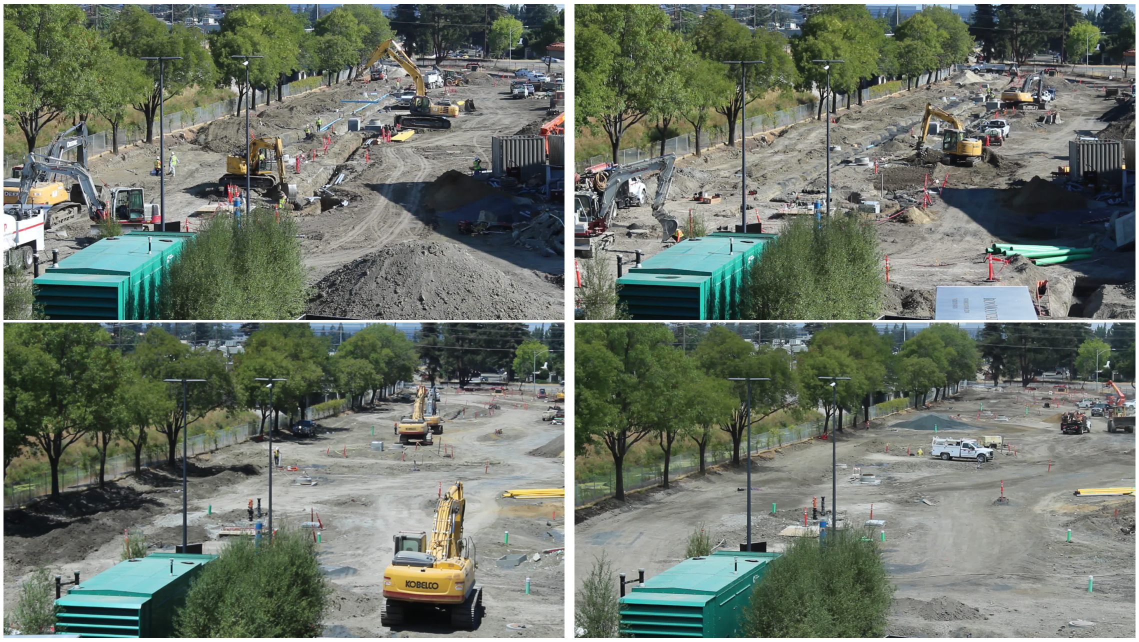


Figure 3.3: Examples frames of Voyager videos. The dataset is recorded within 30 consecutive days. An example of video recorded in the first day is shown at upper-left and the example from the last day appear at the lower-right. Images from the top row are samples from the *trainval* split; and images from the bottom row are samples from the *test* split.

represents 5-10cm in real-world coordinate systems.

TrajNet dataset. Only one subset of TrajNet, *World Plane Human-Human*, is used throughout the experiments. The tracks in this subset are transformed under the world coordinates from several videos in various scenes. Only pedestrians' tracks are recorded. Because the evaluation server is not accessible, only the *train* set is used for training and validation. For the rest of this paper, TrajNet data is referred to as the *train* set of the World Plane Human-Human subset. In total, TrajNet data includes 11448 tracks from 58 scenes, each track has 20 steps, see example tracks in Fig. 3.5. When evaluating Trajnet data, forecasting models use the first eight steps of each test track and general locations of the rest 12 steps. All tracks in TrajNet data are evaluated on the TrajNet evaluation server, all prediction heads' results are averaged.

Implementation details For both datasets, data are normalized by train set coordinates' mean and standard deviation, model predicted locations are denormalized for final predictions. This processing allows for the forecasting model to be applied on non-visual

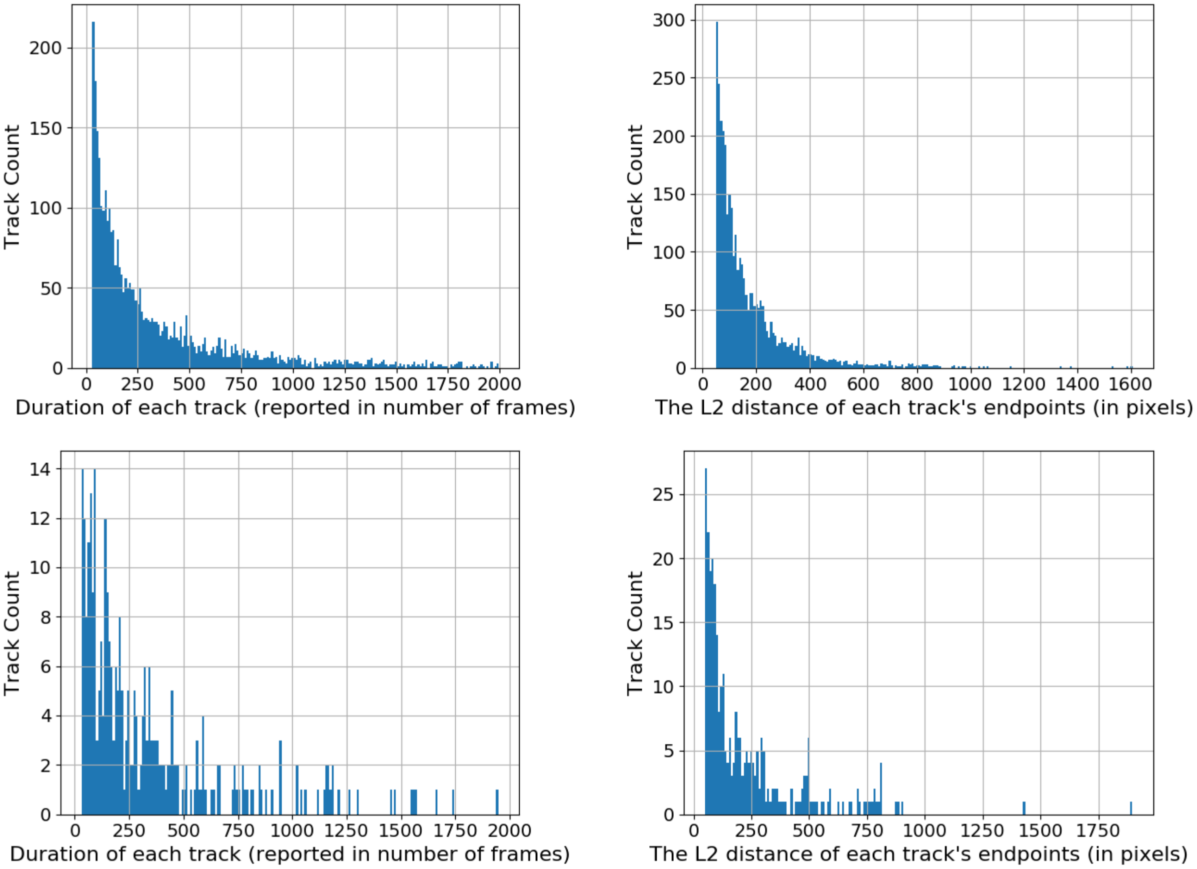
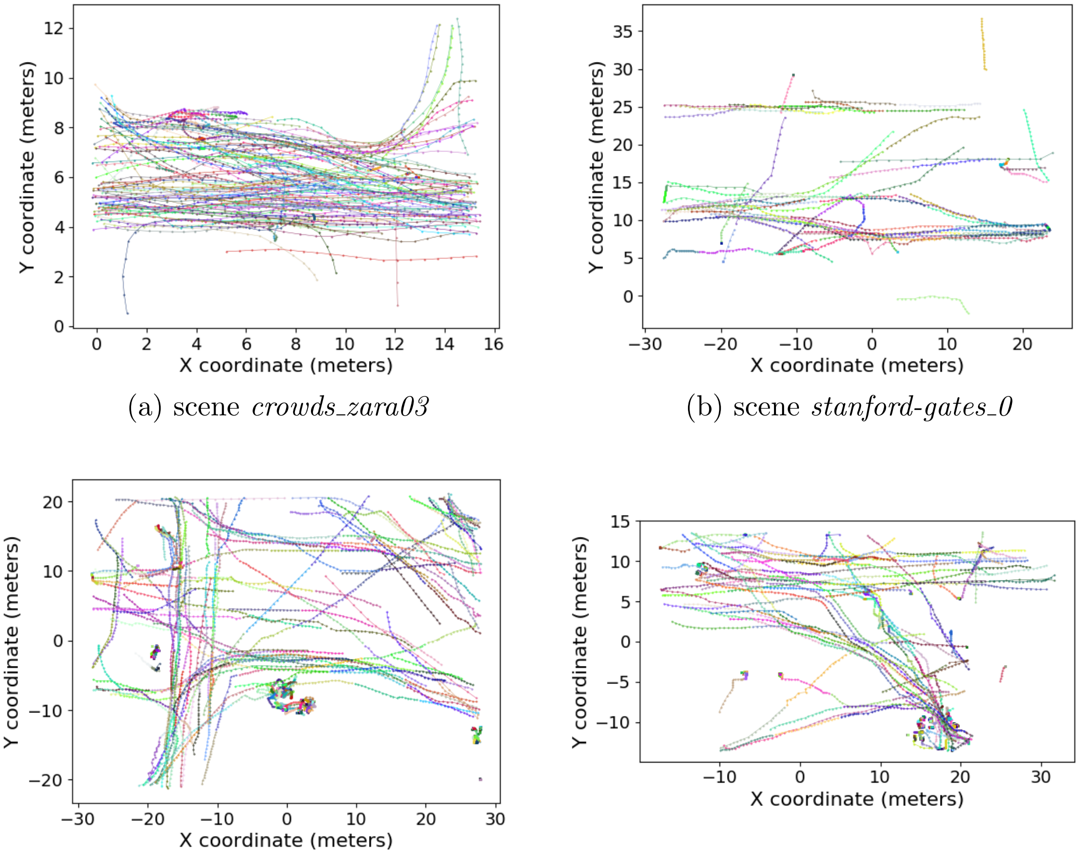


Figure 3.4: The duration and endpoint distance distribution within tracks of the Voyager dataset. Plots in the top row are for the *trainval* set tracks and the bottom row for *test* set tracks. The duration and endpoint distance distributions of the *trainval* set and the *test* set are similar according to admissible track criteria.



(a) scene *crowds_zara03*

(b) scene *stanford-gates_0*

(c) scene *stanford-bookstore_1*

(d) scene *stanford-coupa_3*

Figure 3.5: Examples tracks in TrajNet data by scenes. The unit of measure is meter. Note that even though all track coordinates are transformed to the world plane, they do not necessarily have the same magnitude. Coordinate values can be within 0 to 5 meters or can be around 20 meters. Track patterns also vary between scenes, for example, the *crowds_zara03* tracks are more homogeneous than that of the *standford-bookstore_1*. Such data variance naturally calls to leverage contextual information.

based tracks. Both the encoder and the decoder LSTM have 32 hidden units. For both datasets, the social pooling region is a 2-by-2 square grid centered at the location x_t. In Voyager dataset, this grid has 200 pixels on each side, in the Trajnet dataset, 10 meters is chosen. When choosing the prediction heads, 10, 20, 40 future steps are selected for the Voyager dataset. Note this particular choice does not constrain the proposed method to specific forecasting time-scales, as the choice was simply to make sure verification can represent increasing difficulties for path prediction. Because Voyager dataset video frames are sampled at 20 frames per second, the prediction future times correspond to 0.5, 1.0 and 2.0 seconds, respectively. Here, reducing the video frame sampling ratio increases the actual forecast time. Twelve prediction heads are used for TrajNet data. All models are optimized with Adam optimizer and a learning rate of 0.005. Gradient clipping is applied to 50% of the global gradient norm, this is proved to be essential to stabilize the training. For both models, training tracks are dynamically packed such that one batch of tracks with different lengths can be efficiently trained in one mini-batch. The Voyager model is trained with batch size 128 and 1500 epochs, learning rate decay is set to 0.5 at 600, 900,1100,1300 epochs. The training on the Voyager dataset, which has around one million steps in total, can be finished in one and a half hours. The TrajNet model is trained similarly but the batch size is reduced to 64 and the number of epochs is reduced to 1000. All models are implemented with PyTorch 1.0 and trained with a system with a single Nvidia GeForce 2080Ti GPU, an Intel i7-8700k CPU, and 32 GB RAM.

3.5 Experiments

Baselines. Five different baseline path forecasting models are implemented for comparison with the proposed method. The first is a linear regression model (Linear Reg) parameterized by time t. For each track in the test set, locations from x_0 to x_t are used to regress a function to predict future locations x_{t+s}. Polynomial functions are also tested because they may perform better compared to linear models in smaller datasets. Their results are worse than that of linear regression and thus not reported. The second baseline is the popular time-series analysis method of Vector Autoregression (VAR), here lag order $p = 5$ is chosen based on parameter tuning. VAR combines past observations linearly and is a strong baseline when the subjected data is periodical. One single VAR model is computed for each track in the test, each VAR model takes 5 steps and outputs the next step's location. Errors of all test tracks are averaged. The third baseline does not consider the temporal dependency, each LSTM is replaced by a Multi-Layer Perception (MLP) with the same number of hidden units, this is referred to as MLP+Reg. The fourth is a simplified model where the MDN part is removed (LSTM+Reg.), leaving the LSTM decoder output as the predicted locations. The last baseline is Kim *et al.* (2019) [114], where a retrained Social GAN [122] is introduced. The authors apply the same procedure as in Kim *et al.* (2019) [114] and retrain a Social GAN using the code repository of Gupta *et al.* (2018) [122]. This Social GAN takes the first 8 steps and predicts locations for the next 40 steps. Note that our model is not constrained by observing 8 steps first. The retrained Social GAN model is validated on the Voyager data test set.

Ablation study. The authors exhaustively evaluate each component of the proposed model. Without any contextual cues, LSTM+MDN is a baseline in the ablation study. Multi-head prediction is evaluated versus single-head prediction and evaluations on the contribution of each contextual feature are conducted. Because the gap between experiment results are small and are affected by the randomness of model initialization and optimizer initial state, five rounds of experiments are conducted for each ablation model, the final results are averaged from all experiments. Each ablation model's results are compared with that of LSTM+MDN by student t test. When the p value is smaller than 0.05, the hypothesis is rejected and the difference between the models' results is deemed to be statistically significant.

Error evaluation. Predicted trajectories for the Voyager dataset are evaluated by the average Root of Mean Square Error (RMSE) between actual and predicted locations per prediction head; e.g., in prediction heads for 10, 20, 40 frames their error metrics are RMSE@10, 20, 40, respectively. For the TrajNet dataset, *Mean Disp. L2* is computed by the average RMSE between actual and predicted later 12 steps of all tracks. *Final Disp. L2* is the average RMSE of actual and predicted final locations of all tracks. The average error is the average of the Final error and the Mean error. Lower values in these metrics mean better model performances.

Results and Analysis. Table 3.1 presents all experimental results from the Voyager dataset. First, when prediction performances are compared across all three prediction heads, it is intuitive to see longer forecasting time leads to lower prediction capability. Next, each method's performance is examined.

Section *Baseline* compares all the baseline model performances. Conventional time series models, Linear Reg. and VAR, are significantly worse than the deep learning methods, this shows that the Voyager data is collected in a way that demonstrates real-world difficulties associated with motion trajectory application from existing cameras. The results achieved with MLP+Reg., LSTM+Reg., and LSTM+MDN model are shown in Table 3.1 and validate the effectiveness of using LSTM and MDN when they are used together. By considering the temporal information of trajectories using LSTM, the LSTM+Reg. model reduces nearly 50% localization errors particularly when it is compared with MLP+Reg. model. By using MDN to consider prediction uncertainty, the LSTM+MDN model significantly improves over LSTM+Reg. model, as it directly imitates this behavior from the training data. Quantitative results from the method presented in Kim *et al.* (2019) [114] (ID 4 in Table 3.1) show that the method does not perform better than baseline LSTM+Reg. The underlying intuition is that Social GAN models are designed to predict short-term dense crowd motion trajectories, and the errors propagate significantly as predicting future locations. Hence this method does not offer as a strong baseline for long-term predictions in terms of the number of future steps.

Section *Ours* in Table 3.1 shows the results of the ablation study. In these experiments, in RMSE@40 which the model achieves the best long-term prediction performance is considered as the best model. This is because for the safety application it is intuitive that longer responsive time helps equipment operators, superintendents or managers to react within a reasonable time-frame. Comparing results from multi-head and single-head prediction (i.e., ID 6 and 7 in Table 3.1), the p values for all prediction heads are larger than 0.05, showing there is no statistically significant difference between the two, but the multi-head inference is much faster than that of single-head. Adding $Feat_{social}$ or $Feat_{attribute}$ alone shows statistical significant improvement over the LSTM+MDN baseline (i.e. ID 6 and 8, 6 and 9 in Table 3.1) as well. The best RMSE@10 result is achieved by LSTM+MDN+$Feat_{social}$ model and the

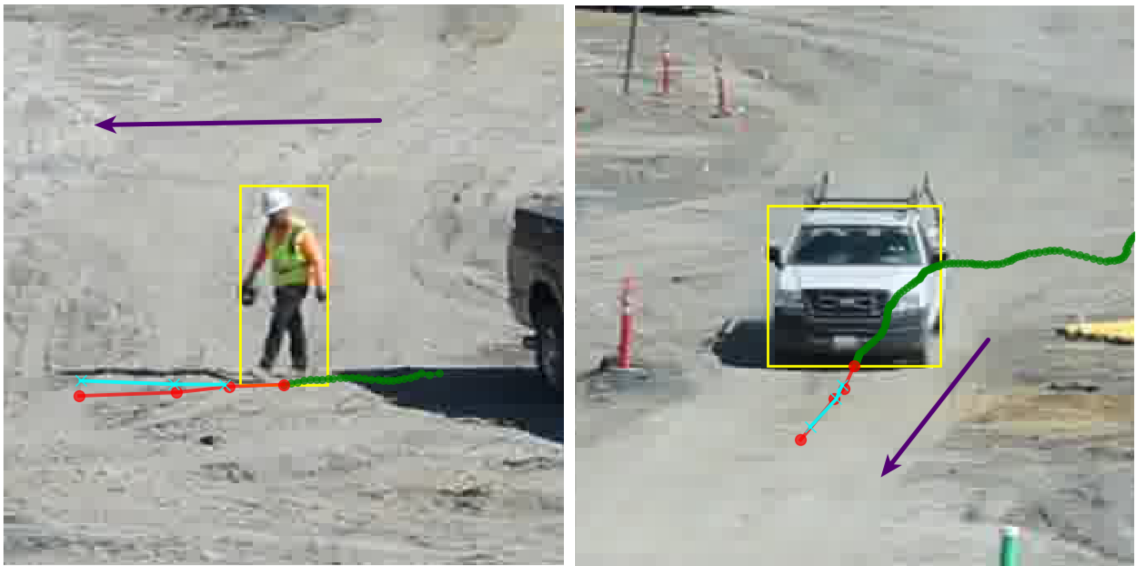


Figure 3.6: Examples of path prediction generated from the final model. Tracks of the lower center of bounding boxes are treated as the object track. The green track represents previously observed locations; the red dot line represents the current location and the actual future locations; the cyan cross line represents the model predictions. The arrows indicate the forward moving directions. Figure best viewed in color.

best RMSE@20 results are from LSTM+MDN+$Feat_{attribute}$. The full model, which combines both social features and attribute features, (ID 10 in Table 3.1) achieves the best long-term result for RMSE@40. Note that all results are reported when models converge. The performance of the model during the training and testing process remains similar and no significant drop is observed. For example, in one of the trials for the LSTM+MDN model, the test set RMSE@40 is 25.99 pixels and train set RMSE@40 is 22.68 pixels. Since the Voyager dataset's *trainval* set and *test* set are split by the date of recording, the close performances between train and test set indicate the proposed method generalizes well on the voyager data and possibly under all similar camera installations on construction sites. Two examples of path prediction from the final model are shown in Fig. 3.6. With a reference object on the near-camera side, a conversion is estimated that a near-camera side pixel measures around 5 cm distance. A new trial for the final model is run using 135 test tracks whose mean y-coordinates do not exceed 400 pixels away from the image button edge. The resulting RSME in meters at 10, 20, and 40 steps ahead are 0.28 m (5.7 pixels), 0.44 m (8.76 pixels), and 0.87 m (17.39 pixels), correspondingly.

Table 3.1: Voyager experiment results. Method ID 1-4 summarizes the baseline models' performances. ID 5-9 shows the ablation study results. Note that number in the parentheses for ID 6-9 are the p values of student t tests with LSTM+MDN performances. When a p value is smaller than 0.05, it means results from the two methods are different with statistical significance.

Group	ID	Model	RMSE@10	RMSE@20	RMSE@40
Baselines	1	Linear Reg.	62.47	68.59	82.51
	2	VAR	46.85	90.27	163.02
	3	MLP+Reg.	14.17	27.08	50.16
	4	Kim *et al.* (2019) [114]	9.33	18.32	36.30
	5	LSTM+Reg.	8.67	14.65	27.39
Ours	6	LSTM+MDN	7.42	13.26	25.25
	7	LSTM+MDN(single-head)	7.51 (0.23)	13.30 (0.54)	25.20 (0.45)
	8	LSTM+MDN+$Feat_{social}$	7.24 (0.02)	**12.70** (0.008)	24.30 (0.01)
	9	LSTM+MDN+$Feat_{attribute}$	**7.22** (0.002)	12.95 (0.01)	24.74 (0.02)
	10	Full Model	7.30 (0.09)	12.71 (0.005)	**24.22** (0.004)

Path prediction may be also affected by trajectory lengths. To investigate this factor, the authors run another trial of the final model and report RMSE conditioned on tracks' end-points L2 distances in pixels. Following the end-point distance distribution shown in Fig. 3.4, the authors consider 209 tracks with the distance less than 250 pixels as short-length tracks, 36 tracks whose distances larger than 500 pixels as long-length tracks. 59 tracks whose distances are between 250 pixels and 500 pixels as mid-length tracks. Results in Table 3.2 suggest the distances traveled is another important factor on path prediction, predictions for the short-length tracks are significantly better than longer tracks. This shows further investigation and research should be done in the future for a continuous long-term path forecasting.

Table 3.2: Voyager test set path prediction as a function of track distances. Results are generated from the final model.

Length	*RMSE@10*	*RMSE@20*	*RMSE@40*
Short-length	5.93	9.45	18.04
Mid-length	8.08	14.87	31.62
Long-length	8.50	16.00	33.10
All	7.08	12.48	24.79

A separate experiment using the final model is run for the extended future time. The actual time whose location forecasted is affected by the track sample rate, and the number of steps ahead the network is configured to predict. The authors conduct an experiment to investigate their effects on predicting locations in the extended future time. In Case 1, the sampling rate is set to 10 frames per second, prediction heads are kept as 20, 40, and 60. In Case 2, sampling rate remains at 20 frames per second, prediction heads are set to 40, 80, 120 steps ahead. Each test case predicts locations in 2, 4, and 6 seconds in the future, respectively. Results reported in Table 3.3 shows Case 1 performs slightly worse at prediction in 2 seconds, but better at predictions for 4 and 6 seconds.

A comparison between the proposed method and Social LSTM on TrajNet data is shown in Table 3.4, Social LSTM implementation are found in an open codebase [144]. The most recent social LSTM performances are reported using the Stanford TrajNet leaderboard, a

Table 3.3: Location predictions for extended future time by different data and model settings. Lower error in pixels means better.

Test Cases	*RMSE for 2 sec.*	*RMSE for 4 sec.*	*RMSE for 6 sec.*
Case 1 (lower freq., same steps)	27.55	58.79	87.12
Case 2 (same freq., extended steps)	27.42	60.11	89.07

generic dataset for short-term path prediction. The proposed path prediction models have been evaluated on the same website. Results in Table 3.4 show we significantly improve over the reported Social LSTM performances and achieve comparable performance to Social GAN [122], which was used in Kim *et al.* (2019) [114].

Table 3.4: Model comparison on TrajNet test set. Other model performances are obtained from the TrajNet leaderboard. Errors reported in meters.

Group	ID	Model	*Mean Error*	*Final Error*	*Average Error*
Social LSTM	1	Occupancy LSTM	1.101	3.12	2.1105
	2	Social LSTM	0.675	2.098	1.3865
Social GAN	3	Social GAN	0.561	2.107	1.334
Ours	4	LSTM+MDN	0.608	1.775	1.1915
	5	LSTM+MDN+$Feat_{social}$	0.574	1.665	1.1195

3.6 Potential use cases of predicted paths and the integration with proactive safety hazard monitoring tools

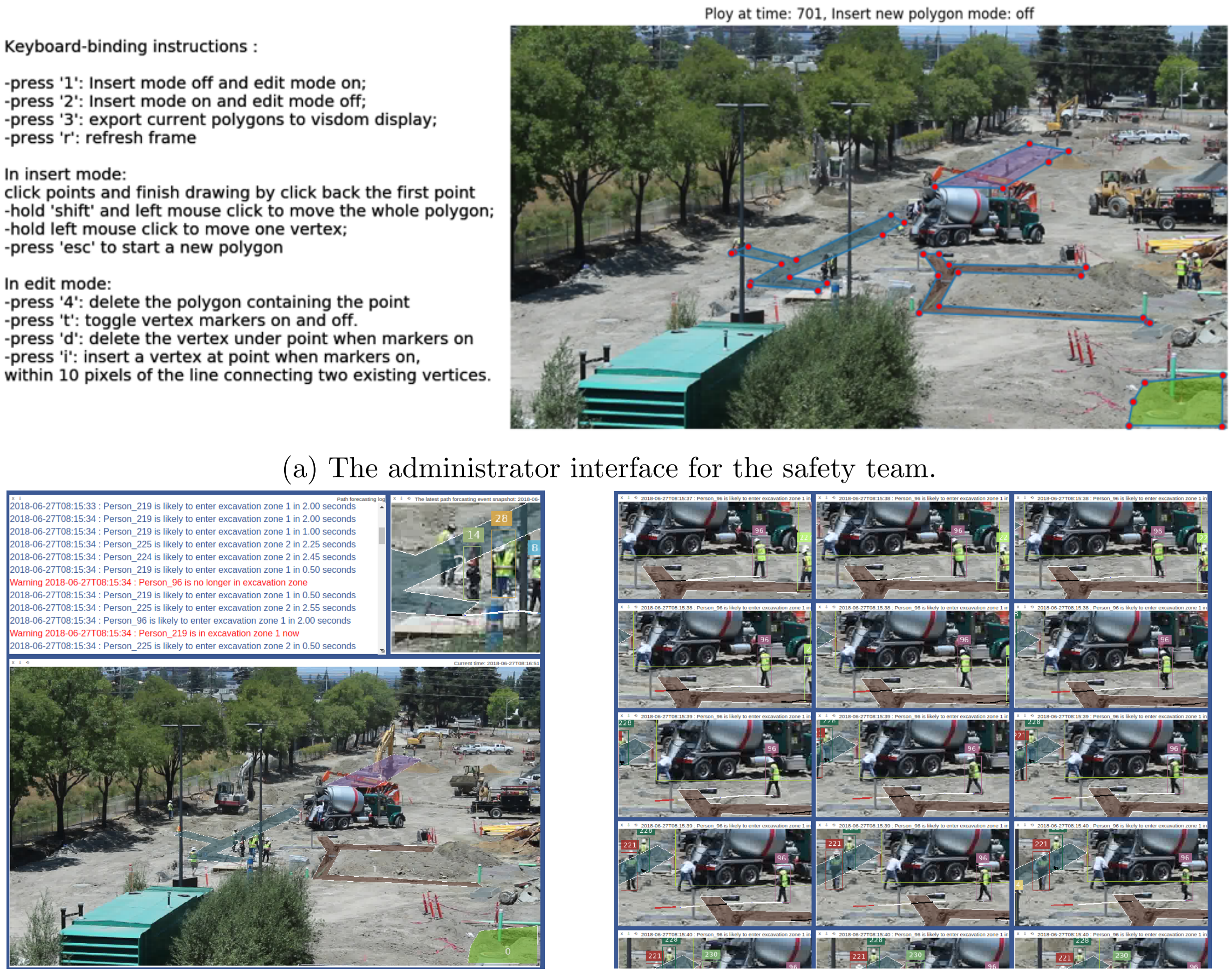
Our goal here is to demonstrate and validate a working prototype and as such a complete interface design to offer actionable items on safety was not considered. Recent commercial solutions such as Smartvid.io and Indus.ai demonstrate user-friendly examples. The trained model presented in this manuscript can be integrated into such visual safety monitoring systems that detect and track workers and equipment. The aforementioned tracking pipeline takes 1080p images and runs at 31 frames per second with 1.4GB GPU memory usage. The path prediction model simply takes in tracked objects' 2D coordinates, computes occupancy maps, and generates predictions. For example, for 91 tracked objects in the image, the forecasting inference for all objects runs at 117 frames per second with 300MB GPU memory. The forecasted locations are outputted along with the tracked locations for each object. As a simple demonstration of the working prototype, the forecasted locations are used to report potential conflicts between workers/equipment and the proximity hazards of predetermined excavation areas. A simple user interface feature is developed to highlight potentially unsafe proximity events, which may go unnoticed in the absence of an automated monitoring tool. A web-based viewer application, which consists of a backend program of two interfaces (Fig. 3.7), is also implemented to detect, forecast and document workers entering excavation areas events. In its current form, the developed prototype serves as an automatic performance update and off-line review program for safety personnel. A complete interface design and development is beyond the scope of this paper.

Specifically, the developed viewer application includes an administrator interface and a web-based viewer interface. On both interfaces, videos and the predicted paths are displayed under a fixed camera viewpoint. A bird-eye view of paths can be more intuitive but it was considered a software development task as opposed to validating the developed prototype. For future improvement, accurate transformation between views can be obtained by finding correspondences and solving a Perspective-n-Point mapping between fixed camera viewpoints and geo-referenced models (e.g., drone-driven orthophotos). Kim *et al.* (2019) [99] shows an interesting example of such an interface. The administrator interface (Fig. 3.7a) shows the streamed video and allows the safety team to add, delete, and modify excavation areas as polygons under the image coordinates. There is no restriction on the shape and the number of polygons to be annotated. The expected time of the flagged events along with the visual evidence will be pushed to the viewer interface. The viewer interface is implemented with Visdom and consists of two panels. The main panel(Fig. 3.7b) visualizes site videos masked by excavation areas (the window at the bottom). A user interface module on the top left corner of the prototyped viewer reports two types of proximity events: (a) forecasted events where a tracked worker will likely enter an annotated region, and (b) an observation that a tracked worker is already in the region. The safety events are reviewed based on image coordinates. Because the excavation polygon areas are drawn close to the ground, the workers' locations and the polygons can be considered on the same plane. By linking the forecasted locations, a predicted path is formed. When a predicted path intersects with a polygon, the estimated time to enter is interpolated based on the intersection's image coordinates and two closest predicted locations' coordinates and timestamps. When a tracked worker is forecasted to enter the zones, this event will be logged in the text box marked by blue color. When a worker is already detected within the zone, the event will be logged and marked by red color. Visual evidence of the latest captured event is shown at the upper right corner window. A separate panel (Fig. 3.7c) stores all visual evidence of captured events, each recognized worker is assigned to a unique ID number not related to their personal information. By querying the ID, all of the captured snapshot sorted by timestamps are presented to the viewer.

The reported events were examined by safety managers and superintendents. For instance, for a 12-min video, all 13 proximity events of six tracked workers were reported by the prototype. One of the safety managers confirmed four forecasted events that workers jump-in/over trench. All four forecasted event times match the actual observation times. The visual snapshot also helped the manager to identify a few PPE compliance issues such as not wearing safety coloring, long pants, and hardhat.

Forecasting future unsafe events can be critical for many intervention actions, for example alerting construction equipment drivers if there is a worker in the way. Also, the prototype design has several practical benefits. First, only images are needed as input, the prototype can take network–connected camera streams as-is. Second, the web-based viewer allows authorized viewers such as executives access real-time site conditions from anywhere on and off a construction site. Third, the current prototype runs live at the frequency of 5 frames per second (5Hz) and applying changes to the excavation areas from the administrator interface incurs a negligible time delay.

The prototype provides abundant visual evidence on how the entire event progresses. However, further studies in some aspects can significantly improve the current framework's practical values in forecasting and preventing imminent accidents. First, robust visual recognition under occlusion is helpful in terms of improving workers and equipment localization and reducing forecasting false negatives. Previous work has proposed approaches such as using a Kalman filter and a multi-view camera system [125]. Second, the proposed forecasting framework assists workers and equipment that are not aware of surrounding proximity hazards and identifies those who willingly take the risk. However, the current formulation does not differentiate those who calibrate their surrounding hazards and act cautiously. Without such differentiation, the current framework is expected to produce a high volume of false alarms, which may lead to stress and annoyance [145]. A likely approach to tell those who being cautious is to capture workers' attention, a recent example following up the proposed framework models attention by using workers' head directions [146] and shows such information improves motion forecasting. Third, although the proposed formulation can produce models forecasting locations in arbitrary future time by setting hyper-parameters, it is recommended to calibrate these parameters with respect to the alarm systems deployed. For instance, a radio system [147] that checks the dangerous distance of an approaching vehicle up to 25km/h. For a vehicle traveling at 10km/h (2.78m/s), such a system takes 21 measurements and 3.6 seconds between the vehicle's first detection to fully stop, allowing 10 meters of total travel since the first detection and 8 meters final safe distance from the vehicle. To double the final safe distance, the forecasting model needs to predict 5 seconds of future locations and the proposed framework's sampling frequency satisfy the requirement of the radio system. Lastly, experimental results show that the presented model and the working prototypes generalize well to different dates for a single construction site and across datasets, though it is still challenging to transfer the trained model to another site with a different camera configuration. Learning a generalized motion prediction model is desired but faces fundamental challenges. One of future work directions is integrating site overlay and job motivation in the forecast model, such that the model can adjust based on the condition of the site and individual trade.


Ploy at time: 701, Insert new polygon mode: off
Path forecasting log
The latest path forcasting event snapshot: 2018-06-
28
14
8
Current time: 2018-06-27T08:16:51
96
221
228
230

Keyboard-binding instructions :

-press '1': Insert mode off and edit mode on;
-press '2': Insert mode on and edit mode off;
-press '3': export current polygons to visdom display;
-press 'r': refresh frame

In insert mode:
click points and finish drawing by click back the first point
-hold 'shift' and left mouse click to move the whole polygon;
-hold left mouse click to move one vertex;
-press 'esc' to start a new polygon

In edit mode:
-press '4': delete the polygon containing the point
-press 't': toggle vertex markers on and off.
-press 'd': delete the vertex under point when markers on
-press 'i': insert a vertex at point when markers on,
within 10 pixels of the line connecting two existing vertices.

(a) The administrator interface for the safety team.

```
2018-06-27T08:15:33 : Person_219 is likely to enter excavation zone 1 in 2.00 seconds
2018-06-27T08:15:34 : Person_219 is likely to enter excavation zone 1 in 2.00 seconds
2018-06-27T08:15:34 : Person_219 is likely to enter excavation zone 1 in 1.00 seconds
2018-06-27T08:15:34 : Person_225 is likely to enter excavation zone 2 in 2.25 seconds
2018-06-27T08:15:34 : Person_224 is likely to enter excavation zone 2 in 2.45 seconds
2018-06-27T08:15:34 : Person_219 is likely to enter excavation zone 1 in 0.50 seconds
Warning 2018-06-27T08:15:34 : Person_96 is no longer in excavation zone
2018-06-27T08:15:34 : Person_219 is likely to enter excavation zone 1 in 0.50 seconds
2018-06-27T08:15:34 : Person_225 is likely to enter excavation zone 2 in 2.55 seconds
2018-06-27T08:15:34 : Person_96 is likely to enter excavation zone 1 in 2.00 seconds
Warning 2018-06-27T08:15:34 : Person_219 is in excavation zone 1 now
2018-06-27T08:15:34 : Person_225 is likely to enter excavation zone 2 in 0.50 seconds
```

(b) The main Viewer Panel

(c) Event snapshots panel

Figure 3.7: Interfaces of the implemented safety application prototype. Figure best viewed in color and when it is zoom-in.

3.7 Conclusion

Vision-based motion trajectory forecasting has the potential to supplement existing proactive safety management systems. An example of these systems was introduced in Teizer *et al.* (2010) [23], where current locations of worker and equipment were being used to prevent struck-by accidents. This paper focus on forecasting workers and equipment's motion trajectory data captured from already available video streams from construction sites. The presented model utilizes an LSTM encoder-decoder structure, multi-head predictions, and embedding of contextual cues for long-term forecasting. To validate the method, a large trajectory dataset, Voyager dataset, is collected and validated and the experimental results prove the benefit of the model design decisions. For 1080p videos, the model forecasts future locations in 10, 20 and 40 frames, corresponding to 0.5, 1.0 and 2.0 seconds in the future, responsively. The final model achieves an average localization error 7.30 pixel to 0.5 seconds, 12.71 pixels to 1.0 second and 24.22 pixels to 2.0 seconds. However, experiment results also suggest forecasting for long distance tracks is more difficult than that of short distance tracks. And reducing data sampling rate is a better option to extend forecast time. The proposed model is also validated and compared in a generic pedestrian path forecasting benchmark TrajNet. As a proof of concept, a prototype is implemented with the presented model for forecasting, documenting and visualization of entering controlled access area events. A simple validation of the prototype shows practical values.

CHAPTER 4: MACHINE LEARNING–BASED RISK ANALYSIS FOR CONSTRUCTION WORKER SAFETY FROM UBIQUITOUS SITE PHOTOS AND VIDEOS[1]

4.1 Introduction

The 2019 National Census of Fatal Occupational Injuries report released by the U.S. Bureau of Labor Statistics (BLS) shows that the private construction industry had 1,061 fatal injuries for the year, up 5% from 2018 and the sector's highest number of worker deaths since 2007 [2]. Exposure to harmful substances or environments caused 167 workers to lose their lives. Another 146 workers lost their lives after making contact with objects or equipment that we deemed unsafe. These statistics are disappointing and makes apparent that more work must be done not only to detect and prevent workers from jobsite hazards, but also improve the total occupational health of the construction workforce.

Having an instant access to safety observations helps project managers identify potential safety issues and allows them to proactively monitor their crew engagements. Thanks to the significant advancement made in deploying imaging technologies such as 360-degree, time-lapse, and drone cameras on construction sites [148, 31], an opportunity has merged to deploy computer vision–based techniques at a scale that matters. Building on the growing availability of these ubiquitous sources of visual data, numerous visual–based safety inspection methods and tools are introduced through academia [24, 5, 25]. Startup companies such as Smartvid.io have also deployed similar capabilities in detecting 50 different safety–related objects (e.g., hardhat, glove, glasses, ladder) from images and videos over 1,000 construction sites. A large number of case studies has already been published that shows the effectiveness of these solution across a variety of residential, commercial and industrial projects. These recent developments fundamentally change the practice of quantitative measurements of

[1] This chapter in whole or in part is accepted by the Journal of Computing in Civil Engineering at the time of writing.

incidents, and safety benchmarks across many projects at the enterprise-level. The wide applicability of these computer vision techniques and the availability of large volume of visual data has built an opportunity for more advanced and frequent safety assessments, particularly when they are compared with traditional labor–intensive weekly or bi-weekly safety reporting methods. Despite the advantages of these solutions in predicting and preventing incidents, a number of scientific problems have yet to be addressed to fully benefit from the potential of the available visual data and current computer vision techniques:

1. Existing computer vision safety inspection tools explicitly detect safety–related objects and their confidence scores, without analyzing (a) worker activities in the context of safety or (2) assessing incident severity under activities, whereas existing construction risk models, such as the Construction Hazard Assessment with Spatial and Temporal Exposure (CHASTE) approach [149, 7], often describe risk as a function of probability and severity of incidents.

2. Many computer vision safety inspection tools count recognized incidents regardless the severity levels of different incident occasions. For example, in a plastering task, wet Portland cement can cause dermal irritation or burns if a worker doesn't wear proper Personal Protective Equipment (PPE), such as safety glasses and gloves. Applying plaster is more likely to cause harm to naked skin than collecting plaster because of the splatter from applications. Understanding this difference requires comprehensive reasoning on a worker's activity, body pose, and interactions with PPE, tools, and material. However, the detection of safety hazards in this rich context of workplace is not fully explored. As such, beyond reporting on the fluctuation of safety incidents and their frequency over project timeline, current computer vision driven methods are unable to provide insights about the severity of the safety incidents.

To enable risk analysis for safety monitoring, modern computer vision techniques such as human–object interaction [30] and object relations [75] are necessary to detect and analyze workers and their interactions with tools, equipment and the work environment. Also safety inspection must cover a wider range of tasks including recognizing unsafe activities, unsafe conditions, and occupational health risks [150, 151, 152, 153], however current computer vision safety inspection methods and tools heavily focus on PPE compliance checking. An underlying technical limitation is that the object detection models introduced in these tools are often incapable of analyzing worker activities to measure risk of exposure to harmful substances or environments, potential fall–related incidents [154], or unsafe body postures [29, 33]. Instead, they primarily focus on pose estimation or activity recognition [155, 156, 157]. There is a need for methods that can jointly detect workers and PPE, as well as their activities, interaction with tools, and work context from site images and videos to allow for modern tools that can minimize the inefficiencies associated with existing isolated safety inspection tools.

Contributions: The main contribution of this study is a new machine learning-based method to predict worker level severity using single workers' visual data captured at a close distance. Experiment results show that a combined worker state, including information of the workers' body pose, the activities being performed, their PPE use, their interactions with tools and material, and the presence of workplace hazards, is the most informative for this prediction task. This study also contributes to the body of knowledge by providing 186,464 bounding box annotations of worker, PPE, tools, and material on a large scale single worker bricklaying and plastering image dataset. In addition to the contributions, this study presents two technical improvements. (a) a joint–task model for recognizing objects, worker activity, and worker body pose from a single image. This model is more efficient and achieves comparable performance than using individual models for each sub-task. (b) a spatio-temporal graph neural network model to refine frame–wise worker activity recognition by taking in per-frame object, activity, and keypoint predictions. These two improvements foreshadow a unification of worker activity recognition, worker PPE compliance checking, and worker ergonomic risk monitoring tasks from videos of workers taken at a close distance. Finally, single worker severity levels are predicted by a trained classifier on a dataset of images of construction workers accompanied with ground truth severity level annotations. The severity level prediction model using the full worker state achieves over 85% cross–validation accuracy on the test data.

The following section reviews the most relevant work on risk analysis models, visual–based unsafe action recognition, safety inspection, and ergonomic risk analysis. The third section introduces the proposed risk analysis model, the breakdown of risk formulation, and the human study procedure. The fourth section introduces the vision model used for joint activity recognition, object detection, and pose estimation. The spatio–temporal model for refining activity recognition is also presented. The fifth section introduces a bricklaying and plastering activity dataset used for model verification. The sixth section presents the details on model specifications and experiment setups. The seventh section analyzes the experiment results. Finally, the findings from this work, the limitations and open research areas are discussed.

4.2 Related Work

Prior research related to this study can be categorized into two buckets: (1) safety risk analysis and (2) computer vision–driven techniques for worker safety inspection, ergonomic assessment, and activity recognition. As the present study covers a broad range of topics and techniques, we offer a list of key features and their code names from previous studies and this work. This list and Table 4.1 summarize previous studies' scopes, objectives, and their main differences with this study. The list of key features in this study are:

- DET: Detect presence of worker or PPE
- HOI: Recognize worker interactions or relations
- 2DWP: Estimate worker pose from images
- 3DWP: Direct reconstruction of 3D worker pose from a single image
- AR: Recognize construction activities
- FAR: Provide frame-level activity labels
- MT: Learn multiple tasks jointly
- SP: Model spatio-temporal dependencies

Table 4.1: A summary to the reviewed literature

Scope	Previous Studies	Objectives	Risk Assmt.	DET	HOI	2DWP	3DWP	AR	FAR	MT	SP
Safety risk analysis models	Hallowell and Gambatese 2010; Rozenfeld *et al.* 2010; Sacks *et al.* 2009	Evaluate accident and process level risks from historical data and observer recordings.	○								
	Paltrinieri *et al.* 2019; Poh *et al.* 2018; Arnes *et al.* 2005	Learning–based methods to predict risks	●								
Computer vision–based safety inspection	Park and Brilakis 2012; Shrestha *et al.* 2015; Park *et al.* 2015; Fang *et al.* 2018b; Nath *et al.* 2020; Wu *et al.* 2019	Recognize personal protective equipment and safety gear non compliant instances by the presence of objects.	◐	●							
	Fang *et al.* 2018a; Xiong *et al.* 2019; Tang *et al.* 2020; Zhang *et al.* 2020b	Recognize personal protective equipment and safety gear non compliant instances by reasoning object relations.	◐	●	●						
Computer vision–based ergonomic risk analysis	Ray and Teizer 2012; Liu *et al.* 2016; MassirisFernández *et al.* 2020	Analyze worker ergonomic risk from a single image, stereo image pairs or image–depth data.	●	●		●					
	Zhang *et al.* 2018; Yu *et al.* 2019; Chu *et al.* 2020	Validate vision–based ergonomic assessment with 3D wearable sensor data; use existing ergonomic assessment systems.	●	●		●	●				
Computer vision based worker activity recognition	Yang *et al.* 2016; Luo *et al.* 2018b; Luo *et al.* 2018a; Ding *et al.* 2018	Recognize worker activity at clip level using stack of frames.	◐	●				●			
	Luo *et al.* 2019a; Cai *et al.* 2019; Roberts *et al.* 2020	Recognize worker activity at frame level; incorporate other visual cues.	◐	●				●	●		

* **DET**: Detect presence of worker or PPE; **HOI**: Recognize worker interactions or relations; **2DWP**: Estimate worker pose from images; **3DWP**: Direct reconstruction of 3D worker pose from a single image; **AR**: Recognize construction activities; **FAR**: Provide frame–level activity labels; **MT**: Learn multiple tasks jointly; **SP**: Model spatio–emporal dependencies.

** A full solid circle (●) means a fully automated process; a half solid circle (◐) marks a user–driven/interactive process with some automation; an empty circle (○) marks a fully user-driven process with no automation.

*** This list of previous studies may be incomplete. Only the most relevant papers are shown.

4.2.1 Safety risk analysis models

A number of prior work on construction risk analysis identify potential area where risk analysis methods can be integrated with safety inspection practices. Hallowell and Gambatese (2010) [158] proposes a safety equilibrium equation, measuring (a) the safety risk as the sum of all risk scores associated with construction activities conducted in a project and (b) the *safety capacity* of a safety program. Here the safety capacity refers to the ability of safety program elements to mitigate a portion of the common safety risks [158]. For a construction activity such as climbing ladder and transporting material, it is associated with a set of frequently observed worker injury causes, such as contact with objects, falls, and overexertion. Risk scores of these causes are summed as the total risk score for that construction activity. Individual risks are measured by the likelihood of injury from the historical records and a pre-determined severity scale. The CHASTE risk analysis method [149, 7] estimates loss-of-control (unwanted, undesired) event risks by the probability of events, expected severity of damage, crew size, and spatial–temporal degree of exposure to hazards during incidents. Real–time risk monitoring [159] estimates the real–time risk as the expected cost of all possible safety states at each time stamp. In practice, the real–time risk formulation can be applied for both online monitoring and offline analysis of time series data. Results aggregated over an event's duration can be used as the the risk score for an activity in Hallowell and Gambatese (2010) [158] and the expected severity level in Sacks *et al.* (2009) [149]. Paltrinieri *et al.* (2019) [160] and Poh *et al.* (2018) [91] validated machine learning approaches to classify risk indexes and categories. Their findings show the practicality of learned classification models for estimating severity from safety observations.

This paper also takes a learning approach to estimate severity levels. However, the proposed model is applied on a single worker level and uses visually recognized information referred to as the worker state as its input features. An approximation heuristic is used to handle a much larger state space than that used by Årnes *et al.* (2005) [159].

4.2.2 Computer vision–based safety inspection

A large body of prior work focuses on detecting PPE (e.g., hard hat, safety vest) using construction site visual data. Early work on this topic uses low-level image features such as edges [161] and histograms of oriented gradients [39]. Classifiers such as the support vector machine and k-nearest neighbors are used on image features [162]. More recently, deep learning–based object detectors have been used to detect hard hats [163, 164, 165, 75, 30, 166], safety vests [164, 30], harnesses [56, 154, 30], and many other protections such as safety glasses [30]. Research has also explored human–object interactions to recognize PPE use in various work conditions including indoors and outdoors [30, 75, 166]. These methods only focus on the recognition task and as such, do not offer any measurements on safety risk. Different from prior work, this paper only use visual–based PPE compliance results as a component of the overall risk analysis by integrating that with an assessment of worker activities, body posture as well as the work context.

4.2.3 Computer vision–based ergonomic risk analysis

Over the last few decades, a number of postural profiling and evaluation systems, such as the Ovako Working Posture Assessment System (OWAS) [167]; the Rapid Upper Limb Assessment (RULA) [168]; the Rapid Entire Body Assessment (REBA) [169]; and the Posture, Activity, Tools, and Handling (PATH) method [170] have been introduced in the literature, however practical deployments and evaluations using these systems are often conducted manually. Ray and Teizer (2012) [43] present one of the earliest methods for automatic worker ergonomic analysis based on visual data. Their method uses Red–Green–Blue–Depth (RGBD) sensors and estimates various body postures of a worker such as raised arms, standing, squatting, bending, and crawling by checking the 2D configurations of predicted worker body keypoint locations. However, the estimated poses are directly used to recognize safety incidents without measuring their severity. More recent works on visual ergonomic assessment feed visually estimated poses and joint angles to the OWAS, REBA and RULA evaluations systems. For instance, MassirisFernández *et al.* (2020) [171] considers the single camera positions to adjust estimated 2D keypoint locations and then estimate

RULA scores. Liu *et al.* (2016) [172] propose a 3D human skeleton extraction method from stereo video camera. Some of the recent worked have reconstructed a set of 3D keypoint locations based on the image-based recognition of the body posture. For instance, Zhang *et al.* (2018) [32], Yu *et al.* (2019) [29], and Chu *et al.* (2020) [33] estimate 2D keypoint locations first and transform the 2D keypoints to 3D coordinates using various pretrained 2D to 3D human pose conversion models. The resulting 3D worker skeletons are used to generate OWAS pose codes or to calculate REBA scores. High inter-observer agreement is achieved between the visual-based REBA scores and REBA scores generated from wearable Inertial Measurement Units (IMU). Their work shows that existing visual-based pose estimation models can be reliability used for localizing workers' 2D and 3D body keypoints and generating pose assessment scores. However, one common limitation of the aforementioned work is that the context of worker activity and workplace is not considered as a part of the ergonomic analysis. Similar to the PATH method –which extends the OWAS approach by also considering workers' activity and tool use– this paper leverages worker pose, activity, PPE, tools, material, and context, simultaneously; however different from PATH, this process is entirely automatic. As a simple illustration of this idea, the OWAS system is used and the pose classification of Zhang *et al.* (2018) [32] is followed. As such, this paper contributes to the body of knowledge in visual-based ergonomic assessment by extending these methods and specifically unifying the recognition with PPE compliance risk analysis as well as activity recognition.

4.2.4 Computer vision–based worker activity recognition

Construction resource activity recognition is a well studied topic [173, 174, 175, 42, 40, 176], a more comprehensive review can be found in Yang *et al.* (2015) [177]. Recent computer vision methods recognize worker activities from video feeds by applying conventional computer vision model architectures and pipelines which rely on hand–selected features. Yang *et al.* (2016) [53] use dense trajectory features to classify worker activities in recorded video clips. Luo *et al.* (2018) [156] apply temporal and spatial streams to classify workers activity in recorded construction site videos. Luo *et al.* (2019) [178] apply RGB image, optical flow, and grey image streams to classify worker activities in a steel reinforcement task. Ding *et al.* (2018) [179] integrate a 2D convolutional neural network and a long short–term memory network to recognize unsafe actions in recorded worker climbing ladder videos. Luo *et al.* (2019) [180] apply a 3D convolution neural network on tracked workers and label activity for each worker tracklet. Luo *et al.* (2019) [181] use a probabilistic graphical model to refine worker activity labels in untrimmed videos. Other more recent works explore worker activity recognition assisted by other high–level visual information. Cai *et al.* (2019) [182] improves group activity recognition by using worker head orientations as attention cues. Roberts *et al.* (2020) [155] simultaneously perform activity recognition and worker pose tracking in untrimmed videos and show that activity recognition can be improved by incorporating recognized pose. This paper builds on the same idea of Roberts *et al.* (2020) [155] to jointly perform detection and activity recognition, but it is different in may ways. First, object features for tools are added along with pose features to assist activity recognition. Second, the present study also performs frame–wise activity recognition, but uses single images instead of stacks of images as input. This allows a unified visual model to simultaneously generate object, activity, and pose predictions, hence much less computation resource is needed. Luo *et al.* (2019) [181] is close to the present study in terms of probabilistic graphical modeling of activity recognition; however, their method does not use pose, tool, equipment, material information as the basis of their recognition.

The present study applies a spatio-temporal graph modeling for activity recognition that is inspired by the structural Recurrent Neural Network (RNN) [183]. The introduced Structural RNN recognizes human activities by modeling humans and objects as nodes in a graph, pairwise human–object and object–object relations at a time step t as spatio-temporal edges, and transition of the same node through time as temporal edges. With the goal of predicting nodes' labels (i.e., human activity), this architecture uses a factor graph formulation to model each node's factor function and each pair-wise edge's factor function. This formulation is tailored to treat worker activity, pose, and interactions with tools and materials at time step t as nodes, and nodes' relations as edges. In the Results and Analysis section, it is shown that this new formulation significantly improves frame-wise activity recognition.

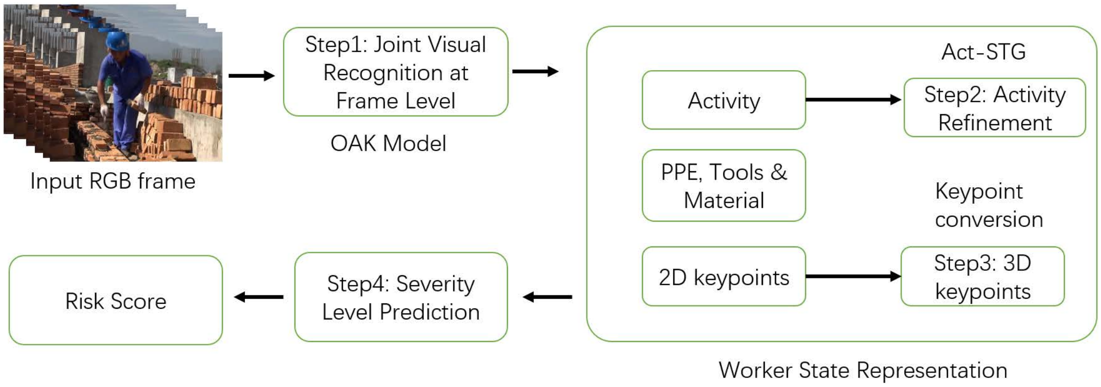

Input RGB frame
Step1: Joint Visual Recognition at Frame Level
OAK Model
Act-STG
Activity
Step2: Activity Refinement
PPE, Tools & Material
Keypoint conversion
2D keypoints
Step3: 3D keypoints
Worker State Representation
Step4: Severity Level Prediction
Risk Score

Figure 4.1: Workflow of the presented risk analysis model. Site image used as input is from the dataset introduced in Yang *et al.* (2016) [53].

4.3 Method

4.3.1 Risk Analysis Model Using Multi-type Visual Information

A new single worker risk analysis model is proposed in this study and the workflow is illustrated in Fig. 4.1. The input data is a stack of consecutive frames of an single worker image taken at a close distance. In step 1, the model first processes every image from the stack and produces initial predictions of worker activity, PPE, tools, material, and 2D body keypoints. In step 2, a spatio–temporal model refines activity recognition from per-frame recognition results. In step 3, the estimated 2D body keypoints are lifted to 3D coordinates using a pretrained keypoint conversion model [184]. In step 4, each single frame's safety severity level per worker is predicted with a trained classifier taking the visual recognized results. With these four steps, worker risk score can be calculated . In this paper, the proposed risk analysis model is validated with single worker outdoor bricklaying and plastering image dataset. To extend to other single worker construction activities, the workflow requires new images taken with similar worker-camera configurations to images used in this study as well as image annotations using the presented format. Models trained in this study can be fine-tuned to other activities to reduce data size requirements.

Risk Analysis Formulation

The risk formulation in Årnes *et al.* (2005) [159] is followed by using a discrete state–space and a function that maps a state to a severity level. The state space describes a single worker's *state* at time t. A worker state is defined as a vector with nine components: (a) three components from the pose codes of worker arms, back, and legs; (b) one component containing worker activity label ; (c) two components for detecting hard hat and gloves as two required worker PPE; (d) two components for tools and material interactions; (e) one component for work context to indicate potential falling to a lower level hazard.

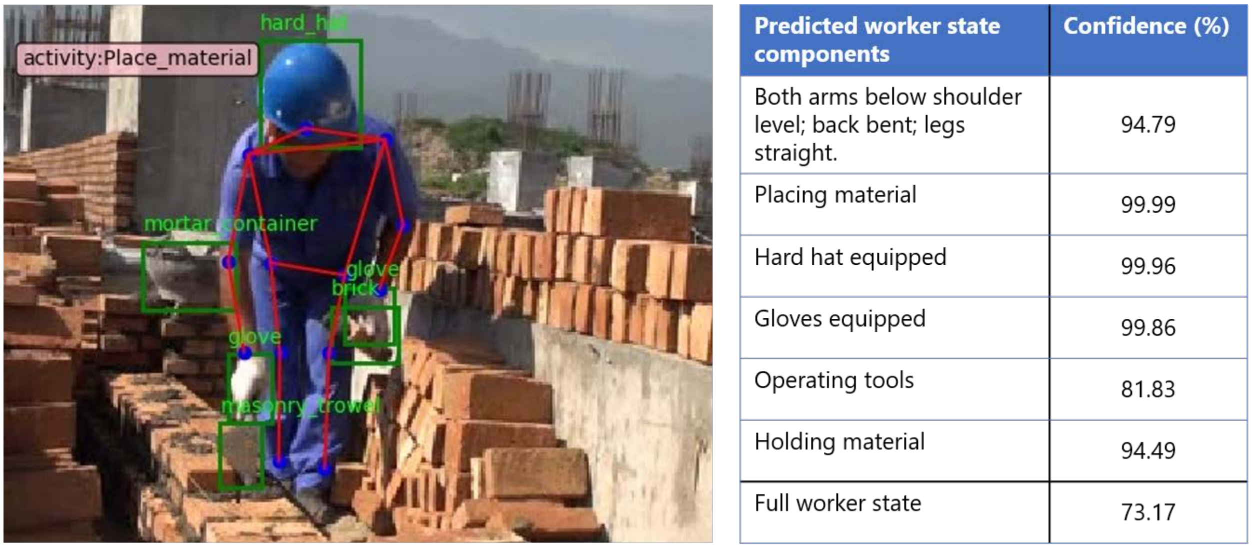
$P(state)$ denotes the probability distribution of worker states when a worker's body pose, and the presence of PPE, tools, and materials are observed in a visual frame. $S(state)$ is the severity level of worker in a state. An assumption is made that components are independent given the input image. This allows a simplification in modeling state probability by products of individual probability instead of a joint probability space. Then $P(state)$ can be evaluated using visual recognition model predictions such that $P(state) = \prod_i P(i)$, and i is a component in the worker state. Thus risk score (R^t) is computed through Eq. 4.1.

$$R^t = \mathbf{E}[P(state)S(state)] = \sum_{j}^{N} S(j) \prod_{i} P(i) \tag{4.1}$$

where j is the jth possible worker state generated by visual models' predictions. Because of the high dimensionality in the worker state space and the high confidence outputs of deep learning models, in practice, the distribution $P(state)$ is peaked at one state. An example of this phenomenon is shown in Fig. 4.2. Since the probability mass is concentrated in a few possible worker states, Eq. 4.1 can be approximated by considering a few top likely worker states, i.e., those who allocate more than 95% probability mass. Computation of a worker state probability is explained in the *Experiment and Implementation Details* section.

Worker Postural Code Using The OWAS System

In this method, a worker's arm, back, and leg poses is classified similar to the OWAS pose profiling [167] and it is later implementations for visual-based ergonomic assessment [32, 28].



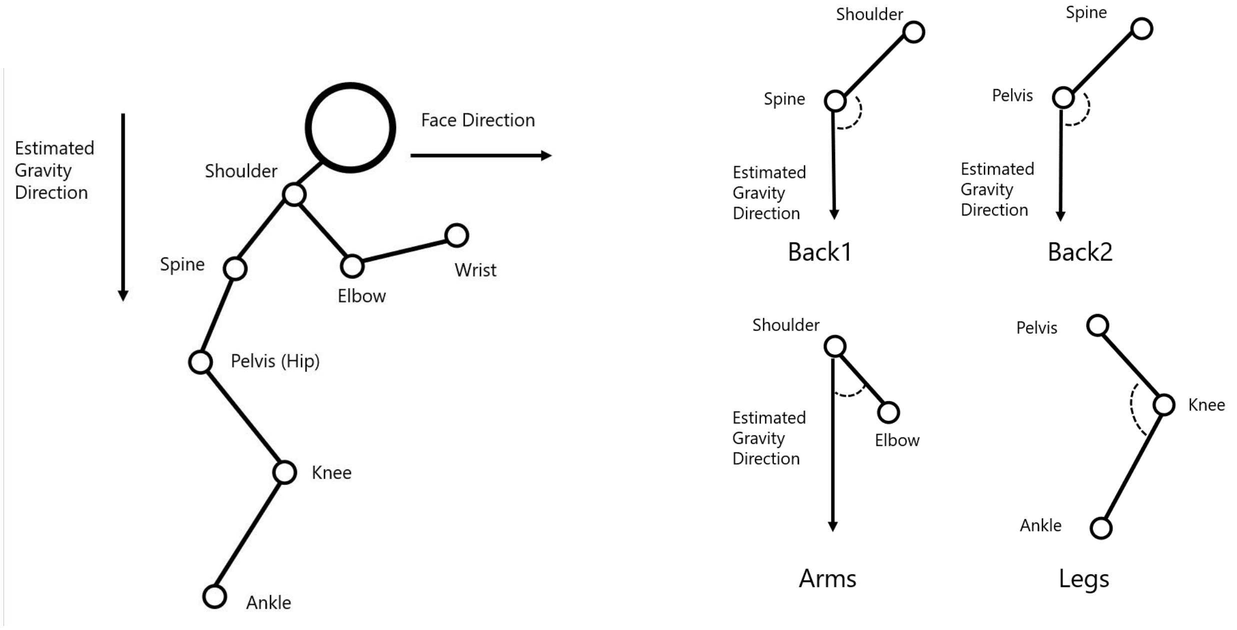
Predicted worker state components	Confidence (%)
Both arms below shoulder level; back bent; legs straight.	94.79
Placing material	99.99
Hard hat equipped	99.96
Gloves equipped	99.86
Operating tools	81.83
Holding material	94.49
Full worker state	73.17

Figure 4.2: An example of the highest confidence worker state predicted by visual models. Left figure shows a visualization of model predictions on objects, activity, and 2D body keypoints that define the mostly likely worker state. Right figure shows prediction confidence scores of the mostly likely worker state's components.

Fig. 4.3 shows the worker's body keypoints arrangement and pose classification angles. Given the estimated gravity direction, poses are classified based on the angles between skeleton (vectors that connect keypoints) and the gravity direction (see Back1, Back2, Arms, Legs angles in Fig. 4.3b). Pose code criteria are listed in Table 4.2, following recommendations in Zhang *et al.* (2018) [32]. Fig. 4.4 shows two examples of arms poses. More details on the estimated gravity direction is presented in the Experiment and Implementation Details section.

Machine Learning–Based Severity Level Classification

The worker–level severity $S(state)$ of a worker state is estimated with a classification model trained on ground truth human ratings. Participants rate one of the four severity levels: *Negligible*, *Low*, *Medium*, and *High*. These severity levels (Table 4.3) are defined similar to OWAS's action categories. A logistic regression model is fitted using worker state features and human severity level ratings. The severity levels can be converted to a normalized severity weight from 1 to 100 for calculating the final risk score, similar to the method proposed in Rozenfeld *et al.* (2010) [7]. Note that the absolute values of severity weights



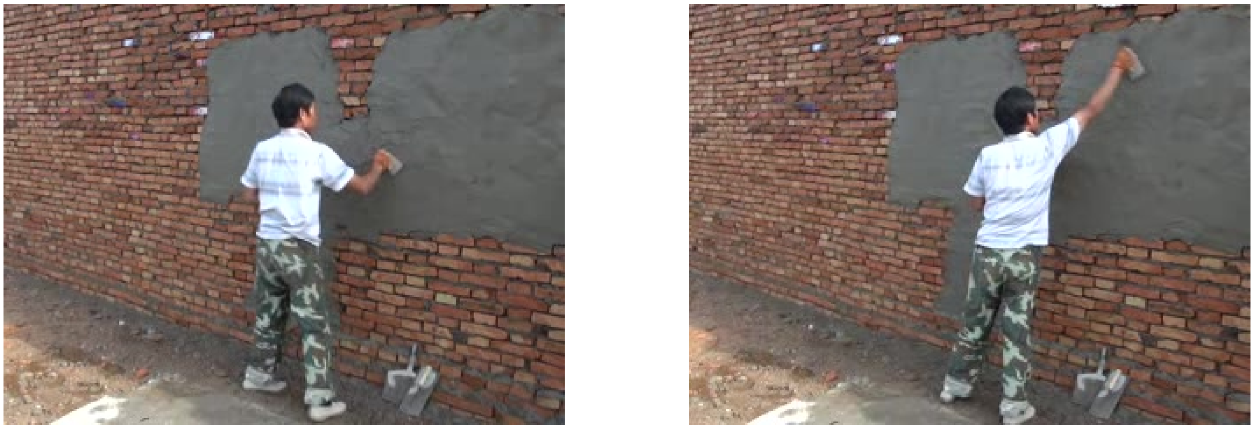
(a) Body keypoint arrangement

(b) pose classification angles

Figure 4.3: Worker body keypoints used for pose classification. The right figure shows four types of angles used to classify pose: the top two types of angles for back, lower left for arms, and lower right for legs.

Table 4.2: Pose classification using the four types of angles

Body Part	Positions	Angle Range
Arms	A:Both arms on or below shoulder level	Both Arms angles $\leq 90°$
	B: One arm on or below shoulder level	One Arms angles $> 90°$
	C: Both arms above shoulder level	Both Arms angles $> 90°$
Back	A: Back straight	Both Back1 and Back2 angles $\geq 160°$
	B: Back bent	Back1 or Back2 angle $\in [120°, 160°)$
	C: Back bent heavily	Back1 or Back2 angle $\in [0°, 120°)$
Legs	A: Standing straight	Both Legs angles $\geq 160°$
	B: Knee bent	At least one Legs angle $\in [60°, 160°)$
	C: Squatting	Both Legs angles $< 60°$



(a) Applying plaster with both arms on or below shoulder level

(b) Applying plaster with one arm above shoulder level

Figure 4.4: Snapshots of arms positions during plastering activity. Image from dataset introduced in [53].

here do not represent any probability or likelihood, they are intended to represent relative importance for reporting instances. Severity weights are interactive parameters that can be adjusted by safety personnel adapting to particular projects and jobsites. It is recommended to adjust the relative values between each severity weights based on the degree of harm and the magnitude of the potential loss. More details in obtaining ground truth severity level annotation is shown in the “ Experiment Setup and Dataset” section.

4.3.2 Visual Recognition Models

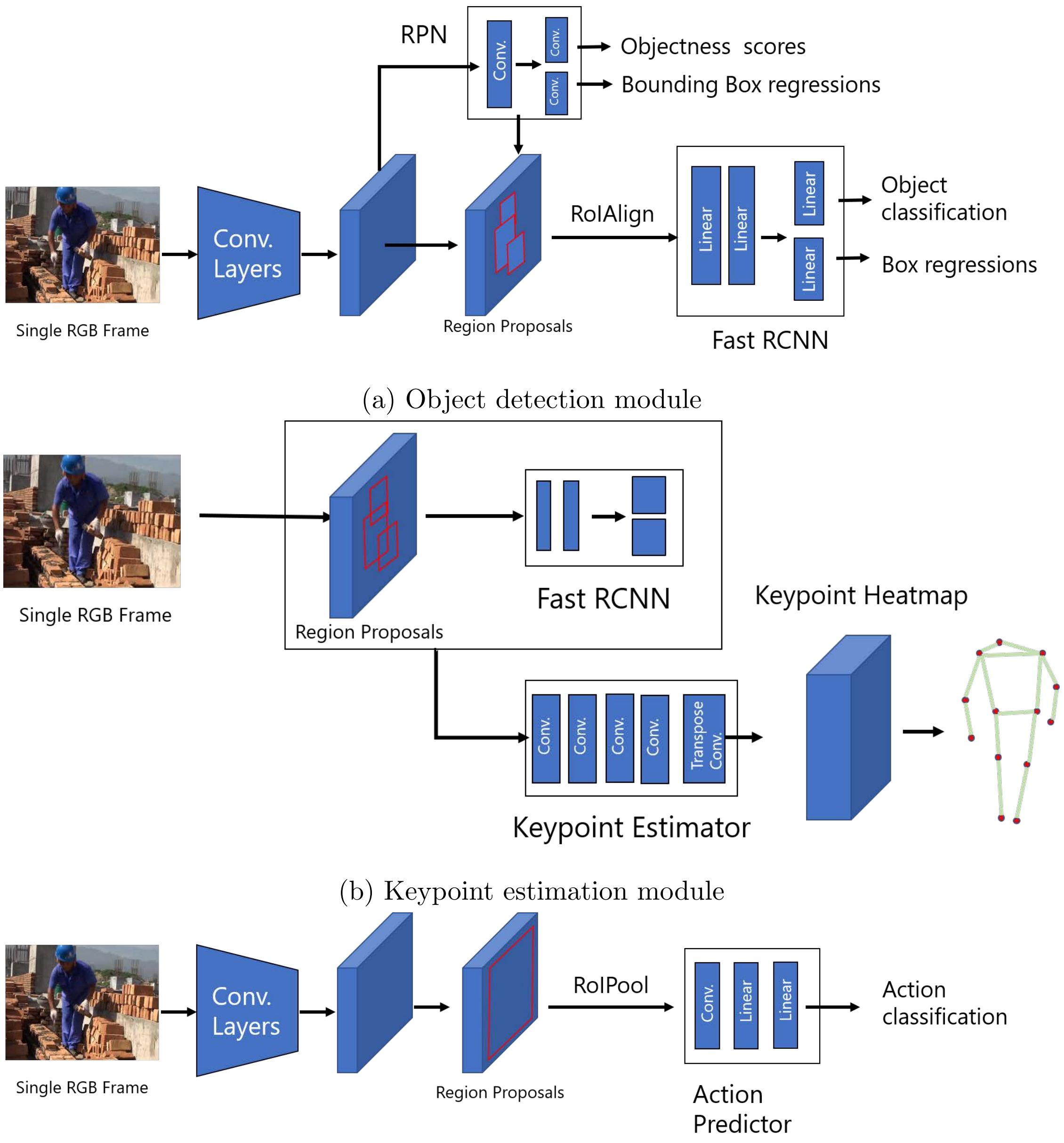
Two types of model are developed to automatically recognize workers’ activities, poses, PPE use, tool and material interactions from video frames. The first model is a modified version of Mask RCNN [46], which in this paper it is referred as the Object–Activity–Keypoint (OAK) RCNN model. This model performs the following tasks simultaneously: detecting worker, material, PPE, and tools; recognizing per–frame worker activity; estimating a worker’s 2D body keypoints’ locations. The second is a spatio–temporal graph neural network model which refines per–frame activity recognition results by connecting recognition results across frames. This model is referred to as Act-STG.

Table 4.3: Severity level definitions

Severity Level	**Severity Weight**	**Definition**
Negligible	1	Activity and workplace context do not cause harm to worker's head, arms, hands, back, or legs. No action required.
Low	5	Activity and workplace context may cause slight but no apparent harm to worker's head, arms, hands, back, or legs. Worker must be checked in the near future.
Medium	25	One of head, arms, hands, back, and legs is subject to acute or chronic harm caused by worker's activity and workplace context. Corrective actions are required as soon as possible.
High	100	Two or more of head, arms, hands, back, and legs are subject to acute or chronic harm caused by worker's activity and workplace context. Immediate corrections are required.

Object-Activity-Keypoint (OAK) RCNN Model

The model's network architecture for per-image object detection, activity classification, and keypoint estimation, is shown in Fig. 4.5. The OAK model is based on a generalized RCNN framework and consists of four parts: the backbone feature maps, an object detector, a keypoint estimator, and an activity predictor. For backbone feature maps, a pretrained Reset50 + FPN [80] convolutional neural network is used as model initial weights; the output image feature (*layer4* feature) is a 3D tensor whose depth is 256 and height and width are proportional to image height and width. The object detector is a Fast RCNN model with region of interest pooling (RoIAlign) layer [46]; the pooled regional features are used to predict object bounding boxes' locations and classes (Fig.4.5a). The activity recognition head pools the regional features from the whole convolutional feature maps (Fig.4.5c). The keypoint estimator is a variation of Mask RCNN with a keypoint estimation head (Fig.4.5b). The keypoint estimation head takes a detected worker bounding box and pools its regional features again. This regional features are fed to a mini-network consisting of four convolutional layers and one transpose convolutional layer. The keypoint head output is a 4D tensor of size $1 \times C \times W \times H$, with C as the number of keypoint classes and W and H being the regional feature maps' height and width; in the present study they are both 56. For each channel in C, the location (w, h) with the highest confidence indicates that keypoint's location in the feature map. A final transformation is used to convert keypoint locations from feature map coordinates to image coordinates. The 3D keypoints' locations are generated using an off-the-shelf keypoint conversion model [184] which can take 13 2D body keypoints to 17 3D body keypoints. Note that the three heads reuse the backbone convolutional features in both training and testing phases. This treatment significantly reduces model size and inference time compared to training three separate models, while keeping a comparative performance. More details are presented in the Results and Analysis section.



(a) Object detection module

(b) Keypoint estimation module

(c) Activity classification module

Figure 4.5: OAK model architecture. All modules share the same input image and backbone convolutional feature map; each task is performed by a specific head. Model input images from dataset introduced in Yang *et al.* (2016) [53].

Spatio-Temporal Graph Model for Activity Recognition (Act-STG)

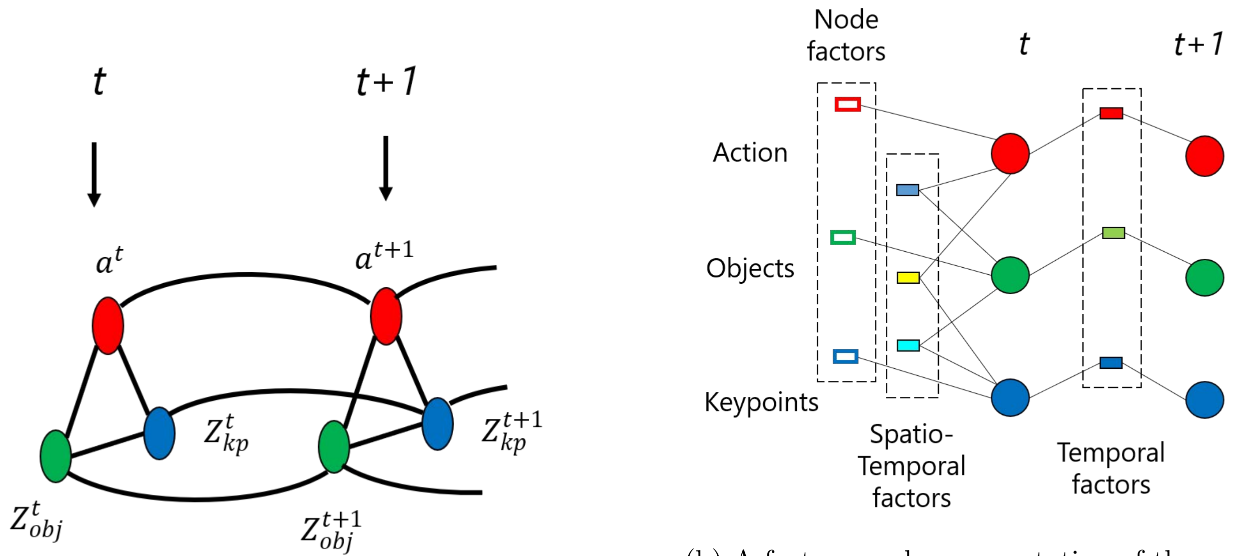
A spatio-temporal graph $G = (V, E_S, E_T)$ is composed of a collection of random variables V that are associated with nodes, a collection of spatio-temporal edges E_S between a pair of nodes at time t, and a collection of temporal edges E_T connecting the same node from time t to time $t+1$. Such a graph models a probability distribution of nodes and is used to predict node labels or real value vectors y at time t [183]. This formulation is considered in the present study to refine activity recognition (Fig. 4.6) with per–frame activity, worker body keypoints, and detected tools and materials as nodes in the graph. Jain *et al.* (2016) [183] apply a factorization of this graph to simplify marginal probability computation with a factor function $\Psi(y_v, x_v)$ for each node, and a factor function $\Psi(y_e, x_e)$ for each edge; x_v and x_e are feature representations for each node and edge, respectively.

As shown in Fig. 4.6a, to construct a model that has a scalable representation of an arbitrary size graph, a pseudo node is introduced representing the summarized features for the detected tools and materials (Eq. 4.2), where f_{MLP} is a single feedforward network to map bounding box regional features to graph node features, and Z_{obj} is the final object feature averaged from all objects. A similar pseudo node is introduced for all estimated worker body 2D keypoints (Eq. 4.3) to generate a summarized keypoint feature Z_{kp} for workers at each frame. Vector a in Fig. 4.6a is the pooled regional activity feature from the OAK model. When Fig. 4.6 is viewed in color, red color refers to activity nodes and factors, green color refers to object nodes and factors, and blue color refers to keypoint nodes and factors. Note that to reduce overfitting, predicted activity scores from OAK model are not used in Act-STG as a feature.

$$Z_{obj} = \frac{\sum n_i}{N}; n_i = f_{MLP}(obj_i) \tag{4.2}$$

$$Z_{kp} = \frac{\sum k_i}{K}; k_i = f_{MLP}(kp_i) \tag{4.3}$$

Similar to structural-RNN [183], a factor graph (see Fig. 4.6b) is used to formulate the spatio-temporal graph (Fig. 4.6a). Nine factors are defined in total: three node factors



(a) Proposed spatio-temporal graph.

(b) A factor graph representation of the proposed spatio-temporal graph.

Figure 4.6: Structure RNN model for spatio-temporal modeling.

for activity (f_{act}), objects (f_{obj}), and keypoints (f_{kp}), three spatial edge factors connecting activity and objects (r_{ao}), activity and keypoints (r_{ak}), and objects and keypoints (r_{ok}), and finally three temporal edge factors that model time propagation of activity (r_a), objects (r_o), and keypoints (r_k). While the spatio-temporal edge factors capture the relations of recognized activity, detected objects, and estimated keypoint locations between frames, the node factors refine per-frame activity using the spatio-temporal features. All factors in the present study are implemented with Gated Recurrent Unit (GRU) [185]. Like many graph neural network models, the temporal update rules from time t to time $t+1$ for Act-STG are defined as in the following:

$$\begin{aligned} e^{t+1}_{ao} &= r_{ao}\left(s^t_a + s^t_o\right) \\ e^{t+1}_{ak} &= r_{ak}\left(s^t_a + s^t_k\right) \\ e^{t+1}_{ok} &= r_{ok}\left(s^t_o + s^t_k\right) \end{aligned} \tag{4.4}$$

And the spatial update rules are:

$$
\begin{aligned}
e_k^{t+1} &= r_k\left(Z_{kp}^{k+1} + e_{ak}^{t+1} + e_{ok}^{t+1}\right) \\
e_o^{t+1} &= r_o\left(Z_{obj}^{k+1} + e_{ok}^{t+1} + e_{ao}^{t+1}\right) \\
e_a^{t+1} &= r_a\left(a^{k+1} + e_{ao}^{t+1} + e_{ak}^{t+1}\right)
\end{aligned}
\tag{4.5}
$$

where e is the output and s is the hidden state of each factor. The final activity prediction F is a summation of all node factor predictions:

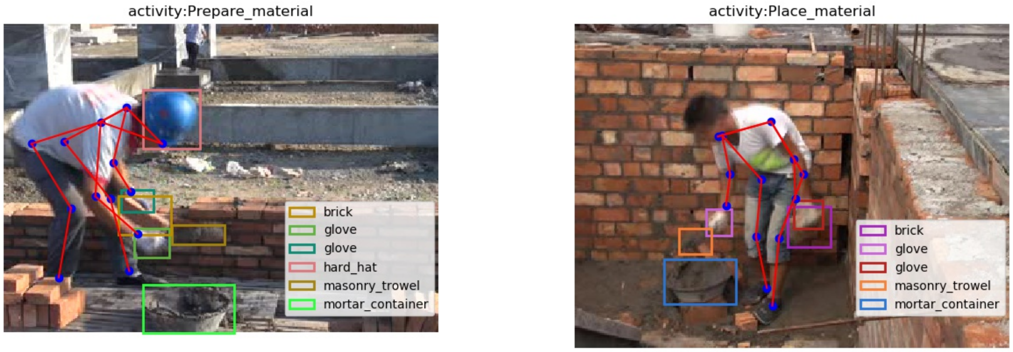
$$
F^{t+1} = f_{act}\left(e_a^{t+1}\right) + f_{obj}\left(e_o^{t+1}\right) + f_{kp}\left(e_k^{t+1}\right) \tag{4.6}
$$

4.4 Experimental Setup and Dataset

The image data used in the present study is originated from a collection of video clips first introduced by Yang *et al.* (2016) [53]. The original video resolution is 240p. The majority of the original video collection captured single worker outdoor construction activities and was taken at a close distance to workers. Yang *et al.* (2016) [53] has provided video–level activity labels, such as bricklaying and plastering. Roberts *et al.* (2020) [155] have augmented a subset of bricklaying and plastering videos and have added frame–wise activity labels and single–worker body keypoints. This work extends the annotations in Roberts *et al.* (2020) [155] by adding per-frame object bounding box annotations of PPE, tools, and material and video level workplace context annotation. Bounding box annotations of PPE, tools and materials were outsourced to Alibaba AI Lab and their quality were thoroughly and rigorously reviewed by construction experts to correct missing and mis-labeled objects.

4.4.1 Dataset Statistics and Characteristics

Images in the presented dataset contains a high inter and intra–class variance in terms of worker appearance, workplace background, and camera viewing angles. In the bricklaying set, 18 workers are captured in 179 videos. In the plastering set, 9 workers are captured in 122 videos. Different workers have the distinguishable appearance, motion, and surrounding


activity:Prepare_material

activity:Place_material

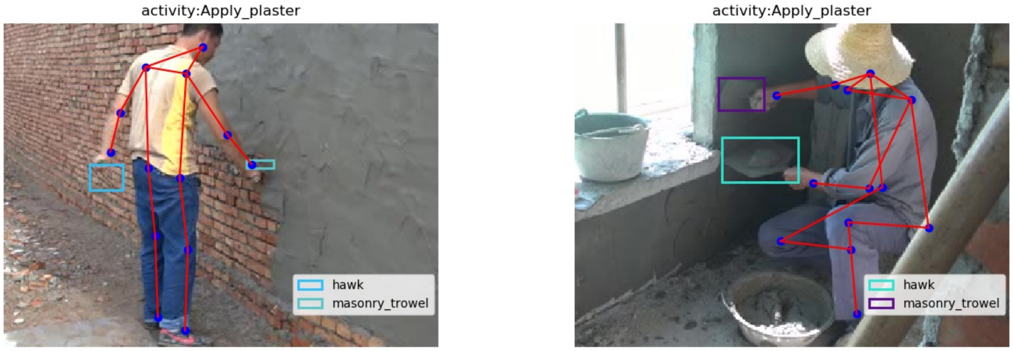
Figure 4.7: Annotation examples of bricklaying task. Each frame is annotated with an activity label, a set of body keypoints connected by line segments, and a set of object annotations on PPE and tools. Image from dataset introduced in Yang *et al.* (2016) [53].

workplace context. Each worker is recorded from more than one camera viewing angles and distances, representing highly diverse motion patterns in images. Images used in this study have a realistic representation of real–world bricklaying and plastering data complexity. Building on previous work, the image dataset presented in this study provides abundant annotations for single worker activity, worker body pose, and PPE, tools, material object bounding boxes, and potential workplace hazard as context labels. Context labels indicate the potential hazard of workers falling to a lower level. In total, this image dataset contains 36,969 worker activity labels, 186,464 object bounding boxes, 480,597 worker body keypoints, and 69 video-level context labels. Table 4.4 shows the definitions of activity, PPE, tools, and material for bricklaying and plastering sets. 13 body keypoints are annotated: *nose*, *shoulders*, *elbows*, *wrists*, *hips*, *knees*, and *ankles*. Qualitative annotation examples are shown in Fig. 4.7 and Fig. 4.8 for bricklaying and plastering, respectively. A detailed annotation breakdown by training and testing sets is presented in the following section.

The majority of images in this dataset focus on outdoor construction activities in commercial projects. However, there are two indoor construction workers in the plastering set, e.g. the right worker in Fig. 4.8. The camera viewpoint constraint due to small and clutter indoor space can be critical to apply the proposed method on indoor construction images because annotating and training on OAKs and severity levels are more challenging when a worker's frontal and side view is not available.

Table 4.4: Activities and objects for bricklaying and plastering.

Task	Activity	Description	Objects
Bricklaying	A: Prepare material	Collect mortar or brick, put mortar on brick, or break brick	Hard hat, gloves, masonry trowel, mortar container, brick
	B: Place material	Place brick or spread mortar on wall	
	C: Consolidate placement	Tap brick into place or remove excess mortar	
Plastering	A: Collect plaster	Collect plaster with hawk or hand board	Hard hat, gloves, masonry trowel, mortar container, hawk
	B: Transfer plaster	Transfer plaster between hawk or hand board and trowel	
	C: Apply plaster	Deposit and spread plaster on the wall	


activity:Apply_plaster

activity:Apply_plaster

Figure 4.8: Annotation examples of plastering task. Each frame is annotated with an activity label, a set of body keypoints connected by line segments, and a set of object annotations on PPE and tools. Image from dataset introduced in Yang *et al.* (2016) [53].

Table 4.5: Annotation numbers in training and testing sets

	Split	Frame–Level Activity Labels	Frame–Level Object Labels	Frame–Level Keypoint Labels	Video–Level Hazards Labels
Bricklaying	train	12,146	68,370	157,898	32
	test	9,327	53,806	121,251	20
Plastering	train	8,829	37,303	114,777	12
	test	6,667	26,985	86,671	5

4.4.2 Data Split

To validate vision models and the proposed risk analysis method, the dataset is split by workers. For each worker conducting that task, (a) when the number of videos for this worker is even, training and testing videos are randomly split by half; (b) when the number of videos is odd, extra videos are placed into training set. Finally, the bricklaying task has 101 videos in the training set, and 78 videos in the testing set. The plastering task has 67 videos in the training set, and 55 videos in the testing set. The dataset is split in such way to avoid almost identical images in training set "bleed" in to testing set while diversifying testing data. Vision model experiments and risk analysis human calibrations are conducted on the testing set videos. Table 4.5 shows a detailed annotation breakdown for traning and testing sets.

4.4.3 Worker Severity Level Annotation

Only testing images of bricklaying and plastering sets are used to obtain ground truth annotations of worker severity level. Since labeling severity level requires substantial safety knowledge and can be subjective to annotators' own experience, a two-stage annotation is applied to obtain consistent severity level labels. In the first stage, a number of safety experts are invited for an online survey to rate severity levels on 10 randomly selected images. Instruction and severity level definitions are provided in the online survey. All annotators

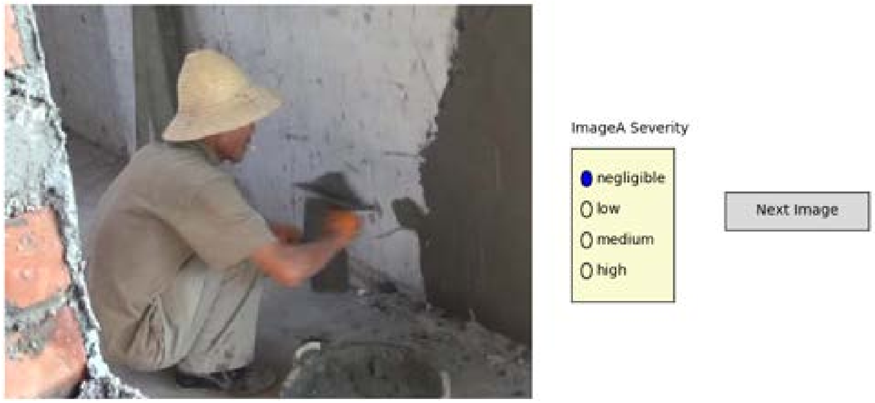


Figure 4.9: The interface used for the second stage severity level annotation. Image from dataset introduced in Yang *et al.* (2016) [53]

Table 4.6: Numbers of ground truth worker severity level annotations

Task	Negligible	Low	Medium	High
Bricklaying	33	88	58	10
Plastering	45	42	96	34

are anonymous and encouraged to disclose their years of experience in construction safety. A multi-participant inter-observer agreement test [186] is conducted to find annotators in close agreements. These annotators are later invited for the second stage annotation where a randomly selected image is presented in an interface (Fig. 4.9). Annotators are asked to select a severity level for the depicted worker and click the “Next Image” button to save their annotations and proceed to the next image. Each image will be annotated at most once. Annotation ends when all test images are annotated or participants close the interface. All severity level annotations collected in the second stage are used as final ground truth severity labels. In the first stage annotation, 7 participants achieved a 0.64 kappa score. This is close to the 0.7 kappa score of safety index prediction from Poh *et al.* (2018) [91]. 4 participants have over 5 years of experience in construction safety and 3 participants have 1 to 3 years of experience in construction safety. In the second stage, the selected participants annotated 189 bricklaying images and 217 plastering images (see severity level examples in Fig. 4.10). Ground truth severity level annotations statistics are shown in Table 5.

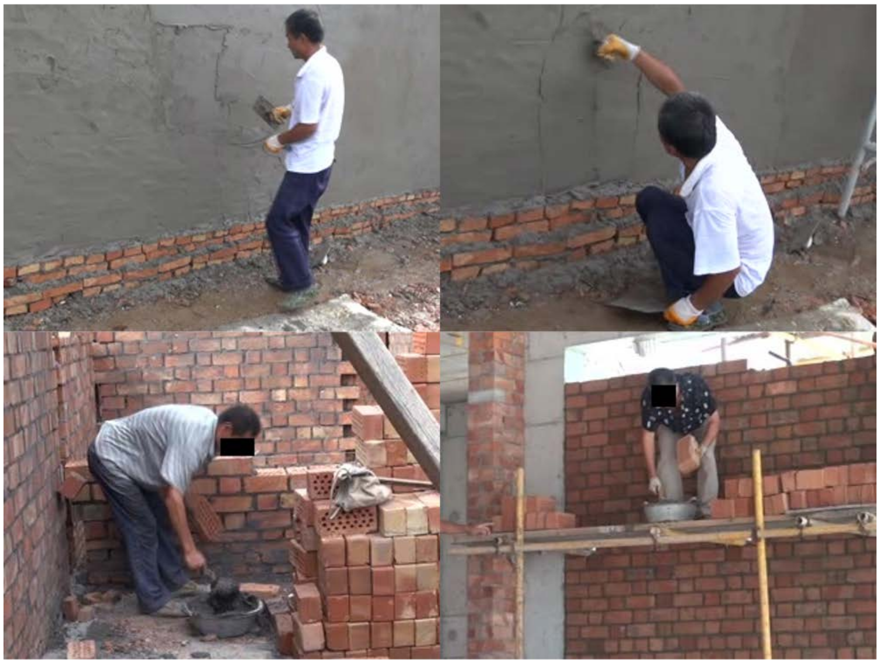


Figure 4.10: Examples of severity level annotations. Upper-left: negligible; Upper-right: low; Bottom-left: medium; Bottom-right: high. Image from dataset introduced in Yang *et al.* (2016) [53].

4.5 Experiment and Implementation Details

4.5.1 Experiment Procedure

For each task, an OAK model is first trained, then an Act-STG model, and apply the trained models for the testing set. Predicted 2D keypoint locations are fed to the 2D-3D keypoint conversion model to generate 3D keypoint locations. Finally, predicted activities from Act-STG, detected PPE, tools, and material from OAK, estimated 3D keypoints, and ground truth work context condition are combined as worker states. A subset of testing images for both tasks is evaluated by human, estimated worker states for these images are used to fit logistic regression models. All training and testing is run using a system with an Intel i7 8700K processor, a Nvidia RTX 2080Ti graphics card, and 32 Gb memory.

4.5.2 Evaluation Metrics

Due to being computationally intensive, the inference speed, memory consumption, and the performance of the OAK model are evaluated first. Per-frame activity prediction of Act-STG and and severity level classification are evaluated as a high–performing visual recognition model is essential for the severity classification. The model performance metrics are defined as in the following.

A correctly recognized instance is called a True Positive (TP). An falsely recognized instance is called a False Positives (FP). A missed ground truth instance is called a False Negative (FN). Precision is defined as the ratio between TP and TP+FP, and recall is defined as the ratio between TP and TP+FN. By adjusting the threshold for prediction confidence, a precision-recall curve of a class is generated. The Average Precision (AP) is defined as the area under the precision-recall curve, mean Average Precision (mAP) is defined as the average of AP for all classes. The mAP metric evaluates the overall model performance across all possible prediction score thresholds; it is a more robust metric focus on the general model performance [67, 45, 46].

For activity recognition, mAP is reported using activity scores of all classes and the ground truth activity class label each frame. Precision and recall averaged over activity classes are reported, in this case a frame wise activity prediction is the class having the highest activity score. mAP for object detection and keypoint estimation should be reported considering localization errors. The Intersection over Union (IoU) is defined as the ratio between intersection area and union area between a detected bounding box and a ground truth bounding box. In the present study, a 0.5 IoU threshold is used to determine whether a ground truth object instances is matched to a detected object instance. So object detection is evaluated by mAP@IoU=0.5. For keypoint estimation, the Object Keypoint Similarity (OKS) [67] is defined in Eq. 4.7.

$$OKS = \frac{1}{N} \sum_{i} \exp\left(\frac{-d_i^2}{2s^2k_i^2}\right) \tag{4.7}$$

where i is the ith keypoint, N is the total number of keypoints, d_i is the Euclidean distance between a ground truth keypoint and its corresponding detected keypoint, s is the person bounding box area in pixels, and k_i are scaling factors for each type of keypoint: 0.026, 0.079, 0.072, 0.062, 0.107, 0.087, and 0.089 for the nose, shoulders, elbows, wrists, hips, knees, and ankles, respectively [67]. In the present study, a worker's pose is correctly detected if its OAK is greater or equal to 0.5; hence the pose evaluation metric is mAP@OAK=0.5.

Regressing severity level is evaluated by the 5-fold cross–validation accuracy. This accuracy is computed by averaging accuracy of each fold while treating the rest 4 folds as training data. Standard error on cross–validation accuracy is used as a measurement of deviation.

4.5.3 Computing Worker State Probability

Worker state probability is computed by the product of components probabilities. For example, the probability of hard hat use is the detected hard hat box's probability p_h if such box's exterior is within 50 pixels of the nose's 2D location; otherwise it is $(1 - p_h)$. The same 50-pixel criterion is applied for gloves, tools, and materials. Note this criterion is a hyperparameter that can be selected by users. Using 50 pixels works well in the presented dataset as all images only contain a single worker and have dimensions of 240-by-320 pixels. For crowded multiple worker images, a learned interaction recognition model is recommended [30]. The probability of work context is zero or one because it is treated as given information apart from the visual model outputs. The probabilities of arms, back, and legs are computed based on their corresponding keypoints' prediction scores. For example, the probability of arms' pose $P(arms)$ is the product of left and right elbows' probability and left and right shoulders' probability. An estimated keypoint's probability is computed as the summed probability over a rectangle box centered at the estimated location, because any keypoint estimation within a certain distance of its ground truth location is not penalized in the OKS keypoint evaluation. The size of the rectangle box is computed by using 10% of the worker box's height and width.

4.5.4 Implementation Details

As previously mentioned, the OAK model is modified based on a Pytorch vision package implementation of Mask RCNN. The joint training of object detection and keypoint estimation is achieved by masking out keypoint targets for non-worker objects. The training process is improved by initiating the backbone feature weights with a pre-trained Microsoft Common Object in Context (MS-COCO) [67] person keypoint estimator. For both tasks, OAK models are trained with Stochastic Gradient Descent (SGD) optimizer with 0.9 momentum. Learning rate is set as 0.001 with 0.1 decay at every five epochs, and weight decay is set as 0.0005 for each epoch. 9 epochs are trained to obtain the best models. Act-STG model is also implemented using Pytorch. For each task, Act-STG model is also trained with SGD optimizer and 0.9 momentum. The learning rate is set as 0.00001. Gradient clipping is set as 50% of the gradient norm to achieve a stable training process. An off-the-shelf 2D-3D person keypoint conversion model from the official code repository of Pavllo *et al.* (2019) [184] is used; the gravity direction is assumed to be $v = (0, 0, -1)$. Two qualitative examples of keypoints conversion are shown in Fig. 4.11.

4.6 Results and Analysis

4.6.1 OAK Model Analysis

First, all models are profiled using the official Pytorch vision metric logger tool. At inference time, OAK model runs at 26.6 frames per second and takes 887 MB memory. Compared to a system running individual models for each sub-task (Bundled Model in Table 4.7), OAK model takes 36.6% less time and 40.1% less memory. Act-STG and Keypoint conversion models add negligible inference overhead. These results show that the proposed workflow has the potential to process images at near real–time speed.

OAK model's improvements in inference speed and space usage are at small or even no cost of model performance. Table 4.8 shows that the OAK model achieves 83.2% mAP at 0.5 OKS for keypoint estimation on the bricklaying task, losing 0.5% mAP to the Keypoint

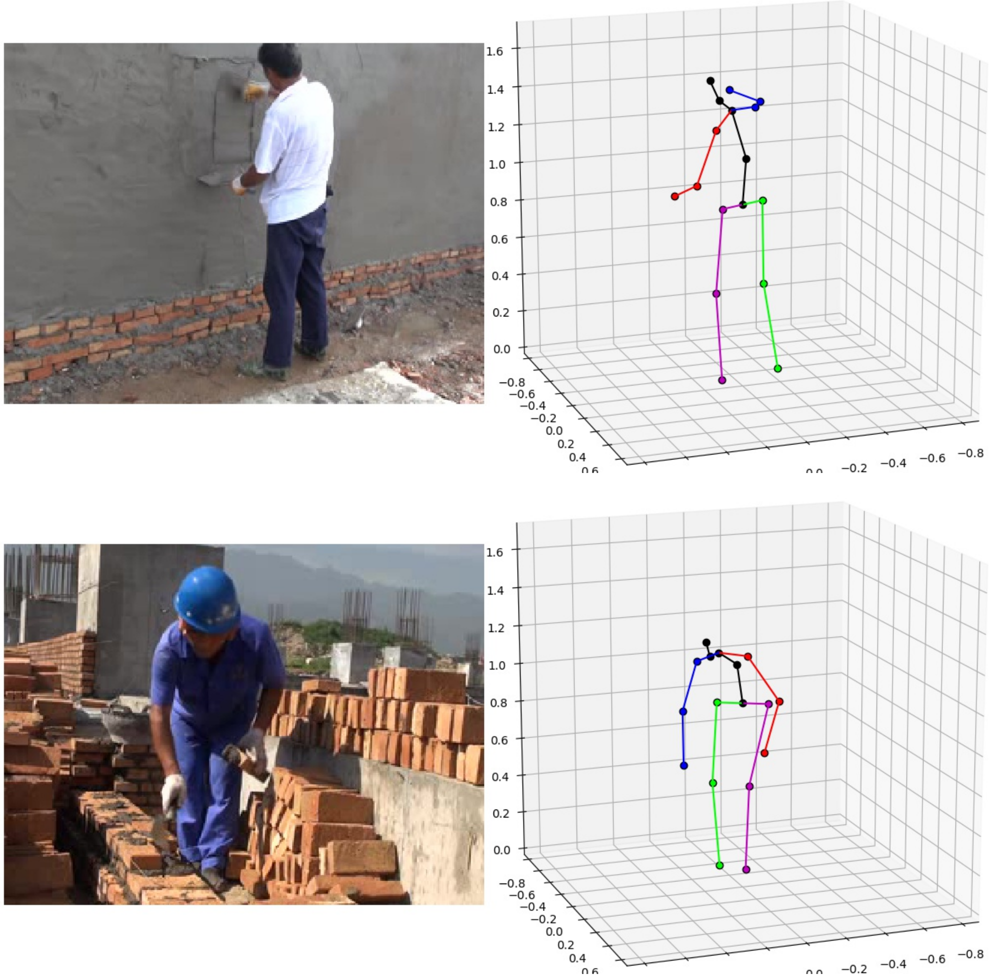


Figure 4.11: Qualitative examples of 2D keypoint conversion to 3D keypoint. Image data from Yang *et al.* (2016) [53].

RCNN model. The OAK model achieves 88.8% mAP at 0.5 OKS for the plastering task, outperforming the Keypoint RCNN model by 1.6% mAP. Even though the OAK model adds one parallel task for activity and detects multiple non-worker object classes, it still achieves comparable keypoint estimation capability to the original Keypoint RCNN on this construction worker dataset. Table 4.9 shows a similar trend for object detection performances. The OAK model's object detection performance is very close to that of the Faster RCNN model. The OAK model achieves 78.7% mAP at 0.5 IoU for the bricklaying task and 69.6% mAP for the plastering task, while the Faster RCNN model achieves 78.1% mAP and 70.6% mAP, respectively. Surprisingly, the OAK model gains small improvements on activity recognition over the activity model. In Table 4.10, the OAK model reaches 62.5% mAP for the bricklaying task, outperforming activity model by 0.7% mAP, it also reaches 75.3% mAP for plastering task, outperforming the Activity model by 0.5% mAP. Precision and recall results between the Activity model and the OAK model are also comparable, although OAK model shows slight advantage in precision or recall. These results show that the proposed strategy of merging low-level image features from different vision tasks into a common set of convolutional feature maps does not significantly reduce model performance. In other words, the three branches in OAK model share a common set of convolutional backbone features, so the low–level image features are informed by three different tasks. However, the high-level task features in their individual classifiers are not aware of how other tasks performs, so their "communication" is indirectly through the low–level convolutional features and backward gradients. This implies two things: the image feature is generalized by three tasks; the convolutional network is over-parameterized such that there are enough space to encode information for three different tasks.

4.6.2 Act-STG Model Analysis

Act-STG model fuses single–frame visual information (i.e., objects, activity, and pose) and refine activity recognition from sequential data. Comparing Act-STG with the best single-frame models in Table 4.10, Act-STG model significantly improves activity recognition performance. For the plastering task, 7.6% increment in mAP, 6.2% increment in precision,

Table 4.7: Model consumed resources

Model	Averaged inference speed (ms)	Model GPU usage (MB)
Activity model	12.3	438
Faster RCNN	22.0	453
Keypoint RCNN	25.0	590
Bundled models	59.3	1481
OAK model	37.6	887

Table 4.8: Keypoint estimation evaluation results

Model	Keypoint mAP(oks=0.5) (%)
Bricklaying	
Keypoint RCNN	83.7
OAK model	83.2
Plastering	
Keypoint RCNN	87.2
OAK model	88.8

and 6.9% increment in recall are gained. For the bricklaying task, mAP, precision, and recall performances are raised by 2.3%, 5.5%, and 2.8%, respectively. These results illustrate the benefits of capturing high-level spatio-temporal relations between different vision information for activity recognition. Act-RNN removes objects and keypoints nodes in ACT-STG, it serves as a baseline that simply models temporal activity. Comparing Act-STG and Act-RNN, it is clear that modeling spatio-temporal relations from objects and keypoints contributes to better activity recognition.

4.6.3 Severity Level Prediction Analysis

A generalized linear model framework is used with two constraints: (1) a ridge regression penalty term; (2) weights are constrained to be non-negative. Table 4.11 shows the severity level prediction results, the best model at the last row uses the full worker state as features.

Table 4.9: Object detection evaluation results

Model	**Object mAP(IoU=.5)**
Bricklaying	
Faster RCNN	77.5
OAK model	78.7
Plastering	
Faster RCNN	70.6
OAK model	69.6

Table 4.10: Activity recognition evaluation results

Model	**mAP (%)**	**Precision (%)**	**Recall (%)**
Bricklaying			
Activity model	61.8	60.6	56.4
OAK model	62.5	60.6	59.3
Act-RNN	62.6	61.5	60.3
Act-STG	**70.1**	**66.8**	**66.2**
Plastering			
Activity model	74.8	65.7	72.0
OAK model	75.3	67.4	71.7
Act-RNN	76.1	68.7	73.4
Act-STG	**77.6**	**72.9**	**74.8**

Table 4.11: Severity level prediction results

	Bricklaying		Plastering	
Features	Acc. (%)	SE (%)	Acc. (%)	SE (%)
whole image	18.4	17.1	34.1	13.0
activity+context	50.5	4.6	54.8	1.3
PPE+tools+material	56.6	2.6	44.2	0.3
arms+back+legs	60.8	1.0	76.5	1.7
Full worker state	**85.7**	3.6	**86.6**	2.6

On the bricklaying task, the best model achieves 85.7% cross–validation accuracy (Acc.) and 3.6% standard error (SE). On the plastering task, the best model achieves 86.6% cross–validation accuracy and 2.6% standard error. These numbers show that regression models can reliably predict human preference using the generated worker states.

In addition, an ablation study is conducted using variations of worker state components and a fine–tuned ImageNet ResNet18 image classifier that directly predicts severity level from whole image. The full comparisons are reported in Table 4.11. *whole image* represents the ResNet18 image classifier; *activity+context* refers to the logistic regression model only using activity and context components as the worker state; *PPE+tools+material* refers to the classifier using PPE, tools, and material components as the worker state; *arms+back+legs* refers to the classifier using worker's pose components as the worker state; *Full worker state* is the proposed method. The proposed severity level prediction model using the full worker state achieves the best result. The reported results also show that severity level prediction cannot be easily learned from pure image classification.

The best model for severity level prediction is not significantly biased towards particular worker OWAS body pose codes and worker activities. However, severity level prediction performance when a worker in the state "Knee bent" or "Transfer plaster" is slightly lower than workers in other states. Detailed performance breakdown is presented in Table 4.12. Note "n/a" under a class in Table 4.12 means no instance of that class in the images with ground truth severity level annotations.

Table 4.12: The best severity level prediction model performance breakdown by poses and activities

	Arms (%)			Back (%)			Legs (%)			Activity (%)		
Task	A	B	C	A	B	C	A	B	C	A	B	C
Plastering	84.5	86.4	n/a	89.3	82.3	85.0	89.2	78.0	90.4	77.8k	89.6	84.8
Bricklaying	86.4	90.0	n/a	84.6	86.8	87.0	91.4	82.3	n/a	81.4	92.4	85.0

For every frame in a given video, worker states are recognized and ranked by their probabilities in descending order. The top worker states are counted until their cumulative probability is over 95%. The severity level of each worker state is predicted using the best model and converted to severity weights in Table 4.3. The proposed risk formulation in Eq. 4.1 is used to generate a risk score for each frame. Fig. 4.12 shows three test set videos' risk scores generated from the proposed method. A dot represents the calculated risk score for a frame computed using, worker images are pointed to their corresponding risk scores and frame numbers. These three videos are not used to train severity level prediction models. The worker in the top video experiences the highest risk among the three likely because of three factors, (i) he stands on an elevated surface with an open edge;(ii) he does not wear any proper PPE; (iii) in the beginning his back was heavily bent. The worker in the middle video is at higher risk in the beginning because of heavily bent back and not wearing gloves, later his risk score increases because of applying plaster with one arm raised above shoulder level and not wearing gloves. Compare three workers in the top middle, and bottom videos, the bottom worker experiences lower risk because of properly worn PPE, even though his back is bent during placing material.

4.7 Limitations and Open Research Areas

The foregoing experiment results and analysis show that automatically obtaining visual knowledge of a worker has a potential to assist worker risk analysis. From a technical standpoint, the present study also shows that abundant visual knowledge of a worker's activity, body pose, PPE use, and interactions with tools and material is beneficial for worker

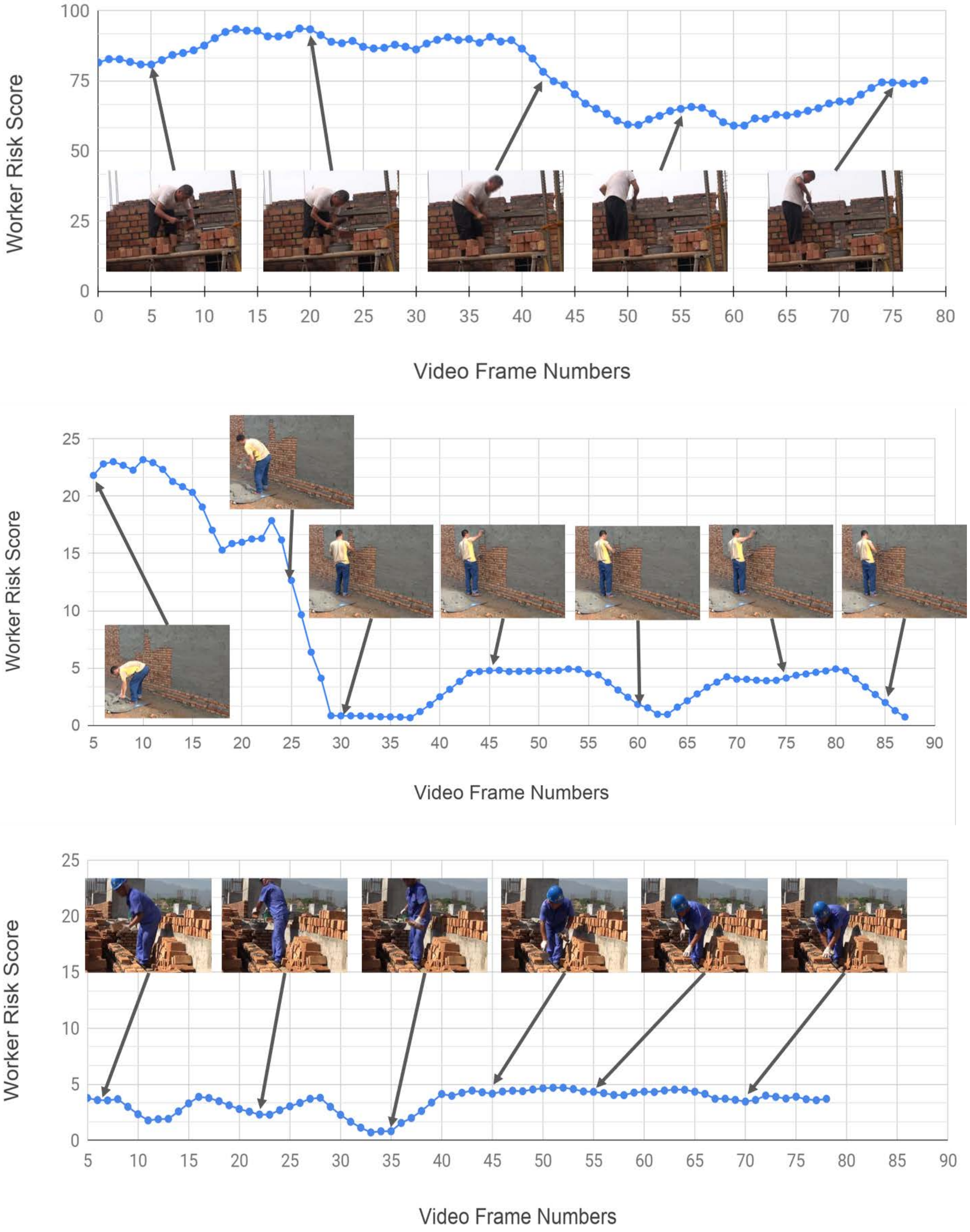


Figure 4.12: Calculated risk scores for three test set videos.

risk analysis and can be efficiently obtained with a multi-tasked visual model framework.

Nevertheless, there are a few limitations that affect the performance, robustness, and applications of the proposed method.

Data Collection: (i) the proposed method is validated mostly on images that capture single worker outdoor bricklaying and plastering construction activities at a close distance to workers. To extend this method to other construction site images, such as indoor construction and far-field images, additional image OAKs annotations are necessary. However, since the proposed OAK model and Act-STG are derived from generic computer vision tasks, they can be fine-tuned and applied even if only partial sub-task data is provided for the new applications. (ii) collecting ground truth severity level labels is still challenging, as substantial safety knowledge and additional inter-observer agreement tests are necessary to obtain high-quality labels. Annotators' agreement can shift between companies, regions, and countries. (iii) it is needed to avoid wasting annotation resources on the most frequent severity levels because annotators recruitment can be expensive and slow.

Model Development: (i) heavy occlusion of body keypoints and objects can greatly affect severity level prediction outcomes. Robust recognition under heavy occlusion is not considered in this work as the dataset used has low occlusion in general. Robust models to image occlusion are often explored in the context of object tracking, this is a work for future development. (ii) worker pose is a crucial cue for safety professionals to assess risks. In this study, an estimated 3D pose is used for predicting severity level. However, the estimated 3D keypoints are not explicitly evaluated in the present study. We entrust that duty to an off-the-shelf model based on its high performance on generic benchmarks and visual examinations of many construction site image examples. 3D keypoint localization can be better evaluated against IMU-based 3D data [29, 33]. Another limitation of the off-the-shelf keypoint conversion model is that camera viewing angles and distances to persons in generic datasets may be different from those in collected construction site images. This difference results in deviations in body pose classification. Future work in this direction will train 3D body keypoint estimation models by construction worker images and their calibrated 3D keypoint locations.

Application: Given the current state of development, the ideal scenario to apply the proposed method is through wearable cameras which are increasingly popular on job sites. This satisfies the requirement of using outdoor, close-ranged worker images. Although the proposed method runs at around 26 frames per second on a single desktop GPU, such a powerful embedded system has not yet been commercially ready. So the proposed method is used offline. Future explorations on system topics, such as hardware acceleration and network communication, are needed to validate real-world case-studies of the proposed method.

4.8 Conclusions

In the present study, a visual–based risk analysis model is proposed for evaluating worker–level risk. The proposed risk model leverages a worker state described by visually obtained worker's activity, worker's body pose, PPE use, worker's interactions with tools and material, and workplace context. A joint learning on object, activity, and keypoints can be efficiently obtained by a unified regional convolutional neural network framework. This strategy significantly reduces resources required and achieves comparable model performance to models that are dedicated for individual tasks. A spatio–temporal graph neural network model that combines per-frame information on objects, activities, and keypoints is beneficial for improving activity recognition. A machine learning–based severity prediction model takes the worker state and predicts worker severity level. An ablation study shows that the full worker state is more informative to predict severity level than any worker state using partial visual information. These findings are validated with a newly introduced large image dataset with exhaustive annotations of single worker activity, body pose, PPE, tools, and materials. The final severity prediction model achieves 85.7% cross-validation accuracy for bricklaying images and 86.6% cross-validation accuracy for plastering images.

CHAPTER 5: CONCLUSION

5.1 Summary

In the past decade, the growth in volume of visual data collected on construction sites and the development of modern computer vision have provided an unprecedented opportunity to automate safety inspection and risk analysis processes on construction sites. These automated processes empowers safety inspection practices with more frequent inspection reports, wider coverage of construction sites, and less expensive labour cost. Despite the fact that commercially available visual–based automatic safety tools are already deployed over 1,000 construction sites in the U.S., many research questions are still need to be addressed for more robust, accurate and explainable methods to assist safety managers. In this dissertation, I presented my research on improving semantic, spatio-temporal visual understanding and in particular how these methods can assist autonomous worker safety inspection and risk analysis. I hope the present study can inspire other to continue exploring systematic approaches to improve automation in safety programs. Specifically, I have studied and presented the areas below:

1. The introduction chapter presented the national level construction safety performance in the U.S. and identified several gaps-in-knowledge between the "zero accident" vision and current safety inspection programs. Limitations of previous research were identified and the research road map for this study was drawn. The succeeding chapters presented the proposed solutions.

2. HOI recognition was used for structuring visual–based safety compliance checking. In particular, a learning–based HOI recognition model was compared with a rule–based HOI recognition model on a newly construction construction image dataset. It was

found that while HOI formulation extends visual–based safety compliance checking from zero–order object level to first–order object relation level, object detection performance is crucial to HOI recognition models. While rule–based HOI recognition methods have been used in previous work and proven to be a strong baseline, the reality is that learning–based HOI is more robust to deal with real–world object detection performance.

3. A new model to forecast motion trajectory of workers and equipment was proposed. The model expanded the research scope of proximity hazard identification on construction sites. The model was validated and outperformed several baseline models using visual object tracks collected from a real–world construction site as well as a generic pedestrian trajectory forecasting benchmark. It was found that acute motion, such as turning, stopping and starting, are the most challenging scenarios to forecast. The proposed forecasting model is general and can be easily adapted to forecast worker activity.

4. A survey from construction safety professionals found a consensus on rating worker severity level from single images can be reached. While a simple image classification model can not predict worker severity level well, using the worker state, a high–level visual summary of a worker's status, significantly improved worker severity level prediction. A generalized object detection framework that unified worker activity recognition, worker body keypoint estimation, and PPE, tools, and material detection, a spatio–temporal graph neural network model for activity label refinement, and a 2D-3D keypoint conversion model were used together to generate high quality worker state.

5.2 Open Research Opportunities

The present study proposed many new research directions for visual–based construction safety management and showed promising results for future exploration. However, there are few critical challenges and opportunities were less addressed in the study. The following sections explains these topics in detail.

5.2.1 Image–Language Models For Construction Safety Inspection

The image–language tasks is a family of machine learning tasks that associate textual data to their visual correspondences. The holy grail of these tasks is often regarded as the Visual Question Answering (VQA) [187]. In VQA, the machine is presented with an image and a textual question, and is expected to answer the question based on image content by free form answers or multiple choices. It is reasonable to imagine that the ultimate form of an automated safety tool will be an "all–knowing" agent that can answer any safety relevant questions that a safety professional may asks. This is close to the problem setting of VQA. Nevertheless, unlike the general VQA, safety questions are likely to be limited to "how many" and "yes or no" questions, and the answers needed to be explained by the machine to convince safety professionals. So the question–answer–explanation format in grounded VQA [188] and visual commonsense [189] are closer to what answering safety questions requires. Future research is expected to explore a VQA–like system that answers natural language safety queries based on a single construction site image or a repository of construction site photologs or with the aid of external safety knowledge. In chapter 2, the HOI formulation is introduced to address safety inspection tasks. HOI provides a basis to address VQA in the construction safety settings, because some questions can be decomposed into HOIs. Questions can also be generated from a scene graph, which is constructed by a set of HOIs. A recent development on this topic in the built environment is the 3D Scene Graph dataset [190], where 3D objects in point clouds are associated based on their relative positions.

5.2.2 Fairness In Computer Vision for Construction Safety

There are times when behavior–based safety programs are being used to place blame on individuals or organizations because its excessive focus on observing worker behaviors. While many safety professionals agree that worker participation is an important factor of any effective safety program, a "justified" blame (and the consequential negative rewards) is detrimental to a company's safety culture. To my best knowledge, existing visual–based safety tools have not consider how their visual recognition models should be used to curb (and not promote) prejudice and systematic bias in current safety management programs.

On the other hand, computer vision models, if unchecked, are known for exploiting dataset biases and correlating individual traits with outcomes. A recent example is how face recognition models are proven to be biased toward people with darker skin colors. Evaluating and ensuring fairness in visual–based safety tools is crucial for the continuous growth of companies that provide these services. It is also a testimony to fulfill the human–centered artificial intelligence vision. Work introduced in chapter 2, chapter 3, and chapter 4 provide a basis for this exploration. High intra and inter–class annotations serve as benchmarks to examine the degree of bias in the trained models. The crowd-sourced data annotation process in chapter 2 and chapter 4 also provide the basis to control annotation distribution in order to build fair datasets.

5.2.3 Low–shot Learning, Generative Models, and Synthesized Dataset for Object Detection in Construction Sites

Building construction site image dataset are often faced with the following challenges: 1) there are only a few open–sourced construction image dataset available and fewer companies are willing to share their image data to the public; 2) annotating construction resources requires professional construction knowledge; 3) there are too many rare classes in construction resources. Since these challenges are not uniquely faced by construction applications, there are many tools that can be borrowed from the general machine learning community. On the learning side, low–shot learning is used to handle extremely low example cases (less than 10 positive examples); domain adaptation is widely applied to assist learning using knowledge from similar tasks. On the data side, generative models, generative adversarial network in particular, learn a high–dimensional distribution of the dataset and generate new data by sampling from the learned distribution; Synthesized dataset, such as those rendered from virtual reality engine, provides supplemental data to assist training. A recent notable development of using large scale synthetic dataset is presented in [191]. In chapter 2 and chapter 3, coarse level object categories are used because of this rare class issue. The PPE and vehicle examples in chapter 2 and chapter 3 can be further annotated and used as a basis for building fine–grained and few–shot recognition models for construction sites.

REFERENCES

[1] OSHA, “Commonly used statistics, occupational safety and health administration,” 2019, (Nov. 1, 2019). [Online]. Available: https://www.osha.gov/data/commonstats

[2] BLS, “Industries at a glance. construction: Naics 23,” 2019, (Nov. 1, 2019). [Online]. Available: https://www.bls.gov/iag/tgs/iag23.htm

[3] G. M. Waehrer, X. S. Dong, T. Miller, E. Haile, and Y. Men, “Costs of occupational injuries in construction in the united states,” *Accident Analysis and Prevention*, vol. 39, no. 6, pp. 1258 – 1266, 2007.

[4] J. Gigstad, “Construction fatalities cost the united states $5 billion per year,” 2017. [Online]. Available: https://midwestepi.org/2017/05/08/construction-fatalities-cost-the-us-5-billion-per-year/

[5] W. Fang, L. Ding, P. E. Love, H. Luo, H. Li, F. Peña-Mora, B. Zhong, and C. Zhou, “Computer vision applications in construction safety assurance,” *Automation in Construction*, vol. 110, p. 103013, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580519301487

[6] M. Namian, A. Albert, C. M. Zuluaga, and E. J. Jaselskis, “Improving hazard-recognition performance and safety training outcomes: Integrating strategies for training transfer,” *Journal of Construction Engineering and Management*, vol. 142, no. 10, p. 04016048, 2016.

[7] O. Rozenfeld, R. Sacks, Y. Rosenfeld, and H. Baum, “Construction job safety analysis,” *Safety science*, vol. 48, no. 4, pp. 491–498, 2010.

[8] A. Albert, M. Hallowell, B. Kleiner, A. Chen, and M. Golparvar Fard, “Enhancing construction hazard recognition with high-fidelity augmented virtuality,” *Journal of Construction Engineering and Management - ASCE*, vol. 140, no. 7, 7 2014.

[9] S. Bahn, “Workplace hazard identification and management: The case of an underground mining operation,” *Safety science*, vol. 57, pp. 129–137, 2013.

[10] R. A. Haslam, S. A. Hide, A. G. Gibb, D. E. Gyi, T. Pavitt, S. Atkinson, and A. R. Duff, “Contributing factors in construction accidents,” *Applied ergonomics*, vol. 36, no. 4, pp. 401–415, 2005.

[11] M. R. Hallowell and J. A. Gambatese, "Construction safety risk mitigation," *Journal of Construction Engineering and Management*, vol. 135, no. 12, pp. 1316–1323, 2009. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/\%28ASCE\%29CO.1943-7862.0000107

[12] R. Salas, M. Hallowell, D. Ph, and M. Asce, "Predictive Validity of Safety Leading Indicators : Empirical Assessment in the Oil and Gas Sector," *Journal of construction engineering and managment*, vol. 142, no. 10, pp. 1–11, 2015.

[13] J. Teizer and T. Cheng, "Proximity hazard indicator for workers-on-foot near miss interactions with construction equipment and geo-referenced hazard areas," *Automation in Construction*, vol. 60, pp. 58–73, 2015.

[14] M. R. Hallowell, J. W. Hinze, K. C. Baud, and A. Wehle, "Proactive construction safety control: Measuring, monitoring, and responding to safety leading indicators," *Journal of Construction Engineering and Management*, vol. 139, no. 10, p. 04013010, 2013.

[15] S. Zhang, J. Teizer, J. K. Lee, C. M. Eastman, and M. Venugopal, "Building Information Modeling (BIM) and Safety: Automatic Safety Checking of Construction Models and Schedules," *Automation in Construction*, vol. 29, pp. 183–195, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.autcon.2012.05.006

[16] S. Zhang, K. Sulankivi, M. Kiviniemi, I. Romo, C. M. Eastman, and J. Teizer, "Bim-based fall hazard identification and prevention in construction safety planning," *Safety Science*, vol. 72, pp. 31 – 45, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925753514001829

[17] D. Oswald, F. Sherratt, and S. Smith, "Problems with safety observation reporting: A construction industry case study," *Safety Science*, vol. 107, pp. 35 – 45, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925753516304581

[18] I. G. Awolusi and E. D. Marks, "Safety Activity Analysis Framework To Evaluate Safety Performance," *Journal of Construction Engineering and Management*, vol. 143, no. 3, p. 05016022, 2016.

[19] J. Hinze, "Safety plus: Making zero accidents a reality," *Construction Industry Institute*, pp. 160–11, 2002.

[20] J. Hinze, M. Hallowell, and K. Baud, "Construction-safety best practices and relationships to safety performance," *Journal of Construction Engineering and Management*, vol. 139, no. 10, p. 04013006, 2013.

[21] N. S. Akroush and I. H. El-Adaway, "Utilizing construction leading safety indicators: Case study of tennessee," *Journal of Management in Engineering*, vol. 33, no. 5, p. 06017002, 2017.

[22] K. Ng, A. Laurlund, G. Howell, and G. Lancos, "lean safety: using leading indicators of safety incidents to improve construction safety," 2012, (Jan. 03, 2020). [Online]. Available: https://www.xlconstruction.com/wp-content/uploads/2012/09/WP_LeanSafety_SafetyLeadingIndicators_060312.pdf

[23] J. Teizer, B. S. Allread, C. E. Fullerton, and J. Hinze, "Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system," *Automation in Construction*, vol. 19, no. 5, pp. 630 – 640, 2010.

[24] J. Seo, S. Han, S. Lee, and H. Kim, "Computer vision techniques for construction safety and health monitoring," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 239–251, 2015.

[25] M. Zhang, R. Shi, and Z. Yang, "A critical review of vision-based occupational health and safety monitoring of construction site workers," *Safety Science*, vol. 126, p. 104658, 2020.

[26] J. Teizer, "Right-time vs real-time pro-active construction safety and health system architecture," *Construction Innovation*, vol. 16, no. 3, pp. 253–280, 2016.

[27] X. Shen and E. Marks, "Near-miss information visualization tool in bim for construction safety," *Journal of Construction Engineering and Management*, vol. 142, no. 4, p. 04015100, 2015.

[28] X. Yan, H. Li, H. Zhang, and T. M. Rose, "Personalized method for self-management of trunk postural ergonomic hazards in construction rebar ironwork," *Advanced Engineering Informatics*, vol. 37, pp. 31 – 41, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034617306213

[29] Y. Yu, X. Yang, H. Li, X. Luo, H. Guo, and Q. Fang, "Joint-level vision-based ergonomic assessment tool for construction workers," *Journal of Construction Engineering and Management*, vol. 145, no. 5, p. 04019025, 2019.

[30] S. Tang, D. Roberts, and M. Golparvar-Fard, "Human-object interaction recognition for automatic construction site safety inspection," *Automation in Construction*, vol. 120, p. 103356, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580520309365

[31] K. K. Han and M. Golparvar-Fard, "Potential of big visual data and building information modeling for construction performance analytics: An exploratory study," *Automation in Construction*, vol. 73, pp. 184 – 198, 2017.

[32] H. Zhang, X. Yan, and H. Li, "Ergonomic posture recognition using 3d view-invariant features from single ordinary camera," *Automation in Construction*, vol. 94, pp. 1 – 10, 2018.

[33] W. Chu, S. Han, X. Luo, and Z. Zhu, "Monocular vision-based framework for biomechanical analysis or ergonomic posture assessment in modular construction," *Journal of Computing in Civil Engineering*, vol. 34, no. 4, p. 04020018, 2020.

[34] S. Chi and C. H. Caldas, "Automated object identification using optical video cameras on construction sites," *Computer-Aided Civil and Infrastructure Engineering*, vol. 26, no. 5, pp. 368–380, 2011.

[35] S. Du, M. Shehata, and W. Badawy, "Hard hat detection in video sequences based on face features, motion and color information," in *2011 3rd International Conference on Computer Research and Development*, vol. 4, March 2011, pp. 25–29.

[36] M. Memarzadeh, M. Golparvar-Fard, and J. C. Niebles, "Automated 2d detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors," *Automation in Construction*, vol. 32, pp. 24 – 37, 2013.

[37] E. R. Azar and B. McCabe, "Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos," *Automation in Construction*, vol. 24, pp. 194 – 202, 2012.

[38] A. Khosrowpour, I. Fedorov, A. Holynski, J. C. Niebles, and M. Golparvar-Fard, "Automated worker activity analysis in indoor environments for direct-work rate improvement from long sequences of rgb-d images," in *Construction Research Congress 2014*, 2014, pp. 729–738.

[39] M.-W. Park, N. Elsafty, and Z. Zhu, "Hardhat-wearing detection for enhancing on-site safety of construction workers," *Journal of Construction Engineering and Management*, vol. 141, no. 9, p. 04015024, 2015.

[40] M. Golparvar-Fard, A. Heydarian, and J. C. Niebles, "Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers," *Advanced Engineering Informatics*, vol. 27, no. 4, pp. 652 – 663, 2013.

[41] J. Gong, C. H. Caldas, and C. Gordon, "Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and bayesian network models," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 771 – 782, 2011, special Section: Advances and Challenges in Computing in Civil and Building Engineering; Last Accessed on March 12th, 2019.

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

[43] S. J. Ray and J. Teizer, "Real-time construction worker posture analysis for ergonomics training," *Advanced Engineering Informatics*, vol. 26, no. 2, pp. 439 – 455, 2012, knowledge based engineering to support complex product design. Last accessed on March 12th, 2019.

[44] E. R. Azar, S. Dickinson, and B. McCabe, "Server-customer interaction tracker: Computer vision-based system to estimate dirt-loading cycles," *Journal of Construction Engineering and Management*, vol. 139, no. 7, pp. 785–794, 2013.

[45] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969250 pp. 91–99.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2980–2988.

[47] K. Liu and M. Golparvar-Fard, "Crowdsourcing construction activity analysis from jobsite video streams," *Journal of Construction Engineering and Management*, vol. 141, no. 11, p. 04015035, 2015.

[48] BLS, "Injuries, illnesses, and fatalities," 2020, (May 1, 2020). [Online]. Available: https://www.bls.gov/iif/#data

[49] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – European Conference on Computer Vision 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[50] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788.

[51] M. Liu, S. Han, and S. Lee, "Potential of convolutional neural network-based 2d human pose estimation for on-site activity analysis of construction workers," in *Computing in Civil Engineering 2017*, 2017, pp. 141–149.

[52] X. Yan, H. Zhang, and H. Li, "Computer vision-based recognition of 3d relationship between construction entities for monitoring struck-by accidents," *Computer-Aided Civil and Infrastructure Engineering*, vol. n/a, no. n/a, 2020.

[53] J. Yang, Z. Shi, and Z. Wu, "Vision-based action recognition of construction workers using dense trajectories," *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 327 – 336, 2016.

[54] J. Cai, Y. Zhang, and H. Cai, "Two-step long short-term memory method for identifying construction activities through positional and attentional cues," *Automation in Construction*, vol. 106, p. 102886, 2019.

[55] J. Kim, S. Chi, and J. Seo, "Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks," *Automation in Construction*, vol. 87, pp. 297 – 308, 2018.

[56] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, and C. Li, "Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment," *Automation in Construction*, vol. 93, pp. 148 – 164, 2018.

[57] S. Tang and M. Golparvar-Fard, "Joint reasoning of visual and text data for safety hazard recognition," in *Computing in Civil Engineering 2017*, 2017, pp. 450–457.

[58] M.-W. Park and I. Brilakis, "Construction worker detection in video frames for initializing vision trackers," *Automation in Construction*, vol. 28, pp. 15 – 25, 2012.

[59] M. Golparvar-Fard, F. Peña-Mora, and S. Savarese, *Monitoring of Construction Performance Using Daily Progress Photograph Logs and 4D As-Planned Models*, 2009, pp. 53–63.

[60] Z. Zhu and I. Brilakis, "Concrete column recognition in images and videos," *Journal of Computing in Civil Engineering*, vol. 24, no. 6, pp. 478–487, 2010.

[61] C. Yuan and H. Cai, "Key nodes modeling for object detection and location on construction site using color-depth cameras," in *Computing in Civil and Building Engineering (2014)*, 2014, pp. 729–736.

[62] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1440–1448.

[63] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. USA: Curran Associates Inc., 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157096.3157139 pp. 379–387.

[64] H. Kim, H. Kim, Y. W. Hong, and H. Byun, "Detecting construction equipment using a region-based fully convolutional network and transfer learning," *Journal of Computing in Civil Engineering*, vol. 32, no. 2, p. 04017082, 2018.

[65] X. Luo, H. Li, D. Cao, F. Dai, J. Seo, and S. Lee, "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*, vol. 32, no. 3, p. 04018012, 2018.

[66] D. Roberts, T. Bretl, and M. Golparvar-Fard, "Detecting and classifying cranes using camera-equipped uavs for monitoring crane-related safety hazards," in *Computing in Civil Engineering 2017*, 2017, pp. 442 – 449.

[67] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – European Conference on Computer Vision 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[68] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun 2010.

[69] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, Oct 2009.

[70] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 17–24.

[71] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 835–848, April 2013.

[72] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015, january 11th, 2018. [Online]. Available: https://arxiv.org/abs/1505.04474

[73] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 8359–8367.

[74] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2018, pp. 381–389.

[75] R. Xiong, Y. Song, H. Li, and Y. Wang, "Onsite video mining for construction hazards identification with visual relationships," *Advanced Engineering Informatics*, vol. 42, p. 100966, 2019.

[76] X. Wei, Y. Qi, J. Liu, and F. Liu, "Image retrieval by dense caption reasoning," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, Dec 2017, pp. 1–4.

[77] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 1270–1279.

[78] E. Hollnagel, "Chapter 6 - cream — a second generation hra method," in *Cognitive Reliability and Error Analysis Method (CREAM)*, E. Hollnagel, Ed. Oxford: Elsevier Science Ltd, 1998, pp. 151 – 190.

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

[80] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[81] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[82] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, “Visual relationship detection with language priors,” in *Computer Vision – European Conference on Computer Vision 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 852–869.

[83] D. Hoiem, Y. Chodpathumwan, and Q. Dai, “Diagnosing error in object detectors,” in *Computer Vision – European Conference on Computer Vision 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 340–353.

[84] M. S. Bajjou, A. Chafi, and A. En-Nadi, “The potential effectiveness of lean construction tools in promoting safety on construction sites,” in *International Journal of Engineering Research in Africa*, vol. 33, 12 2017, pp. 179–193.

[85] D. Zhao and J. Lucas, “Virtual reality simulation for construction safety promotion,” *International Journal of Injury Control and Safety Promotion*, vol. 22, no. 1, pp. 57–67, 2015.

[86] CII, “Cii safety summary report - summary of cii 2018 safety rates,” 2019, (Jan. 03, 2020). [Online]. Available: https://www.construction-institute.org/securefile?filename=dpc2019_2.pdf\#page=3f

[87] J. Hinze, S. Thurman, and A. Wehle, “Leading indicators of construction safety performance,” *Safety science*, vol. 51, no. 1, pp. 23–28, 2013.

[88] C. Sheehan, R. Donohue, T. Shea, B. Cooper, and H. D. Cieri, “Leading and lagging indicators of occupational health and safety: The moderating role of safety leadership,” *Accident Analysis & Prevention*, vol. 92, pp. 130 – 138, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457516300896

[89] B. H. W. Guo and T. W. Yiu, “Developing leading indicators to monitor the safety conditions of construction projects,” *Journal of Management in Engineering*, vol. 32, no. 1, p. 04015016, 2016. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/\%28ASCE\%29ME.1943-5479.0000376

[90] R. Salas and M. Hallowell, “Predictive validity of safety leading indicators: empirical assessment in the oil and gas sector,” *Journal of construction engineering and management*, vol. 142, no. 10, p. 04016052, 2016.

[91] C. Q. Poh, C. U. Ubeynarayana, and Y. M. Goh, “Safety leading indicators for construction sites: A machine learning approach,” *Automation in construction*, vol. 93, pp. 375–386, 2018.

[92] N. Xia, P. X. Zou, X. Liu, X. Wang, and R. Zhu, "A hybrid bn-hfacs model for predicting safety performance in construction projects," *Safety science*, vol. 101, pp. 332–343, 2018.

[93] P. Jafari, E. Mohamed, E. Pereira, S. Kang, and S. AbouRizk, "Leading safety indicators: Application of machine learning for safety performance measurement," in *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 36. IAARC Publications, 2019, pp. 501–506.

[94] O. Golovina, J. Teizer, and N. Pradhananga, "Heat map generation for predictive safety planning: Preventing struck-by and near miss interactions between workers-on-foot and construction equipment," *Automation in construction*, vol. 71, pp. 99–115, 2016.

[95] X. Luo, H. Li, T. Huang, and M. Skitmore, "Quantifying hazard exposure using real-time location data of construction workforce and equipment," *Journal of Construction Engineering and Management*, vol. 142, no. 8, p. 04016031, 2016. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/\%28ASCE\%29CO.1943-7862.0001139

[96] Y. Li, Y. Hu, B. Xia, M. Skitmore, and H. Li, "Proactive behavior-based system for controlling safety risks in urban highway construction megaprojects," *Automation in Construction*, vol. 95, pp. 118 – 128, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518302723

[97] H. Chen, X. Luo, Z. Zheng, and J. Ke, "A proactive workers' safety risk evaluation framework based on position and posture data fusion," *Automation in Construction*, vol. 98, pp. 275 – 288, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518308227

[98] M. Arslan, C. Cruz, and D. Ginhac, "Semantic enrichment of spatio-temporal trajectories for worker safety on construction sites," *Procedia Computer Science*, vol. 130, pp. 271 – 278, 2018, the 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918303909

[99] D. Kim, M. Liu, S. Lee, and V. R. Kamat, "Remote proximity monitoring between mobile construction resources using camera-mounted uavs," *Automation in Construction*, vol. 99, pp. 168 – 182, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518304102

[100] A. Carbonari, A. Giretti, and B. Naticchia, "A proactive system for real-time safety management in construction sites," *Automation in Construction*, vol. 20, no. 6, pp. 686 – 698, 2011, selected papers from the 26th ISARC 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580511000756

[101] Y. Yu, H. Guo, Q. Ding, H. Li, and M. Skitmore, "An experimental study of real-time identification of construction workers' unsafe behaviors," *Automation in Construction*, vol. 82, pp. 193 – 206, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580517304223

[102] W. Umer, H. Li, W. Lu, G. P. Y. Szeto, and A. Y. Wong, "Development of a tool to monitor static balance of construction workers for proactive fall safety management," *Automation in Construction*, vol. 94, pp. 438–448, 2018.

[103] J. Teizer, "Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 225 – 238, 2015, infrastructure Computer Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034615000336

[104] N. Soltanmohammadlou, S. Sadeghi, C. K. Hon, and F. Mokhtarpour-Khanghah, "Real-time locating systems and safety in construction sites: A literature review," *Safety Science*, vol. 117, pp. 229 – 242, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092575351831556X

[105] M. Hallowell, B. Esmaeili, and P. Chinowsky, "Safety risk interactions among highway construction work tasks," *Construction Management and Economics*, vol. 29, no. 4, pp. 417–429, 2011. [Online]. Available: https://doi.org/10.1080/01446193.2011.552512

[106] J. Wang and S. Razavi, "Spatiotemporal network-based model for dynamic risk analysis on struck-by-equipment hazard," *Journal of Computing in Civil Engineering*, vol. 32, no. 2, p. 04017089, 2018. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/\%28ASCE\%29CP.1943-5487.0000732

[107] M. Zhang, T. Cao, and X. Zhao, "Applying sensor-based technology to improve construction safety management," *Sensors*, vol. 17, no. 8, p. 1841, 2017.

[108] I. Awolusi, E. Marks, and M. Hallowell, "Wearable technology for personalized construction safety monitoring and trending: Review of applicable devices," *Automation in Construction*, vol. 85, pp. 96 – 106, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580517309184

[109] Y. Ham, K. K. Han, J. J. Lin, and M. Golparvar-Fard, "Visual monitoring of civil infrastructure systems via camera-equipped unmanned aerial vehicles (uavs): a review of related works," *Visualization in Engineering*, vol. 4, no. 1, p. 1, 2016.

[110] Z. Zhu, X. Ren, and Z. Chen, "Visual tracking of construction jobsite workforce and equipment with particle filtering," *Journal of Computing in Civil Engineering*, vol. 30, no. 6, p. 04016023, 2016.

[111] H. Kim, K. Kim, and H. Kim, "Vision-based object-centric safety assessment using fuzzy inference: Monitoring struck-by accidents with moving objects," *Journal of Computing in Civil Engineering*, vol. 30, no. 4, p. 04015075, 2016. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/\%28ASCE\%29CP.1943-5487.0000562

[112] Z. Zhu, M.-W. Park, C. Koch, M. Soltani, A. Hammad, and K. Davari, "Predicting movements of onsite workers and mobile equipment for enhancing construction site safety," *Automation in Construction*, vol. 68, pp. 95 – 101, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580516300802

[113] M.-W. Park and I. Brilakis, "Continuous localization of construction workers via integration of detection and tracking," *Automation in Construction*, vol. 72, pp. 129–142, 2016.

[114] D. Kim, M. Liu, S. Lee, and V. R. Kamat, "Trajectory prediction of mobile construction resources toward pro-active struck-by hazard detection," in *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 36. IAARC Publications, 2019, pp. 982–988.

[115] Indus.ai, "Construction intelligence," 2020, april. 10, 2020. [Online]. Available: https://www.indus.ai/

[116] Oxblue, "Site camera firm uses ai to measure virus impact," Apr 2020, april. 16, 2020. [Online]. Available: https://www.theconstructionindex.co.uk/news/view/site-camera-firm-uses-ai-to-measure-virus-impact

[117] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3464–3468.

[118] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *European Conference on Computer Vision*. Springer, 2012, pp. 201–214.

[119] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 961–971.

[120] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani, "Forecasting interactive dynamics of pedestrians with fictitious play," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 4636–4644.

[121] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 336–345.

[122] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[123] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013, (Nov. 1, 2019).

[124] J. Yang, M.-W. Park, P. A. Vela, and M. Golparvar-Fard, "Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 211 – 224, 2015, infrastructure Computer Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034615000233

[125] I. Brilakis, M.-W. Park, and G. Jog, "Automated vision tracking of project related entities," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 713 – 724, 2011, special Section: Advances and Challenges in Computing in Civil and Building Engineering. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034611000048

[126] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942

[127] D. Roberts and M. Golparvar-Fard, "End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level," *Automation in Construction*, vol. 105, p. 102811, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518308525

[128] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint arXiv:1806.11230*, 2018, (Nov. 1, 2019).

[129] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 3308–3318.

[130] B. T. Morris, M. M. Trivedi et al., "A survey of vision-based trajectory learning and analysis for surveillance," 2008.

[131] T. Hirakawa, T. Yamashita, T. Tamaki, and H. Fujiyoshi, "Survey on vision-based path prediction," in *International Conference on Distributed, Ambient, and Pervasive Interactions.* Springer, 2018, pp. 48–64.

[132] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision.* Springer, 2016, pp. 549–565.

[133] R. Hug, S. Becker, W. Hübner, and M. Arens, "Particle-based pedestrian path prediction using lstm-mdl models," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC).* IEEE, 2018, pp. 2684–2691.

[134] F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo, "Context-aware trajectory prediction," in *2018 24th International Conference on Pattern Recognition (ICPR).* IEEE, 2018, pp. 1941–1946.

[135] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012, (Nov. 1, 2019).

[136] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[137] H. Li, M. Lu, S.-C. Hsu, M. Gray, and T. Huang, "Proactive behavior-based safety management for construction safety improvement," *Safety Science*, vol. 75, pp. 107 – 117, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925753515000144

[138] T. Li and M. Gong, "A review on unsafe behavior of employees based on big data," in *International Conference on Applications and Techniques in Cyber Security and Intelligence.* Springer, 2019, pp. 650–657.

[139] Triax, "Better manage your safety, risk and efficiency with spot-r wearable technology," Apr 2020, April 20, 2020. [Online]. Available: https://www.triaxtec.com/

[140] J. J. Lin and M. Golparvar-Fard, "Visual data and predictive analytics for proactive project controls on construction sites," in *Workshop of the European Group for Intelligent Computing in Engineering.* Springer, 2018, pp. 412–430.

[141] J. Kanner, "Can we predict construction incidents before they happen?" 2018, (Nov. 1, 2019). [Online]. Available: https://www.smartvid.io/industrial-video-blog/can-we-predict-construction-incidents-before-they-happen

[142] A. Khosrowpour, A. Sadeghi, and M. Golparvar-Fard, "Detecting and analyzing hard-hats and safety vests from jobsite video cameras," *University of Illinois RAAMAC Lab*, vol. 1, no. 1, pp. 1–10, 2013.

[143] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi, "Trajnet: Towards a benchmark for human trajectory prediction," *arXiv preprint*, 2018, (Nov. 1, 2019).

[144] B. Nama, "Social lstm implementation in pytorch," 2019, (Nov. 1, 2019). [Online]. Available: https://github.com/quancore/social-lstm

[145] V. Vaillancourt, H. Nélisse, C. Laroche, C. Giguère, J. Boutin, P. Laferrière et al., "Comparison of sound propagation and perception of three types of backup alarms with regards to worker safety," *Noise and Health*, vol. 15, no. 67, p. 420, 2013.

[146] J. Cai, Y. Zhang, L. Yang, H. Cai, and S. Li, "A context-augmented deep learning approach for worker trajectory prediction on unstructured and dynamic construction sites," *Advanced Engineering Informatics*, vol. 46, p. 101173, 2020.

[147] J. Li, J. Carr, and C. Jobes, "A shell-based magnetic field model for magnetic proximity detection systems," *Safety science*, vol. 50, no. 3, pp. 463–471, 2012.

[148] J. J. Lin, K. K. Han, and M. Golparvar-Fard, *A Framework for Model-Driven Acquisition and Analytics of Visual Data Using UAVs for Automated Construction Progress Monitoring*, 2015, pp. 156–164. [Online]. Available: https://ascelibrary.org/doi/abs/10.1061/9780784479247.020

[149] R. Sacks, O. Rozenfeld, and Y. Rosenfeld, "Spatial and temporal exposure to safety hazards in construction," *Journal of Construction Engineering and Management*, vol. 135, no. 8, pp. 726–736, 2009.

[150] H. W. Heinrich et al., "Industrial accident prevention. a scientific approach." *Industrial Accident Prevention. A Scientific Approach.*, 1941.

[151] A. Luttmann, M. Jäger, B. Griefahn, G. Caffier, and F. Liebers, "Preventing musculoskeletal disorders in the workplace," p. 32 p., 2003.

[152] P. Mitropoulos, G. Cupido, and M. Namboodiri, "Cognitive approach to construction safety: Task demand-capability model," *Journal of Construction Engineering and Management*, vol. 135, no. 9, pp. 881–889, 2009.

[153] R. M. Choudhry, "Behavior-based safety on construction sites: A case study," *Accident Analysis and Prevention*, vol. 70, pp. 14 – 23, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457514000736

[154] W. Fang, L. Ding, H. Luo, and P. E. Love, "Falls from heights: A computer vision-based approach for safety harness detection," *Automation in Construction*, vol. 91, pp. 53 – 61, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580517308403

[155] D. Roberts, W. T. Calderon, S. Tang, and M. Golparvar-Fard, "Vision-based construction worker activity analysis informed by body posture," *Journal of Computing in Civil Engineering*, vol. 34, no. 4, p. 04020017, 2020.

[156] X. Luo, H. Li, D. Cao, Y. Yu, X. Yang, and T. Huang, "Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks," *Automation in Construction*, vol. 94, pp. 360 – 370, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580517311019

[157] X. Yan, H. Zhang, and H. Li, "Estimating worker-centric 3d spatial crowdedness for construction safety management using a single 2d camera," *Journal of Computing in Civil Engineering*, vol. 33, no. 5, p. 04019030, 2019.

[158] M. R. Hallowell and J. A. Gambatese, "Population and initial validation of a formal model for construction safety risk management," *Journal of Construction Engineering and Management*, vol. 136, no. 9, pp. 981–990, 2010.

[159] A. Årnes, K. Sallhammar, K. Haslum, T. Brekne, M. E. G. Moe, and S. J. Knapskog, "Real-time risk assessment with network sensors and intrusion detection systems," in *Computational Intelligence and Security*, Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, and Y.-C. Jiao, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 388–397.

[160] N. Paltrinieri, L. Comfort, and G. Reniers, "Learning about risk: Machine learning for risk assessment," *Safety Science*, vol. 118, pp. 475 – 486, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925753518311184

[161] K. Shrestha, P. P. Shrestha, D. Bajracharya, and E. A. Yfantis, "Hard-hat detection for construction safety visualization," *Journal of Construction Engineering*, vol. 2015, no. 1, pp. 1–8, 2015.

[162] M.-W. Park and I. Brilakis, "Construction worker detection in video frames for initializing vision trackers," *Automation in Construction*, vol. 28, pp. 15 – 25, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580512001136

[163] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, T. M. Rose, and W. An, "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in Construction*, vol. 85, pp. 1 – 9, 2018.

[164] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," *Automation in Construction*, vol. 112, p. 103085, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580519308325

[165] J. Wu, N. Cai, W. Chen, H. Wang, and G. Wang, "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset," *Automation in Construction*, vol. 106, p. 102894, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092658051930264X

[166] M. Zhang, M. Zhu, and X. Zhao, "Recognition of high-risk scenarios in building construction based on image semantics," *Journal of Computing in Civil Engineering*, vol. 34, no. 4, p. 04020019, 2020.

[167] O. Karhu, P. Kansi, and I. Kuorinka, "Correcting working postures in industry: A practical method for analysis," *Applied Ergonomics*, vol. 8, no. 4, pp. 199 – 201, 1977. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0003687077901648

[168] L. McAtamney and E. Nigel Corlett, "Rula: a survey method for the investigation of work-related upper limb disorders," *Applied Ergonomics*, vol. 24, no. 2, pp. 91 – 99, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000368709390080S

[169] S. Hignett and L. McAtamney, “Rapid entire body assessment (reba),” *Applied Ergonomics*, vol. 31, no. 2, pp. 201 – 205, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003687099000393

[170] B. Buchholz, V. Paquet, L. Punnett, D. Lee, and S. Moir, “Path: A work sampling-based approach to ergonomic job analysis for construction and other non-repetitive work,” *Applied Ergonomics*, vol. 27, no. 3, pp. 177 – 187, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000368709500078X

[171] M. MassirisFernández, J. Álvaro Fernández, J. M. Bajo, and C. A. Delrieux, “Ergonomic risk assessment based on computer vision and machine learning,” *Computers and Industrial Engineering*, vol. 149, p. 106816, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835220305192

[172] M. Liu, S. Han, and S. Lee, “Tracking-based 3d human skeleton extraction from stereo video camera toward an on-site safety and ergonomic analysis,” *Construction Innovation: Information, Process, Management*, vol. 16, pp. 348–367, 2016.

[173] E. R. Azar and B. McCabe, “Automated visual recognition of dump trucks in construction videos,” *Journal of Computing in Civil Engineering*, vol. 26, no. 6, pp. 769–781, 2012.

[174] M. Memarzadeh, A. Heydarian, M. Golparvar-Fard, and J. C. Niebles, *Real-Time and Automated Recognition and 2D Tracking of Construction Workers and Equipment from Site Video Streams*, 2012, pp. 429–436.

[175] T. Cheng and J. Teizer, “Real-time resource location data collection and visualization technology for construction safety and activity monitoring applications,” *Automation in Construction*, vol. 34, pp. 3 – 15, 2013, information Technologies in Safety Management.

[176] A. Khosrowpour, J. C. Niebles, and M. Golparvar-Fard, “Vision-based workface assessment using depth images for activity analysis of interior construction operations,” *Automation in Construction*, vol. 48, pp. 74 – 87, 2014.

[177] J. Yang, M.-W. Park, P. A. Vela, and M. Golparvar-Fard, “Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future,” *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 211 – 224, 2015, infrastructure Computer Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034615000233

[178] H. Luo, C. Xiong, W. Fang, P. E. Love, B. Zhang, and X. Ouyang, “Convolutional neural networks: Computer vision-based workforce activity assessment in construction,” *Automation in Construction*, vol. 94, pp. 282 – 289, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518305855

[179] L. Ding, W. Fang, H. Luo, P. E. Love, B. Zhong, and X. Ouyang, "A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory," *Automation in Construction*, vol. 86, pp. 118 – 124, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580517302650

[180] X. Luo, H. Li, H. Wang, Z. Wu, F. Dai, and D. Cao, "Vision-based detection and visualization of dynamic workspaces," *Automation in Construction*, vol. 104, pp. 1 – 13, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580518312706

[181] X. Luo, H. Li, X. Yang, Y. Yu, and D. Cao, "Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and bayesian nonparametric learning," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 4, pp. 333–351, 2019.

[182] J. Cai, Y. Zhang, and H. Cai, "Two-step long short-term memory method for identifying construction activities through positional and attentional cues," *Automation in Construction*, vol. 106, p. 102886, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0926580519302316

[183] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5308–5317.

[184] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7745–7754.

[185] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014. [Online]. Available: https://arxiv.org/abs/1412.3555

[186] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and psychological measurement*, vol. 33, no. 3, pp. 613–619, 1973.

[187] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[188] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6700–6709.

[189] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[190] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5664–5673.

[191] W. T. Calderon, D. Roberts, and M. Golparvar-Fard, "Synthesizing pose sequences from 3d assets for vision-based activity analysis," *Journal of Computing in Civil Engineering*, vol. 35, no. 1, p. 04020052, 2021.