



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Master of Science in Engineering

**Automatic Multi-Label Image  
Classification Model for  
Construction Site Images**

August 2019

Department of Architecture & Architectural Engineering

The Graduate School

Seoul National University

**Ye Seul Kim**

**Automatic Multi-Label Image  
Classification Model for Construction  
Site Images**

**by**

**Ye Seul Kim**

**A dissertation submitted to the Graduate School of  
Seoul National University  
in partial fulfillment of the requirements  
for the degree of  
Master of Science in Engineering**

**July 2019**

# Automatic Multi-Label Image Classification Model for Construction Site Images

지도교수 박 문 서

이 논문을 공학석사 학위논문으로 제출함

2019년 7월

서울대학교 대학원

건축학과

김 예 슬

김예슬의 공학석사 학위논문을 인준함

2019年 7月

위 원 장 \_\_\_\_\_ 이 현 수 \_\_\_\_\_ (인)

부 위 원 장 \_\_\_\_\_ 박 문 서 \_\_\_\_\_ (인)

위 원 \_\_\_\_\_ 지 석 호 \_\_\_\_\_ (인)

# **Automatic Multi-Label Image Classification Model for Construction Site Images**

**July 2019**

**Approved by Dissertation Committee:**

---

**Hyun-Soo Lee**

---

**Moonseo Park**

---

**Seok ho Chi**

## **Abstract**

# **Automatic Multi-Label Image Classification Model for Construction Site Images**

YeSeul Kim  
Department of Architecture  
The Graduate School  
Seoul National University

Activity recognition in construction performs as the prerequisite step in the process for various tasks and thus is critical for successful project management. In the last several years, the computer vision community has blossomed, taking advantage of the exploding amount of construction images and deploying the visual analytics technology for cumbersome construction tasks. However, the current annotation practice itself, which is a critical preliminary step for prompt image retrieval and image understanding, is remained as both time-consuming and labor-intensive. Because previous attempts to make the process more efficient were inappropriate to handle dynamic nature of construction images and showed limited performance in classifying construction activities, this research

aims to develop a model which is not only robust to a wide range of appearances but also multi-composition of construction activity images. The proposed model adopts a deep convolutional neural network model to learn high dimensional feature with less human-engineering and annotate multi-labels of semantic information in the images. The result showed that our model was capable of distinguishing different trades of activities at different stages of the activity. The average accuracy of 83% and maximum accuracy of 91% holds promise in an actual implementation of automated activity recognition for construction operations. Ultimately, it demonstrated a potential method to provide automated and reliable procedure to monitor construction activity.

**Keyword : Multi-label Image Classification, Construction Site Image Data Management, Convolutional Neural Network, Deep Learning**

**Student Number : 2017-27421**

# Table of Contents

<b>Chapter 1. Introduction.....</b>	<b>1</b>
1.1. Research Background .....	1
1.2. Research Objectives and Scope.....	5
1.3. Research Outline .....	7
<b>Chapter 2. Preliminary Study .....</b>	<b>9</b>
2.1. Challenges with Construction Activity Image Classification Task .....	10
2.2. Applications of Traditional Vision-based Algorithms in Construction Domain .....	13
2.3. Convolutional Neural Network-based Image Classification in Construction Domain .....	18
2.4. Summary .....	21
<b>Chapter 3. Development of Construction Image Classification Model .....</b>	<b>22</b>
3.1. Customized Construction Image Dataset Preparation .....	23
3.1.1. Construction Activity Classification System.....	23
3.1.2. Dataset Collection.....	24
3.1.3. Data Pre-Processing .....	25
3.2. Construction Image Classification Model Framework.....	27
3.2.1. Multi-label Image Classification .....	27
3.2.2. Base CNN Model Selection.....	28



3.2.3. Proposed ResNet Model Architecture .....	29
3.3. Model Training and Validation.....	33
3.3.1. Transfer Learning.....	33
3.3.2. Loss Computation and Model Optimization .....	33
3.3.3. Model Performance Indicator .....	35
3.4. Summary .....	37
<b>Chapter 4. Experiment Results and Discussion.....</b>	<b>38</b>
4.1. Experiment Results.....	38
4.2. Analysis of Experiment Results .....	42
4.3. Summary .....	44
<b>Chapter 5. Conclusion .....</b>	<b>45</b>
5.1. Research Summary .....	45
5.2. Research Contributions.....	46
5.3. Limitations and Further Study .....	47
<b>References .....</b>	<b>49</b>
<b>Appendix.....</b>	<b>57</b>
<b>Abstract in Korean .....</b>	<b>63</b>

## List of Tables

<b>Table 2-1. Previous vision-based classification methods in the construction field .....</b>	<b>15</b>
<b>Table 3-1. Dataset Composition.....</b>	<b>25</b>
<b>Table 3-2. Comparison of CNN Models' Performances.....</b>	<b>28</b>
<b>Table 4-1. Revised Dataset Composition.....</b>	<b>39</b>
<b>Table 4-2. Examples of Correct Test Example .....</b>	<b>41</b>
<b>Table 4-3. Examples of Incorrect Test Example .....</b>	<b>43</b>
<b>Table 5-1. Examples of Misclassified Early-phased Images .....</b>	<b>47</b>

## List of Figures

<b>Figure 1-1. Overview of Keyword-based Digital Image Database System .....</b>	<b>3</b>
<b>Figure 1-2. Overview of Classification Keyword Categories.....</b>	<b>6</b>
<b>Figure 1-3. Overview of the Research .....</b>	<b>8</b>
<b>Figure 2-1. Examples of high intra-variability of masonry work with wide-ranging appearances and configuration .....</b>	<b>11</b>
<b>Figure 2-2. Examples of low inter-class variability among tile, plaster, and masonry works.....</b>	<b>11</b>
<b>Figure 2-3. Illustration of Traditional Vision-Based Algorithms Process.....</b>	<b>13</b>
<b>Figure 2-4. Illustration of the Deep CNN-Based Algorithms Process .....</b>	<b>18</b>
<b>Figure 3-1. Example of Construction Activity Classification System .....</b>	<b>23</b>
<b>Figure 3-2. Overall Performance of ResNet.....</b>	<b>29</b>
<b>Figure 3-3. Model Result Confusion Matrix.....</b>	<b>29</b>
<b>Figure 3-4. Illustration of Residual Learning with shortcut connection .....</b>	<b>30</b>
<b>Figure 3-5. Illustration of ResNet 18 architecture .....</b>	<b>31</b>
<b>Figure 3-6. Illustration of Model Framework.....</b>	<b>32</b>
<b>Figure 3-7. Graph of sigmoid function.....</b>	<b>34</b>

<b>Figure 4-1. Examples of Multi-label construction images.....</b>	<b>38</b>
<b>Figure 4-2. Experiment Result: Cross-entropy Loss .....</b>	<b>40</b>
<b>Figure 4-3. Experiment Result: Test Accuracy .....</b>	<b>40</b>

# Chapter 1. Introduction

## 1.1. Research Background

In the context of the construction industry, a significant amount of image data is produced throughout the entire life cycle of the construction project. In particular, the advent and development of digital photographing equipment such as cameras and unmanned aerial vehicles have helped construction project practitioners readily acquire visual records of construction sites daily (K. K. Han & Golparvar-Fard, 2017). Thus, an ever-increasing amount of construction activities are captured in the forms of still images, time-lapse images, and videos (Hamledari, McCabe, & Davari, 2017). As a result, construction activities are now easily and periodically documented at a low cost.

As such visual data explicitly captures the exact state of construction job-site, they contain various essential project-related information including 1) the type of equipment and worker trades of on-going operations, 2) the number of equipment or workers, and 3) the states of construction activities (Zhu, Ren, & Chen, 2017). Based on the information retrieved from visual content of construction images, researchers have demonstrated an opportunity to alleviate construction project practitioners from such cumbersome tasks such as progress tracking and control (Golparvar-Fard, Peña-Mora, Arboleda, & Lee, 2009; Omar,

Mahdjoubi, & Kheder, 2018), productivity analysis and improvement (J. Kim, Chi, & Seo, 2018; Yang, Park, Vela, & Golparvar-Fard, 2015), surveillance of construction operation for safety and quality control (Ding et al., 2018; Dung & Anh, 2019; S. Han & Lee, 2013), resource management (Jog, Brilakis, & Angelides, 2011), supporting contractual claim documents (Kangari, 1995) and better communication among stakeholders (Golparvar-Fard et al., 2009; Teizer, 2009), and education and training (Azar, 2017).

Despite their availability and effectiveness, visual resources are, however, not used to their full potential. Instead, most of the visual data are likely to be unutilized soon because image search and information retrieval are challenging due to unorganized and scattered images in the system. Current information retrieval systems are mostly built on keyword-based content representation and query processing techniques (Lv & El-Gohary, 2016). Thus, it is very difficult for practitioners to search and identify the target image of interest through the large collections of project images unless an image is archived with adequate categorical descriptions or keywords annotated to the image. In other words, organizing construction images into operational-level categories that are meaningful to the project team is extremely useful and essential for proper and prompt image information retrieval.

Yet, the current annotation process heavily relies on manual observation and analysis (Brilakis & Soibelman, 2005). Taking consideration of the exploding

amounts of images that are regularly generated in the construction projects, even a seemingly trivial task of manual annotation can pose a burden on the project practitioners. Due to the time-consuming and labor-intensive process to analyze and label each image, the majority of valuable resources are instead remained unutilized, leaving room for better exploitation of visual resources.

In this regard, several image annotation tools and methods were proposed to support automating the annotation process. Unfortunately, those approaches remained as time-consuming and tedious tasks. Some degree of users' actions is still required to manually analyze the visual context of the image and provide proper annotations (Soltani, Zhu, & Hammad, 2016).

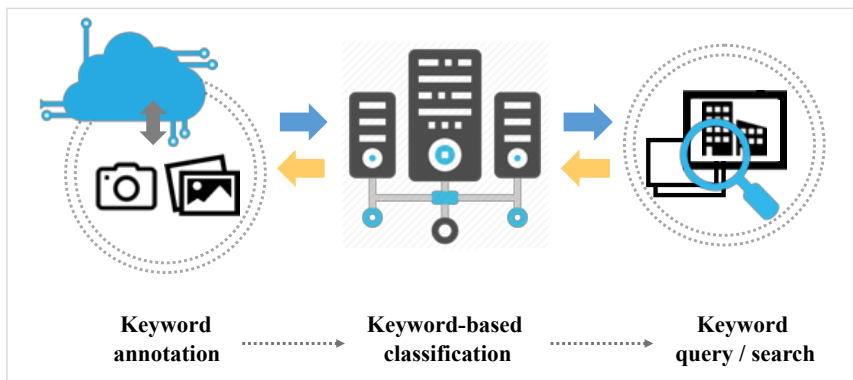


Figure 1-1. Overview of Keyword-based Digital Image Database System

Another approach to tackle this issue is to use image processing techniques, namely image classification, to facilitate annotation. Image classification refers to the task of identifying the target entity and assigning into one of the predefined semantic categories and is usually the preliminary step for understanding images

and assigning adequate annotations automatically. In this dissertation, a visual analytics approach is proposed to support automatic annotation process for construction image classification. This paper can help construction project practitioners fully utilize project image data for various laborious project management tasks by proposing an efficient way to manage a large volume of onsite project image data.



## 1.2. Research Objective and Scope

To accommodate the shortcomings of the current image classification process, the goal of this dissertation is to validate the performance of the current state-of-the-art computer vision technology, deep convolutional neural network, to automatically classify construction activities into semantic categories.

To achieve this goal, the following objectives are proposed:

- (1) To validate the feasibility of an end-to-end deep convolutional neural network model for construction image classification, making the annotation procedure more efficient and minimizing human intervention
- (2) To develop the multi-label classification model to produce associated multiple class labels for a single input image
- (3) To optimize an image classification model that is robust for both high intra-variability and generic characteristics across the appearance of different image classes

The proposed model is optimized for all activity classes and performs solely based on the input image without any external information. The scope of this study is thoroughly selected for a set of six work trades from architectural and structural activities: concrete, steel, masonry, tile, drywall, and curtainwall. They are reasonable representations of the dynamic nature of construction activities which possess intra-variability of wide-ranging appearances as well as common

features across different trades.

This study also assumes that the scope of image annotation is provided for activity trade keywords at WBS Level 1 and 2. The detailed descriptions of each category are as followed.

WBS 1	WBS 2
1) Concrete	1-1) Rebar work 1-2) Concrete Pour 1-3) Formwork
2) Masonry work	2-1) Red Brick 2-2) Concrete Block
3) Drywall work	3-1) Framing and insulation 3-2) Board installation
4) Tile work 5) Steel work 6) Curtainwall	

Figure 1-2. Overview of Classification Keyword Categories

It deals with higher operational level activity description and material types only, and any further elaboration of construction activities and entities such as pose of worker are not considered in this study.

### **1.3. Research Outline**

This dissertation consists of five chapters. The brief content of the following chapters is described as follows:

Section 2 examines the overview of the use of image data and the applications of computer vision algorithms in the construction domain. In particular, it describes the challenges related to construction image classification task. Then it introduces the previous applications of computer vision algorithms in the construction domain and investigates relevant issues of traditional algorithms in the context of construction activity classification. Finally, other researches using deep Convolutional Neural Network model for construction image classification were examined.

Section 3 explains the proposed architecture of image classification model and describes the framework for the proposed research, consisting of (1) customized image dataset preparation, (2) image classification model architecture selection, and (3) model training and validation in detail.

Section 4 further elaborates and evaluates the results of the experiments.

Section 5 summarizes the research finding, expected research contribution, and research limitation, and finally proposes future works.

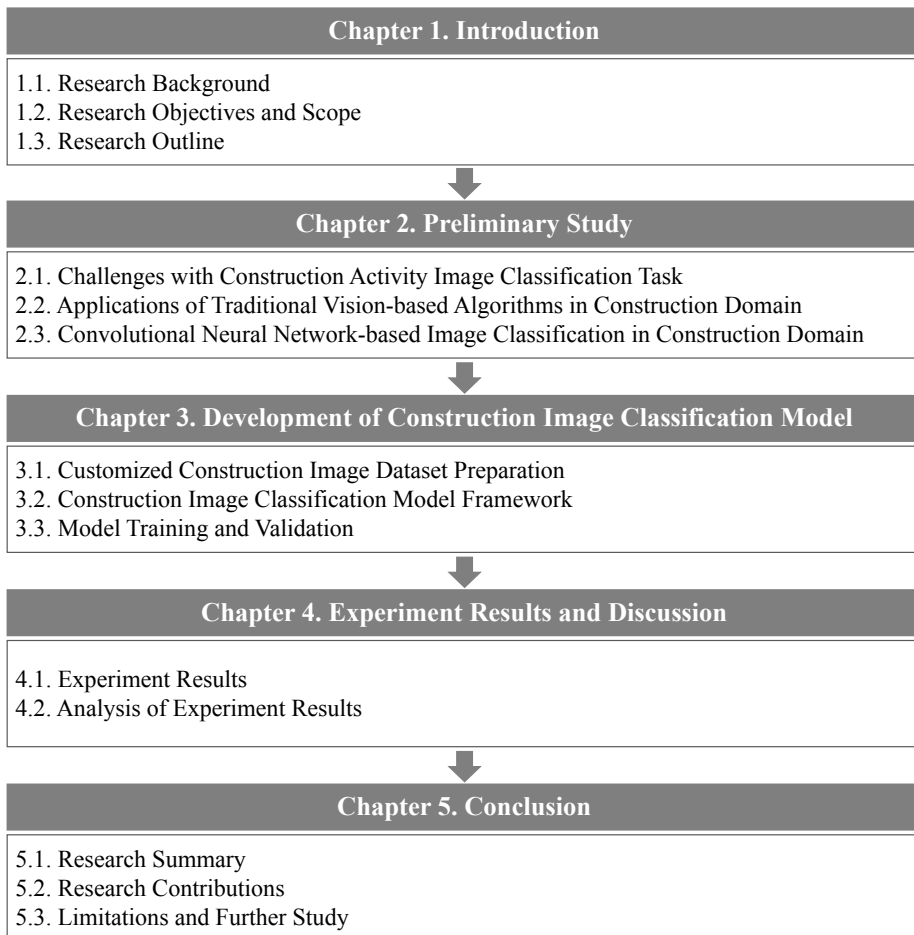


Figure 1-3. Overview of the Research

## **Chapter 2. Preliminary Study**

In the construction domain, most of the visual resources are unutilized due to the lack of efficient annotation methods, leaving room for better exploitation of visual data. In this chapter, previous annotation approaches as well as traditional computer vision algorithms applied in the construction domain are examined. Then, the limitations of these previous computer vision algorithms for construction activity classification tasks are scrutinized. After highlighting the need for improving model capacity to more robustly recognize high dimensional visual representations, the last part of this chapter describes the preceding applications of convolutional neural networks in the construction domain. By addressing the shortcomings of the CNN models, this research proposes a multi-label CNN model for construction activity classification.

## 2.1. Challenges of Construction Image Classification Task

Over the past decade, there have been an increasing number of researches that attempt to use visual analytics techniques in the construction domain as a result of two major forces: the prevalent construction image data generated in a cost-efficient way as well as the continuous development of computer vision algorithms. Nevertheless, early researches suffered from several challenges associated with construction images. Images taken from actual construction sites possess both intrinsic and extrinsic factors that preclude the high-performance rate which is easily observed in other benchmark datasets.

Inherently, construction images express high intra-class variability because a single construction activity class can have a wide range of variances in the appearance across different projects and even within a single project. Under the same work trade, the material texture and size can vary much from one to another because every construction project is unique. And more importantly, one activity can be presented in diverse configurations of construction entities as depicted in Figure 1. Each image will have different composite and interaction among workers, materials, equipment, and tools (Khosrowpour, Niebles, & Golparvar-Fard, 2014). Therefore, if the feature extractor is optimized at a particular project or a condition, the algorithms will not perform consistently in other projects which have different appearances (H. Kim, Kim, Hong, & Byun, 2018).



Figure 2-1. Examples of high intra-variability of masonry work with wide-ranging appearances and configuration

At the same time, algorithms are also required to deal with relatively low inter-class variability among different construction activities. Similar visual features can be shared among different activity classes. For instance, if the work processes are related or materials are similar in two work trades, such as concrete block wall and tile wall installation, the images captured from those trades will look very similar. Thus, the computer vision algorithms are required to learn distinct enough features for each trade while generalized enough to learn high intra-class variability simultaneously.

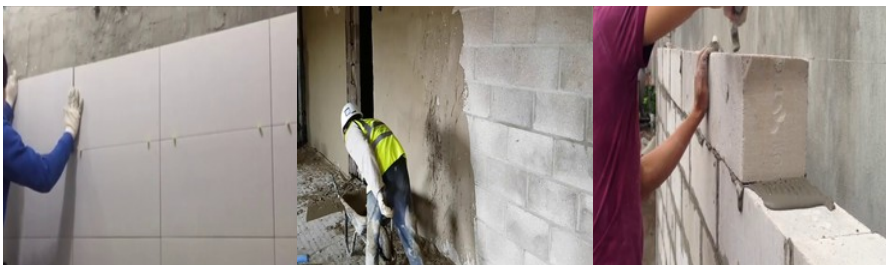


Figure 2-2. Examples of low inter-class variability among tile, plaster, and masonry works

Another unique challenge related to construction image classification task is the highly dynamic external factors associated with the construction

environment. Each project is surrounded by a unique yet continuously changing environment, which is subject to changes to lighting, viewpoints, and backgrounds. Under these dynamic conditions, images display construction entities that are often randomly cropped objects, partially self-occluded or occluded by other objects. As pointed by most of the previous researches, occlusion is still a major challenge for visual analytic task (Yang, Shi, & Wu, 2016).

As a result of construction images' intrinsic and extrinsic issues, it is very challenging to exploit generalized feature representations to classify vastly dynamic construction activities images for all projects. In Section 2.2., the limitations of early vision-based methods for construction activity classification task are scrutinized.



## 2.2. Applications of Traditional Vision-based Algorithms in Construction Domain

Over the past decade, machine-learning techniques have blossomed. Several pieces of research in the construction domain leveraged on computer vision-based algorithms in support of construction entity recognition – namely, construction workers, equipment and/or building components. Early vision algorithms were based on human-designed feature representations like shape, color, texture, gradient, and motion characteristics. They manually generate optimal feature descriptors based on the set of input data and learn the underlying pattern of the object appearance (Gong & Caldas, 2011; Zhu et al., 2017). Feature representations are then passed onto classifiers for classification and evaluated for the accuracy of the method.

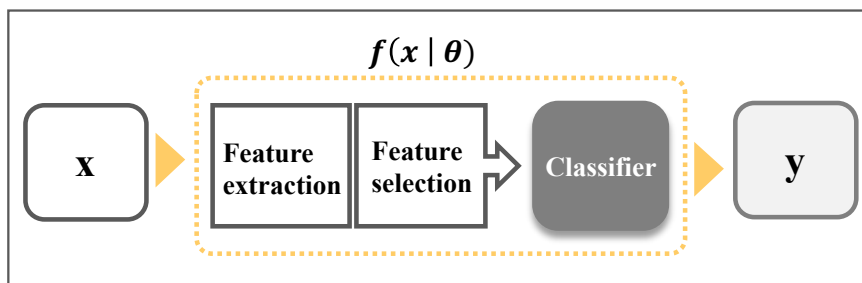


Figure 2-3. Illustration of Traditional Vision-Based Algorithms Process

The common algorithms are Harris detector (Harris & Stephens, 1988), scale-invariant feature transform (SIFT) (Lowe, 2004), histogram of oriented gradients (HOG) (N. Dalal & Triggs, 2005), histogram of oriented optical flow

(HOF) (Navneet Dalal, Triggs, & Schmid, 2006), and deformable part-based model (DPM) (Felzenszwalb, Girshick, & McAllester, 2010).

In the construction domain, several researchers have facilitated on the aforementioned computer vision-based algorithms for construction entity recognition task. For example, Gong et al. (2011) proposed classification module to classify worker and heavy equipment from video using Harris detector as the feature detector, local histograms as the feature representation, Bag-of-Words as the feature model, and Bayesian network models as the learning mechanism for action learning and classification (Gong, Caldas, & Gordon, 2011). Park and Brilakis (2012) detected construction workers wearing safety vests based on the histograms of color features after background subtraction. (Park & Brilakis, 2012). Memarzadeh et al. (2013) detected construction equipment and workers from construction site images with HOG descriptor and SVM classifier by extracting features from the histograms of oriented gradients and colors (Memarzadeh, Golparvar-Fard, & Niebles, 2013). Khosrowpour et al. (2014) detected and tracked workers' body skeleton from a sequence of image and then classified the stage of interior wall activities with a bag-of-worker pose (Khosrowpour et al., 2014). Park et al. (2015) also detected workers wearing a hardhat using a histogram of oriented gradients (HOG) and geometric relationships of the human body (Park, Elsafty, & Zhu, 2015). Hamledari et al. (2017) detected four partition components of indoor partition works with each

extracted visual feature and SVM and then infer the state of under-construction activities (Hamledari et al., 2017).

Table 2-1. Previous classification methods in the construction field

Article	Feature	Entity of Interest			Condition	
		ppl	bldg	eqmt	int	ext
Gong et al. (2011)	Harris detector; local histograms and Bag-of-Words			o		o
Park and Brilakis (2012)	Histograms of color features	o			o	o
Memarzadeh et al. (2013)	Histograms of oriented gradients and color features and SVM	o		o		o
Khosrowpour et al. (2014)	Bag-of-worker poses of spatio-temporal features	o			o	
Park et al. (2015)	Histogram of oriented gradients	o			o	
Hamledari et al. (2017)	Extracted visual feature and SVM		o		o	

In most of these studies, features were thoroughly selected based on the target problems and conditions because a particular feature is more appropriate for certain types of applications. Although they demonstrated acceptable performance rate for a specific task, these algorithms embody limited effectiveness for more generic tasks like identifying varied construction activities. Because these algorithms only learn low-level features instead of high dimensional features, they will not consistently perform on activity classification

due to the wide range of appearances and configuration of construction images.

To address these challenges of traditional human-engineered algorithms, researches have incorporated considerable domain knowledge and meticulous engineering to better define the problem. Nevertheless, it still does not show consistent performance when the designated visual cues are jeopardized. For feature extractors in which color plays the key role, the performance level is largely hampered by the presence of color of entities in the image, such as workers' clothes and backgrounds. If workers were not wearing fluorescent safety vests, wearing hardhat color that was not shown during training, or background color was similar to that of workers' clothes, the model accuracy can be undermined. Similarly, for orientation-feature extractors, the performance level is affected by the site's topography and spatial conflicts. The worker detection algorithms usually assume that the background is static and workers are at a certain posture like standing or walking. Thus, the image classification model will have an acceptable result only if the worker's full body is clearly presented.

In short, traditional approaches learn independent classifiers for each category and optimize for the specific classification task, and they do not show consistent performance for construction activity classification. Due to the limitation of the existing human-engineering methods whose performance is constrained to a particular task, early studies suffered from the low performance

in other tasks as increasing accuracy in one task may decrease the accuracy in other tasks (Zhu et al., 2017). Thus, it is necessary to employ more than one simple rule that learns low-level features to tackle construction image classification problem. In other words, the proposed classification model needs to acquire a predominant capacity to distinguish features in a high dimensional space of construction images. In the following section, a deep convolutional neural network model, which is well-known for its superior capacity for image classification, was introduced to cope with intrinsic and extrinsic issues of construction site images.

## 2.3. Convolutional Neural Network-based Image Classification in Construction Domain

Deep Neural Network models for image classification have been rapidly developed to the level of human recognition capability over the past years. Among deep neural network models, since the introduction of Le-Net in 1998 (Lecun, Bottou, Bengio, & Haffner, 1998), Convolution Neural Network (CNN) model has continuously proven its exceeding capacity for image classification to the level of human recognition capability (He, Zhang, Ren, & Sun, 2015; Krizhevsky, Sutskever, & Hinton, 2012, 2017; Simonyan & Zisserman, 2014; Szegedy et al., 2014; Zeiler & Fergus, 2013). Unlike the traditional human-designed feature algorithms, CNN models do not require explicit feature engineering.

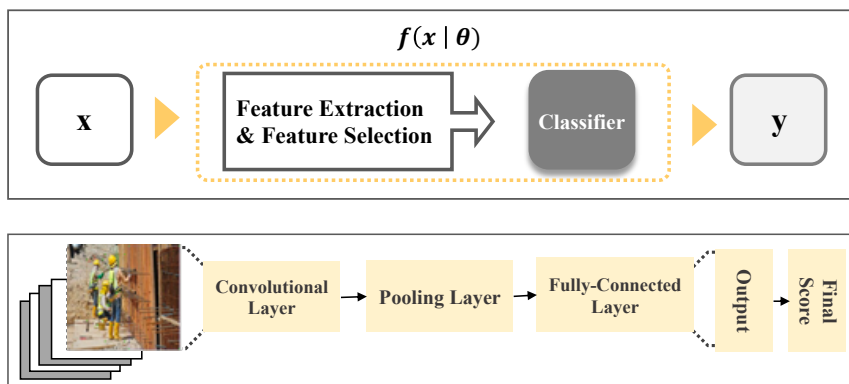


Figure 2-4. Illustration of the Deep CNN-Based Algorithms Process

Instead, it automatically learns the relationship between the underlying

representational features and high-level image semantics by conducting convolution operations on all pixels of the input image with learnable filters. After the convolutional operation, a feature map is produced for each operation and then activated by a nonlinear function. It helps with preserving spatial information as well as effectively discovering hidden visual features within high-dimensional datasets. A model can achieve even higher representation capacity by stacking the convolutional layers (Simonyan & Zisserman, 2014; Zeiler & Fergus, 2013).

As the most dominant model for visual recognition tasks, CNN models have been applied to automate various applications in the construction domain, as well. Ding et al. (2018) proposed a CNN-based model for safety control to detect unsafe behaviors of construction workers (Ding et al., 2018) and to detect the presence of personal safety protection like harness (Fang, Ding, Luo, & Love, 2018). Other researches also proposed CNN-based detection models for quality assessment such as automatic visual assessment for concrete defect detection (Beckman, Polyzois, & Cha, 2019; Cha, Choi, & Büyüköztürk, 2017; Dung & Anh, 2019) and fastener defect detection (Chen, Liu, Wang, Núñez, & Han, 2018). In terms of activity monitoring task, Son et al. (2019) used a state-of-the-art CNN model, Res-Net, for construction worker detection exposed to various poses (Son, Choi, Seong, & Kim, 2019) and Luo et al. (2018) monitored construction activities for steel reinforcement work by proposing an improved

CNN model that integrates RGB, optical flow and gray stream (Luo et al., 2018). Azar et al. (2017) also applied a convolutional neural network to the extracted keyframes of video data to automatically monitor heavy-equipments (Azar, 2017). These researches adopting CNN models demonstrated that they achieved improved performance rates for the given tasks compared to the early hand-crafted feature engineering methods.

However, only single label image classification model has been extensively studies over the past decades. There have not been enough researches in construction domain to detect more than one entity type or to extend the scope of classification to various trades of construction activities in the image. One reasonable explanation for the gap is that the CNN model suffers from its inability to handle multi-composition and multi-interaction of a single activity. CNN architectures handle each input image as one instance and encode an image as a dense one-dimensional vector through the final fully-connected (FC) layer.

Images taken from the construction sites are, however, likely to capture multiple activities, and they are required to be described by more than one semantic label. Thereby multi-label classification problem for construction image dataset is more useful yet challenging than the single label classification task. Thus, this study proposes to adopt multi-label image classification to get more semantic categorical labels for construction images.



## **2.4. Summary**

Due to the complex nature of construction activities, the previous vision-based approaches which learn low-level features failed to comprehensively understand construction image data. Thus, deep convolutional neural network models gained attention as an alternative computer vision algorithm in classifying construction site images. However, the preceding researches using CNN models focused on single-label classification, calling for a need for more practical model to be implemented in the actual site. In addressing the gap, a multi-label CNN model that can deal with classification tasks of the complex construction image dataset is proposed in the following section.

## **Chapter 3. Development of Construction Image Classification Model**

This dissertation aims to examine the feasibility of a CNN model for a multi-label image classification task for construction image dataset. As CNN algorithms are continuously developed, the models are better trained with deeper networks and better generalized with generalization methods. With appropriate methods, theoretically, CNN models are capable of representing more than one type of features, and the model can learn a multi-label representation of the image content (Nguyen, Yosinski, & Clune, 2016). In this study, an multi-label image classification model is proposed to classify an input image of structural and architectural activities without any additional sub-model. In this chapter, the model framework of data preparation, model selection, and model validation will be elaborated in detail.

### 3.1. Customized Construction Image Dataset Preparation

#### 3.1.1. Construction Activity Classification System

The proposed classification model aims to classify a construction image into the corresponding activity categories. In this study, the construction activity class label was determined according to a typical Work Breakdown Structure (WBS). WBS is a hierarchical structure which scopes and defines work activity as a manageable unit for planning estimating scheduling, and monitoring of activities, where Level 1 refers to upper-division like work trade and Level 2 refers to sub-division like work activities. MasterFormat is one of the international standards that is widely used to establish WBS for building trades and methods (Li & Lu, 2017). In this study, the dataset was composed of thirteen structural and architectural activity classes of WBS – six categories at WBS Level 1 and seven categories at Level 2, based on MasterFormat.

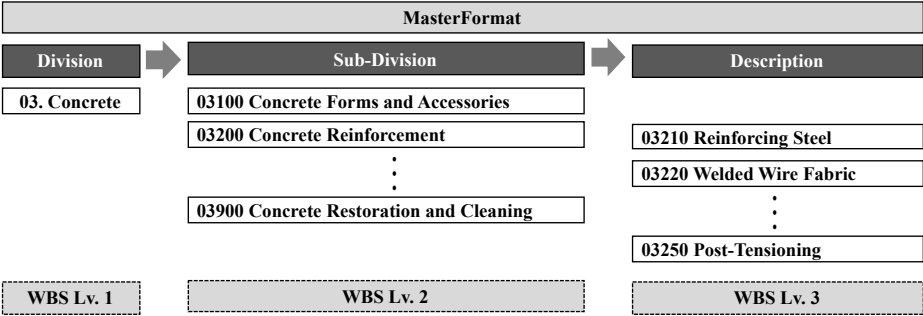


Figure 3-1. Example of Construction Activity Classification System

For each class, this research maintained a similar number of images – approximately 500 per class for WBS Level 1 and Level 2.

### **3.1.2. Dataset Collection**

Under a supervised image classification task, a model classifies images based on a set of labeled data of predefined classes. Because the performance of deep neural network model is highly dependent on the dataset, a customized dataset has carefully collected in six construction activity trades– concrete, steel, masonry, tile, drywall, and curtainwall. To assure as close to the actual construction project conditions as possible, this set of trades which exhibits a wide range of visual contents was chosen to properly demonstrate the inherently complex nature of construction images. Each image also contains a random composite of a worker, equipment and materials to demonstrate high intra-variability of each trade.

The main data sources are private construction project documents as well as open-source images search engine. For project-based data, images were acquired from project documents including, but not limited to, daily report, weekly/monthly progress report, meeting minutes, etc. For open-source data, both video clips and images were crawled from Google image, Flickr, Youtube, and other search engines. Keywords which used to search construction images are descriptions of construction activity such as structural steel lifting, steel erection, steel installation, etc. All crawled data were manually validated for its

appropriateness and any images which were not taken from an actual construction job site were excluded. The final completed dataset is a fair representation of construction activities of the wide range of variety in appearance worldwide.

Table 3-1. Dataset Composition

No.	WBS Level 1		No.	WBS Level 2	
	Activity Category	No. of data		Activity Category	No. of data
1	Concrete	3947	1-1	Formwork	135
			1-2	Rebar	280
			1-3	Concrete Pouring	210
2	Steel	3320			
3	Curtainwall	3012			
4	Masonry	3062	4-1	Red Brick	
			4-2	Concrete Block	
5	Tile	3248			
6	Drywall	3185	6-1	Framing and insulation	
			6-2	Board installation	
<b>Total No. of data :</b>		<b>23,714</b>	<b>Total No. of data :</b>		<b>625</b>

Finally, the dataset was split into training and validation sets randomly before model training. The training set was used to train the model, while the validation set was then used to tune model parameters. To evaluate the performance of the model, a new set of test set was prepared for all ten class.

### 3.1.3. Data Pre-Processing

After collecting a sufficient amount of image dataset, the dataset was pre-processed prior to model training. Since the dataset was collected from different sources, all multimedia data (MP4) were converted into an image format – JPG,

JPEG, PNG - to make the dataset into the same format. Then, they were resized into the same 256\*256 size with an identical color channel, RGB.

Because the performance of deep neural network models is highly related to the amount of dataset, data augmentation techniques were implemented in order to secure a suitable number of training data. The existing dataset was transformed by adding noise and applying affine transformations such as translation, zoom, flips, shear, mirror, color perturbation, and random crops.

Lastly, each image was assigned with correct labels according to the predefined activity class categories.

## **3.2. Construction Image Classification Model Framework**

In this dissertation, the proposed model was based on the use of a graphics processing unit (GPU) mode and CUDA 10.0 and was developed in the Linux Operating System (Ubuntu).

### **3.2.1. Multi-label Image Classification Model**

Real-world images are often associated with multiple labels than a single label. Especially, construction images are more likely to have more than one activity or attribute within a single image because construction activities are co-occurring simultaneously in its highly dynamic environment. Thus, multi-label classification can be more practical in the context of construction image classification. Therefore, in this study, a multi-label classification model is proposed to capture rich semantic information of construction images, such as the state of activity, the types of materials, and their interactions.

Similar to single label classification, multi-label image classification task also learns independent classifier for each category. Unlike single label classification, however, each image can belong to more than one class in the multi-label image classification task. The output of each class is not affected by other output values, and the overall classification result is determined by ranking or thresholding values. In this study, multi-label classification problem is transformed into multiple single-label classification problems, and converts the

result to a multi-label representation.

### 3.2.2. Base CNN Model Selection

The proposed model aims to examine the feasibility of the convolutional neural network model for classifying a set of distinguished activities of a wide range of various object configurations and appearances. From the modeling perspective, the framework attempts to train the algorithm to learn the multi-faceted representation of a single activity without any other external context information. Leveraging the state-of-the-art deep learning models, a supervised Convolutional Neural Network was implemented to classify the predefined set of activities presented in various circumstances. In order to select the most suitable CNN model for construction activity recognition, the currently available CNN models, namely AlexNet, VGGNet, ResNet, and Inception were all examined for their expected performance for a single-label classification task.

Table 3-2. Comparison of CNN Models' Performances

	WBS 1								WBS 2			
	Avg	1	2	3	4	5	6	7	Avg	CP	FW	RB
<b>Alexnet</b>	0.995	0.987	1.000	0.998	0.957	0.998	1.000	1.000	0.735	0.676	0.700	0.791
<b>Vggnet</b>	0.979	0.975	0.998	0.990	0.884	0.981	1.000	0.961	0.715	0.764	0.400	0.812
<b>ResNet</b>	0.964	0.948	1.000	0.974	0.824	0.959	1.000	0.941	0.637	0.617	0.250	0.812
<b>Inception</b>	0.944	0.936	0.987	0.928	0.727	0.964	1.000	0.912	0.696	0.735	0.100	0.916

Although most of the CNN models demonstrated acceptable classification performance, ResNet was selected for its acceptable performance in both WBS 1 and 2 dataset, with the overall accuracy of 81.5%.



$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = 81.5\%$	TP : correctly classified image FP : wrongly classified as true FN : wrongly not classified as true
$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = 81.5\%$	

Figure 3-2. Overall Performance of ResNet

Figure 3-3 shows the metrics that evaluated the performance of the proposed model based on average accuracy.

		Predicted Label							
		Precision	0.78	0.88	0.84	0.64	0.84	0.90	0.82
Ground Truth Label	Recall	0.77	327	13	11	24	26	13	10
	0.84	11	494	18	13	17	10	25	
	0.83	13	14	516	15	11	10	40	
	0.61	37	9	13	157	15	9	17	
	0.92	9	8	15	12	647	7	9	
	0.91	10	11	13	12	9	522	7	
	0.83	13	14	27	11	45	8	503	

Figure 3-3. Model Result Confusion Matrix

### 3.2.3. Proposed ResNet Model Architecture

In this study, ResNet, or Residual Neural Network, model was selected as the basic architecture for its superior performance in the single label

classification task. Among other CNN models, ResNet is especially powerful dealing with overfitting issue, which is a common problem of deep learning models as their network goes deeper. In general, the information passed throughout the network often cannot be directly propagated from the deeper layers to shallow layers. ResNet architecture, however, handles this degrading problem by introducing residual learning with shortcut connection, or identity mapping (He et al., 2015).

$$y = F(x, \{W_i\}) + x \quad (1)$$

The identity shortcut is an additional function that allows direct connection to the next block; thus, it successfully extracts feature maps via very deep residual networks.

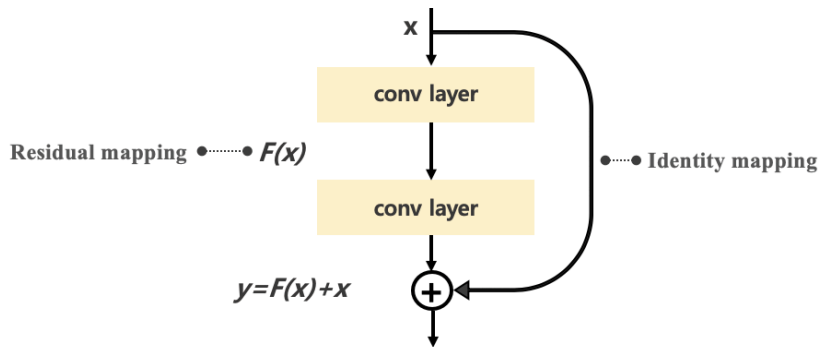


Figure 3-4. Illustration of Residual Learning with shortcut connection

Details of our model architecture were then carefully selected via an

extensive experiments and trial and error. The result showed that the model performance had positive correlation with the model complexity and negative correlation with the number of datasets in general. Due to the limited availability of the dataset, the model complexity was determined in relation to the characteristics of the dataset.

Finally, by arranging each block of CNN layers, ResNet 18 architecture which is consisted of a series of a convolutional layer, pooling layer followed by the activation function, as described in Figure 3-5, was finally chosen for its optimal model performance.

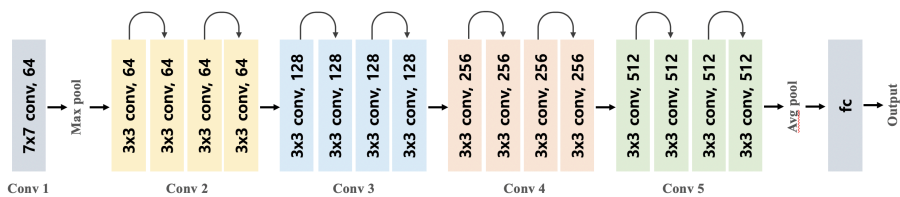


Figure 3-5. Illustration of ResNet 18 architecture

The overall framework for the proposed model is as followed.

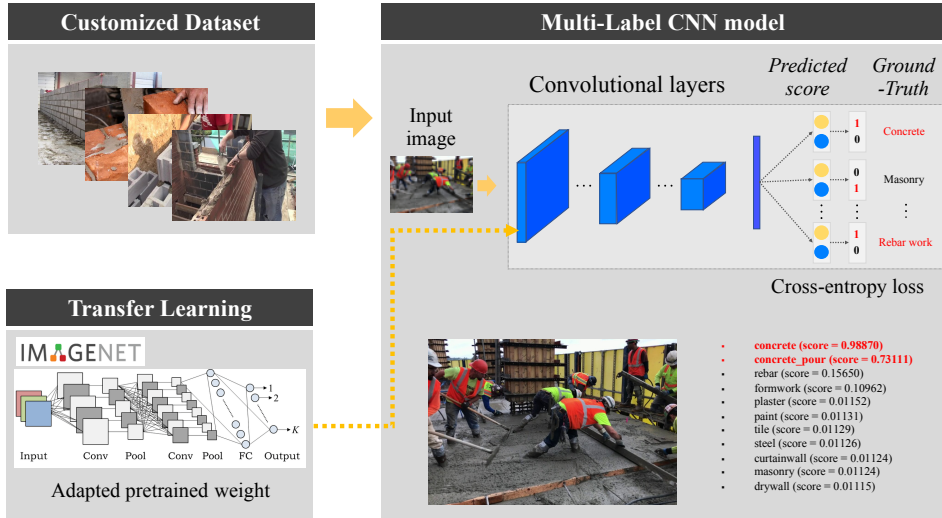


Figure 3-6. Illustration of Model Framework

## **3.3. Model Training and Validation**

### **3.3.1. Transfer Learning**

In addition to selecting the appropriate model architecture, transfer learning technique was implemented because the customized dataset has relatively smaller number of images. The performance of deep CNN models is highly dependent on the volume of dataset and their superior performance is guaranteed when there is abundant amount of data for training. With smaller dataset, therefore, transferring can be helpful by employing pre-learned knowledge. It usually refers to feature vector extracted from the last convolutional layer of a pre-trained model.

In this study, the proposed model was initialized with the pre-trained model weight which was trained on the ImageNet, an open-source large visual dataset designed for image processing, and applied fine-tuning strategies on the customized dataset. In this way, the model can avoid overfitting issue, which is a common issue for deep learning models with a relatively small number of the dataset.

### **3.3.2. Loss Computation and Model Optimization**

During model training phase, model loss computation and optimization were conducted as followed. In this model, the input image label is represented in k-dimensional binary vector, where  $y_i = 1$  if the label is the correct instance

and  $y_i = 0$  otherwise.

$$Y_i = (y_1, y_2, \dots, y_k) \quad (2)$$

As the proposed model is multi label classification problem, each output is a valid target label. The output vector is going through sigmoid activation function to squash each vector in the range between 0 and 1. This expresses the class probability of how much the image belongs to a class with the probability.

$$f(s_i) = \frac{1}{1 + e^{-s_i}} \quad (3)$$

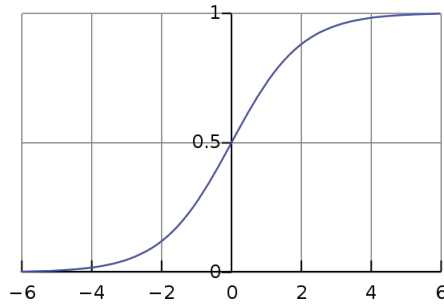


Figure 3-7. Graph of sigmoid function

Since this study translated a multi label classification problem into a set of single label classification problems, binary cross entropy loss is used to compute loss. The Cross-entropy loss is defined as:

$$L(w) = -y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n) \quad (4)$$

Each label output was compared with the ground-truth image label. The

cross-entropy loss increases when the predicted probability diverges from the actual label. During test time, labels above the given threshold can be chosen for the correct label.

Finally, the model was fine-tuned via weight update for optimization. In general, there are a number of parameter-updating strategies such as Stochastic gradient descent (SGD) combined with the momentum method, AdaDelta, AdaGrad, Nesterov. In this study, SGD showed the best validation performance by updating parameters using a portion of the sample parameters at one time (Wang et al., 2019). In the weight update process, hyperparameters of momentum ( $\gamma$ ) and learning rate ( $\eta$ ), or the step size of the weight update, are also required to decide to update velocity ( $v_t$ ), or the gradient of the loss function ( $\nabla_{\theta} J(\theta)$ ).

$$v_t = \gamma v_{t-1} - \eta \nabla_{\theta} J(\theta) \quad (5-1)$$

$$\theta = \theta - v_t \quad (5-2)$$

The model was trained at the hyperparameters of learning rate 0.001 and momentum 0.9 via trial-and-error.

### 3.3.3. Model Performance Indicator

Evaluating the multi-label prediction performance requires standardized measures and metrics. The performance of the proposed classification model was

assessed quantitatively for two main performance indicator, precision, and recall.

The precision and recall rates are defined as follows:

$$\begin{aligned} \textit{Precision} &= \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}} \\ \textit{Recall} &= \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}} \end{aligned} \quad (6)$$

True Positive refers to the number of correctly classified prediction, in which the target entity is correctly detected as the ground-truth class, whereas False Positive refers to the number of target entity incorrectly detected as the ground-truth class. FN indicates the number non-target entity incorrectly detected as the ground-truth class. In other words, TP represents a target activity is detected when it actually occurs, FN represents that target activity is not detected even when it actually occurs, and FP represents the target activity does not occur, but other activities are detected as the target activity. High recall rate indicates that the majority of activities were correctly recognized by the algorithm.



### **3.4. Summary**

In this chapter, the framework of the proposed model was elaborated in details in the following order: 1) the customized dataset preparation, 2) ResNet-based model selection, and 3) model training and validation. Based on the ResNet model, real world images generated from the actual construction job site containing multiple labels are tested for classification tasks.

## Chapter 4. Experiment Results and Discussion

### 4.1. Experiment Results

This study constructed extensive experiments to validate the feasibility of the multi-label classification model. Figure 4-1 illustrates some sample images of multi-labels in this dataset.



Figure 4-1. Examples of Multi-label construction images:  
1) Steel and concrete works, 2) Masonry and tile works,  
3) Steel and masonry works, and 4) Curtainwall and concrete works

The initial model correctly identified the given construction activity class with the error rate of 21.9% during the validation phase. The error rate is much higher than that of the benchmark image classification models – approximately

less than 5%. By assessing the results, it was found that most of the misclassified error occurred with WBS Level 2 activities, in failing to distinguish images with smaller dataset. To resolve this discrepancy, the dataset was reassigned to WBS Level 2 categories and some of the image data in the WBS Level 1 categories was discarded as follows.

Table 4-1. Revised Dataset Composition

No.	WBS Level 1		No.	WBS Level 2	
	Activity Category	No. of data		Activity Category	No. of data
1	Concrete	549	1-1	Formwork	180
			1-2	Rebar	182
			1-3	Concrete Pouring	187
2	Steel	535			
3	Curtainwall	503			
4	Masonry	521	4-1	Red Brick	266
			4-2	Concrete Block	255
5	Tile	595			
6	Drywall	516	6-1	Framing and insulation	259
			6-2	Board installation	257
<b>Total No. of data :</b>		<b>3,219</b>	<b>Total No. of data :</b>		<b>1,586</b>

Consequently, the experiment result was improved, achieving final test accuracy of 91.7% as illustrated in Figure 4-2 and Figure 4-3. Although multi-label image classification tasks included both WBS Level 1 and 2 categories, the result for multi-label image classification showed a superior performance overall.

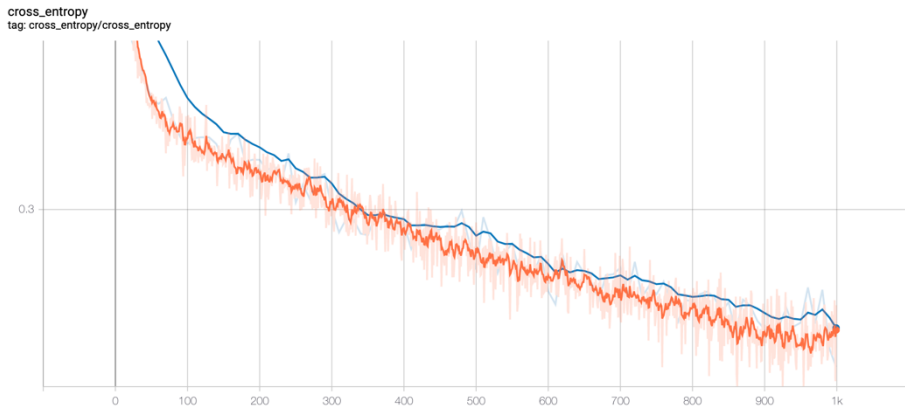


Figure 4-2. Experiment Result: Cross-entropy Loss

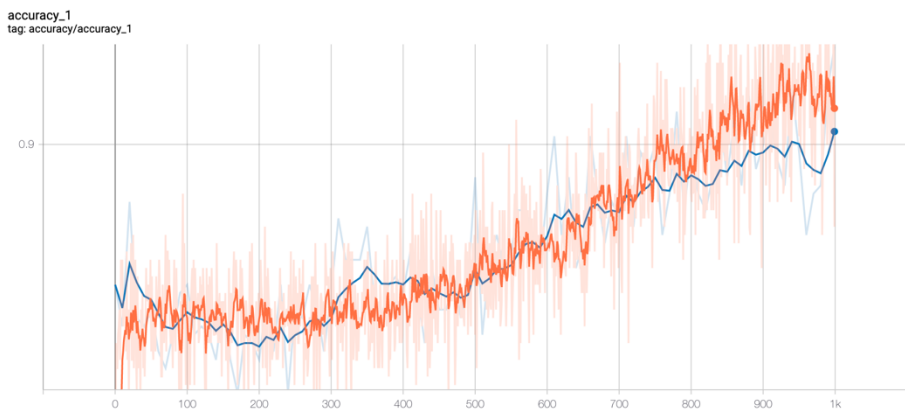




Figure 4-3. Experiment Result: Test Accuracy

With the enhanced model, the model was able to improve construction image classification performance in both single label and multi label classification, as depicted in Table 4-2.

Table 4-2. Examples of Correct Test Example

Input Image	Activity trade	Score
	<b>concrete</b> <b>steel</b> masonry formwork concrete_pour rebar red_brick conc_block curtainwall tile drywall frame_insulation gypsum_board	<b>0.41825</b> <b>0.24857</b> 0.18955 0.09871 0.09840 0.09650 0.08567 0.08347 0.07597 0.03703 0.03400 0.03234 0.02360
	<b>concrete</b> <b>concrete_pour</b> masonry conc_block rebar tile curtainwall red_brick steel formwork drywall frame_insulation gypsum_board	<b>0.48652</b> <b>0.19819</b> 0.19033 0.10578 0.09314 0.09173 0.06258 0.05781 0.05265 0.03307 0.03203 0.02658 0.01960

In short, the model was able to successfully classify the given input images correctly, in both single label and multi label classification.

## 4.2. Analysis of Experiment Results

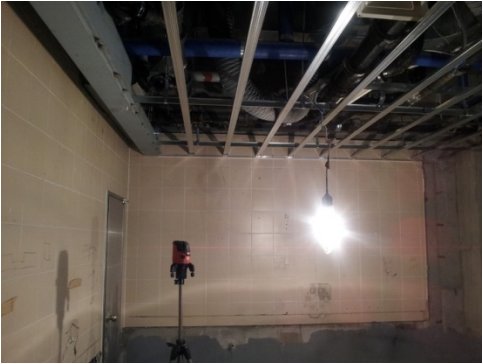
In this section, the experiment result is further discussed. In comparing the result of each class, tile work, concrete block of masonry work, red brick of masonry work, and drywall work showed higher performance, with the highest accuracy of 97.9%. On the other hand, rebar of concrete work, steel work, and concrete pour of concrete work showed lower performance with the lowest accuracy of 85.0%. Although the overall accuracy of the classification model was acceptable, some of the incorrectly classified examples exhibited the shortcomings of the proposed model.

First of all, some of the incorrect results revealed that the proposed model suffered from the lack of understanding the hierarchical structure of the construction activities. The proposed model structure treats each label independently and is incapable of learning the correlations or dependency among multiple labels. Therefore, it fails to distinguish different activities possessing similar visual features in spite of their distinctive conditions.

For example, as shown in Table 4-3, the model misunderstood the framing work for internal gypsum board installation as structural steel work, even after it correctly classified drywall label with the highest probability. If the model learned the hierarchical interaction among activities and understood the fact that steel structure occurred with drywall was not structural steel work but framing

work for drywall, the model could demonstrate higher performance.

Table 4-3. Examples of Incorrect Test Example

Input Image	Activity trade	Score
	<p><b>drywall</b> steel <b>frame_insulation</b> concrete <b>tile</b> gypsum_board masonry curtainwall conc_block concrete_pour red_brick formwork rebar</p>	<p><b>0.29861</b> 0.17720 <b>0.13960</b> 0.11028 <b>0.10594</b> 0.10197 0.09238 0.08544 0.06334 0.06040 0.05751 0.05093 0.04689</p>

### **4.3. Summary**

In this chapter, the proposed multi-label image classification model performance was assessed and the result showed that multi-label classification task performed as reliable as the single label classification task. In the following section, the research summary will be elaborated.



## **Chapter 5. Conclusion**

### **5.1. Research Summary**

This dissertation presented a Convolutional Neural Network model to automatically understand the visual content of construction site images and assign into relevant categories accordingly. Most of the dataset included actual construction site photos which composed random composites of a worker, equipment and materials in six different work activities— concrete, steel, masonry, tile, drywall, and curtainwall. Due to the inherently complex nature of the construction site, the dataset has imposed difficulties that challenged activity recognition such as occlusion, randomly cropped images with multi-viewed and multi-scaled representations. In order to address the challenges posed to construction image classification task, this study conducted a series of experiment to select the model architecture. As a result, the experiment result showed an accuracy of 91%. Although this result still underperforms compared to the current state-of-art computer vision model in other domains, it satisfies the minimum acceptable range for image classification and demonstrated a reasonable performance for classifying construction activity image dataset with a wide range of appearance variance.

## 5.2. Contribution

The contributions of this paper can be summarized as follows:



- To the best of my knowledge, this study proposed the first multi-label image classification model for construction image dataset.
- The feasibility of the state-of-the-art deep convolutional neural network models for comprehensive construction activity recognition was validated as a promising way of automatically classifying construction activities
- The customized dataset can be used as a reference dataset for future projects.
- From a practical standpoint of the construction industry, it can provide a more efficient and reliable way to classify and annotate construction activities, alleviating cumbersome tasks.
- It will ultimately allow construction projects to quickly retrieve project-related information for various construction management tasks, enhancing the usability of visual data.

### 5.3. Limitations and Further Study

Despite the effort, there are still several challenges to be further addressed in this study. First of all, our model structure treats each label independently and is incapable of learning the correlations or dependency among multiple labels. Some of the incorrect results revealed that the proposed model suffered from strong label co-occurrence dependencies. To deal with this issue, our model can learn the hierarchical structure among semantic elements in images based on the WBS. For future study, this study proposes to leverage the interactions and correlations among construction entities and activities by learning the embedded hierarchical structure of construction image dataset.

Another limitation of the proposed model is that it failed to recognize certain images taken in the early phase because they did not embed with sufficient features to be correctly classified. For instance, early plastering work are in fact more like masonry work than plastering work. In order to improve this drawback, sequential information should be additionally provided at each stage.

Table 5-1. Examples of Misclassified Early-phased Images

Misclassified as plaster work	Misclassified as tile work
 <p data-bbox="422 1532 577 1628">Tile work with plastered wall presented</p>	 <p data-bbox="875 1524 1085 1653">Plaster work with no plastering yet (only masonry work presented)</p>

In addition, the model can provide more detailed information by analyzing the visual contents of construction images. In future, it is planned to further extend the proposed method to more diverse construction entities and attributes at WBS Level 3.

## References

- Azar, E. R. (2017). Semantic Annotation of Videos from Equipment-Intensive Construction Operations by Shot Recognition and Probabilistic Reasoning. *Journal of Computing in Civil Engineering*, 31(5), 04017042. doi:doi:10.1061/(ASCE)CP.1943-5487.0000693
- Beckman, G. H., Polyzois, D., & Cha, Y.-J. (2019). Deep learning-based automatic volumetric damage quantification using depth camera. *Automation in Construction*, 99, 114-124. doi:<https://doi.org/10.1016/j.autcon.2018.12.006>
- Brilakis, I., & Soibelman, L. (2005). Content-Based Search Engines for construction image databases. *Automation in Construction*, 14(4), 537-550. doi:<https://doi.org/10.1016/j.autcon.2004.11.003>
- Cha, Y.-J., Choi, W., & Büyüköztürk, O. (2017). Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361-378. doi:doi:10.1111/mice.12263
- Chen, J., Liu, Z., Wang, H., Núñez, A., & Han, Z. (2018). Automatic Defect Detection of Fasteners on the Catenary Support Device Using Deep Convolutional Neural Network. *IEEE Transactions on Instrumentation and Measurement*, 67(2), 257-269. doi:10.1109/TIM.2017.2775345
- Dalal, N., & Triggs, B. (2005, 20-25 June 2005). *Histograms of oriented gradients for human detection*. Paper presented at the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).

- Dalal, N., Triggs, B., & Schmid, C. (2006, 2006//). *Human Detection Using Oriented Histograms of Flow and Appearance*. Paper presented at the Computer Vision – ECCV 2006, Berlin, Heidelberg.
- Ding, L., Fang, W., Luo, H., Love, P. E. D., Zhong, B., & Ouyang, X. (2018). A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Automation in Construction*, 86, 118-124. doi:<https://doi.org/10.1016/j.autcon.2017.11.002>
- Dung, C. V., & Anh, L. D. (2019). Autonomous concrete crack detection using deep fully convolutional neural network. *Automation in Construction*, 99, 52-58. doi:<https://doi.org/10.1016/j.autcon.2018.11.028>
- Fang, W., Ding, L., Luo, H., & Love, P. E. D. (2018). Falls from heights: A computer vision-based approach for safety harness detection. *Automation in Construction*, 91, 53-61. doi:<https://doi.org/10.1016/j.autcon.2018.02.018>
- Felzenszwalb, P. F., Girshick, R. B., & McAllester, D. (2010, 13-18 June 2010). *Cascade object detection with deformable part models*. Paper presented at the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
- Golparvar-Fard, M., Peña-Mora, F., Arboleda, C. A., & Lee, S. (2009). Visualization of Construction Progress Monitoring with 4D Simulation Model Overlaid on Time-Lapsed Photographs. *Journal of Computing in Civil Engineering*, 23(6), 391-404. doi:doi:10.1061/(ASCE)0887-3801(2009)23:6(391)
- Gong, J., & Caldas, C. H. (2011). An object recognition, tracking, and contextual

reasoning-based video interpretation method for rapid productivity analysis of construction operations. *Automation in Construction*, 20(8), 1211-1226. doi:<https://doi.org/10.1016/j.autcon.2011.05.005>

Gong, J., Caldas, C. H., & Gordon, C. (2011). Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. *Advanced Engineering Informatics*, 25(4), 771-782. doi:<https://doi.org/10.1016/j.aei.2011.06.002>

Hamledari, H., McCabe, B., & Davari, S. (2017). Automated computer vision-based detection of components of under-construction indoor partitions. *Automation in Construction*, 74, 78-94. doi:<https://doi.org/10.1016/j.autcon.2016.11.009>

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

Han, S., & Lee, S. (2013). A vision-based motion capture and recognition framework for behavior-based safety management. *Automation in Construction*, 35, 131-141. doi:<https://doi.org/10.1016/j.autcon.2013.05.001>

Harris, C., & Stephens, M. (1988). *A combined corner and edge detector*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv e-prints*. Retrieved from <https://ui.adsabs.harvard.edu/#abs/2015arXiv151203385H>

- Jog, G. M., Brilakis, I. K., & Angelides, D. C. (2011). Testing in harsh conditions: Tracking resources on construction sites with machine vision. *Automation in Construction*, 20(4), 328-337. doi:<https://doi.org/10.1016/j.autcon.2010.11.003>
- Kangari, R. (1995). Construction Documentation in Arbitration. *Journal of Construction Engineering and Management*, 121(2), 201-208. doi:doi:10.1061/(ASCE)0733-9364(1995)121:2(201)
- Khosrowpour, A., Niebles, J. C., & Golparvar-Fard, M. (2014). Vision-based workplace assessment using depth images for activity analysis of interior construction operations. *Automation in Construction*, 48, 74-87. doi:<https://doi.org/10.1016/j.autcon.2014.08.003>
- Kim, H., Kim, H., Hong, Y. W., & Byun, H. (2018). Detecting Construction Equipment Using a Region-Based Fully Convolutional Network and Transfer Learning. *Journal of Computing in Civil Engineering*, 32(2), 04017082. doi:doi:10.1061/(ASCE)CP.1943-5487.0000731
- Kim, J., Chi, S., & Seo, J. (2018). Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks. *Automation in Construction*, 87, 297-308. doi:<https://doi.org/10.1016/j.autcon.2017.12.016>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. Paper presented at the Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Lake Tahoe, Nevada.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6), 84-90.



doi:10.1145/3065386

- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. doi:10.1109/5.726791
- Li, D., & Lu, M. (2017). Automated Generation of Work Breakdown Structure and Project Network Model for Earthworks Project Planning: A Flow Network-Based Optimization Approach. *Journal of Construction Engineering and Management*, 143(1), 04016086. doi:doi:10.1061/(ASCE)CO.1943-7862.0001214
- Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91-110. doi:10.1023/B:VISI.0000029664.99615.94
- Luo, H., Xiong, C., Fang, W., Love, P. E. D., Zhang, B., & Ouyang, X. (2018). Convolutional neural networks: Computer vision-based workforce activity assessment in construction. *Automation in Construction*, 94, 282-289. doi:<https://doi.org/10.1016/j.autcon.2018.06.007>
- Lv, X., & El-Gohary, N. M. (2016). Semantic Annotation for Supporting Context-Aware Information Retrieval in the Transportation Project Environmental Review Domain. *Journal of Computing in Civil Engineering*, 30(6), 04016033. doi:doi:10.1061/(ASCE)CP.1943-5487.0000565
- Memarzadeh, M., Golparvar-Fard, M., & Niebles, J. C. (2013). Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction*, 32, 24-37.

doi:<https://doi.org/10.1016/j.autcon.2012.12.002>

Nguyen, A., Yosinski, J., & Clune, J. (2016). Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks. *arXiv e-prints*. Retrieved from <https://ui.adsabs.harvard.edu/#abs/2016arXiv160203616N>

Omar, H., Mahdjoubi, L., & Kheder, G. (2018). Towards an automated photogrammetry-based approach for monitoring and controlling construction site activities. *Computers in Industry*, 98, 172-182. doi:<https://doi.org/10.1016/j.compind.2018.03.012>

Park, M.-W., & Brilakis, I. (2012). Construction worker detection in video frames for initializing vision trackers. *Automation in Construction*, 28, 15-25. doi:<https://doi.org/10.1016/j.autcon.2012.06.001>

Park, M.-W., Elsafty, N., & Zhu, Z. (2015). Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction Workers. *Journal of Construction Engineering and Management*, 141(9), 04015024. doi:doi:10.1061/(ASCE)CO.1943-7862.0000974

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*. Retrieved from <https://ui.adsabs.harvard.edu/#abs/2014arXiv1409.1556S>

Soltani, M. M., Zhu, Z., & Hammad, A. (2016). Automated annotation for visual recognition of construction resources using synthetic images. *Automation in Construction*, 62, 14-23. doi:<https://doi.org/10.1016/j.autcon.2015.10.002>

Son, H., Choi, H., Seong, H., & Kim, C. (2019). Detection of construction

workers under varying poses and changing background in image sequences via very deep residual networks. *Automation in Construction*, 99, 27-38. doi:<https://doi.org/10.1016/j.autcon.2018.11.033>

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv e-prints*. Retrieved from <https://ui.adsabs.harvard.edu/#abs/2014arXiv1409.4842S>

Teizer, J. S. B. a. J. (2009). Benefits and Barriers of Monitoring Construction Activities Using Hi-Resolution Automated Cameras *Building a Sustainable Future* (pp. 21-30).

Wang, N., Zhao, X., Zhao, P., Zhang, Y., Zou, Z., & Ou, J. (2019). Automatic damage detection of historic masonry buildings based on mobile deep learning. *Automation in Construction*, 103, 53-66. doi:<https://doi.org/10.1016/j.autcon.2019.03.003>

Yang, J., Park, M.-W., Vela, P. A., & Golparvar-Fard, M. (2015). Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. *Advanced Engineering Informatics*, 29(2), 211-224. doi:<https://doi.org/10.1016/j.aei.2015.01.011>

Yang, J., Shi, Z., & Wu, Z. (2016). Vision-based action recognition of construction workers using dense trajectories. *Advanced Engineering Informatics*, 30(3), 327-336. doi:<https://doi.org/10.1016/j.aei.2016.04.009>

Zeiler, M. D., & Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *arXiv e-prints*. Retrieved from

<https://ui.adsabs.harvard.edu/#abs/2013arXiv1311.2901Z>

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

# Appendix

## A. Test Results



**concrete (score = 0.41825)**  
**steel (score = 0.24857)**  
masonry (score = 0.18955)  
formwork (score = 0.09871)  
concrete\_pour (score = 0.09840)  
rebar (score = 0.09650)  
red\_brick (score = 0.08567)  
conc\_block (score = 0.08347)  
curtainwall (score = 0.07597)  
tile (score = 0.03703)  
drywall (score = 0.03400)  
frame\_insulation (score = 0.03234)  
gypsum\_board (score = 0.02360)



**concrete (score = 0.40656)**  
**steel (score = 0.25776)**  
curtainwall (score = 0.19534)  
**rebar (score = 0.07826)**  
formwork (score = 0.06738)  
concrete\_pour (score = 0.04545)  
masonry (score = 0.03349)  
conc\_block (score = 0.02379)  
frame\_insulation (score = 0.01883)  
red\_brick (score = 0.01679)  
drywall (score = 0.01568)  
tile (score = 0.01034)  
gypsum\_board (score = 0.00735)



**drywall (score = 0.40170)**  
**tile (score = 0.22111)**  
**gypsum\_board (score = 0.20956)**  
**frame\_insulation (score = 0.07460)**  
 steel (score = 0.06819)  
 masonry (score = 0.06654)  
 conc\_block (score = 0.04752)  
 curtainwall (score = 0.04623)  
 concrete (score = 0.04292)  
 concrete\_pour (score = 0.04126)  
 red\_brick (score = 0.03122)  
 formwork (score = 0.01830)  
 rebar (score = 0.01663)



**masonry (score = 0.44017)**  
**tile (score = 0.28679)**  
**conc\_block (score = 0.19354)**  
 drywall (score = 0.16995)  
 red\_brick (score = 0.15141)  
 frame\_insulation (score = 0.08943)  
 concrete (score = 0.08831)  
 gypsum\_board (score = 0.08215)  
 steel (score = 0.04489)  
 rebar (score = 0.04486)  
 formwork (score = 0.04368)  
 concrete\_pour (score = 0.04259)  
 curtainwall (score = 0.02856)



**tile (score = 0.33733)**  
**masonry (score = 0.16704)**  
**drywall (score = 0.11186)**  
**conc\_block (score = 0.07686)**  
 curtainwall (score = 0.05583)  
 red\_brick (score = 0.05566)  
 gypsum\_board (score = 0.05124)  
 frame\_insulation (score = 0.04537)  
 concrete (score = 0.03986)  
 steel (score = 0.03503)  
 concrete\_pour (score = 0.02869)  
 rebar (score = 0.01912)  
 formwork (score = 0.01467)





**concrete (score = 0.49792)**  
**concrete\_pour (score = 0.22276)**  
 masonry (score = 0.21888)  
 steel (score = 0.12300)  
 curtainwall (score = 0.12252)  
 red\_brick (score = 0.11350)  
 formwork (score = 0.10908)  
 conc\_block (score = 0.09902)  
 drywall (score = 0.09591)  
 rebar (score = 0.07315)  
 frame\_insulation (score = 0.06921)  
 gypsum\_board (score = 0.04615)  
 tile (score = 0.04264)



**concrete (score = 0.74265)**  
**rebar (score = 0.27767)**  
 concrete\_pour (score = 0.10837)  
 curtainwall (score = 0.07548)  
 steel (score = 0.06284)  
 masonry (score = 0.06031)  
 formwork (score = 0.04333)  
 red\_brick (score = 0.03566)  
 conc\_block (score = 0.03158)  
 tile (score = 0.03087)  
 drywall (score = 0.02391)  
 frame\_insulation (score = 0.02175)  
 gypsum\_board (score = 0.01256)



**concrete (score = 0.48652)**  
**concrete\_pour (score = 0.19819)**  
 masonry (score = 0.19033)  
 conc\_block (score = 0.10578)  
 rebar (score = 0.09314)  
 tile (score = 0.09173)  
 curtainwall (score = 0.06258)  
 red\_brick (score = 0.05781)  
 steel (score = 0.05265)  
 formwork (score = 0.03307)  
 drywall (score = 0.03203)  
 frame\_insulation (score = 0.02658)  
 gypsum\_board (score = 0.01960)



**concrete (score = 0.64867)**  
**formwork (score = 0.13908)**  
 curtainwall (score = 0.12077)  
 concrete\_pour (score = 0.11303)  
 masonry (score = 0.08766)  
 steel (score = 0.06859)  
 rebar (score = 0.06656)  
 red\_brick (score = 0.04554)  
 drywall (score = 0.03783)  
 conc\_block (score = 0.03758)  
 frame\_insulation (score = 0.03353)  
 gypsum\_board (score = 0.01452)  
 tile (score = 0.01314)



**drywall (score = 0.61944)**  
**gypsum\_board (score = 0.29135)**  
**frame\_insulation (score = 0.24611)**  
 tile (score = 0.18861)  
 masonry (score = 0.14780)  
 curtainwall (score = 0.10933)  
 red\_brick (score = 0.09487)  
 conc\_block (score = 0.08942)  
 steel (score = 0.08179)  
 concrete (score = 0.07050)  
 rebar (score = 0.05674)  
 concrete\_pour (score = 0.05427)  
 formwork (score = 0.05023)



**masonry (score = 0.79909)**  
**red\_brick (score = 0.48250)**  
 conc\_block (score = 0.23826)  
 concrete (score = 0.14031)  
 tile (score = 0.09036)  
 steel (score = 0.07966)  
 formwork (score = 0.06930)  
 concrete\_pour (score = 0.05592)  
 curtainwall (score = 0.04663)  
 rebar (score = 0.04632)  
 frame\_insulation (score = 0.03271)  
 drywall (score = 0.03251)  
 gypsum\_board (score = 0.02477)





**tile (score = 0.43469)**  
 masonry (score = 0.29774)  
 red\_brick (score = 0.14484)  
 concrete (score = 0.11136)  
 conc\_block (score = 0.09482)  
 concrete\_pour (score = 0.05056)  
 drywall (score = 0.04592)  
 curtainwall (score = 0.04362)  
 steel (score = 0.04195)  
 rebar (score = 0.04118)  
 frame\_insulation (score = 0.03573)  
 formwork (score = 0.03119)  
 gypsum\_board (score = 0.02812)



**curtainwall (score = 0.56194)**  
 steel (score = 0.21573)  
 concrete (score = 0.13461)  
 drywall (score = 0.08032)  
 frame\_insulation (score = 0.07896)  
 rebar (score = 0.07217)  
 concrete\_pour (score = 0.06331)  
 conc\_block (score = 0.05596)  
 formwork (score = 0.04681)  
 masonry (score = 0.04587)  
 tile (score = 0.03744)  
 gypsum\_board (score = 0.03305)  
 red\_brick (score = 0.02452)



**curtainwall (score = 0.31085)**  
 concrete (score = 0.12945)  
 masonry (score = 0.10183)  
 steel (score = 0.07173)  
 concrete\_pour (score = 0.04663)  
 conc\_block (score = 0.04363)  
 red\_brick (score = 0.03372)  
 tile (score = 0.02736)  
 formwork (score = 0.02424)  
 rebar (score = 0.02306)  
 drywall (score = 0.02262)  
 frame\_insulation (score = 0.01783)  
 gypsum\_board (score = 0.01192)



**steel (score = 0.66028)**

concrete (score = 0.33782)  
formwork (score = 0.22596)  
curtainwall (score = 0.21089)  
rebar (score = 0.09731)  
concrete\_pour (score = 0.08706)  
masonry (score = 0.07868)  
red\_brick (score = 0.06937)  
frame\_insulation (score = 0.06039)  
conc\_block (score = 0.05527)  
drywall (score = 0.04955)  
gypsum\_board (score = 0.03149)  
tile (score = 0.02379)



**steel (score = 0.42482)**

curtainwall (score = 0.36910)  
concrete (score = 0.24988)  
formwork (score = 0.11609)  
concrete\_pour (score = 0.08329)  
masonry (score = 0.05326)  
red\_brick (score = 0.04416)  
rebar (score = 0.04257)  
conc\_block (score = 0.03251)  
drywall (score = 0.02734)  
frame\_insulation (score = 0.02664)  
tile (score = 0.02315)  
gypsum\_board (score = 0.02129)

## 국 문 초 록

# 건설 현장 이미지 기반 다중 레이블 분류 자동화

최근 이미지 분석 기술이 발전함에 따라 건설 현장에서 다양한  
방면에서 현장에서 수집된 사진을 활용하여 건설 프로젝트를  
관리하고자 하는 시도가 이루어지고 있다. 특히 촬영 장비의  
발전되자 건설 현장에서 생산되는 사진의 수가 급증하여 건설 현장  
사진의 잠재적인 활용도는 더욱 더 높아지고 있다. 하지만 이렇게  
생산되는 많은 양의 사진은 대부분 제대로 분류되지 않은 상태로  
보관되고 있기 때문에 현장 사진으로부터 필요한 프로젝트 정보를  
추출하는 것은 매우 어려운 실정이다. 현재 현장에서 사진을  
분류하는 방식은 사용자가 직접 개별 사진을 검토한 뒤 분류하기  
때문에 많은 시간과 노력이 요구되고, 이미지 분류를 위한 특징을  
직접적으로 추출하는 기존의 이미지 분석 기술 역시 복잡한 건설  
현장 사진의 특징을 범용적으로 학습하는 데는 한계가 있다.

이에 본 연구에서는 건설 현장 사진의 모습이 매우 다양하고,  
동적으로 변하는 것에 대응하기 위해 이미지 분류에서 높은 성능을

보이고 합성곱 신경망(Dep Convolutional Neural Network) 알고리즘을 적용하여 개별 건설 현장 사진에 적합한 키워드를 자동으로 할당할 수 있는 모델을 개발하고자 한다. 합성곱 신경망 모델은 모델 구조가 깊어짐에 따라 높은 차원의 항상성(invariant) 특징도 효과적으로 학습할 수 있는 특징이 있기 때문에 복잡한 건설 현장 사진 분류 문제에 적합하다.

따라서 본 연구에서는 합성곱 신경망 모델을 토대로 현장에서 필요한 사진을 빠르고 정확하게 찾을 수 있도록 각 사진에 적합한 키워드를 자동으로 할당하는 모델을 개발하였다. 특히, 건설 현장 사진의 대부분이 하나 이상의 레이블과 연관이 있다는 점에 기반하여 다중 레이블 분류 모델을 적용하였다. 이를 통해 일차적으로는 건설 사진에서 프로젝트와 관련된 다양한 정보를 추출하여 건설 현장 사진의 활용도를 개선하고, 나아가 사진 데이터를 활용하여 효율적인 건설 관리를 도모하고자 한다.

본 연구의 진행 순서는 다음과 같다. 우선 모델을 학습시키기 위해서 실제 건설 현장 및 오픈소스 검색엔진을 통하여 총 6개 공종의 사진을 수집하고, 하위 분류 범위를 포함한 총 10개 레이블의 데이터셋을 구성하여 학습을 진행했다. 또한 구체적인 모델 선택을 위해 대표적인 합성곱 신경망 모델을 비교 검토하여 가장

우수한 성능을 보인 ResNet 18을 최종 모델로 선택했다. 실험 결과 평균 91%의 정확도를 보이며 건설 현장 사진을 자동으로 분류할 수 있는 가능성을 확인하였다.

또한 본 연구는 최근 타 분야 이미지 분석에서 좋은 성과를 보인 합성곱 신경망을 활용하여 건설 현장 사진을 자동으로 분류할 수 있다는 가능성을 확인했다는 점과, 건설 현장 사진 분류 문제에 다중 레이블 분류를 적용한 첫 연구라는 점에서 의의가 있다. 실제 현장에서는 사진을 자동으로 분류할 수 있게 됨에 따라 기존에 번거로운 수동 사진 분류 작업을 줄이고, 건설 현장 사진의 활용도를 높일 수 있을 것으로 기대된다.

하지만 본 연구는 각 레이블 간에 연관성이나 의존성을 고려하지 않기 때문에 추후 연구에서는 각 사진 간의 계층적 관계를 모델에 추가적으로 학습시켜 정확도를 높이고, 학습 레이블도 더 낮은 단계의 키워드까지 포함하여 현장 사진으로부터 보다 다양한 정보를 얻을 수 있도록 모델을 개선하는 것을 목표로 하고 있다.

**키워드:** 다중 레이블 이미지 분류, 현장 사진 데이터 관리, 합성곱 신경망 모델, 딥러닝

**학 번:** 2017-27421