Semantic annotation of images from outdoor construction sites

Layan Farahat[1] and Ehsan Rezazadeh Azar[1]

[1] Faculty of Engineering and Architectural Science, Toronto Metropolitan University, Canada
layan.farahat@torontomu.ca, ehsan.azar@torontomu.ca

Abstract –

Valuable information is embedded in construction images which can be used for different construction engineering and management purposes. The availability of low-cost cameras and robust artificial intelligence methods has increased the use of imaging technology in construction sites. However, these rich data sources are not often used to their full potential due to subjective documentation, leading to potentially overlooking valuable content. This study proposes an ensemble approach that utilizes deep learning techniques for object recognition, pixel-level segmentation, and text classification to annotate images from outdoor construction scenes at medium (ongoing activities) and high (project type) levels. Experimental results demonstrate the potential of this approach by achieving a 70% overall recall rate.

Keywords –

Annotation; construction images; construction management; deep learning; object detection; semantic segmentation.

1 Introduction

The construction industry has vastly employed image and video recording technologies, and a growing number of research projects investigate methods for better utilization of this valuable data [1]. This trend has been facilitated by the emergence of low-cost capturing systems, robust computer vision methods, and the flexibility in imaging offered by the UAV systems [1], [2]. The resulting image databases, typically organized through manual labels and metadata, could serve various project management purposes, such as progress tracking, quality inspection, safety audits, and training [1]. However, the unrestrained accumulation of these visual data poses challenges in annotation and retrieval, which potentially results in underutilizing valuable information [3]. Unlike captioning of generic images, the technical complexities of construction images demand expert knowledge for a practical annotation [4]. Past research focused on feature extraction and object detection to enhance image annotation [5]. Some explored equipment poses [6] and interactions [7], while others provided semantic annotation for construction videos through spatiotemporal data interpretation of equipment motion [4]. However, current annotations mainly cover appearing resources and their interactions, lacking identification of ongoing activities without visible actors. For examples, methods were developed to caption an image as “a dozer is pushing the soil on the ground” [7] or “the excavator is loading dirt to the truck”. [14]. But there is a gap to provide useful annotation where there is no actor, i.e. equipment and workers, in the image.

This research introduces an innovative ensemble method utilizing deep learning for low-level, medium-level, and high-level annotations of outdoor construction images, defining objects and materials as low-level, activities as medium-level, and project types as high-level annotations. The proposed approach integrates deep learning-based object detection, semantic segmentation, and text classification, focusing on outdoor construction images to demonstrate the potential advancements of this approach. Since indoor and outdoor construction scenery includes distinct elements and resources, this research only focuses on outdoor settings.

2 Literature Review

The construction industry has experienced widespread application of digital imaging since the 1990s, leading to a substantial increase in image-making rates [8]. Initial efforts focused on feature extraction for image retrieval based on material patterns and shapes, incorporating metadata like location and date [9]. Feature-based object detection techniques were used to detect construction equipment and workers, but they had mediocre performance and faced limitations in multiclass detections [10]. Deep convolutional neural networks (DCNN) addressed these limitations, offering multiclass recognition, automated feature extraction, and improved detection performance [11]. Some of the DCNN models, like Region-Based CNN (R-CNN) [12] and You Only Look Once (YOLO) [13], were employed in construction contexts for different tasks like progress monitoring [14]

and safety management [15]. DCNN models were also employed for semantic segmentation at pixel-level recognition. Methods like Mask R-CNN [16] and DeepLab V3+ were used for progress estimation [17], and understanding of construction visual data [18]. Efforts were made to produce semantic annotations by combining object recognition and Bayesian belief networks [4]. Long Short-Term Memory (LSTM) architecture [19] could be employed in conjunction with a CNN model for extracting semantic information from images. This approach contributed to safety management [20] and descriptive caption generation for equipment activities and interactions [7] by analyzing the interaction between two objects. Recent developments to caption construction images focus on actors and their action(s), and generally produce a caption as "an equipment/worker is doing something" [7, 14, 28]. Many construction images, however, do not contain equipment/workers and only capture a snapshot of the progressing construction work. This research seeks to address this gap by integrating detected objects and materials/surfaces to annotate construction images with ongoing activities and project types, even in cases where the main actors, such as equipment and workers, are not present in the images.

3 Research Methods

This study proposed an ensemble model, integrating three machine-learning models to predict general construction activities and project types in outdoor construction site images. The approach involved two deep-learning models for object detection and surface segmentation, and a third text classifier using a neural network for predicting construction activities and project type/phase. In this approach, the results of object detection and semantic segmentation, i.e. low-level annotations, are fed to the text classifier to predict medium-level and high-level annotations. Each model was trained and tested on separate datasets. Figure 1 shows a flowchart of the developed approach with detailed steps outlined in the subsequent sections, including results, discussion, and conclusion.

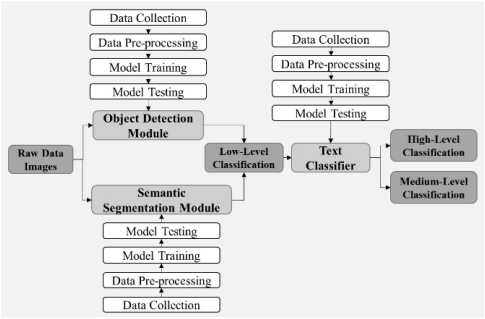

Figure 1. Flowchart of the Proposed Framework

3.1 Data Collection

This study used images of diverse construction projects collected from open online platforms like Google Images, YouTube videos, Pexels, and Pixabay to train and test the developed models. Some of the images were taken by the authors from construction sites using smartphones. Separate datasets were created for each model, tailored to their specific objectives. For the object detection model, 5,260 images featuring various construction equipment were collected. A dataset of 321 images was collected for the semantic segmentation model. The text classifier dataset comprised 545 images representing different types and phases of outdoor construction projects.

3.2 Object Detection Module

The object detection module was trained using 4,990 images (95% training and 5% validation) containing various construction equipment and workers. The training process involved preparing raw data, resizing images, and labelling objects using the Image Labeler tool by MATLAB [24]. Eleven classes were manually labelled with bounding boxes, including nine construction equipment types (excavator, bulldozer, scraper, off-road truck, truck, loader, compactor, grader, concrete mixer), in addition to two generic classes of humans and regular cars. The YOLO v4 with DarkNet53 as the backbone, pre-trained on the COCO public dataset, was chosen for its performance and processing time. The training was conducted on a desktop computer with 32 GB RAM, a 4.7 GHz Intel Core CPU, and an NVIDIA GeForce RTX 3060 GPU. Hyperparameters included a gradient decay factor of 0.9, a learning rate 0.001, and data augmentation with random horizontal reflection and scaling. Batches of 8 images were processed over 85 epochs using the MATLAB® 2022a software Deep Learning toolbox [24] for training and testing.

3.3 Semantic Segmentation Module

The second DCNN model employed semantic segmentation to classify image elements, like materials and construction surfaces, that might be impractical to be recognized by the object detection. This model classifies objects at the pixel level, with training and test datasets manually annotated using the same Image Labeler tool utilized for the object detection module. Twenty-one classes, including concrete, formwork, glass, tower crane, human, bin, dirt, equipment, lumber, asphalt, steel, rebar, scaffold, bitumen, aggregate, rail, waterproofing, pipe, curing blanket, brick, and other (such as sky, mountains, trees), were defined for labeling major elements in construction sites. The DeepLab V3+ ResNet50 [25] model was retrained using a labeled dataset of 220 images. The Deep Learning toolbox of MATLAB®

2022a software was used for training [24], adjusting stochastic gradient descent with a learning rate of 0.01 and a momentum value of 0.9. Batches of six images and 100 epochs were utilized, with data augmentation involving random X and Y translations and right/left pixel reflections. The training process was carried out on the same desktop computer mentioned earlier.

3.4 Text Classification Module

The last model in the system is the text classifier, which analyzes the outcomes of the object detection and semantic segmentation modules to predict medium-level (activities) and high-level (project types) annotations. This module was trained using a dataset of 385 images from a diverse set of construction projects. The objects, associated activities, and project phases/types were manually extracted from these images based on the authors' construction expertise. Then, these extracted textual data were manually converted into binary format, with 1 representing the presence and 0 indicating the absence of objects, surfaces, activities, or project types in the images. The dataset encompassed seven types of construction projects: building sub-structure, building super-structure, bridge construction, road construction, heavy construction (i.e., tunnel construction and subway construction), railway construction, and pipeline construction, and fifteen activities, including rebar installation, steel erection (structural), formwork shuttering and removal, concrete work, lumber work, earth hauling, paving work, material lifting, earthwork (i.e., loading, compacting, and removing dirt), excavation, masonry work, glazing, rail work, waterproofing, and pipework.

Since the neural network (NN) method has shown promising results in various construction-related analysis, such forecasting labor productivity [21], analyzing accidents [22], and project delay risks [23], it was also used in this research. NN models with multiple hidden layers and varying neuron numbers were developed using RapidMiner Studio [27], which offers various operators for data retrieval, model evaluation, and algorithms. The study trained and tested the model across nine configurations, with the most effective performance observed using two hidden layers, each with five neurons, and a Rectifier activation function over 100 epochs. The module training was conducted on a laptop with 16 GB RAM, 2.3 GHz Intel Core i7-11800H, and an NVIDIA GeForce RTX 3060 GPU.

3.5 Ensemble Model

The ensemble model combines all the mentioned modules, in which the trained object detection and semantic segmentation models extract low-level data, i.e. detected objects and material/surfaces, and pass them as input to the text classifier for medium-level and high-level annotations.

4 Experimental Results

The developed modules and the ensemble model were assessed in four phases: 1) the object detection module evaluation, 2) the semantic segmentation module evaluation, 3) the text classifier evaluation, and 4) the ensemble model evaluation, which encompasses all the modules and possible propagation of errors.

4.1 Object Detection Module Results

The object detection module was evaluated with 270 images from various construction site scenery gathered from the same sources as the training dataset. Model evaluation metrics included precision and recall rates. Precision is denoted as the proportion of correctly predicted positive instances (true positives) out of all predicted positives (true positives + false positives). Recall as the proportion of correctly predicted positive instances (true positives) out of all actual positive instances in the dataset. The 11 object classes were evaluated individually in addition to the overall performance. Table 1 presents the performance of each class by the trained YOLO V4 – DarkNet50 classifier.

Table 1. Performance metrics for each class

Classes	Recall	Precision
Excavator	80.28%	93.44%
Bulldozer	91.43%	87.67%
Scraper	76.92%	78.95%
Off truck	87.93%	72.86%
Truck	71.83%	77.27%
Loader	61.36%	57.45%
Human	54.10%	94.29%
Compactor	76.36%	66.67%
Grader	91.43%	78.05%
Car	84.00%	72.41%
Concrete Mixer	78.57%	61.11%

The precision rates for the human and excavator classes exceeded 90%, while other equipment classes like grader, truck, scraper, and bulldozer achieved precision rates higher than 75%. The object detection classifier demonstrated an overall recall rate of 77.7% and an overall precision rate of 76.4% in detecting construction equipment and workers.

4.2 Semantic Segmentation Module Results

The semantic segmentation model's performance was assessed using a test dataset of 101 images collected from open online sources for an outdoor construction setting. The test images were manually labeled with the 21

previously mentioned classes to develop the ground truth file for model evaluation. The mean Intersection over Union (mIoU) metric was the evaluation metric used, which calculates the overlap between the ground truth and the classifier output, divided by the area of their union. The Deeplab V3+ with ResNet50 has achieved a weighted mIoU of 54.3%. Table 2 presents the results for the 21 classes in test images. Some classes, such as tower crane, dirt, human, steel, curing blanket, and equipment, achieved mIoU rates higher than or equal to 50%. Others, including concrete, rebar, lumber, and brick, had mIoU rates ranging from 40% to 50%. Classes like formwork, glass, asphalt, pipe, scaffold, bitumen, aggregate, and bin exhibited mIoU values between 17% and 39%. However, the model exhibited poor performance in classes like rail and waterproofing, potentially due to their low number of sample pixels in the training dataset. These results show the model's strengths and limitations, which are further detailed in the discussion section.

Table 2. Semantic segmentation classes' mIoU

Classes	mIoU	Classes	mIoU
Concrete	45%	Steel	51.3%
Formwork	27.6%	Rebar	46.6%
Glass	26.8%	Scaffold	16.8%
Tower crane	56.2%	Bitumen	13.1%
Human	50%	Aggregate	25%
Other	69.3%	Rail	10.6%
Bin	20.7%	Brick	49.1%
Dirt	67.5%	Waterproofing	2.9%
Equipment	60%	Pipe	17%
Lumber	43.7%	Curing_Blanket	68.2%
Asphalt	39%	Steel	51.3%

Despite the varying mIoU rates across the 21 classes, the semantic segmentation model mainly aimed to identify major objects/surfaces in construction images for input to the text classifier. In this study, the top 7 detected classes were selected for the text classifier based on their pixels counts. A "commonality percentage" assessment method determined the accuracy of the top 7 segmented classes in representing the actual classes in test images. Figure 2 illustrates how the commonality percentage of actual classes was calculated among the top 7 detected classes. The average commonality percentage across 101 test images was 85.9%.

Ground truth classes: Concrete, Formwork, Rebar, Curing_Blanket, Human, Lumber

Top 7 detected classes: Concrete, Formwork, Other, Dirt, Rebar, Curing_Blanket, Human

This image shows 5 out of 6 actual classes among the top 7 detected classes. Therefore, the percentage $= \frac{5 \times 100\%}{6} = 83.3\%$

Figure 2. Commonality percentage of existing actual classes

4.3 Text Classifier Module Results

The text classifier model was evaluated on 160 test images using precision and recall rates as performance metrics. The aim of this test setup was to assess the performance of the text classifier alone to assess how well it can annotate activities and project phase/type based on the correct appearing objects and surfaces. Thus, the ground truth objects and surfaces were given to the models. In other words, the object detector and semantic segmentation classifiers were not used to feed the text classifier. The text classifier achieved an overall precision rate of 92.1% and an overall recall rate of 86.7%. Table 3 and Table 4 present the results for medium-level and high-level predictions of the NN text classifier. Most activity predictions achieved high precision and recall rates exceeding 90%. Road, rail, and pipe construction showed higher results than other project types among the high-level predictions.

Table 3. Medium-level (activities) prediction results

Class	Precision	Recall
Rebar Installation	98.1%	100%
Steel Erection	88.2%	93.8%
Formwork Shuttering and Removal	98.2%	100%
Concrete Work	95.3%	99%
Lumber Work	75%	81.8%
Earth hauling	100%	86%
Paving work	75%	60%
Material Lifting	100%	100%
Earthwork	100%	86.5%
Excavation	100%	100%
Masonry Work	66.7%	100%
Glazing	80%	57.1%
Rail Work	100%	86.7%
Waterproofing	100%	87.5%
Pipework	92.9%	100%
Overall Medium-Level	95.9%	94.6%

Table 4 High-level (project type) prediction results

Class	Precision	Recall
Building Sub-structure	51.6%	53.3%
Building Super-structure	86.5%	59.3%
Bridge Construction	33.3%	8.3%
Road Construction	90.3%	96.6%
Heavy Construction	83%	50%
Railway Construction	100%	93.3%
Pipeline Construction	100%	90%
Overall High-Level	80.2%	65.6%

4.4 Ensemble Model Results

The same 160 images were used to test the NN text classifier to evaluate the entire system (ensemble model). The object detection and segmentation classifiers processed the images, and their results were passed to the text classifier. The ensemble model achieved an overall precision of 60.23% and a recall of 70%. Table 5 and Table 6 detail the medium-level and high-level prediction results. Figure 3 shows a sample image from the test dataset. It encompasses a building superstructure project, formwork activity, concrete work, lumber, earth hauling, and material lifting. The model correctly predicted the project type and four activities. However, rebar installation was also classified, while lumber work was missed, resulting in 5 true positive instances, 1 false positive instance, and 1 false negative. Thus, the ensemble model had a precision and recall rates of 83.33% in this sample.

gtruth: Building Superstructure, formwork shuttering and removal, concrete work, lumber work, earth hauling, material lifting
Predicted project type: Building Superstructure
Predicted activities: Rebar work, formwork shuttering and removal, concrete work, earth hauling, material lifting

Figure 3. Ensemble model test image sample

Among the medium-level classifications, there were different performance trends. Concrete work, material lifting, and earthwork exhibited high-performance rates. Some activities, such as rail work, paving work, pipework, and earth hauling, demonstrated higher precision than recall, meaning that the model made fewer predictions than the activities in the ground truth dataset, although most were correct. On the contrary, rebar installation, steel erection, formwork shuttering and removal, and lumber exhibited recall rates surpassing precision, which means that the model made many predictions for these activities; however, not all were correct. Waterproofing and glazing showed lower performance, with precision and recall rates below 33%, and masonry work was not detected. Among high-level classifications, railway and pipeline constructions demonstrated precision values of 100%, yet their recall values were only 33.3% and 10%, respectively. Road construction and building super-structure followed, achieving precision rates of 95.65% and 63.89%, respectively. Building sub-structure, bridge, and heavy construction obtained less than 36% precision values, indicating a lower performance than other project types.

Table 5. Medium-level (activities) prediction results

Class	Precision	Recall
Rebar Installation	60.98%	94.34%
Steel Erection	40%	75%
Formwork Shuttering and Removal	51.85%	100%
Concrete Work	78.57%	97.1%
Lumber Work	30.95%	59.1%
Earth hauling	76.92%	47.62%
Paving work	60%	30%
Material Lifting	80.85%	95%
Earthwork	87.1%	72.97%
Excavation	66.67%	51.28%
Masonry Work	0.00%	0.00%
Glazing	28.57%	28.57%
Rail Work	100%	46.67%
Waterproofing	33%	25%
Pipework	85.71%	46.15%
Overall Medium-Level	63.77%	78.2%

Table 6 High-level (project type) prediction results

Class	Precision	Recall
Building Sub-structure	35.59%	70%
Building Super-structure	63.89%	42.59%
Bridge Construction	11.54%	25%
Road Construction	95.65%	75.9%
Heavy Construction	12.5%	10%
Railway Construction	100%	33.33%
Pipeline Construction	100%	10%
Overall High-Level	48.1%	47.5%

The ensemble model exhibited a noticeable performance reduction due to errors from the object detector and semantic segmentation compared to standalone text classifier. A sensitivity analysis was conducted in two scenarios to assess the impact of each DCNN module. In the first scenario, only the object detection classifier was used with the ground truth segmentations, while in the second scenario, only the semantic segmentation classifier processed images.

Based on the ground truth dataset for the text classifier, the analysis revealed that semantic segmentation had a more pronounced effect on the performance of the overall system than the object detection. Scenario two exhibited a 6% drop in precision and a 19.38% drop in recall compared to scenario one, as shown in Figure 4.

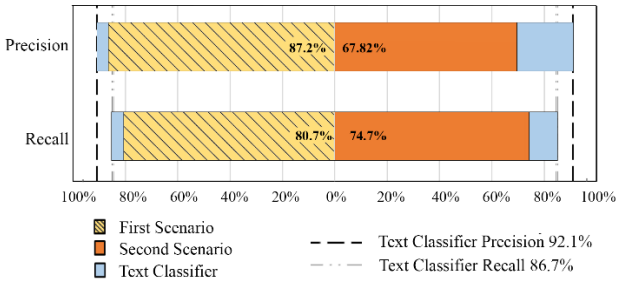

Figure 4. The difference between the two scenarios and the base scenario

5 Discussion

The outcomes obtained from the three modules showed promising results in improving the annotation of outdoor construction images, by providing appearing objects, ongoing activities, and project types. However, these results also showed challenges and limitations encountered by the classifiers in specific instances. The subsequent four subsections discuss these challenges and propose potential solutions to enhance the practical application of the proposed approach.

5.1 Object Detection Module

The object detection classifier resulted in recall and precision rates of 77.7% and 76.6% across 11 object classes, respectively. Despite this promising performance, there were instances of missed or misclassified objects. For example, truck, off-truck, and concrete mixer classes exhibited lower precision rates than major equipment like excavators and bulldozers. Some misclassifications included regular dump trucks identified as off-road dump trucks and concrete mixers classified as regular trucks. Similarities in the front view of this equipment from different manufacturers contributed to such misclassifications. Misclassifications were observed for loaders, roller compactors, and graders due to visual similarities, specifically from their rearview. Moreover, image quality factors, such as resolution and occlusion, which are common in busy construction sites, contributed to increased false negatives, notably in the human class. Enhancing the training dataset with more images featuring diverse equipment types and poses could potentially improve recall and precision.

5.2 Semantic Segmentation

The mIoU metric was utilized to evaluate the performance of the semantic segmentation classifier in detecting elements in outdoor construction sites. Across 101 test images, the model achieved an mIoU rate of 54.3% in segmenting 21 classes. However, misclassifications occurred, especially when elements shared similar visual features. The model confused lumber, formwork, and wooden scaffolding systems due to material and texture similarities. Waterproofing was misclassified as asphalt or bitumen, and rail was often detected as steel due to similarities in their material composition. Additional misclassifications included aggregate as dirt and bitumen as asphalt. Pipe was detected as dirt or equipment. This module's performance relies highly on the training dataset; thus, it is suggested to expand the training dataset and enhance image quality.

5.3 Text Classifier

The standalone NN classifier achieved an overall precision rate of 92.1% and a recall rate of 86.7%. However, among high-level classifications, heavy construction, building sub-structure, and bridge construction exhibited lower precision and recall rates than other project types. These misclassifications are due to visual element similarities with other project types, as illustrated in Figure 5 (a) and Figure 5 (b), where a heavy construction project (road underpass) and a sub-structure construction of a high-rise shared visual elements like excavators, concrete, and dirt, leading to confusion during classification. The text classifier faced challenges in differentiating instances of these classes, leading to higher classification error rates. Heavy construction was frequently misclassified as building sub-structure, building super-structure, or road construction. Similarly, building sub-structure was often confused with bridge construction, heavy construction, building super-structure, and road construction. Moreover, bridge construction was mixed up with building super-structure and building sub-structure classes.

(a) (b)

Figure 5. (a) Heavy Construction (underpass construction) and (b) Building sub-structure

The model struggled to predict paving, masonry, and lumber work in the medium-level classification. For example, the classifier tended to overpredict lumber work,

particularly when lumber was stored on the construction site (e.g., for formwork or lagging) without apparent lumber work in the image. Paving work was mainly classified when bitumen or asphalt was present in road project images. However, when a road made of asphalt appears in the background of a project, as seen in Figure 6 in a building sub-structure project, the classifier predicts paving work. Furthermore, the presence of common road construction equipment, including an excavator and dump trucks, in this image further confused the classifier. Similarly, some images included adjacent buildings made of bricks (and other building elements) in the project's background, which is not part of the project. The classifier could mistakenly identify masonry work in such cases, leading to more false positives and reducing the precision rate.

Figure 6. Asphalt roads beside a building sub-structure project

5.4 Ensemble Model

The integration of the three modules had impacted the overall performance of the ensemble model. For example, it struggled to predict pipeline construction projects, achieving a 10% recall rate, due to the poor performance of semantic segmentation in classifying "pipes" which achieved an mIoU of 16.9%. Consequently, the text classifier could not classify a pipeline project type due to the inaccurate low-level classification. Similarly, masonry work predictions were affected as the text classifier predicted masonry work where the "brick" class was present among the data, but buildings made of bricks in the project surroundings led to such misclassifications. Thus, the precision rate for masonry work was consequently impacted. The semantic segmentation modules also achieved mIoU of 49.1% for "brick" detection, further contributing to the ensemble model's overall low performance.

The sensitivity analysis highlighted the substantial impact of semantic segmentation on the ensemble model's performance; however, developing a robust semantic segmentation model was not the primary goal of this study. It should be mentioned that a properly trained model can result in mIoUs of up to 65%, as shown in recent studies [18]. Inherent issues in annotating project types in images exist, as an image may include limited visual information, making it challenging even for experts to accurately determine the project type without additional context, as shown in Figure 7.

Figure 7. A retaining wall under construction with an unclear project type

6 Conclusion

This study introduces an ensemble model utilizing computer vision and machine learning to annotate outdoor construction images with activities and project types. Two DCNN classifiers for object detection and semantic segmentation were trained to detect key elements and surfaces in outdoor construction scenes. These classifiers initially process construction images to detect various construction elements, materials, and equipment. The identified objects and surfaces are then fed to a trained NN text classifier to predict construction activities and project phases/types. The results showed promising performance, achieving a precision rate of 60.23% and a recall rate of 70% for predicting 15 construction activities and seven project types. The ensemble model offers potential improvements in automating image documentation and retrieval in the construction industry, with suggestions to enhance performance by increasing training datasets. However, one of the limitations of this study is to predict detailed activities due to the absence of a temporal dimension, relying on still images only. Future research may explore dynamic datasets, such as videos, to address this limitation. Additionally, a multitask vision language pre-training approach, such as Bootstrapping Language-Image Pre-training (BLIP) [27], could be explored to develop image and text classifiers simultaneously in future work.

7 References

[1] Paneru, S., & Jeelani, I. Computer vision applications in construction: Current state, opportunities & challenges. *Autom. Constr.*, 132,

103940, 2021.

[2] Lopes Amaral Loures, L., & Rezazadeh Azar, E. Condition Assessment of Unpaved Roads Using Low-Cost Computer Vision–Based Solutions. *Journal of Transportation Engineering, Part B: Pavements*, 149(1), 04022066, 2023.

[3] Ma, J. W., Czerniawski, T., & Leite, F. An application of metadata-based image retrieval system for facility management. *Adv. Eng. Inf.*, 50, 101417, 2021.

[4] Rezazadeh Azar, E. Semantic annotation of videos from equipment-intensive construction operations by shot recognition and probabilistic reasoning. *J. Comput. Civ. Eng.*, 31(5), 04017042, 2017.

[5] Xiao, B., & Kang, S. C. Development of an image data set of construction machines for deep learning object detection. *J. Comput. Civ. Eng.*, 35(2), 05020005, 2021.

[6] Kim, J., Chi, S., & Kim, J. 3D pose estimation and localization of construction equipment from single camera images by virtual model integration. *Adv. Eng. Inf.*, 57, 102092, 2023.

[7] Wang, Y., Xiao, B., Bouferguene, A., Al-Hussein, M., & Li, H. Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Adv. Eng. Inf.*, 53, 101699, 2022.

[8] Brilakis, I., & Soibelman, L. Content-based search engines for construction image databases. *Automation in Construction*, 14(4), 537-550, 2005.

[9] Brilakis, I., & Soibelman, L. Multimodal image retrieval from construction databases and model-based systems. *J. Constr. Eng. Manage.*, 132(7), 777-785, 2006.

[10] Rezazadeh Azar, E., & McCabe, B. Automated visual recognition of dump trucks in construction videos. *J. Comput. Civ. Eng.*, 26(6), 769-781, 2012.

[11] Pal, A., & Hsieh, S. H. Deep-learning-based visual data analytics for smart construction management. *Automation in Construction*, 131, 103892, 2021.

[12] Girshick, R., Donahue, J., Darrell, T., & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE CVPR conference*, pages 580-587, 2014.

[13] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE CVPR conference,* pages 779-788, 2016.

[14] Chen, C., Xiao, B., Zhang, Y., & Zhu, Z. Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition. *Autom. Constr.*, 146, 104702, 2023.

[15] Wu, S., Hou, L., Zhang, G. K., & Chen, H. Real-time mixed reality-based visual warning for construction workforce safety. *Autom. Constr.*, *139*, 104252, 2022.

[16] He, K., Gkioxari, G., Dollár, P., & Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961-2969, 2017.

[17] Wang, Z., Zhang, Q., Yang, B., Wu, T., Lei, K., Zhang, B., & Fang, T. Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase. *J. Comput. Civ. Eng., 35(1)*, 04020056, 2021.

[18] Wang, Z., Zhang, Y., Mosalam, K. M., Gao, Y., & Huang, S. L. Deep semantic segmentation for visual understanding on construction sites. *Computer‐Aided Civil and Infrastructure Engineering*, 37(2), 145-162, 2022.

[19] Hochreiter, S., & Schmidhuber, J. Long short-term memory. Neural computation, 9(8), 1735-1780. 1997.

[20] Ding, L., Fang, W., Luo, H., Love, P. E., Zhong, B., & Ouyang, X. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Automation in construction* 86, 118-124, 2018.

[21] Heravi, G., & Eslamdoost, E. Applying artificial neural networks for measuring and predicting construction-labor productivity. *J. Constr. Eng. Manage.*, 141(10), 04015032, 2015.

[22] Gerassis, S., Martín, J. E., García, J. T., Saavedra, A., & Taboada, J. Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. *J. Constr. Eng. Manage.*, 143(2), 04016093, 2017.

[23] Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. Machine learning algorithms for construction projects delay risk prediction. *J. Constr. Eng. Manage.*, 146(1), 04019085, 2020.

[24] The MathWorks Inc. MATLAB R2022b: 9.13, Natick, Massachusetts: *The MathWorks Inc.* https://www.mathworks.com, 2022.

[25] He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR conference*, pages 770-778, 2016.

[26] Mierswa, I, and Klinkenberg, R. "RapidMiner." *RapidMiner Inc.,* https://rapidminer.com, 2018.

[27] Li, J., Li, D., Xiong, C., & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of International Conference on Machine Learning*, pages 12888-12900, 2022.

[28] Xiao, B., Wang, Y., & Kang, S. C. Deep learning image captioning in construction management: a feasibility study. *J. Constr. Eng. Manage.*, 148(7), 04022049, 2022.