buildings

MDPI

Article

Applying Deep Learning and Single Shot Detection in Construction Site Image Recognition

Li-Wei Lung and Yu-Ren Wang *

Department of Civil Engineering, National Kaohsiung University of Science and Technology, Kaohsiung 80778, Taiwan
* Correspondence: yrwang@nkust.edu.tw; Tel.: +886-7-381-4526

Abstract: A construction site features an open field and complexity and relies mainly on manual labor for construction progress, quality, and field management to facilitate job site coordination and productive results. It has a tremendous impact on the effectiveness and efficiency of job site supervision. However, most job site workers take photos of the construction activities. These photos serve as aids for project management, including construction history records, quality, and schedule management. It often takes a great deal of time to process the many photos taken. Most of the time, the image data are processed passively and used only for reference, which could be better. For this, a construction activity image recognition system is proposed by incorporating image recognition through deep learning, using the powerful image extraction ability of a convolution neural network (CNN) for automatic extraction of contours, edge lines, and local features via filters, and feeding feature data to the network for training in a fully connected way. The system is effective in image recognition, which is in favor of telling minute differences. The parameters and structure of the neural network are adjusted for using a CNN. Objects like construction workers, machines, and materials are selected for a case study. A CNN is used to extract individual features for training, which improves recognizability and helps project managers make decisions regarding construction safety, job site configuration, progress control, and quality management, thus improving the efficiency of construction management.

Keywords: construction image; artificial intelligence; deep learning; object detection; single shot multibox detector (SSD)

Citation: Lung, L.-W.; Wang, Y.-R. Applying Deep Learning and Single Shot Detection in Construction Site Image Recognition. *Buildings* **2023**, *13*, 1074. https://doi.org/10.3390/buildings13041074

Academic Editors: Maxim A. Dulebenets and Saeed Banihashemi

Received: 20 January 2023
Revised: 28 March 2023
Accepted: 9 April 2023
Published: 19 April 2023

1. Introduction

Construction work is tedious and subject to delays, and its quality may be compromised by many factors, such as construction equipment, workers, and materials. Therefore, it is necessary to improve construction quality and progress in today's increasingly competitive market by considering good job site management and meeting construction costs. At a job site currently, a job site manager oversees everything construction-related, including workers, machines, and materials [1–5]. The manager has to take care of virtually everything at the job site [6]. The improvement of management methods using innovative technology helps to not only accelerate the development of the construction industry but also improve a company's competitiveness in the market.

Most general contractors deploy imaging devices, such as photo and video cameras, to document the progress of construction activities throughout the entire process. The image data are collected, in general, by filming with a mobile camera operated by a worker or a video camera set up at a fixed location. Most image data collected are used passively for reference or even just shelved. The others are used to prepare quality documents or demonstrate construction status and progress. Suppose artificial intelligence (AI) is introduced to recognize objects in the images and help job site management identify and tag things in the images. In that case, these image data may serve as an essential basis

for decision-making within construction activities, including construction planning and design, job site safety, automated equipment management [7–10], and quality monitoring and maintenance. For example, suppose a specific machine is tagged in video footage of construction activities [11]. In that case, the project team may exploit the captured data for project decisions of route management, machine setup, and site safety [12–15].

When recognizing and classifying objects in many images, a deep learning model may be introduced to accelerate the extraction of high-value digital information crucial for construction management. The mainstream in developing the deep neural network is the convolution neural network (CNN) which extracts critical feature information by including one or more convolution layers and pooling layers through a combination of algorithms and multi-layer computation of convolution neurons as the images are converted into data [16]. The feature information is fed to the neural network for training in a fully connected manner until identical or similar features in the same class of images are identified and documented. The relative locations and features digitally arranged during the recognition of new images are systematically computed and processed to identify the similarities between images for successful image judgment [17,18].

AI is having revolutionary impacts on construction engineering [19]. Thanks to the powerful capability of AI in data processing, analysis, and searching for massive digitization, a model to recognize construction objects at a job site can be built to rapidly and accurately identify workers [20–22], machines [23,24], and materials [25] in job site footage while tagging their relative locations in the images to provide more site-related information for project management, which is a rising topic in the industry in the pursuit of breakthroughs and innovation.

2. Literature Review

A construction project has unique complexity. The completion of a project involves an engineering lifecycle consisting of many links, from design and construction to final acceptance. In an era in which the development of technical information evolves at the speed of light, the innovative technologies and management systems used in construction management help not only maintain control over safety and health as the construction work progresses but also facilitate the successful completion of construction projects by reducing uncertainties while focusing on the goal of sustainable development [26].

Artificial intelligence, or AI, is an engineering study focusing on researching and developing intelligent entities. AI includes the use of programs and big data to make computers and machines mimic human thinking and simulate the "intelligent" behaviors of a human being; when AI is the object of study, machine learning (ML) is a model to improve the performance of specific algorithms while learning from experiences, i.e., learning from data collected [27]. However, data learning is based on massive data processed using a multi-layer neural network. A self-learning method is found after linear or nonlinear conversion via multiple processing layers, which automatically extracts features representative of data characteristics in place of the long time taken for traditional feature engineering. Deep learning is a technology that evolved from machine learning [28].

The applications of deep learning in computer vision in recent years are in the following classes [29], as shown in Figure 1: (1) classification: putting an image in one of the established classes by its nature and type; (2) semantic segmentation: identifying pixel blocks by event type instead of classifying into "instances"; (3) classification + localization: tagging a message to a single object with its location and size (w, h); (4) object detection: tagging multiple objects with their locations and sizes; and (5) instance segmentation: tagging "instances"; the objects of the same class are identified by individual locations and sizes, particularly when they are overlapping.

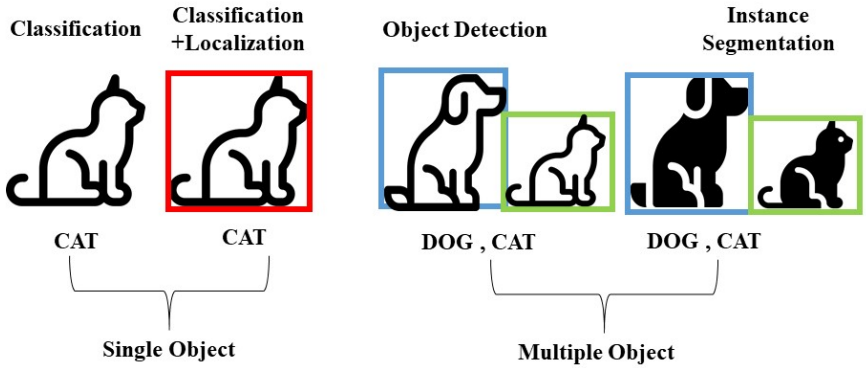


Figure 1. Applications in computer vision.

Most recent object detection studies are focused on the use of a CNN for typical model applications in which a matching object is identified before determining in which area a matching thing exists and tagging the location of highest probability with a box, as shown in Figure 2. Two fully connected layers are connected behind the CNN, one for classification and the other for tagging the matching area. There are three algorithms to organize an area: sliding window, region proposal, and grid-based.

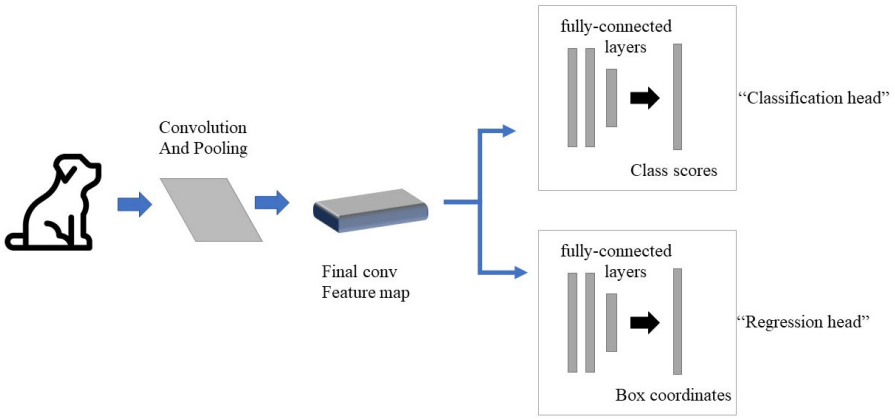


Figure 2. Locating algorithm model.

1. Sliding window: a simple but time-consuming method based on the method of exhaustion. It works by establishing windows of various sizes for image scanning and extracting the feature information of every image window. Next, the data is fed to a classifier for object recognition to determine if the probability of the window matching the object to be detected is accurate. This method is the simplest but most time-consuming [30], as presented in Figure 3.

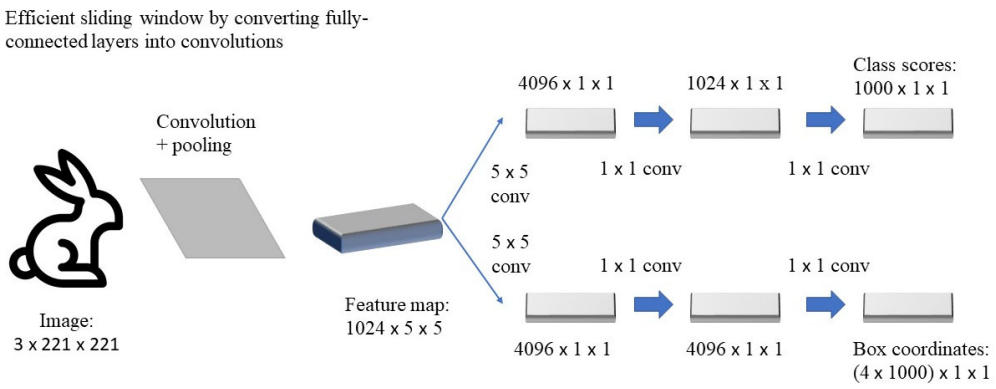


Figure 3. Sliding window algorithm.

2. Region proposal: information in the image, such as texture, edges, and color, are used to predetermine the regions of interest (ROI) containing the object and determine the probability of these regions for matching. The high recall is maintained by filtering thousands of regions per second. Similar algorithms are R-CNN, Fast R-CNN, and Faster R-CNN [31–34], as shown in Figure 4.

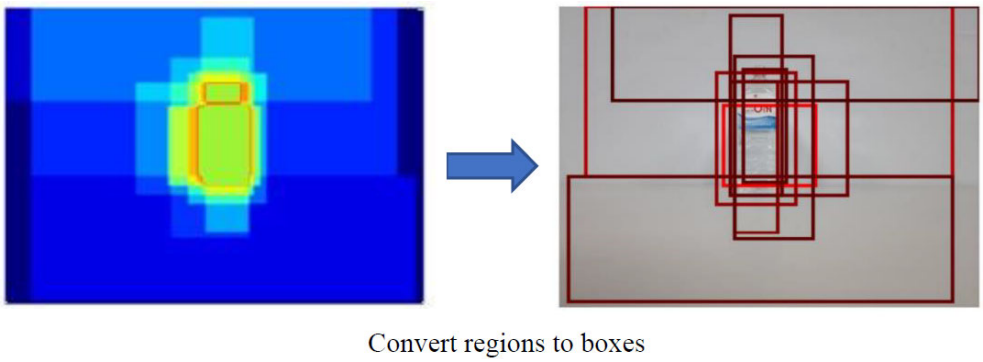


Figure 4. Region proposals algorithms.

3. Grid-based regression: a picture is divided into grids, and regions of various sizes are selected with the grids as centers. Regression determines the probability that every bounding box contains the target. This approach is suitable for real-time detection. Similar algorithms are you only look once (YOLO) and single shot multibox detector (SSD) [35], as shown in Figure 5.

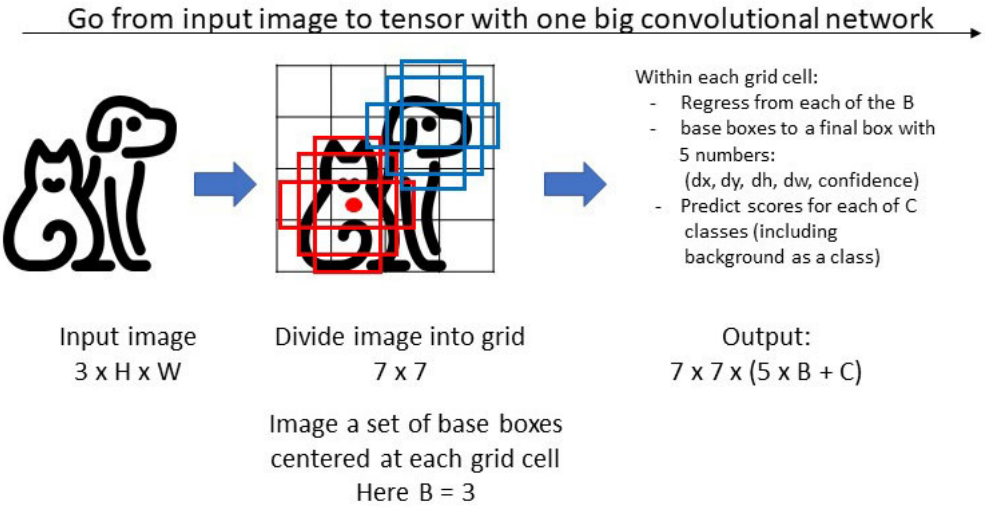


Figure 5. Region Proposal algorithms.

You only look once (YOLO) predicts multiple bounding boxes and types of CNNs, realizing end-to-end target detection and identification. This algorithm avoids the weakness that object detection must be trained separately and accelerates the computation dramatically [36], as indicated in Figure 6.

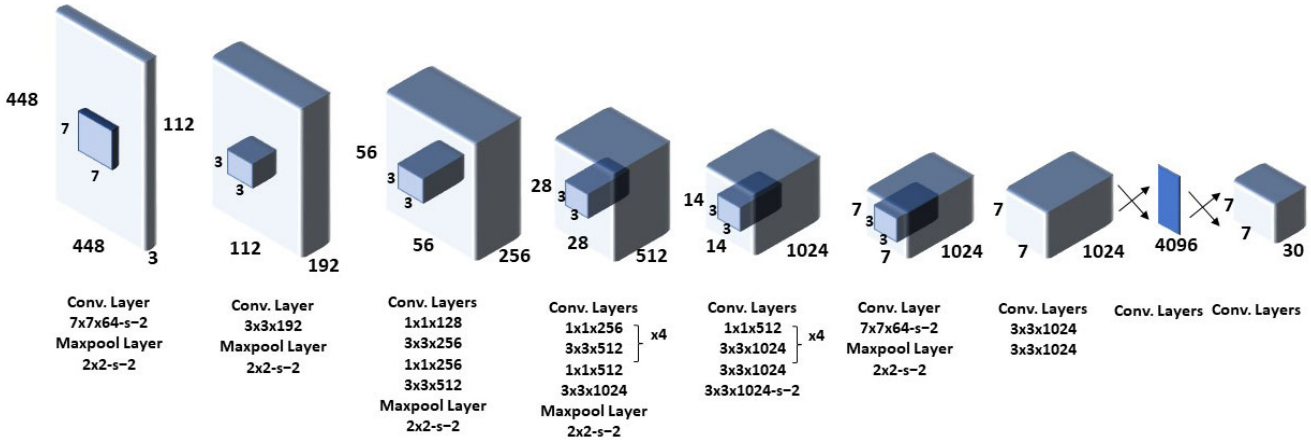


Figure 6. Structure of a YOLO model.

The single shot multibox detector (SSD) is based on a feed-forward CNN that generates bounding box sets and scores of different types on the boxes, followed by non-maximum value suppression to complete the final detection process. This explains the incorporation of both the regression concept in YOLO and the anchor mechanism in Faster-CNN in single shot multibox detector (SSD), as regression is performed on the multi-dimensional region features of every location in the entire picture, which retains YOLO's characteristics of being fast while ensuring the window prediction is as accurate as Faster-RCNN [37], as shown in Figure 7.

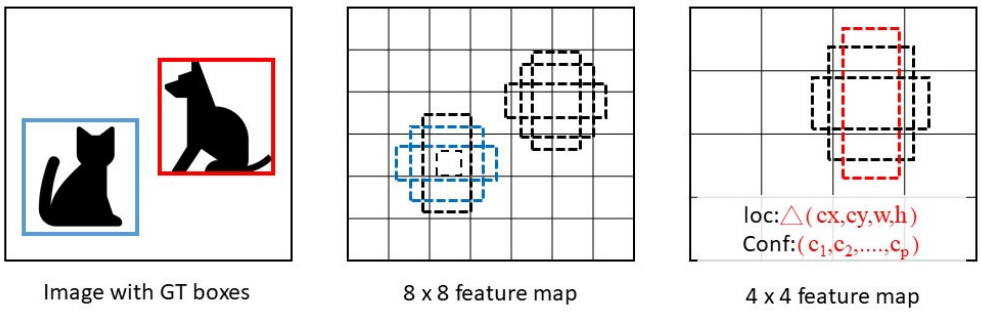


Figure 7. Default boxes in the single shot multibox detector model.

Liu et al. (2016) tested the speed and accuracy of different object detection methods. The test results are shown in Table 1:

Table 1. Object detection algorithm speed and accuracy comparison.

Method	FPS	Boxes	mAP
Faster R-CNN	7	6000	73.2
Faster YOLO	155	98	52.7
SSD300	29	8732	74.3

A fast YOLO has faster processing speed but poor mAP. Although Faster R-CNN has a higher accuracy rate (73.2% mAP), it is not significantly more accurate at determining the number of images. In contrast, a single shot multibox detector (SSD) not only has a high accuracy rate but also a fast image detection speed [36].

Single shot multibox detector (SSD) object recognition has been used in many engineering applications. For example, Yudin and Slavioglo [38] used the single shot multi-box detector (SSD) to test how well the model identifies a traffic light, producing good results. Wang et al. [39] proposed an improved single shot multibox detector (SSD) capable of detecting a ship in a noisy background. The results were compared with those from Faster

R-CNN, and it was found that the enhanced single shot multibox detector (SSD) improved detection accuracy.

Much research on image recognition using deep learning has accumulated in recent years. Many people use deep learning technology in artificial intelligence to let computers handle more complex image recognition problems. Table 2 shows the development of deep learning in the construction industry in the past five years of applied research on image recognition.

Table 2. Research on the application of deep learning in construction image recognition.

Author (Year)	Abstract
Dorafshan, S., Thomas, R. J., and Maguire, M. (2018) [40]	Compares the performance of deep convolutional neural networks and edge detection algorithms for image-based crack detection in concrete, finding that the neural network approach outperforms traditional edge detection methods.
Spencer Jr, B. F., Hoskere, V. and Narazaki, Y. (2019) [41]	Recent advances in computer vision-based civil infrastructure inspection and monitoring techniques, including object detection, semantic segmentation, and deep learning methods, highlight their benefits and challenges.
Dung, C. V. (2019) [42]	Proposes an autonomous system for concrete crack detection using a deep, fully convolutional neural network, achieving high accuracy and efficiency compared to traditional manual inspection methods.
FANG, Weili, et al. (2020) [43]	A review and discussion of future directions of computer vision for behavior-based safety in construction.
Li, Y., Lu, Y. and Chen, J. (2021) [25]	A deep learning approach based on the YOLOv3 detector is proposed for real-time rebar counting on construction sites, which can effectively improve construction efficiency and safety.
Chou, J. S. and Liu, C. H. (2021) [24]	An automated system for recognizing trucks in real-time in river dredging areas using computer vision and deep learning.
Li, X., Chi, H., Lu, W., Xue, F., Zeng, J., and Li, C. Z. (2021) [44]	An intelligent work packaging system that preserves construction workers' personal image information using federated transfer learning.
DEL SAVIO, Alexandre Almeida, et al. (2021) [45]	Artificial intelligence (AI) and computer vision are used to identify objects and equipment on a construction site and how they can improve safety and efficiency.
LIN, Chih-Lung, et al. (2022) [22]	Presents a gait-based pedestrian automatic detection and recognition system using a deep learning neural network.
Greeshma, A. S. and Edayadiyil, J. B. (2022) [10]	An automated system that uses machine learning and image processing to monitor construction project progress.
Del Savio, A., Luna, A., Cárdenas-Salas, D., Vergara, M., and Urday, G. (2022) [11]	A manually classified dataset of construction site images containing 1046 images of eight object classes that can be used to develop computer vision techniques in the engineering and construction fields.
Yeşilmen, S. and Tatar, B. (2022) [16]	The efficiency of using convolutional neural networks (CNN) for image classification in monitoring construction-related activities, with a case study on aggregate mining for concrete production.

Source: This study collated.

Past studies used deep learning algorithms to recognize three postures of construction workers, including standing, bending over, and squatting [20–22]. They provide engineering professionals with comprehensive deep learning solutions for detecting construction vehicles [23,24]. Only single objects, such as people, materials, or engineering vehicles, were seen in the above studies; therefore, the shapes and boundary types recognized were relatively pure. This study uses image automation to simultaneously identify workers, machinery, and materials in the current construction situation, assist the construction site manager in making safety judgments on the location of construction equipment, safety

protection measures, and material stacking, and monitor the construction status and maintenance of the construction site to reduce environmental hazards and control progress.

With the continuous evolution of technology, combining big data and artificial intelligence machine learning/deep learning can maximize the value of data. Therefore, this research collects the construction site image data set, imports the object detection system, uses it as artificial intelligence and machine learning training data, and builds AI to automatically identify the personnel, materials, and equipment on the construction site. In the future, continuous learning, modification, and technical improvement can reduce or avoid labor accidents on the construction site, thereby improving construction efficiency and schedule management.

3. Methodology

3.1. Study Setup

The CNN in single shot multibox detector (SSD) method of deep learning required massive training images for learning. Firstly, image files of a construction site were collected and converted to matrices by regulating the size of images before data pre-processing, such as optimization. Next, features were extracted using CNN and fed into the fully connected neural network to predict and identify classes. Finally, the trained model was verified by feeding it the test data. The model's learning rate setting would affect the weight adjustment, so this study set the learning rate = 0.00002, epoch = 100, step per epoch = 320, and optimizer type = sgd. The model structure of this proposed method is presented in Figure 8.

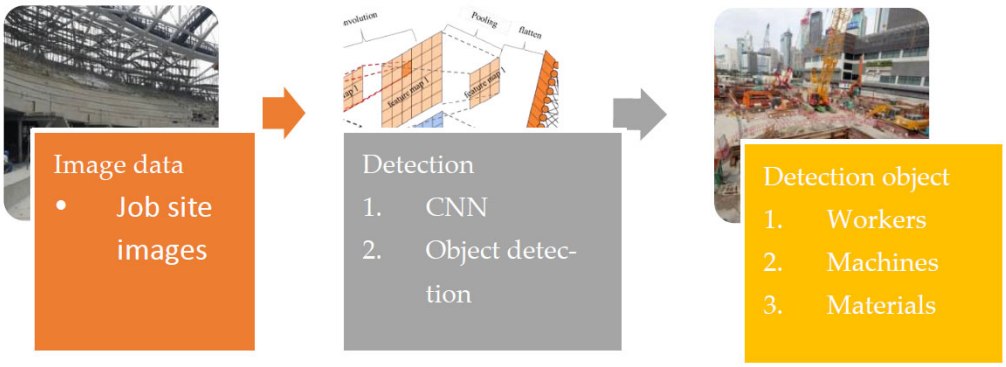


Figure 8. Structure of job site image object detection model.

3.2. Collection of Job Site Images for a Construction Project

The multi-class classification in the image classification was selected for the study. Data sets were classified as rebar, worker, and machine. The deep learning model required massive amounts of information for training to improve its recognition accuracy, and the size of the data set was a critical factor for the experiment's success. Data came from three sources, as follows:

1. Legal and free job site pictures obtained from Google under "Creative Commons";
2. Free databases provided by computer vision institutes, such as ImageNet and Labelme of MIT; and
3. Photos of construction job sites taken for the study.

Four hundred sixty-one job site images were collected from the above sources (Figure 9). Feasible data were extracted from the images in the preliminary classification. The job site images collected were manually tagged for workers, machines, and rebar using the image tool provided in "LabelImg." In addition, movements were selected and tagged for classes.

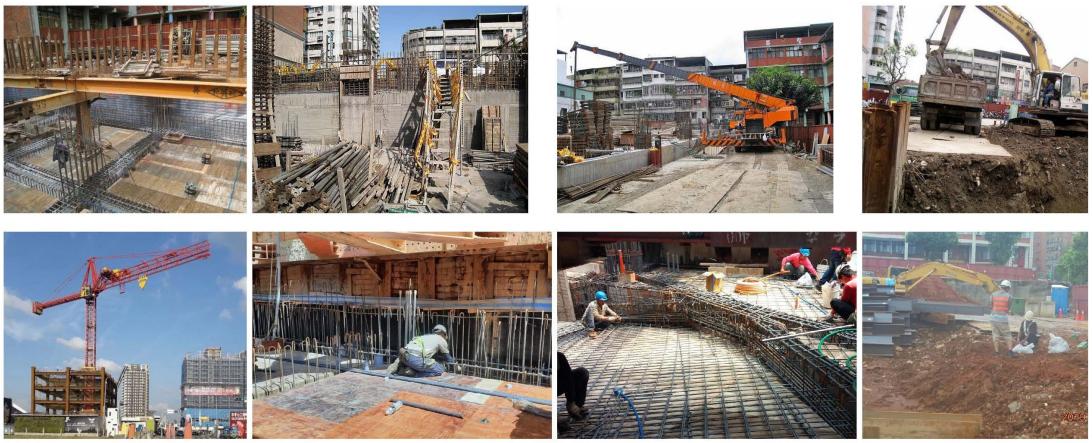


Figure 9. Collection of construction site image files.

Two types of files were generated after tagging with LabelImg; one was the image files themselves, and the other was the XML files with image locations tagged. In Figure 10, for example, workers, rebars, and machines are tagged and given specific names in the image. Figure 11 provides an example of the contents of the XML file, including dimensions such as image coordinates. The single shot multibox detector (SSD) deep learning model was established and tested as all images were tagged.

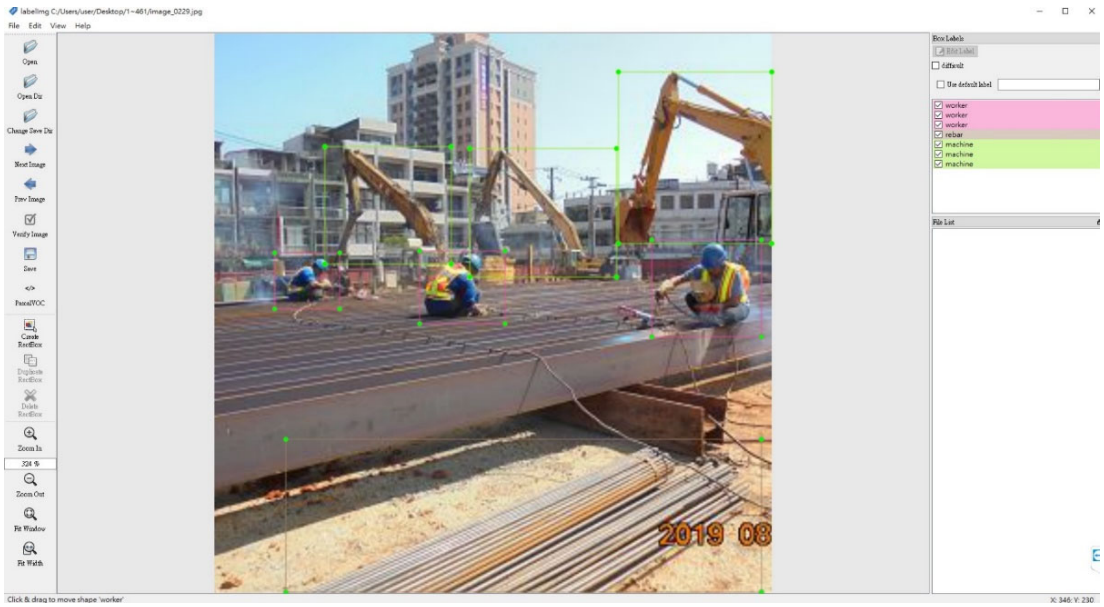


Figure 10. LabelImg tagging of a job site photo.

3.3. Method of Object Detection (SSD)

Wei Liu [36] devised the single shot multibox detector (SSD), a one-stage method in which a neural network (VGG-16) is used to extract feature maps for classification and regression before the target objects are tested. It incorporates the regression concept in YOLO and identifies the location of the target class in regression. Similar to the anchor mechanism in Faster-RCNN, prior boxes are established and features are extracted from the backbone network. Feature maps of various dimensions are used for prediction, with large feature maps to detect small targets and small maps to detect large targets. Convolution kernel is applied on the feature maps to predict the classes and coordinate offsets of a series of default bounding boxes.

VGG-16 serves as the backbone model for the single shot multibox detector (SSD) structure. The fully connected layer of VGG, fc6, is modified and converted into a 3 × 3 convolution layer, Conv6, and fc7 into a 1 × 1 convolution layer, Conv7, while the pooling layer, pool5, is changed from originally 2 × 2 with stride = 2 to 3 × 3 with stride = 1. 4; convolution layers are added; the test module layer of the 1st feature map is Conv4_3, followed by Conv8_2, Conv9_2, Conv10_2, and Conv11_2 [36,39]. Their sizes are shown in Figure 12.

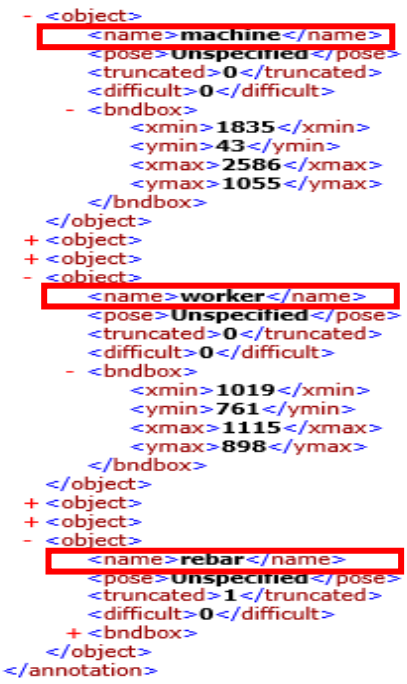


Figure 11. Contents of the XML file of a tagged job site image.

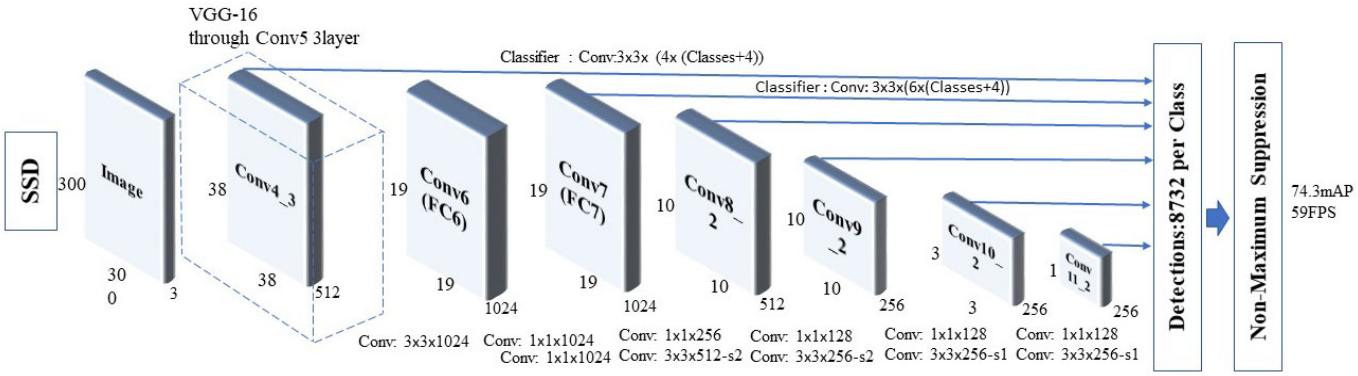


Figure 12. Single shot multibox detector model structure.

The size and length–width ratio require consideration for testing the box on a feature map. Every grid on the feature map is scanned to generate corresponding testing boxes (Figure 13). During the training, the ground truth in the picture is checked to match the testing box. The best-fit box is filtered based on intersection over union (IOU). The exact positive and negative sample ratio is close to 1:3. The loss function depends on the weights of location error and confidence error. Data enhancement is carried out via horizontal flipping, random cutting, color twisting, and random sampling of block regions. Top-k prediction boxes with high confidence levels are reserved during the prediction before the object detection algorithm of non-maximum suppression (NMS) is used to filter prediction terms with significant overlapping. The prediction term left at the end is the result [36].

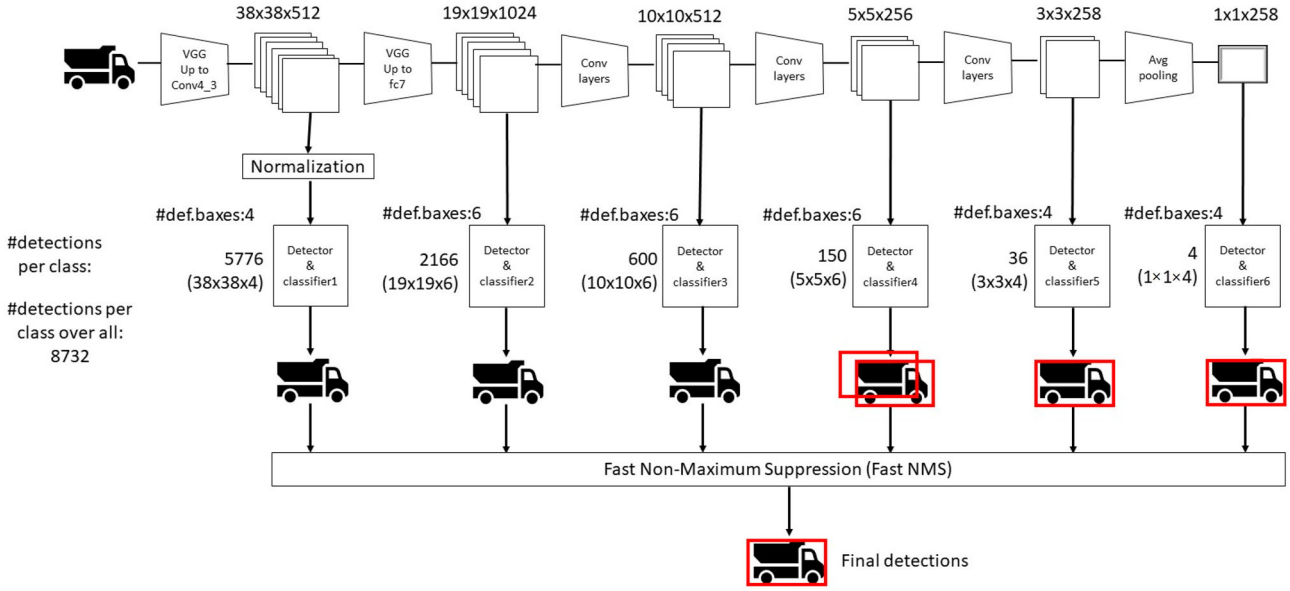


Figure 13. Single shot multibox detector target feature detection process.

4. Study Contents and Outcomes

4.1. Establishment and Testing of Single Shot MultiBox Detector Model

The main feature extraction program used to establish a single shot multibox detector (SSD) model was vgg.py. Features were extracted using 9 module computati9on feature layers in the sizes of 38 × 38, 19 × 19, 10 × 10, 5 × 5, 3 × 3, and 1 × 1 (Figure 13). At the first convolution computation feature layer, the image fed was 300 × 300 in size. Randomly generated 3 × 3 filters were used at the convolution layer to extract 64 features, and the activation function of ReLu was adopted to eliminate negative values. Batch normalization was introduced next to improve the stability of data distribution. After two rounds of convolution feature extraction, the pooling layer shrank the image down to 150 × 150 in size for the convolution computation of the second set. The filters extracted 128 features at the second set convolution computation feature layer. The same applied to the rest of the computation. Ultimately, the pooling layer reduced the images to 1 × 1 in size.

The detect_image feature in the ssd.py program was used for predicting and testing the results. The height and width of the picture were determined after the photo was fed. However, the picture was converted into RGB format to improve detection for the pre-training weight of the image and convenience of color setup in the box. The letterbox_image feature was used to identify the resized image without distortion. The image was normalized based on the batch_size attribute before being fed into the model for regression and type prediction.

Data sets needed to be imported into classes_path while the image training program train.py parameters were established to identify the image classes of rebar, worker, and machine. The pre-training weight, weight_path, was established, and the shape was selected to be 300 × 300. The prior box size was defined as anchors_size = [30, 60, 111, 162, 213, 264, 315]. The image training consisted of 2 stages, "freeze" and "unfreeze." The feature extraction network experienced no change during the freezing stage but minor network tuning. Thus, 50 generations were established. The number of data samples captured for one training run was 16. The backbone and feature extraction network experienced changes during the unfreezing stage. Ample memory was used, and, therefore, 100 generations were established. The number of training samples was 8.

The single shot multibox detector (SSD) program selected the pattern to be detected during the establishment test on the training outcome prediction program predict.py. The parameter setting patterns during the detection were single pictures, pre-recorded footage,

or images captured directly from the camera. For this study, images were used for the prediction model.

4.2. Model Training Data Analysis

In machine learning and deep learning, a loss function is frequently used to evaluate the error between predictions and valid values. The smaller the value, the closer the prediction to the actual value and the more accurate the model. Loss functions commonly used are mean square error (MSE) and cross-entropy; the former is usually used for regression and the latter for classification.

Data outcomes were evaluated based on the performance of the two accuracy indicators, F1 measure and overall accuracy, on the model. Both indicators above were determined using the four factors of the confusion matrix, and they were true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The F1 measure was the harmonized average between accuracy and recall. It was used as an indicator of model performance and expressed as:

$$\text{F1 Measure} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (1)$$

The overall accuracy was defined as the ratio of correct prediction of positive and negative samples in the models over all samples and expressed in Equation (2):

$$\text{Overall Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \quad (2)$$

The single shot multibox detector (SSD) was deployed to identify the classes of rebar, worker, and machine in all images collected in the data set. A total of 461 images were collected, including 400 photos of job site activities as machine learning samples, with 80% images for training. In addition, 40 images, accounting for 10% of the data set, served as the test samples during the training; another 40 were used as verification samples, accounting for 10%. In the end, 61 photos the model had not seen were brought in for recognition, and a 1×1 confusion matrix was generated, as shown in Table 3.

Table 3. Confusion matrix generated by single shot multibox detector model.

TP	30	FN	18
FP	3	TN	10

A calculation was performed for the two accuracy evaluation indicators based on the four factors generated in the confusion matrix. It was found that the F1 measure was 64%, and the oval accuracy was 66%. The details are provided in Table 4.

Table 4. The two accuracy evaluation indicators of the single shot multibox detector model.

Indicators	**Value**
F1 Measure	64%
Overall Accuracy	66%

The process mentioned above reveals that an SSD-based job site activity image recognition system is built by combining the job site image data collected and deep learning in AI. This system can identify and tag essential objects in a job site image, such as workers, machines, and construction materials. With more job site activity information gained from image recognition, the proposed system may help project managers develop project decisions regarding construction safety, job site configuration, progress control, and quality management, thus improving industrial competitiveness.

4.3. Single Shot MultiBox Detector Deep Learning Model Training Outcomes

Three hundred twenty job site activity images, accounting for 80% of the data set, were selected as the training sample for the SS-based job site activity image recognition system proposed herein. In addition, 40 images, or 10% of the data set, were chosen as the test samples during the training. In the end, 61 images the model had not seen were used for recognition; thus, 461 images were collected and used. The visualization outcomes after recognition are presented in Table 5.

Table 5. Outcomes of single shot multibox detector image recognition model test.

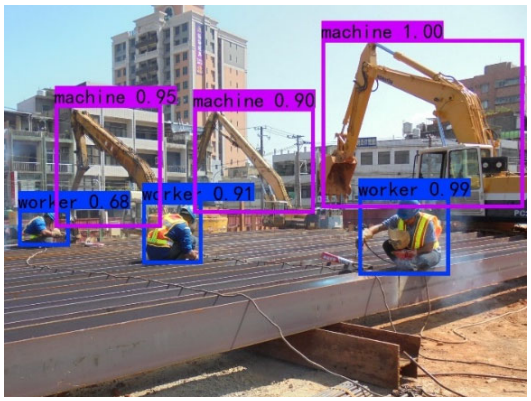
Originals1	Outcomes1
	

Image Data Form 1

Object Name	confidence level	Pixel Coordinates Ymin	Xmin	Ymax	Xmax	Image Number	Time Record
worker	0.99	733	1872	1036	2223	img/image_0229.jpg	2023/3/20 09:15
worker	0.91	750	1043	992	1274	img/image_0229.jpg	2023/3/20 09:15
worker	0.68	781	566	924	765	img/image_0229.jpg	2023/3/20 09:15
machine	1.00	129	1730	778	2508	img/image_0229.jpg	2023/3/20 09:15
machine	0.95	379	707	852	1118	img/image_0229.jpg	2023/3/20 09:15
machine	0.90	393	1238	799	1705	img/image_0229.jpg	2023/3/20 09:15

Originals 2	Outcomes 2
	

Image Data Form 2

Object Name	confidence level	Pixel Coordinates Ymin	Xmin	Ymax	Xmax	Image Number	Time Record
rebar	0.89	992	257	1772	1320	img/image_0158.jpg	2023/3/20 09:15
worker	0.94	959	431	1163	626	img/image_0158.jpg	2023/3/20 09:15
machine	0.97	0	1749	1336	2505	img/image_0158.jpg	2023/3/20 09:15
machine	0.77	1049	1448	1280	1709	img/image_0158.jpg	2023/3/20 09:15

Automated generation of EXCEL forms for the recognized results included object names, confidence level, pixel coordinates, and time record. The timestamp was based on the computer time when the form was generated, which could be used as the basis for specific management items (Table 5):

1. Monitoring the operation status of construction site personnel and equipment: real-time monitoring of the operation status of construction site personnel and equipment, including entry and exit times, the number of construction personnel, and the number of equipment appearing at that time, thereby effectively improving construction safety and efficiency.
2. Ensuring the supply of construction site materials: effectively monitoring the entry and exit of construction site materials and inventory status, ensuring the timely use of materials, and ensuring the adequate and timely supply of materials on site.
3. Improving the efficiency of construction site management: automatically recording the entry and exit time, location, and other information of construction site personnel and equipment, reducing the cost and risk of manual management, and improving the efficiency and accuracy of site management.
4. Optimizing construction site scheduling: using image recognition technology to record construction logs and monitor the progress of various works at the construction site, adjusting the schedule promptly, improving construction efficiency, and reducing construction delays.

Construction activities at a job site vary widely. The machines subject to image recognition are excavators, loaders, dump trucks, cranes, and concrete mixer trucks, and the recognition accuracy is 69%, on average. The workers are wearing work clothing and reflective vests without a uniform standard, and they are at various locations within the job site performing various tasks, resulting in difficulties in recognition due to the bright side, dark side, and body position, and the recognition accuracy is 53%. The accuracy is 28% for the rebar. The reason for the low recognition accuracy could be that they are similar materials divided into two different classes; also, there are more than civil work activities at the job site; for example, there are plumbing and electrical tasks at a job site, and their materials, such as pipes and cables, may affect the recognition results, as shown in Table 6.

Table 6. Model performance indices.

	mAP	Recall (Threshold = 0.5)	Precision (Threshold = 0.5)	F1-Score (Threshold = 0.5)
Rebar	0.29	0.09	1.00	0.17
Worker	0.53	0.37	0.86	0.52
Machine	0.69	0.62	0.95	0.75

This study uses automatic identification of construction site workers, material locations, and construction environment conditions of equipment. The resulting photos can identify more than two items simultaneously, providing site supervisors with active warnings of potential occupational safety hazards and increasing construction efficiency through image automation.

5. Conclusions and Suggestions

A construction job site covers the building footprint, work area, or material storage. With the simultaneous recognition of objects, such as workers, machines, and materials using a single shot multibox detector (SSD) in this case, it was found that the recognition performed better for large machines, including excavators, cranes, dump trucks, and concrete mixer trucks, with recognition accuracy close to 70%. Recognition accuracy was 53% for workers, and rebar was the least accurately identified of the three.

This study used the single shot multibox detector model with the VGG-16 neural network as its backbone network and VGG-16 is a 16-layer convolutional neural network, including 13 convolutional layers and 3 fully connected layers. A total of 320 construction

site construction images (80%) were trained, and the results could mark personnel, machinery, and materials simultaneously. The complexity of each on-site construction image was different; therefore, the time required for each image recognition was also different, but the average single image recognition time was 6 s. The object detection process encountered the following problems:

1. Regarding detection personnel: For construction personnel, posture changes, construction site brightness changes, and object occlusion these problems would lead to false detections.
2. Regarding detection materials: densely packed rebar would produce different degrees of joint and section difficulties; in addition, in the single target detection algorithm, the stacking between the background and the foreground was different, which may have led to a decrease in the sensitivity of the model to the sample. It resulted in false detections.
3. Detection of equipment: Construction equipment detection items included excavators, shovel loaders, dump trucks, cranes, concrete mixer trucks, etc. There were more data sets than construction personnel and materials, and their identification performance was better. But to enhance the training of another project may have led to further overfitting.

Based on the above, this study proposes future research directions regarding technology application, database construction, and algorithm optimization to enhance the accuracy and applicability of detection items:

1. The evolutionary many-objective optimization algorithm with new techniques, such as domain decomposition and multi-objective optimization decomposition can improve the efficiency and accuracy of construction site management and enhance image recognition in construction engineering [46].
2. Optimizing truck scheduling through algorithms can improve the efficiency and accuracy of material transportation and scheduling at construction sites, leading to intelligent and automated material transportation and ultimately enhancing construction efficiency and quality [47].
3. Multi-objective optimization algorithms can significantly enhance the efficiency and accuracy of construction sites management tasks, such as material transportation, equipment scheduling, and personnel management. Integrating image recognition applications with these algorithms enables the intelligent and automated monitoring and control of construction sites, improving construction efficiency and quality [48].
4. Image recognition technology can monitor the construction site in real time, detect potential risk factors, and determine the direction of improvement. At the same time, efficient dock scheduling algorithms can optimize construction materials and equipment logistics, reduce waiting time, and improving overall productivity [49].
5. The direction is to combine image recognition technology to monitor the safety of construction sites in real time, detecting potential safety hazards early, and using NSGA-II and MOPSO algorithms for ambulance routing to improve rescue efficiency and emergency response capabilities [50].
6. Applying the augmented self-adaptive parameter control method to a broader range of construction scenarios can improve construction efficiency and safety. Further research will explore combining the technique with other optimization algorithms to enhance its effectiveness and reduce construction costs [51].
7. To enhance the simultaneous detection of personnel, equipment, and materials, upcoming methods will include feature pyramid, complete intersection over union (Ciou) loss, focal loss, and bag of freebies target detection optimization [52].

Construction engineering is characterized by complexity; therefore, image recognition technology at construction sites enhances the safety and efficiency of construction site management. This technology enables more detailed identification and improvement of

production efficiency and quality in the construction industry, thereby providing more significant development opportunities for the future of construction engineering.

Author Contributions: Conceptualization, Y.-R.W.; Methodology, L.-W.L.; Investigation, L.-W.L.; Resources, L.-W.L.; Writing—original draft, L.-W.L.; Writing—review & editing, L.-W.L.; Supervision, Y.-R.W.; Project administration, Y.-R.W.; Funding acquisition, Y.-R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Image in the research content is all owned by private companies, so they cannot be published.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, J.; Arif, O.; Vela, P.A.; Teizer, J.; Shi, Z. Tracking multiple workers on construction sites using video cameras. *Adv. Eng. Inform.* **2010**, *24*, 428–434. [CrossRef]
2. Riveiro, B.; Lourenço, P.B.; Oliveira, D.V.; González-Jorge, H.; Arias, P. Automatic morphologic analysis of quasi-periodic masonry walls from LiDAR. *Comput.-Aided Civ. Infrastruct. Eng.* **2016**, *31*, 305–319. [CrossRef]
3. Thakar, V.; Saini, H.; Ahmed, W.; Soltani, M.M.; Aly, A.; Yu, J.Y. Efficient Single-Shot Multi-Box Detector for Construction Site Monitoring. In Proceedings of the 2018 IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA, 16–19 September 2018; pp. 1–6.
4. Zhu, Z.; Ren, X.; Chen, Z. Visual tracking of construction jobsite workforce and equipment with particle filtering. *J. Comput. Civ. Eng.* **2016**, *30*, 04016023. [CrossRef]
5. Wang, Q.; Cheng, J.C.; Sohn, H. Automated estimation of reinforced precast concrete rebar positions using colored laser scan data. *Comput.-Aided Civ. Infrastruct. Eng.* **2017**, *32*, 787–802. [CrossRef]
6. Nimmo, J.; Green, R. Pedestrian avoidance in construction sites. In Proceedings of the 2017 International Conference on Image and Vision Computing New Zealand (IVCNZ), Christchurch, New Zealand, 4–6 December 2017; pp. 1–6.
7. Alizadehslehi, S.; Yitmen, I. A Concept for Automated Construction Progress Monitoring: Technologies Adoption for Benchmarking Project Performance Control. *Arab. J. Sci. Eng.* **2018**, *44*, 4993–5008. [CrossRef]
8. Fang, W.; Ding, L.; Luo, H.; Love, P.E. Falls from heights: A computer vision-based approach for safety harness detection. *Autom. Constr.* **2018**, *91*, 53–61. [CrossRef]
9. Mahami, H.; Nasirzadeh, F.; Ahmadabadian, A.H.; Esmaeili, F.; Nahavandi, S. Imaging network design to improve the automated construction progress monitoring process. *Constr. Innov.* **2019**, *19*, 386–404. [CrossRef]
10. Greeshma, A.S.; Edayadiyil, J.B. Automated progress monitoring of construction projects using Machine learning and image processing approach. *Mater. Today Proc.* **2022**, *65*, 554–563.
11. Del Savio, A.; Luna, A.; Cárdenas-Salas, D.; Vergara, M.; Urday, G. Dataset of manually classified images obtained from a construction site. *Data Brief* **2022**, *42*, 108042. [CrossRef]
12. Fang, Q.; Li, H.; Luo, X.; Ding, L.; Luo, H.; Rose, T.M.; An, W. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Autom. Constr.* **2018**, *85*, 1–9. [CrossRef]
13. Fang, W.; Ding, L.; Zhong, B.; Love, P.E.; Luo, H. Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. *Adv. Eng. Inform. Rmatics* **2018**, *37*, 139–149. [CrossRef]
14. Kim, Y.; Choi, Y. Smart Helmet-Based Proximity Warning System to Improve Occupational Safety on the Road Using Image Sensor and Artificial Intelligence. *Int. J. Environ. Res. Public Health* **2022**, *19*, 16312. [CrossRef] [PubMed]
15. Buniya, M.K.; Othman, I.; Sunindijo, R.Y.; Kashwani, G.; Durdyev, S.; Ismail, S.; Antwi-Afari, M.F.; Li, H. Critical Success Factors of Safety Program Implementation in Construction Projects in Iraq. *Int. J. Environ. Res. Public Health* **2021**, *18*, 8469. [CrossRef] [PubMed]
16. Yeşilmen, S.; Tatar, B. Efficiency of convolutional neural networks (CNN) based image classification for monitoring construction related activities: A case study on aggregate mining for concrete production. *Case Stud. Constr. Mater.* **2022**, *17*, e01372. [CrossRef]
17. Lee, H.; Grosse, R.; Ranganath, R.; Ng, A.Y. Convolutional deep belief networks for calable unsupervised learning of hierarchical representations. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 609–616.
18. Makantasis, K.; Protopapadakis, E.; Doulamis, A.; Doulamis, N.; Loupos, C. Deep convolutional neural networks for efficient vision based tunnel inspection. In Proceedings of the 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 3–5 September 2015; pp. 335–342.
19. Pan, Y.; Zhang, L. Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Autom. Constr.* **2021**, *122*, 10357. [CrossRef]

20. Yan, X.; Li, H.; Wang, C.; Seo, J.; Zhang, H.; Wang, H. Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion. *Adv. Eng. Inform.* **2017**, *34*, 152–163. [CrossRef]
21. Sepas-Moghaddam, A.; Etemad, A. Deep Gait Recognition: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 264–284. [CrossRef]
22. Lin, C.-L.; Fan, K.-C.; Lai, C.-R.; Cheng, H.-Y.; Chen, T.-P.; Hung, C.-M. Applying a Deep Learning Neural Network to Gait-Based Pedestrian Automatic Detection and Recognition. *Appl. Sci.* **2022**, *12*, 4326. [CrossRef]
23. Arabi, S.; Haghighat, A.K.; Sharma, A. A deep-learning-based computer vision solution for construction vehicle detection. *Comput.-Aided Civ. Infrastruct. Eng.* **2020**, *35*, 753–767. [CrossRef]
24. Chou, J.-S.; Liu, C.-H. Automated Sensing System for Real-Time Recognition of Trucks in River Dredging Areas Using Computer Vision and Convolutional Deep Learning. *Sensors* **2021**, *21*, 555. [CrossRef]
25. Li, Y.; Lu, Y.J.; Chen, J. A deep learning approach for real-time rebar counting on the construction site based on YOLOv3 detector. *Autom. Constr.* **2021**, *124*, 103602. [CrossRef]
26. Cha, Y.J.; Choi, W.; Büyüköztürk, O. Deep learning-based crack damage detection using convolutional neural networks. *Comput.-Aided Civ. Infrastruct. Eng.* **2017**, *32*, 361–378. [CrossRef]
27. Chang, C.W.; Lin, C.H.; Lien, H.S. Measurement radius of reinforcing steel bar in concrete using digital image GPR. *Constr. Build. Mater.* **2009**, *23*, 1057–1063. [CrossRef]
28. CS231n Convolutional Neural Networks for Visual Recognition, Stanford. 2016. Available online: http://cs231n.stanford.edu/ (accessed on 27 March 2023).
29. Martinez, P.; Al-Hussein, M.; Ahmad, R. A scientrometic analysis and critical review of computer vision applications for construction. *Autom. Constr.* **2019**, *107*, 102947. [CrossRef]
30. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
31. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
32. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef]
33. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
35. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
36. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot Multi-box detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
37. Yudin, D.; Slavioglo, D. Usage of fully convolutional network with clustering for traffic light detection. In Proceedings of the 2018 7th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 10–14 June 2018; pp. 1–6.
38. Wang, Y.; Wang, C.; Zhang, H. Combining a single shot Multi-box detector with transfer learning for ship detection using sentinel-1 SAR images. *Remote Sens. Lett.* **2018**, *9*, 780–788. [CrossRef]
39. Deshpande, A. A Beginner's Guide to Understanding Convolutional Neural Networks. Retrieved March 2017; Volume 31. Available online: https://adeshpande3.github.io/A-Beginner\T1\textquoterights-Guide-To-Understanding-Convolutional-Neural-Networks/ (accessed on 27 March 2023).
40. Dorafshan, S.; Thomas, R.J.; Maguire, M. Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. *Constr. Build. Mater.* **2018**, *186*, 1031–1045. [CrossRef]
41. Spencer, B.F., Jr.; Hoskere, V.; Narazaki, Y. Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering* **2019**, *5*, 199–222. [CrossRef]
42. Dung, C.V. Autonomous concrete crack detection using deep fully convolutional neural network. *Autom. Constr.* **2019**, *99*, 52–58. [CrossRef]
43. Fang, W.; Love, P.E.; Luo, H.; Ding, L. Computer vision for behaviour-based safety in construction: A review and future directions. *Adv. Eng. Inform.* **2020**, *43*, 100980. [CrossRef]
44. Li, X.; Chi, H.L.; Lu, W.; Xue, F.; Zeng, J.; Li, C.Z. Federated transfer learning enabled smart work packaging for preserving personal image information of construction worker. *Autom. Constr.* **2021**, *128*, 103738. [CrossRef]
45. Del Savio, A.A.; Luna, A.; Cárdenas-Salas, D.; Vergara Olivera, M.; Urday Ibarra, G. The use of artificial intelligence to identify objects in a construction site. In Proceedings of the International Conference on Artificial Intelligence and Energy System (ICAIES) in Virtual Mode, Jaipur, India, 12–13 June 2021.
46. Zhao, H.; Zhang, C. An online-learning-based evolutionary many-objective algorithm. *Inf. Sci.* **2020**, *509*, 1–21. [CrossRef]
47. Dulebenets, M.A. An Adaptive Polyploid Memetic Algorithm for scheduling trucks at a cross-docking terminal. *Inf. Sci.* **2021**, *565*, 390–421. [CrossRef]

48. Pasha, J.; Nwodu, A.L.; Fathollahi-Fard, A.M.; Tian, G.; Li, Z.; Wang, H.; Dulebenets, M.A. Exact and metaheuristic algorithms for the vehicle routing problem with a factory-in-a-box in multi-objective settings. *Adv. Eng. Inform.* **2022**, *52*, 101623. [CrossRef]
49. Dulebenets, M.A. A novel memetic algorithm with a deterministic parameter control for efficient berth scheduling at marine container terminals. *Marit. Bus. Rev.* **2017**, *2*, 302–330. [CrossRef]
50. Rabbani, M.; Oladzad-Abbasabady, N.; Akbarian-Saravi, N. Ambulance routing in disaster response considering variable patient condition: NSGA-II and MOPSO algorithms. *J. Ind. Manag. Optim.* **2022**, *18*, 1035. [CrossRef]
51. Kavoosi, M.; Dulebenets, M.A.; Abioye, O.F.; Pasha, J.; Wang, H.; Chi, H. An augmented self-adaptive parameter control in evolutionary computation: A case study for the berth scheduling problem. *Adv. Eng. Inform.* **2019**, *42*, 100972. [CrossRef]
52. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.