Comparative Study of Automatic Multi-class Object Detection Algorithms with Transfer Learning based on a Dataset from Construction Sites

Jing Zhang[a], Carl Haas[a], Sean Hanna[b]
[a]University of Waterloo, Canada, [b]University College London, United Kingdom
j2546zhang@uwaterloo.ca

Abstract. Advanced technologies, such as Computer Vision, are helping to transform the traditional Architecture, Engineering, and Construction (AEC) industry. Although this cutting-edge artificial intelligence technology has begun to enter more construction sites, automated methods of data collection for construction site management have room for further development. Therefore, this paper compares two vision-based automatic detection algorithms for multiple categories of moving target objects and their improvement based on the construction site dataset (MOCS). The study methodology used includes two stages: (1) Basic Model: The first stage compares the detection performance of mainstream target detection algorithms (Faster R-CNN and YOLOv7) on the same database, and (2) Transfer Learning: The second phase added the pre-training network by transfer learning strategy, anticipating that it may improve detection performance in the initial algorithms. This digital route relies on an HD camera installed on UAVs to achieve monitor automation in the construction process to support supervision engineers.

1. Introduction

As one of the significant issues in the study of computer vision, the work of object detection is to find out nearly all the targets, also called objects, in the given image. After that, the location and scale of these targeted objects also need to be acquired. Since all varieties of objects have diverse appearances, shapes, postures, and interference from factors in the different images, object detection may have become one of the trickiest matters in the research field of advanced computer vision. Therefore, in essence, target detection includes two main tasks: object image recognition and object location in the image. At present, target detection is mainly used in pedestrian detection, vehicle detection, face recognition, medical image detection, and so on.

The core problems of target detection can be summarized into three points: (1) the distinction of object categories in the image; (2) the determination of the position of the target in the image; (3) the consideration of the size and shape of the target. According to whether the algorithm needs to generate region proposals in the middle, the current target detection algorithm can be roughly divided into two categories, namely two-stage and one-stage. Additionally, the multi-stage algorithm has also been proposed before. But, because the calculation speed of the multi-stage algorithm is slow and the detection accuracy has not been significantly improved, it is rarely used in actual scenarios. Consequently, the current mainstream target detection algorithms can be mainly divided into two types: one-stage as well as two-stage.

(1) The two-stage algorithm includes two stages to solve this object detection problem. First, the candidate area is generated, and then the candidate area is classified after the location is refined. The two-stage detection algorithm has a low recognition error rate and a low miss-recognition rate, but the speed is slow and cannot meet real-time detection scenarios, such as video target detection.

(2) The one-stage algorithm merely includes one stage to solve this object detection problem, which directly produces the category possibility and location coordinate value of the target and

can directly obtain the final detection result, thus its detection speed is much faster, but the general recognition accuracy is worse than the two-stage algorithm.

Depending on application, speed or accuracy could be more important, consequently, this paper selected two mainstream object detection algorithms from the one-stage algorithm YOLO as well as the two-stage algorithm R-CNN. After that, these two chosen algorithms-YOLOv7 (Wang, Bochkovskiy and Liao, 2022) and Faster R-CNN (Girshick, 2015) respectively are utilized to achieve object recognition and detection for moving objects, such as workers, tower cranes, pump trucks, and pile driving, on construction sites. Then, they are improved by means of a transfer learning approach (Weiss, Khoshgoftaar and Wang, 2016). From the background cited in the previous introduction and based on experiments in other applications and on the structure of the two algorithms, one would expect R-CNN to be more accurate while YOLOv7 to be faster. And, after training and testing, the experiment results may show that transfer learning, based on these two algorithms, improves target detection.

2. Background and Related Work

Many previous kinds of research about object detection methods on construction sites were relying on manual work. However, due to the developments in AI technologies in recent years, researchers introduced machine learning approaches into the AEC industry. Consequently, solutions for monitoring construction sites based on automatic vision detection schemes have become popular. The significant reason is that this method increases the range of categories for object recognition on construction sites, from the original only being able to recognize people and appliances to more categories, such as tower cranes, workers, excavators, mixers, trucks, etc. After that, when Memarzadeh, Golparvar-Fard and Niebles (2013) distinguished between worker characteristics and equipment characteristics, they proposed that a Histogram of Oriented Gradient (HOG) and a new multi-binary SVM classifier could be used to identify important objects on construction sites. Based on the previous SVM idea, one-vs-all multi-class SVM realizes semi-automated detection of more than 20 kinds of moving objects on construction sites. This was a significant breakthrough, however relatively large differences in detection accuracy remained for different objects. The average detection accuracy of these three types of main objects-workers, dump trucks, and excavator-on construction sites are 98.83%, 84.88%, and 82.10% separately. Later, more detailed detection of machinery and personnel was proposed, and experiments were conducted by many researchers. Nevertheless, the detected categories of objects on construction sites obtained by this method are still small and does not have generalization. Other problems are that the speed of automatic processing is still relatively low, and it remains easy to make mistakes.

Consequently, more cutting-edge deep learning algorithms appeared. The introduction of deep networks, specifically convolution, produced a jump in the detection accuracy. For example, the emergence of the R-CNN family, Single Shot Multibox Detector (SSD) and You Only Look Once (YOLO) in recent years has gradually been applied to various research fields, and this deep learning approach has also attracted widespread attention from civil engineering experts together with construction researchers. Thus, the region-based convolutional neural networks (IFaster R-CNN) were then proposed based on Faster R-CNN to detect workers and excavators on construction sites (Fang et al., 2018). Although only two types of target objects can be detected this way, their detection accuracy is quite high: 91% and 95% respectively. YOLOv5 was then used to detect whether construction workers in different postures were wearing helmets on the construction site when they were working. In addition, the algorithm that combines CNN with the long short-term memory (LSTM) model (Hochreiter and Schmidhuber,

1997) to identify unsafe behaviors on construction sites has also been proven to have good results. Subsequently, orientation-aware feature fusion single-stage detection (OAFF-SSD) was proposed by Guo, Xu and Li (2020). This new end-to-end neural network is suitable for detecting abundant densely gathering vehicles which are utilized to transport construction materials on construction sites. This intelligent learning method currently used can improve accuracy and speed to a certain extent. Besides, the trained network may be a reference for other models and scenarios.

Nonetheless, since these mainstream learning algorithms and related research used to identify target objects on construction sites belong to supervised learning in machine learning, it is necessary to obtain a large number of training sets from actual scenes so as to train neural networks. However, a huge construction site project under construction is fundamental for obtaining the data set as the training input. In addition, the collection of pictures and videos also requires the support of related hardware equipment, such as a drone with good performance and a high-definition camera that requires a certain degree of definition. Most importantly, these data sets need to be manually annotated after obtaining pictures or videos. This annotation process will consume a lot of manpower and time. Therefore, Yabuki, Nishimura and Fukuda (2018) used two experiments to compare the performance of algorithms that: (1) used transfer learning and did not use transfer learning, and (2) that used transfer learning and traditional deep learning. They were based on the SSD algorithm and thus do not show different experimental results of transfer learning on different algorithms. Meanwhile, they also did not achieve more classes instead of only several categories. Afterward, Nath and Behzadan (2020) compared the transfer learning performance based on YOLOv2 and YOLOv3 partly and the precision of the latter is much better than that of the former. However, the model is only suitable for inspection scenes on well-lit construction sites, so it has certain limitations. In a paper recently published, Xiao et al. (2017) implemented the automatic detection of multi-class objects on the construction site scene based on semi-supervised learning by transfer learning. Although this data set contains 10,000 labeled photos, and it can feed many training sets to the neural network, it only targets the Faster R-CNN algorithm and does not achieve the effect of other algorithms through the semi-supervised method. Thus, this paper compares the mainstream algorithm YOLO in the one-stage algorithm and the mainstream algorithm Faster R-CNN in the two-stage algorithm and their respective algorithm effects based on migration learning, which proves that it can be adapted to the classification and localisation of more than 10 common target objects on the construction site.

3. Methodology

There are five key processes in the implementation of object detection algorithms. To begin, photographic data was used that was collected from nearly two hundred construction sites, and the data format was modified to fit the needs of the project. Object detection models were created in the second step using the PyTorch platform. The three primary responsibilities in this step were configuring the computer programming environment, the deep learning platform (PyTorch), and the object identification API. Third, to increase the generalizability of the object detection technique to unseen data, models based on chosen object detection algorithms were trained on the training data set and changed based on validation data sets. On the tagged images, two models based on Faster R-CNN and YOLOv7 were trained. The number of photos used, the algorithm used, and the training steps all had an impact on detection accuracy. Following that, the PyTorch platform is utilized to achieve the fine-tuning of the pre-training model and retrain the network. The final stage was to examine the performance-trained models using the metrics that had been chosen.

In total, 43,000 photographs were taken at 174 distinct construction sites from MOCS (An et al., 2021). Five distinct imaging devices were used to take images with varying views, heights, and illumination. Meanwhile, the MOCS photos span a wide range of construction projects and jobs. The kind of bounding box and mask were precisely marked in 222,861 instances of 13 classes. The ultimate border boxes were designed by searching the boundaries of the masks instead of using the coarse boxes recommended by four experts. The annotation style of the MOCS dataset is similar to the COCO dataset (Lin at al., 2014). There is a single JSON file for each divided data set to store the annotation information, covering the training set, validation set, and test set. There are over 1000 photographs in each category, with "worker" having the most instances and images. This is in line with the reality of building sites. Training (19,404 images), validation (4000 photos), and testing (18,264 pictures) were divided into three components in the MOCS dataset. MOCS dataset.

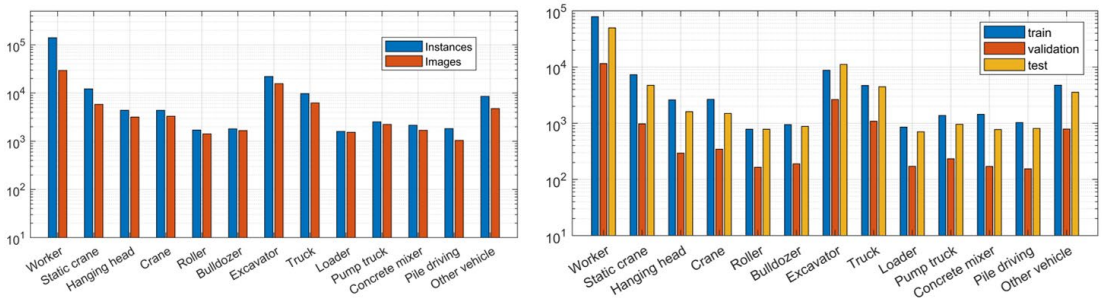

Figure 1: Classifications of moving objects from the construction site and its images' number and instance number in a) total dataset b) training set, validation set together with a test set

Later, transfer learning mainly comes through the usage of pre-trained models in the field of computer vision. A pre-trained model is a model that has been trained to address comparable issues on a large benchmark. Thus, the author adopted COCO due to its similarity to MOCS in the dataset format and object category. Because of the high computing costs of training, importing the published findings and using the appropriate model is a typical technique. Hence, the use of ResNetX101 and Darknet53 backbones based on training COCO is applied respectively to two algorithms-Faster R-CNN and YOLO.

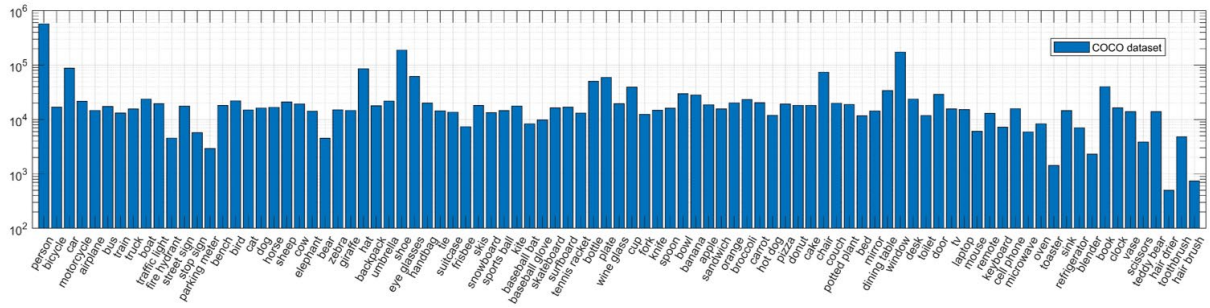

Figure 2: The object categories in COCO

Transfer learning in this project principally adopts the following fine-tuning method. Initially, pre-train the ResNetX101 and Darknet53 network models separately on COCO. Then, copy all the model designs and their parameters on the source model. That is, the output of the last convolutional layer of the trained source model is used as the CNN feature, and then the Softmax classifier as activation function is directly used for solving multi-class classification problem. After that, the output layer whose output size is the number of MOCS categories is added to the target model, and the model parameters of this layer are randomly initialized. Finally, train Faster R-CNN and YOLOv7 network on MOCS. That is to say, the output layer

will be trained from scratch, and the parameters of all remaining layers will be fine-tuned according to the parameters of the source model.

This research is based on the Pytorch platform (Pytorch 1.8.0+cu111), which includes numerous deep learning algorithm packages for programmers' convenience. Two GPUs are utilized to train the network and are the GeForce RTX 3090. The system platform is Linux and the language is Python 3.7.10. After Pytorch is installed, Faster R-CNN and YOLOv7 models are trained and validated. Instead of training from scratch, COCO was used to pre-train Faster R-CNN and YOLOv7 models. To minimize the time during training and compensate for the lack of a large data set, a pre-trained model was used. The usage of the COCO data set led to a simple fine-tuning of the training set's parameters.

4. Experiment Results

It is usually the case that training set curves continue to go down perpetually. However, test set curves hit a minimum and then begin to rise, as the network begins to over-fit to the data. The minimum point is the actual minimum error.

The output of Cross-Entropy is the logarithm of the likelihood of the correct label, which has a certain relationship with the accuracy, but the value range is larger. The Cross-Entropy loss formula is as follows:

$$H\left(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)}\right) = -\sum_{j=1}^{q} y_j^{(i)} \log \hat{y}_j^{(i)} \tag{1}$$

When the training sample is n, the cross-entropy loss function is as follows:

$$\ell(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} y_j^{(i)} H\left(\boldsymbol{y}^{(i)}, \hat{\boldsymbol{y}}^{(i)}\right) \tag{2}$$

The definition of accuracy in the classification problem is seeing if the predicted category is compatible with the true category. The result is an N-dimensional vector for the classification problem. The category of each position of the vector and its value denotes the likelihood that the projected goal corresponds to that category. Take the label with the highest probability as the final anticipated result label when the predicted result is produced.

The loss curves are drawn based on all training datasets from MOCS. In the following Figure 3, YOLO's loss value is 8 times more than Faster R-CNN's when they arrive at a similar iteration in the first, which may demonstrate YOLOv7 needs more training time than Faster R-CNN. However, the loss significantly declines with the number of training iterations in both basic and transfer learning models in YOLOv7. Importantly, in the training process, the loss value of the transfer learning model is almost always lower than the loss value of the basic model. This also indicates that the method of transfer learning may shorten the training time to reach the same error.

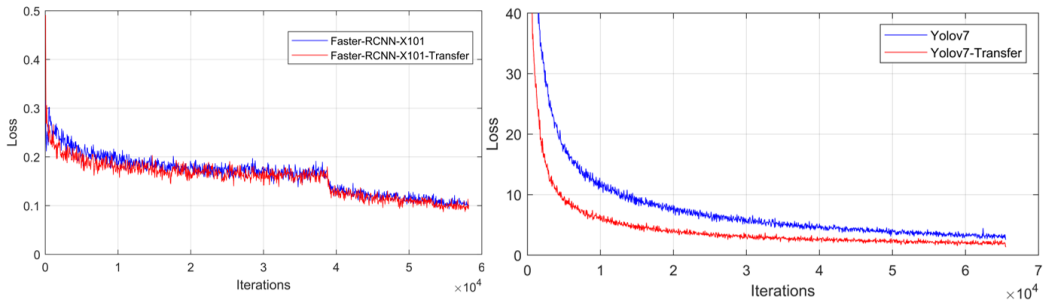

Figure 3: Loss changes of a) Faster R-CNN and b) YOLOv7 in the training data

The mAP-50 curves are calculated based on all test data from MOCS. In Figure 4 below, Faster R-CNN's initial accuracy is around 5 times higher than YOLO's. However, with the increase of Epoch, in terms of the final test precision, although both are a little more than 70%, YOLOv7's is slightly higher than Faster-RCNN's. Moreover, as predicted, the mAP significantly increases and then levels off with the number of training Epochs in both basic and transfer learning models. Interestingly, in the training process, the mAP value of the transfer learning model is sometimes lower than the mAP value of the basic model. However, the approach of transfer learning can perhaps improve the test accuracy of neural networks in the end.

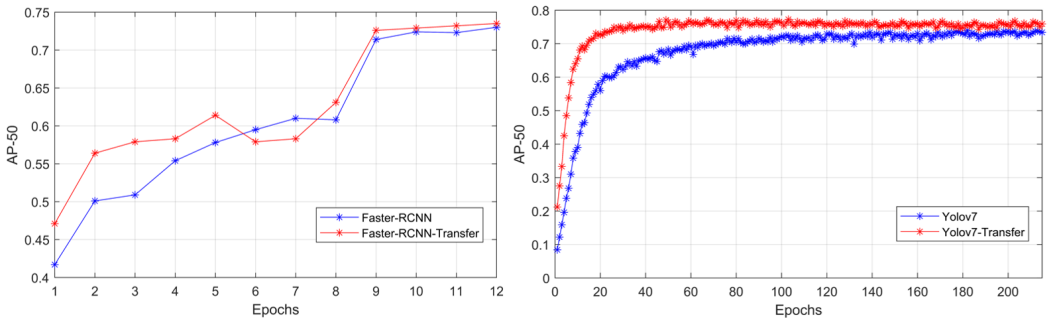

Figure 4: mAPs of a) Faster R-CNN and b) YOLOv7 in the test data

In summary, the Faster R-CNN algorithm is much better applied for this task than YOLOv7 based on the beginning comparison while the performance difference is distinct in the end. Meanwhile, this analysis also proves that the method of transfer learning could improve the training effect of neural networks. Hence, the lead author recommends the Transfer model and prefers the Yolov7 algorithm.

Considering the evaluation metric, the COCO target detection evaluation index is selected because the MOCS dataset is more like the COCO dataset rather than the PASCAL VOC. This set of indexes includes six metrics, those are mAP, mAP_50, mAP_75, mAP_s, mAP_m, and mAP_l. Table 1 includes the precise definition of those metrics. Table 2 presents the precision and speed indexes separately from 'basic' Faster R-CNN and YOLOv7.

Table 1: Evaluation Metric explanation of Faster R-CNN and YOLOv7

Metrics	Definition
mAP	mean Average Precision over all classes
mAP_50	AP at Intersection over Union (IoU) = 50%
mAP_75	AP at Intersection over Union (IoU) = 75%
mAP_s	AP for small objects (area of object < 32*32 pixels)
mAP_m	AP for medium objects (32*32 pixels < area of object < 96*96 pixels)
mAP_l	AP for large objects (area of object > 96*96 pixels)
speed	average iterations of frame image each second

Table 2: Evaluation Metric Result of Faster R-CNN and YOLOv7

	mAP	mAP_50	mAP_75	mAP_s	mAP_m	mAP_l	speed
Faster_rcnn	0.493	0.73	0.545	0.205	0.383	0.697	6.23
Faster_rcnn-trans	0.508	0.735	0.558	0.186	0.414	0.62	6.26
Yolov7	0.502	0.734	0.528	0.225	0.384	0.565	27.63
Yolov7-trans	0.515	0.759	0.561	0.243	0.412	0.593	27.82

In the index of detection mean precision of every image when IOU is set as 50%, the accuracy of YOLOv7 is 0.004 higher than that of Faster R-CNN. However, after utilizing the transfer learning methods, the accuracy of YOLOv7 is 0.024 higher than that of Faster R-CNN. It also means the accuracy rate of YOLOv7 is apparently improved after the transfer learning, while that of Faster R-CNN is slightly improved. This comparison shows the effect of transfer learning on the accuracy index based on the object algorithms increased to a degree.

In the index of detection time of frame image every second, the speed of YOLOv7 is more than four times higher than that of Faster R-CNN. However, after utilizing the transfer learning method, whether Faster R-CNN or YOLOv7, its detection speed has not improved much, like test speed results in similar experiments from other papers. Hence, this comparison suggests the effect of transfer learning on the speed index based on the object algorithms slightly increased.

Considering both indexes of detection precision and speed, this may be likely a comparison process of balance. For models that use transfer learning and those that do not use transfer learning, overall, whether it is detection accuracy or detection speed, the training effect of the model using the transfer learning method is better. Then in the model that uses the transfer learning for comparison, if it is from the perspective of detection accuracy, YOLOv7 based on transfer learning has higher detection accuracy than Faster R-CNN based on transfer learning; if it is from the perspective of detection speed, YOLOv7 based on transfer learning has also higher detection speed than Faster R-CNN based on transfer learning. However, in the actual construction site application scenario, the importance of the measurement index of detection accuracy is far greater than that of detection speed. Therefore, considering both detection accuracy and detection speed, the author supposes that the YOLOv7 model based on transfer learning has quite a good effect in classifying and localising various types of moving objects on construction sites.

In Figure 5, the different kinds of moving objects in the test images and videos from construction sites in real life can be well detected. However, the YOLO algorithm is a little easier to ignore some common objects that are visible to the naked eye than Faster R-CNN. Nonetheless, for instance, neither of the two algorithms detected one worker with an unusual pose in the test. But the applied detection algorithms are still able to detect most tiny objects to some extent.

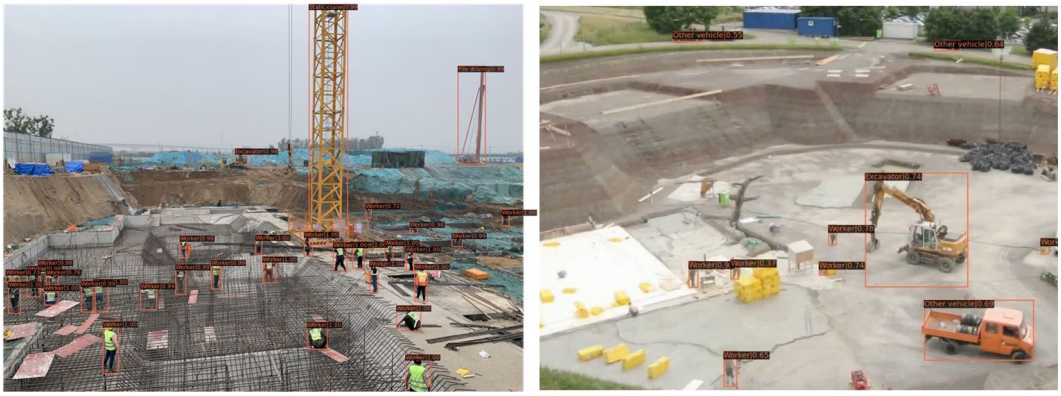

Figure 5: Examples of object recognition image effect in construction sites

5. Discussion

This research first compared various types of multi-target detection algorithms including

YOLOv7 as well as Faster R-CNN. Additionally, our study has also proposed and tested a transfer learning detection approach based on these two algorithm frameworks. The aim of those experiments is for vision-based automatic monitoring of many varieties of moving objects from construction sites. Finally, the results of those experiments show that the test scheme realized the study goal of improving the detection performance of moving multiple classes of targets from construction sites. However, the improvement of detection performance is also different because it is based on different object detection algorithms. Several improvements are possible.

The proposed transfer learning scheme offers a dependable backbone network for vision-based object recognition and detection on construction sites. The backbone network is quite important for the detection of target objects because it can influence the detection precision. The suggested improvement scheme realized a mAP@50 of 75.9% in those experiments, which was the better performance during the process of training the network with the MOCS dataset. The automatic detection algorithms may be integrated into automated construction engineering management processes. For instance, this approach is perhaps able to address issues including the production schedule of construction workers, the collision monitoring of construction equipment, as well as identification of construction process activities on the construction site. It is generally considered that the MOCS dataset has a significant impact on the comparison and improvement of different kinds of detection algorithms. The MOCS dataset contains the largest number of images by now in recent studies of the civil engineering field. In detail, those images are captured in various construction projects, such as building engineering as well as bridge construction, and so on, and work tasks, including foundation work and decoration work, and other works, with different photo angles using HD cameras that are installed on drones. Furthermore, the moving objects on the construction site in the MOCS dataset contain a variety of features about vision information covering size, color, angle, as well as position. Therefore, a relatively higher variety of the training dataset leads to better performance of robustness as well as the generalizability of different kinds of detection algorithms. However, if the number of images together with instances in MOCS were a little more, the detection effect based on various kinds of algorithms will be further improved. Additionally, the universal suitability of the suggested approach perhaps could be enhanced by means of improving the annotation quality in MOCS. In addition, this dataset is collected from Mainland, China. Thus, the detection performance of this dataset would also be better if it could be also collected from other countries and regions.

The proposed YOLOv7 approach realized better detection precision performance than Faster R-CNN. Furthermore, this paper supplemented an improvement experiment in which the transfer learning scheme was trained on the MOCS training set and obtained mAP@50 of 73.6% (Faster R-CNN-Trans) and 75.9% (YOLOv7-Trans) on the validation set. The pre-training model used in mainstream target detection algorithms is effective and efficient for improving multi-target detection performance in construction engineering scenarios. Transfer learning could be defined as the expansion of training data quantities in order to improve detection performance. The main reason is that this training process is more advanced than before due to the basis of prior knowledge provided. The strong pre-training model provides the detection algorithms with the architecture as well as parameters of the backbone network which are training well by means of similar datasets. In this study, the paper has proposed a transfer learning strategy combining the pre-training model including ResNetX101 as well as darknet53 and traditional detection algorithm frameworks. In experiments, the proposed approach outperforms separately the traditional detection method of YOLOv7 as well as Faster R-CNN by 2.59% and 3.04% on mAP evaluation metrics, which demonstrates the feasibility of the transfer learning method for improving the multi-class object detection in construction

engineering scenes. The method proposed by the authors has two limitations that may be addressed in later research. Firstly, the transfer learning approach is limited to the two detection algorithms - YOLOv7 and Faster R-CNN in this paper. However, it does not mean that the detection accuracy and speed of almost all target algorithms will be significantly improved merely through the adoption of a transfer learning strategy in those two algorithms. Nonetheless, transfer learning means could further improve the performance of most multi-target detection algorithms. This observation can be further investigated in future experiments.

6. Conclusion

With the development of advanced technologies, such as Computer Vision, the Internet of Things (IoT), and Mobile Robotics, the traditional AEC industry is also transforming and upgrading in recent years. Although these cutting-edge artificial intelligence and robotics technologies have begun to enter more and more construction sites, the automation methods of building construction site supervision have yet to be developed. Therefore, this project compares a transfer learning method based on a one-stage algorithm (YOLOv7) and a two-stage algorithm (Faster R-CNN) for target detection of moving objects on construction sites. The proposed improvement method introduces a pre-training model structure. The network is mainly composed of two main modules. The first part is the pre-training model, and the second part is the algorithm structure of target detection. For the YOLOv7 algorithm, the darknet53 network trained on the COCO data set was used for fine-tuning, and then its network structure and weights were replaced by the darknet53 network in the original YOLOv7 structure for re-training; for the Faster R-CNN algorithm, COCO was used. The ResNetX101 network trained on the data set is fine-tuned, and then its network structure and weights are replaced with the ResNetX101 network in the original YOLOv7 structure for re-training. The proposed improved method has been trained, validated, and tested based on the MOCS data set related to the recognition of moving objects on construction sites. Meanwhile, fine-tuning for specific hyper-parameters for the mentioned architectures alongside the transfer learning approach contributes to an overall improvement in the evaluation metrics used for the assessment of the trained models without impacting inference speeds. Therefore, the proposed improved method of transfer learning has a significant effect on the mainstream one-stage algorithm or the mainstream two-stage algorithm in target detection. However, through algorithm comparison experiments, it is demonstrated that although transfer learning can effectively increase the effect of training, it has a more significant improvement effect on Faster R-CNN's mAP than YOLO's mAP. Nevertheless, the proposed method provides a more effective solution for infrastructure project applications that rely on vision-based moving object detection.

The contribution of this research has three main aspects. Firstly, it compared two kinds of classical YOLO and R-CNN target detection algorithms with different structures on the automatic monitoring effect of different types of moving objects on the construction site. Transfer learning was observed to have a greater benefit in R-CNN than in YOLO, as expected. In addition, the pre-training model can be integrated into the framework of deep learning algorithms with better recognition effects in the future, which is also what is expected by the lead author. Secondly, the proposed method effectively improves the training and testing effects of the neural network and can more accurately detect the target moving objects on the construction site, thereby improving the effect of visual recognition and detection based on deep learning detection on the actual construction site and application feasibility in the project. Thirdly, this study also compared different algorithm frameworks based on transfer learning. Among the compared transfer learning strategies, compared with YOLOv7, the pre-trained model can significantly improve the detection accuracy of the Faster R-CNN algorithm; and in

terms of detection speed, The detection speed of YOLOv7 after using transfer learning is higher than that of Faster R-CNN.

In summary, contrary to the lead author's prediction that the two-stage algorithm may be more accurate while the one-stage detects faster, Yolov7 has a slightly better detection performance than Faster R-CNN. The reason why this paper focuses more on accuracy is that the detection precision measure index is more appropriate for determining good use on construction sites. Furthermore, due to both algorithms being relatively fast to test, the accuracy metrics seem to be more essential. Consequently, the author suggests selecting a One-stage algorithm: Yolov7 (darknet53) based on the transfer learning approach in the experiment to achieve multi-class object detection on a typical building construction site. In the automatic object detection scheme, later, this digital technology relies on an HD camera which is installed on UAVs to achieve monitor automation of various categories of moving objects in the construction process instead of supervision engineers. Through the combination of traditional construction process management and frontier vision detection technologies, the improvement of automatic site monitoring will bring many benefits to the construction engineering industry.

References

Wang, C.Y., Bochkovskiy, A. and Liao, H.Y.M., 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.

Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

Weiss, K., Khoshgoftaar, T.M. and Wang, D., 2016. A survey of transfer learning. Journal of Big data, 3(1), pp.1-40.

Memarzadeh, M., Golparvar-Fard, M. and 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, pp.24-37.

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

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

Guo, Y., Xu, Y. and Li, S., 2020. Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network. Automation in Construction, 112, p.103124.

Yabuki, N., Nishimura, N. and Fukuda, T., 2018. Automatic object detection from digital images by deep learning with transfer learning. In Advanced Computing Strategies for Engineering: 25th EG-ICE International Workshop 2018, Lausanne, Switzerland, June 10-13, 2018, Proceedings, Part I 25 (pp. 3-15). Springer International Publishing.

Nath, N.D. and Behzadan, A.H., 2020. Deep convolutional networks for construction object detection under different visual conditions. Frontiers in Built Environment, 6, p.97.

Xiao, B., Zhang, Y., Chen, Y. and Yin, X., 2021. A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation. Advanced engineering informatics, 50, p.101372.

An, X., Li, Zhou, L., Liu, Z., Wang, C., Li, P. and Li, Z., 2021. Dataset and benchmark for detecting moving objects in construction sites. Automation in Construction, 122, p.103482.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.