A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation

Xiao, Bo; Zhang, Yuxuan; Chen, Yuan; Yin, Xianfei

Published in:
Advanced Engineering Informatics

Published: 01/10/2021

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1016/j.aei.2021.101372

Publication details:
Xiao, B., Zhang, Y., Chen, Y., & 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, Article 101372. https://doi.org/10.1016/j.aei.2021.101372

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation

Bo Xiao a, Yuxuan Zhang a, Yuan Chen b, Xianfei Yin c, * a Department of Civil and Environmental Engineering, University of Alberta, Edmonton T6G 2R3, Canada b College of Management and Economics, Tianjin University, Tianjin 300072, PR China c Department of Construction Management and Engineering, University of Twente, Enschede, The Netherlands

ARTICLE INFO

Keywords: Deep learning Object detection Teacher-student networks Data augmentation Vision-based monitoring Construction sites

ABSTRACT

Recently, deep-learning detection methods have achieved huge success in the vision-based monitoring of construction sites in terms of safety control and productivity analysis. However, deep-learning detection methods require large-scale datasets for training purposes, and such datasets are difficult to develop due to the limited accessibility of construction images and the need for labor-intensive annotations. To address this problem, this research proposes a semi-supervised learning detection method for construction site monitoring based on teacher–student networks and data augmentation. The proposed method requires a limited number of labeled data to achieve high detection performance in construction scenarios. Initially, the proposed method trains the teacher object detector with labeled data following weak data augmentation. Next, the trained teacher object detector generates pseudo-detection results from unlabeled images that have been weakly augmented. Finally, the student object detector is trained with the pseudo-detection results and unlabeled images that have been both weakly and strongly augmented. In our experiments, 10,000 annotated construction images from the Alberta Construction Image Dataset (ACID) have been divided into a training set (70%) and a validation set (30%). The proposed method achieved a 91% mean average precision (mAP) on the validation set while only requiring 30% of the training set. In comparison, the existing supervised learning method ResNet50 Faster R-CNN achieved a mAP of 90.8% when training on the full training set. These experimental results show the potential of the proposed method in terms of reducing the time, effort, and costs spent on developing construction datasets. As such, this research has explored the potential of semi-supervised learning methods and increased the practicality of vision-based monitoring systems in the construction industry.

1. Introduction

Continuous monitoring is an efficient way for project managers to follow the progress of projects, to evaluate crew productivity (i.e., direct and indirect costs), and to identify any safety risks on worksites (e.g., potential collisions between workers and equipment) [1]. Monitoring construction sites offers the potential to deliver high-quality projects without unexpected delays. The traditional, manual, way of monitoring construction sites is for project managers to visit their sites in person. Given that modern construction sites are complex and dynamic, manual monitoring is both time-consuming and error-prone [2]. To avoid such limitations, considerable research effort has been put into developing automated monitoring systems for construction sites by adopting cameras, sensors, drones, etc.

Compared with other high-tech solutions such as global positioning systems (GPS) [3] and laser scanners [4], cameras have the advantages of being low-cost, easy to deploy, and offering a large monitoring range, making them suitable for monitoring on construction sites. After investigating 142 construction professionals in the US, Bohn and Teizer [5] suggested the cameras can efficiently reduce travel, safety, communication, and documentation costs in project management. Further, automatic analysis of construction videos using vision-based methods is beneficial for productivity analysis and safety control. For example, Xiao and Kang [6] have proposed a vision-based method to calculate excavator productivity in earthmoving operations by continuously identifying excavators and dump trucks. Son et al. [7] have...
developed a real-time warning system to avoid potential collisions by video-tracking workers and construction machines.

The detection of construction objects (e.g., workers, machines, and materials) from images or videos is the fundamental step in developing such applications. Recently, deep-learning detection methods [8] that build upon convolutional neural networks (CNN) have proven their ability to robustly detect predefined classes of construction objects, even under challenging scenarios such as occlusions, illumination variations, and motion blurs [9]. Furthermore, deep-learning detection methods are able to achieve near real-time detection speeds by adopting parallel computation on advanced graphic cards, something that is difficult for non-deep learning methods [10]. Consequently, deep-learning detection methods have become a standard component of many state-of-the-art automation studies [11-14] on the efficient monitoring of construction sites.

A large image dataset that includes labeled construction objects is essential to apply deep-learning detection methods in construction scenarios. However, annotating the collected construction images in terms of object classes and the corresponding pixel locations is time-consuming, labor-intensive, and costly [15]. Further, this annotating work needs to be conducted by specialists who have sound knowledge in both construction management and deep learning. To address this problem, semi-supervised learning has been introduced from the computer vision community to explore the hidden information within unlabeled images. Semi-supervised learning refers to the concept of combining a small number of labeled data with a larger number of unlabeled data to train machine-learning algorithms [16]. Compared with supervised learning methods, semi-supervised learning methods are able to achieve similar or even better performance while requiring significantly fewer labeled data.

The main objective of the research reported in this paper is to propose a novel semi-supervised learning detection method for vision-based construction site monitoring. In the proposed method, the weak data augmentation has been firstly applied on both labeled and unlabeled images. Then, a teacher object detector that is pre-trained on labeled images is used to detect construction objects from unlabeled images and so generate pseudo-detection results. Following this, strong data augmentation is applied to the unlabeled images that are combined with pseudo-detection results to train the student object detector by optimizing the consistency loss. This student object detector then forms the output of the proposed method for the detection of construction objects from images. The proposed method is expected to reduce the costs and efforts of annotating construction-specific image datasets for training deep-learning methods. Further, the proposed method can be integrated in advanced applications in the field of construction automation such as machine idling analysis, dirt-loading cycle calculation, and potential collision detection.

2. Literature review

Monitoring construction sites using vision-based methods fulfills an essential need in modern construction management. By adopting semi-supervised learning methods, the volume of data required for training vision-based methods can be efficiently reduced and the performance of vision-based systems improved. The relevant literature is summarized below under three sub-headings: vision-based monitoring of construction sites, semi-supervised learning methods, and semi-supervised learning methods in construction.

2.1. Vision-based monitoring of construction sites

Researchers in the construction community have put considerable effort into developing vision-based solutions for automatically monitoring construction sites. Object detection is an important technique in indicating the presence of pre-defined classes of objects and locating their pixel locations from images [17]. The detection of construction objects is the first and fundamental step in most vision-based monitoring systems aimed at productivity analysis, safety management, or carbon footprint monitoring. Rezazadeh Azar et al. [18] have proposed a hybrid system to calculate the dirt-loading cycles in earthmoving operations by detecting excavators and dump trucks from videos. Yang et al. [19] have developed a crane activity analyzing system based on the detection of crane jibs. To improve site safety, Chi and Caldas [20] have developed a vision-based system that identifies potential collision risks by detecting and tracking construction machines. Gualdi et al. [21] have proposed a safety hat detection method to ensure workers correctly wear such hats while on construction sites. To minimize the environmental impacts of construction, Heydarian et al. [22] have benchmarked the carbon footprint of construction machines in earthmoving projects using a vision-based detection method.

Recently, deep-learning detection methods have attracted considerable attention in the construction sector because of their robust detection performance when dealing with complex conditions (e.g., occlusions and illumination variations). For example, Chen et al. [23] have employed the faster region-based convolutional neural networks (Faster R-CNN [24]) detector to recognize the activities of excavators in order to calculate crew productivity. Similarly, Kim et al. [25] have integrated the region-based fully convolutional neural networks (R-FCN [26]) detector with simulation for productivity analysis in tunnel earthmoving constructions. Fang et al. [12] have proposed improved region-based convolutional neural networks (IFaster R-CNN) for detecting construction resources, which have achieved a detection accuracy of 91% with workers and 95% with excavators.

Training deep-learning detection methods requires large construction image datasets, and annotating these construction images is a time- and labor-intensive process requiring a sound knowledge base in terms of both construction and deep learning. A previous study [27] has indicated that the average cost of annotating a single construction image can be as high as $0.51, while training deep-learning methods usually requires hundreds of thousands of labeled images. As such, annotating construction images becomes an uneconomic process.

2.2. Semi-supervised learning methods

Semi-supervised learning methods can efficiently reduce the annotation efforts by exploring the potential information hidden in unlabeled images, an area that has been studied in the computer vision community for several years. The developed semi-supervised learning methods can be categorized into self-training methods and consistency regularization methods.

Self-training methods first train a model on labeled data and then detect objects in unlabeled data to produce pseudo-labels. These pseudo-labeled data are then combined with the labeled data and the supervised training process repeated. Wang et al. [28] utilized visual tracking results as pseudo-labeled data for training convolutional neural networks (CNN) for object detection, and achieved a mean average precision (mAP) of 52% on the PASCAL Visual Object Classes (VOC) dataset [17]. Doersch et al. [29] have investigated various CNN architectures for self-training methods and proposed a joint network for semi-supervised object detection. Wang et al. [30] have developed a principled self-supervised sample mining method for object detection. Their proposed method improves the detection performance by integrating self-learning and active learning techniques, and this achieved a mAP of 62.9% on the VOC dataset when using 30% of the training dataset. However, the training process with self-training methods is time consuming, and the detection performance may be unstable if the training dataset is unbalanced.

Consistency regularization methods involve two steps: 1) applying perturbations to an image x to produce an image x; and 2) forcing the model to generate consistent predictions on x and x by optimizing their joint loss. Antti and Harri [31] proposed the Mean Teacher model that...
defines the consistency loss as the expected distance between the prediction of the student model and the prediction of the teacher model to improve the semi-supervised learning results by over 20% relative to the ImageNet benchmark. Jeong et al. [32] have proposed a semi-supervised learning method for object detection that applies the consistency loss not only for object classification but also for localization.

The effectiveness of data augmentation (e.g., color and geometric transformations) [33–35] has recently been studied as a way to improve the generalization and robustness of semi-supervised learning methods. Sohn et al. [36] have proposed the FixMatch model that matches the prediction of the strongly-augmented unlabeled data and the weakly-augmented counterpart using consistency regularization, and this FixMatch model has achieved an accuracy of 94.93%. For the task of object detection, Sohn et al. [37] have proposed a simpler semi-supervised learning detection method by integrating strong data augmentation (i.e., box-level geometric transformations and cutout). Compared with supervised learning methods, semi-supervised learning methods require significantly fewer labeled data to achieve similar performance in terms of accuracy and robustness.

2.3. Semi-supervised learning methods in construction

In the construction research community, researchers have worked on reducing the efforts involved in labeling images for training deep-learning detectors to identify construction workers, machines, and materials. Liu and Golparvar-Fard [38] have investigated the feasibility of crowdsourcing for annotating construction data on the Amazon mTurk platform, and achieved an 85% labeling accuracy in experiments with a time-efficient labeling speed. However, additional manual efforts are then required to amend incorrect annotations resulting from the crowdsourcing. Another direction is to generate synthetic images and annotations for training deep-learning object detectors. Soltani et al. [39] have developed a synthetic dataset using 3D models of excavators and obtained promising results. However, deep-learning methods learn features from datasets, and training with synthetic datasets may not prove effective because synthetic images have different visual characteristics (e.g., textures and colors) than real construction images.

Although semi-supervised learning methods have achieved success in the computer vision community, only a limited number of studies have taken place in the construction industry. Addressing structural safety inspection, Guo et al. [16] have developed a supervised learning method based on CNN and an uncertainty filter for façade defect classification that achieved an accuracy of 76.74% in experiments using only 5% of the labeled data. For construction machine detection, Kim et al. [40] have proposed a semi-supervised learning method based on an active learning mechanism that evaluates the uncertainty of unlabeled images and then trains the deep-learning detector using data sampling and user-interactive labeling. Compared to crowdsourcing and adopting synthetic datasets, the advantages of semi-supervised learning methods are two-fold: 1) a reduction in the quantity of training data required; and 2) better accuracy and robustness than supervised learning methods. However, none of existing semi-supervised learning methods used in construction have applied data augmentation to improve the performance of vision-based site monitoring.

3. Methodology

The objective of this research was to develop a semi-supervised learning detection method based on teacher–student networks and data augmentation for monitoring construction sites. The proposed method is expected to successfully recognize construction resources (i.e., machines, workers, and materials) from videos using only a limited number of labeled images. The overall framework of the proposed methodology is first introduced. Following this, the three main modules involved in the methodology that address data augmentation, object detection, and consistency regularization are illustrated in detail.

3.1. Overall framework

The overall framework of the proposed method (depicted in Fig. 1) can be divided into two stages: a supervised training stage and a semi-supervised training stage. The main purpose of the supervised training stage is to train the teacher object detector based on labeled data. The labeled data are denoted as $D_L = \{(x_i, y_i)\}_{i=1}^N$, where $x_i$ is an image in the labeled dataset, $y_i$ is its corresponding object-level label, and $N$ is the number of images in the labeled dataset. First, weak data augmentation...
is applied to images and their corresponding labels in order to increase the volume of training data. The augmented data are denoted as \( \{(w(x_i), y_i)\}_{i=1}^N \), where \( w \) refers to the weak augmentation function. Next, the augmented data and labeled data are combined to form jointed labeled data \( D_{UL} = \{(x_i, y_i), w(x_i)\}_{i=1}^N \). Finally, \( D_{UL} \) is then employed to train the teacher object detector.

The semi-supervised training stage takes in unlabeled data and the trained teacher object detector with the main purpose of training the student object detector that forms the output of the proposed method. The unlabeled data are denoted as \( D_U = \{x_i\}_{i=1}^N \), where \( x_i \) is an unlabeled image. As in the previous stage, the unlabeled data are processed using weak data augmentation, and the weakly augmented data are combined with the original unlabeled data to build the jointed unlabeled dataset \( D_{UW} = \{x_i, w(x_i)\}_{i=1}^N \). Furthermore, strong data augmentation is applied to the jointed unlabeled data \( D_{UW} \) to obtain the strongly augmented jointed unlabeled data \( D_{SW} = \{s(x_i), s(w(x_i))\}_{i=1}^N \), where \( s \) represents the strong augmentation function.

In the semi-supervised training stage, a batch of images contained in \( D_{UW} \) are processed with the teacher object detector to produce the detection results. A confidence threshold is adopted to produce pseudo-labels from the teacher detection results. The supervised loss can then be calculated as the difference between the pseudo-labels and the teacher labels from the teacher detection results. The supervised loss can then be updated by optimizing the consistency loss through back-propagation. The trained model of the student object detector is the final output of the proposed method, which will be used for detecting construction machines from images. Note that the teacher object detector is frozen and not optimized in the semi-supervised training stage. As Fig. 1 illustrates, three main modules addressing data augmentation, object detection, and consistency regularization are involved in the proposed method and these are now presented in the following subsections.

The novelty of the proposed method is the strategy used for training the student object detector. The development of deep learning object detection methods contains two stages, which are the training stage and the deployment stage. The training stage is aiming to train the object detection networks to optimize parameters in each CNN layer, while the trained object detection networks are employed to detect objects from images in the deployment stage. In the proposed method, a novel strategy has been proposed by integrating the data augmentation, teacher-student networks, and consistency regularization for training the student object detector. As such, the proposed method is expected to achieve better performance by using a much smaller number of annotated data than the original student object detector. To be noted, the CNN architectures of the teacher and student object detectors are proposed in the previous study, and the deployment stage of the proposed method remains the same as the original student object detector.

### 3.2. Data Augmentation

Data augmentation is the key module in the proposed method. In vision-based construction site monitoring, developing construction image datasets usually requires several months of effort, while annotating construction images is a costly and time-consuming process. Data augmentation can effectively increase the volume of training data by applying transformations to images and the corresponding labels. Furthermore, the extent of background changes in construction images is less than in images of everyday life, which means that data augmentation can have larger impacts on construction images. Applied to semi-supervised learning, the larger impacts refer to the increased learning ability in the consistency regularization process. As such, data augmentation for semi-supervised learning methods is considered to be especially suitable in construction scenarios.

In this research, two types of augmentation strategy have been proposed: weak data augmentation and strong data augmentation. Weak data augmentation has been applied in both the supervised and the semi-supervised learning stages to increase the quantity of training data. For example, 1000 weakly augmented images can be generated from the labeled dataset \( D_L \) that has 1000 labeled images, meaning that the jointed labeled dataset \( D_{UL} \) can provide 2000 images for training. The weak data augmentation involves one of the following five geometric transformations on each image (the corresponding label if needed): 1) randomly translate the image by \(-10\%\) to \(+10\%\) on the x-axis; 2) randomly shear the image by \(-15\) degrees to \(+15\) degrees on the x-axis; 4) randomly shear the image by \(-15\) degrees to \(+15\) degrees on y-axis; or 5) randomly rotate the image by \(-15\) degrees to \(+15\) degrees.
noted, the values of translate, shear, and rotate are determined by experiments, where the detailed experimental process and discussions are demonstrated in Appendix A. Fig. 2 shows example images of applying such weak data augmentations.

Strong data augmentation has been applied in the semi-supervised learning stage with the aim of forcing the student object detector to make predictions consistent with the teacher object detector when the jointed unlabeled data $D_{UL}$ are strongly augmented. In the strong data augmentation, each image is first processed with one of the following five transformations: 1) blur the image with a gaussian kernel with a sigma of 3.0; 2) sharpen the image and overlap the result with the original image using an alpha between 0.0 and 1.0; 3) randomly add a value of $-10$ to $10$ to each pixel on the image; 4) randomly change the brightness of the image by $50\%$ to $150\%$; or 5) randomly change the contrast of the image by $50\%$ to $200\%$. Note that the parameters used in the above five operations are based on suggestions in [34]. Following this, a cutout operation [41] at multiple locations is applied to all the processed images output from the strong data augmentation. Fig. 3 shows sample images of applying strong data augmentations. In the proposed method, the weak data augmentation and strong data augmentation should be considered as an integral augmentation strategy instead of a simple combination of weak and strong data augmentations.

3.3. Object detection

The object detector identifies construction objects from images and as output provides the pixel locations of construction objects and their corresponding categories. In this research, the object detection module contains two independent object detectors: a teacher object detector and a student object detector. The teacher object detector is firstly trained using the jointed labeled data in the supervised training stage, and the weights obtained from the teacher object detector are then used in the semi-supervised training stage. Here, each unlabeled image $x_i$ from the jointed unlabeled data $D_{UL} = \{x_i, w(x_i)\}_{i=1}^N$ is input to the teacher object detector to obtain detection results. Meanwhile, the corresponding
A strongly augmented image \( s(x_i) \) is processed with the student object detector. The consistency regularization model will then force the student object detector to make predictions consistent with the teacher object detector. As such, the knowledge held within the teacher object detector can be transferred to the student object detector. Since the student object detector utilizes the strongly augmented images, it will eventually outperform the teacher object detector.

In this research, the teacher and student object detectors both adopt the same deep-learning object detection method (Faster R-CNN). Faster R-CNN is one of the most robust and accurate detectors and has achieved outstanding performance in vision-based construction site monitoring studies [23,42-44]. As illustrated in Fig. 4, the Faster R-CNN detector contains three main modules: a feature extractor, a region proposal network (RPN), and a classifier network (CLS).

First, an input image is resized into \( 224 \times 224 \) and fed into the feature extractor module to generate feature maps that represent the original image. The ResNet50 [45] network has been integrated into the feature extractor module. The ResNet50 is a state-of-the-art neural network for feature extraction in the computer vision community, which consists of 50 layers of neural networks for implementing the residual block. As illustrated in Fig. 5, the concept of residual block can be understood as a “shortcut connect” that adds the convolutional output \( \mathcal{F}(x) \) with the input weights \( X \) to produce the final output weights of the residual block. The residual block has been proven to have the feasibility of solving the gradient vanishing problem in deep neural networks. Detail architecture of ResNet50 has been summarized in the Table 1, where \([\cdot]\) refers to a residual block. In ResNet50, the image is firstly processed by a \( 7 \times 7 \) convolutional layer (conv1) and a \( 3 \times 3 \) max-pooling layer. Then, four stacks of residual blocks (conv2, x to conv5, x) are implemented after the max-pooling layer. Finally, the ResNet50 feature extractor produces the feature maps with the size of \( 7 \times 7 \times 2048 \) to represent the original image.

Then, the feature maps are forwarded to the RPN module, where \( n \times n \) spatial windows are slid over the input feature maps. As a default, 12 anchor boxes are initialized for each sliding window as the regions of interest (ROIs), with the proposed anchor boxes defined by three aspect ratios (1:1, 1:2, and 2:1) and four scales (32\(^2\), 64\(^2\), 128\(^2\), 512\(^2\)). Each sliding window is processed by a three-layer convolutional network, by a fully connected layer for box classification, and by a fully connected layer for box regression. The box classification layer computes the possibility of an ROI being an object and the box regression layers correct the pixel location of the ROIs. The top 300 ROIs, according to their likelihood of being an object, are then selected as the output of the RPN module. Furthermore, the ROI pooling technique [46] is adopted to generate fixed-shape features from the feature maps generated by ResNet50. These fixed-shape features are input to the CLS module which comprises a fully connected box classification layer and a fully connected box regression layer. In the CLS module, the box classification layer computes the confidence of each ROI belonging to one of the predefined classes and the box regression layer predicts the pixel coordinates of objects. The outputs of the CLS module are the detection results of the Faster R-CNN detector which comprises pixel locations and categorical information on predefined classes of construction objects.

### Table 1

| Layer name | Output size | Architecture (kernel size, depth) |
|------------|-------------|-----------------------------------|
| conv1      | 112 × 112   | 7 × 7; 64; stride 2               |
| conv2, x   | 56 × 56     | 3 × 3max pool; stride 2          |
| conv3, x   | 28 × 28     | 1 × 1.128; 3 × 3; 1 × 1.256      |
| conv4, x   | 14 × 14     | 1 × 1.128; 3 × 3; 1 × 1.512      |
| conv5, x   | 7 × 7       | 1 × 1.2048; 3 × 3; 1 × 1.512      |

Note: The contents of Table 1 are obtained from [45].

### 3.4. Consistency regularization

The main purpose of the consistency regularization module is to calculate the jointed loss of the teacher object detector and the student object detector. In the proposed method, the jointed loss is defined as the combination of supervised and semi-supervised losses, which can be calculated in three steps: 1) apply confidence thresholds to the teacher detection results to produce the pseudo-labels; 2) calculate the supervised loss using the pseudo-labels and teacher model detection results; and 3) calculate the semi-supervised loss using pseudo-labels and student detection results. Once the jointed loss has been defined, the student object detector can be readily updated using back-propagation.

In the semi-supervised learning process, it was initially found that the detection results of the teacher object detector could not be directly used as pseudo-labels because the detected bounding boxes were prone to repetition and inaccuracies. In order to eliminate incorrect bounding boxes, it was necessary to apply confidence thresholds. Following a strategy adopted in a previous study [37], detected bounding boxes with a confidence level below 0.9 are removed. The results of the processed teacher object detector are the pseudo-labels, which are then used to calculate the loss function.

Calculating the supervised loss involves the variances between the pseudo-labels and the teacher detection results. As introduced in Section 3.3, Faster R-CNN has two heads (RPN and CLS) on top of the ResNet50 networks, and each head has a box classification layer and a box regression layer. The supervised loss is defined as the sum of the RPN and CLS losses (Equation (1)):

\[
\mathcal{L} = \mathcal{L}_{RPN} + \mathcal{L}_{CLS}
\]  

\( \mathcal{L}_{RPN} \) is the loss for the RPN head (Equation (2)), which is the sum up of losses of the box classification layer and the box regression layer.

\[
\mathcal{L}_{RPN} = \mathcal{L}_{cls} + \mathcal{L}_{box}
\]

In a mini-batch, the RPN loss becomes:

\[
\mathcal{L}_{RPN} = \mathcal{L}(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_{i} L_{mult}(p_i, t_i) + \frac{1}{N_{box}} \sum_{i} L_{reg}(t_i - \tilde{t}_i)
\]

\text{(3)}

where \( i \) is the index of an anchor in the mini-batch. \( N_{cls}, N_{box} \), and \( \tilde{t} \) are
normalization terms to balance the weights of the box classification and box regression. \( p_i \) is the predicted probability of anchor \( i \) being an object. \( p_i^t \) is the ground truth binary label of anchor \( i \) with respect to ground truth boxes. \( t_i \) is the parameterized coordinates of the anchor \( i \), and \( t_i^t \) is the ground truth coordinates of the object associated with anchor \( i \). Further, \( L_{\text{box}} \) is the smooth \( L_1 \) loss as defined in [46].

Similar to the definition of RPN loss, the CLS loss is defined in Equations (4) and (5):

\[
\mathcal{L}_{\text{CLS}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} \tag{4}
\]

\[
\mathcal{L}_{\text{CLS}} = \mathcal{L}(p_i; \{t_i\}) = \frac{1}{N_{\text{cls}}} \sum_i \text{categorical}(p_i; p_i^t) + \frac{\lambda}{N_{\text{box}}} \sum_i p_i^t L_{\text{box}}^\text{smooth}(t_i - t_i^t) \tag{5}
\]

where \( \text{categorical} \) is the categorical cross entropy loss among \( K \) (number of pre-defined classes) + 1 (background) classes. All other parameters are the same as in Equation (3).

Since the teacher and student object detectors apply the same detection method, the semi-supervised loss \( \mathcal{L}_{\text{ss}} \) can be defined in the same way as the supervised loss \( \mathcal{L} \). The only difference is that \( \mathcal{L}_{\text{ss}} \) calculates the variances between the pseudo-labels and the detection results of the student object detector rather than the teacher object detector. Following this, the jointed loss \( \mathcal{L} \) can be obtained by summing the supervised loss and the semi-supervised loss (Equation (6)).

\[
\mathcal{L} = \mathcal{L}_s + \alpha \mathcal{L}_{\text{ss}} \tag{6}
\]

where \( \alpha \) is a constant term to balance the weights of the supervised and semi-supervised losses. In this research, \( \alpha \) is set to 1.

In the backpropagation, the stochastic gradient descent with momentum (SGDM) algorithm [47] is employed to optimize the joint loss \( \mathcal{L} \) for training the student object detector. The optimizing process of SGDM is shown in Equation (7) and Equation (8). Compare with the stochastic gradient descent, the SGDM can achieve faster converging speed when training with a large quantity of data. Detailed configuration of SGDM in the training process is provided in Section 4.1. The trained student object detector is the final output of the proposed method, which can be used for detecting construction objects from images.

\[
x_{t+1} = x_t + \Delta x_t \tag{7}
\]

\[
\Delta x_t = \rho \Delta x_{t-1} - \eta g_t \tag{8}
\]

where \( \eta \) is the weights of deep neural networks at step \( t \), \( \eta \) is learning rate, \( g_t \) is the gradient descent, and \( \rho \) is the momentum.

4. Experimental design and results

The proposed method has been tested on construction scenarios to assess its feasibility for vision-based monitoring of construction sites. This section presents the experimental design and results in four subsections: 1) implementation details of the proposed method; 2) experimental design and evaluation metrics; 3) experimental results; and 4) testing results on a real construction video.

4.1. Implementation details of the proposed method

The overall framework of the proposed method was programmed in the Python language and run under an Ubuntu 18.04 operating system on a computer with two NVIDIA GTX 1080Ti graphic cards (11 GB memory), two 32 GB memory cards, and an Intel Core i9-7920X @2.90 Hz CPU. The OpenCV and NumPy libraries were adopted for image I/O. The data augmentation module was implemented using the imgaug library. In the object detection module, the Faster R-CNN are implemented in the Pytorch library.

In the proposed method, the teacher object detector is trained in the supervised learning stage by using the original loss of Faster R-CNN, where the student object detector is trained in the semi-supervised learning stage by using the joint loss \( \mathcal{L} \). In the training process, both object detectors adopt the SGDM optimizer in the backpropagation. The training configurations including a batch size of 8, a training iteration of 180,000, a learning rate of 0.0001, and a momentum of 0.9.

4.2. Experimental design and evaluation metrics

To validate the performance of the proposed method, comprehensive
experiments were conducted using the Alberta Construction Image Dataset (ACID) [27]. ACID is an open dataset developed in the construction research community, which contains construction machine images specifically annotated for deep-learning object detection and can be accessed through its project website [48]. The experimental results of the present study can be reproduced on ACID. Testing a proposed method with ACID can indicate its feasibility for vision-based monitoring of construction sites. ACID comprises 10,000 construction images and 15,767 labeled machine objects belonging to ten types of common construction machines (excavator, compactor, dozer, grader, dump truck, concrete mixer truck, wheel loader, backhoe loader, tower crane, and mobile crane).

The raw images in ACID were collected from various sources including photo-sharing websites (i.e. Google Image and Naver), the video-sharing website (i.e. YouTube), and site visits (some by drones and mobile phones). Then, 10,000 images were carefully selected from collected raw images of ACID by removing the duplicated images, undersized and oversized images, and low-resolution images. Fig. 6 shows some examples of images in the dataset. After the image selection process, all construction images in ACID were manually labeled by annotators and crowdsourced workers with two types of information about machine objects on the image, which are machine category and pixel location information. The annotation results have been checked twice by an annotator and the authors of ACID to ensure the dataset quality. All images in ACID were converted to JPEG format, and the annotations were stored in XML format. Fig. 7 shows an example of annotation XML and the corresponding image in ACID. As such, because of its wide diversity, using ACID can avoid the overfitting problem associated with training deep-learning object detection methods.

In this research, the ACID dataset is divided into a training set (70% of the images) and a validation set (30%). To explore the feasibility of the proposed method, nine different proportions of the training set (from 10% to 50% in 5% intervals) were initially used as the labeled data for training the proposed method. For example, a 10% data proportion refers to 10% of the training set being used as labeled data and the remaining 90% treated as unlabeled data for training the proposed method. Furthermore, this experiment was supplemented with three extreme cases with 1%, 3%, and 5% of the training set as labeled data. With each data proportion, we also trained two detection methods for comparison purposes: 1) the supervised learning method, which is the original Faster R-CNN detector without any special training strategies (e.g., data augmentations and teacher-student networks); and 2) the supervised learning method with weak data augmentation (WDA), which means applying the weak data augmentation proposed in the present study on the original Faster R-CNN detector. The supervised learning method with WDA is identical to the teacher object detector in the proposed method trained in the supervised learning stage. To be noted, these detection methods have adopted the same backbone object detector (Faster R-CNN) with different training strategies compared with the proposed method. Finally, for evaluation purposes, the trained models for the proposed method and the above two detection methods were tested on the validation set. Note that the unlabeled data have only been used for training the proposed method and excluded when training supervised learning methods.

The mean average precision (mAP) [49] metric has been used to evaluate the detection performance in this research, which is a homogeneous metric calculated based on precision and recall. The mAP metric is a standard metric for evaluating the performance of deep learning object detection methods [17,50,51] in recent studies. To obtain mAP, the precision and recall are firstly calculated as in Equation (9) and Equation (10), respectively. Then, the average precision (AP) can then be calculated by averaging the precision at different levels of recall for each class of object (Equation (11)). Finally, mAP is calculated as the mean of the AP of all the predefined classes (Equation (12)), with a higher mAP indicating better detection performance. To be noted, the precision and recall are not presented in the experimental results in this study since the main usage of precision and recall is to calculate AP and mAP.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{9}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{10}
\]

\[
AP = \frac{1}{\text{recall}} \sum \text{Precision@recall} \tag{11}
\]

\[
mAP = \frac{1}{K} \sum AP \tag{12}
\]
where TP (true positive) is the number of correct detections. A correct detection occurs when the bounding box has an IoU (Intersection over Union) larger than 0.5 with its corresponding ground truth bounding box. FP (false positive) is the number of negative objects that have been incorrectly detected as positive objects. FN (false negative) is the number of positive objects that have been identified as negatives. K is the number of pre-defined object classes.

### 4.3. Experimental results

Table 2 indicates the detection performance, in terms of mAP, of the three methods when trained with different proportions of the training set. Fig. 8 shows some sample detection results produced by the proposed method. As shown in Table 2, the proposed method achieved the highest mAP (92.7%) among the three methods tested when training with a 50% data proportion (3500 labeled images) from the construction image dataset (ACID). In comparison, the supervised learning method achieved a mAP of 89.1%, and the supervised learning method with WDA a mAP of 90.6%, both with the same data proportion (50%). Further, the proposed method achieved a mAP above 90% even when

| Data proportion | Supervised learning method | Supervised learning method + WDA | Proposed method |
|-----------------|---------------------------|---------------------------------|-----------------|
| 1%              | 36.7%                     | 51.9%                           | 63.3%           |
| 3%              | 60.2%                     | 71.4%                           | 78.5%           |
| 5%              | 66.1%                     | 76.2%                           | 81.7%           |
| 10%             | 75.9%                     | 82.6%                           | 85.5%           |
| 15%             | 80.3%                     | 84.7%                           | 87.0%           |
| 20%             | 83.0%                     | 86.7%                           | 88.5%           |
| 25%             | 83.8%                     | 88.0%                           | 90.1%           |
| 30%             | 84.9%                     | 88.9%                           | 91.0%           |
| 35%             | 86.9%                     | 89.3%                           | 91.6%           |
| 40%             | 87.5%                     | 90.1%                           | 92.0%           |
| 45%             | 88.3%                     | 90.4%                           | 92.4%           |
| 50%             | 89.1%                     | 90.6%                           | 92.7%           |

Note: WDA refers to weak data augmentation.

**Table 2**

Detection performance (mAP) of the three methods with different data proportions.

**Table 3**

The performance of the proposed method in terms of construction machine classes.

**Fig. 8.** Sample detection results using the proposed method trained with 50% data proportion.
using only 25% of the training data. To achieve a similar performance, the supervised learning method requires over 50% of the data for training and the supervised learning method with WDA needs 40% of the training data. The experimental results demonstrate that the proposed method achieves a better detection performance than two supervised learning methods applied in vision-based monitoring of construction sites. Also, the experimental results prove the effectiveness of the proposed training strategy of integrating data augmentation, teacher-student networks, and consistency regularization in the detection of construction objects.

Overall, the proposed method has, on average, outperformed the supervised learning method and the supervised learning method with WDA by 9.3% and 3.6% respectively in terms of mAP when training with various data proportions. Further the performance gap increases when only small proportions of the training data are used. In our most extreme cases (1%, 3%, and 5% data proportions), the proposed method achieved on average a 20.2% higher mAP than the supervised learning method, whereas the advantage was only 5.7% when averaged across the other data proportions (10% to 50%). With the minimal 1% data proportion, the proposed method still achieved a mAP of 63.3%, while the supervised learning method only achieved 36.7%. Therefore, the proposed method particularly shows good performance when training is only possible with a small number of labeled data, highlighting its practicality in vision-based construction applications.

The supervised learning method with WDA involves the Faster R-CNN routine trained with the jointed labeled data (combining labeled data and weakly augmented data), as such reflecting the teacher object detector in the proposed method. On average, the supervised learning method with WDA outperforms the basic supervised learning method in terms of mAP by 5.7% across the different data proportions. As in the previous comparison, in extreme situations, with very low data proportions, the difference in performance was more marked, with the supervised learning method with WDA achieving an average improvement in mAP of 12.2%. These results show that the weak data augmentation proposed in this research can effectively improve the detection performance of the supervised learning method in construction scenarios.
Table 3 shows the performance of the proposed method in terms of construction machine classes when training with different proportions of the training set. The mean of the AP for construction machines under different data proportions is 86.2%, while the standard deviation is 9.24%. As illustrated in Table 3, the proposed method has achieved the better efficiency in detecting graders, backhoe loaders, and compactors with an average AP of 96.1%, 94.2%, and 92.3%, respectively. Detection of tower cranes, mobile cranes, and dump trucks is relatively difficult for the proposed method, while the average AP for these three machine classes are 64.2%, 78.1%, and 79.2%, respectively. It is found that the analysis results between construction machine classes are similar to that reported in the original ACID paper [27]. In experiments, the proposed method has achieved over 80% average AP on seven types of construction machines under training with different proportions of the training set, which demonstrates the efficiency of the proposed method in construction scenarios.

4.4. Testing results on a real construction video

In the current practice of vision-based monitoring of construction sites, object detection methods are usually applied on construction videos captured by engineers or fixed-position cameras, while the detection results can be used to automatically calculate crew productivities, identify safety risks, and record project progress. Therefore, testing the proposed method in continuous construction videos is important for evaluating its feasibility of dealing with challenges in real construction scenarios. In this experiment, the proposed method that trained on 50% of the training set has been employed to detect construction machines from a 42.8 mins construction video. This testing video contains dirt-loading operations of one excavator and several dump trucks in an earthmoving site, which has been captured by a construction engineer using a smartphone. The testing video has a resolution of 1280 × 720, a framerate of 24 frames per second (fps), and 61,593 frames in total, where Fig. 9 shows an example image of the testing video. It is noticed that the proposed method has faced the challenges that frequently appeared in real construction scenarios including appearance changes, scale variations, and occlusions due to the movements of cameras.

Fig. 10 shows the example detection results produced by the proposed detection method trained on 50% of the training set. The testing results show the proposed method can overcome the frequent challenges that appeared in real job site videos. For example, the appearance of the dump trucks changes significantly in frames #8690, #9360, #21140, and frame # 36,240 due to the differences in camera viewpoints, and the proposed method can successfully detect dump trucks in these frames. The proposed method also demonstrates reliable performance when dealing with scale variations in construction machine detection. For example, the scale of the excavator changes largely in frames #18792, #30000, and #58512, and the proposed method produced precise detection results of the excavator under different object scales. In frames #25920 and #44400, the excavator was occluded by the dump truck, which is commonly happened in earthmoving videos. The proposed method has successfully detected the occluded excavators in these frames. The testing results prove the proposed method can be used to continuously detect construction machines in real construction scenarios. The present study provides a robust backbone detector for advanced vision-based applications in construction for enhancing crew productivities and reducing safety risks in the construction site.

5. Discussion

Our research has proposed and tested a semi-supervised learning detection method for vision-based monitoring of construction sites. The experiments indicate the proposed method has achieved the research objective of increasing detection performance of construction objects. The research findings and areas that need to be further improved are now discussed:

1. The proposed method provides a reliable backbone object detector for vision-based monitoring on construction sites. In our experiments, the proposed method achieved a mAP of 92.7% when training with 50% of the ACID training set. The proposed method can be integrated into automated construction management processes addressing issues such as crew productivity analysis, machine collision monitoring, and idle time observations as a reliable backbone object detector. It should be noted that the ACID dataset plays an important role in the proposed method. The ACID dataset is very comprehensive with images captured in various scenarios (e.g. indoor and outdoor construction) with different camera views and levels of illumination. Further, the construction objects in the ACID dataset cover a range of visual characteristics in terms of size, color, shape, and position. A higher diversity of the training dataset results in the better robustness and generalizability of the proposed semi-supervised learning detection method. Meanwhile, the quantity of training dataset also influences the performance of the proposed method. If the number of images and annotations in ACID is increased, the detection performance of the proposed method will be further improved. The annotation quality (i.e. precision and correctness) of the training dataset is also positively correlated to the performance of the proposed method. Improving the annotation quality in ACID will enhance the general applicability of the proposed method.

2. The proposed semi-supervised learning method achieved higher detection performance than the supervised learning method. Additionally, we have supplemented an experiment that training the supervised learning method on the full ACID training set and obtained the mAP of 90.8% on the validation set. Actually, the proposed method achieved higher mAP of 92.7% with only 50% of the training set. Given that the proposed method employed the same backbone detector, the Faster R-CNN, as the supervised learning method, the proposed semi-supervised learning method is able to improve the peak detection performance of the supervised learning detector by exploring the potential of unlabeled data. This is a useful advantage because, in construction management, it is much easier to collect unlabeled than labeled data. Further, semi-supervised learning detectors are considered more suitable for use in construction scenarios.

3. The data augmentation strategy adopted in the proposed method is effective for improving the detection performance of object detection in construction scenarios. The weak data augmentation expands the quantity of training data to improve detection performance. The strong data augmentation enlarges the input differences of teacher and student object detectors, which improves the learning performance of the student object detector. In previous studies [33-35,37], most existing augmentation strategies simply applied consistent data augmentations to all image data, while few efforts have been conducted to categorize data augmentations into different types. In this research, the authors have proposed an integral data augmentation strategy that combining weak and strong data augmentations is a cascade structure. In experiments, the proposed method outperforms the supervised learning method + WDA by 3.6% on mAP averagely, which demonstrates the feasibility of our data augmentation strategy for improving the semi-supervised learning detection in construction scenarios.

4. The proposed method efficiently reduces the manual efforts and costs involved in labeling construction images. The proposed method achieved a mAP of 91.0% with a data proportion of 30%, which is similar to the level achieved by the supervised learning method (90.8%) trained with full training set. Therefore, the proposed method can avoid the effort involved in labeling 70% of the ACID training set (4,900 images) while still achieving a similar detection performance. According to ACID research [27], labeling a
In all experiments, the full training set is used to train the Faster R-CNN and the trained model is evaluated by the validation set. Data augmentation plays important roles in the proposed method. The proposed method introduces a combined data augmentation strategy that employs both weak and strong data augmentation. Employing weak data augmentation doubled the volume of training data and improved the detection performance of the supervised learning method (as discussed in Section 4.3). The strong data augmentation approach enabled knowledge to be more readily transferred from the teacher object detector to the student object detector. In typical teacher–student architectures [16,31], a more complicated teacher model is required to transfer knowledge to the student model. In our research, we adopted the same deep-learning detector for both the teacher and the student networks, while the knowledge distillation was conducted through the strong data augmentation. By adopting strong data augmentation, the proposed method forces the student object detector to make predictions consistent with the teacher object detector while processing more difficult data. In this way, the student object detector can achieve higher detection performance than the teacher object detector.

The proposed method has three limitations that need to be improved in the future. Firstly, the data augmentation strategy is limited to the image level augmentation in this research, which means all transformations (e.g., translate, shear, and sharpen) are applied to the construction images. Box-level augmentation could further improve the performance of the proposed method, which needs to be investigated in the future. Secondly, the parameters of strong data augmentation are determined by the suggestions from literatures, which need to be fine-tuned by experiments in the future. Thirdly, the proposed method adopted the same object detector for teacher and student networks in the current implementation. By replacing the teacher object detector with other advanced object detectors, the learning ability of student networks will increase and the detection performance will be further improved.

6. Conclusions and future works

This paper introduces a semi-supervised learning detection method for vision-based monitoring of construction sites. The proposed method has adopted the teacher–student architecture, consisting of three main modules that address data augmentation, object detection, and consistency regularization. Further, the proposed method has been trained and validated using the ACID dataset. In experiments, the proposed method achieved a mAP of 92.7% utilizing only 50% of the training set as labeled data. Even with a data proportion of only 5% (350 labeled images), the proposed method achieved a mAP of over 80%. Compared with the supervised learning method, the proposed method is able to reduce labeling efforts and costs by 70% and still achieve a similar detection performance of 90% mAP. As such, the proposed method can efficiently reduce the efforts involved in annotating construction image datasets for training deep-learning detectors. The proposed method provides a strong backbone detector for advanced construction applications that rely on vision-based object detection.

The contributions of this research are threefold. First, the proposed method involves a novel technical framework that relies on a data augmentation strategy, teacher–student networks, and consistency regularization to enable the semi-supervised learning detection of construction objects. Further, alternative deep-learning object detectors could be integrated into the proposed framework in the future. Second, the proposed method efficiently reduces the costs and time involved in manually labeling construction images, thereby improving the practical feasibility of vision-based monitoring of construction sites based on integrating deep-learning detection. Third, this research proposes a new data augmentation strategy that combines weak and strong data augmentations in a cascade structure. In the proposed strategy, weak data augmentation can significantly improve the detection performance of a supervised learning method, and strong data augmentation is effective for knowledge distillation.

Future works will focus on improving the data augmentation strategy for semi-supervised learning detection method in construction. For example, the box-level data augmentations will be implemented in the proposed method. More experiments will be conducted to optimize the parameters in the strong data augmentation. Meanwhile, the advanced object detectors will be implemented in the proposed method as the teacher network in the future. Currently, the teacher object detector uses the ResNet50 Faster R-CNN algorithm. If, for example, the ResNet50 was replaced by the ResNet101 algorithm, the teacher object detector would achieve better detection performance. As a consequence, the student object detector will be able to achieve better performance because the teacher object detector makes more precise predictions.

Appendix A. Parameter finetuning in weak data augmentation

In this study, the parameters in weak data augmentation, including translate, shear, and rotate values, are determined from the experiment base. The previous study [34] suggests that the range of translate, shear, and rotate values is between 10% and 30%, which has been adopted in this research. Therefore, five values (from 10% to 30% in 5% interval) have been selected as the candidate parameter for each kind of weak data augmentation (i.e. translate, shear, and rotate). The parameters of translate, shear, and rotate have been independently decided by three experiments. In all experiments, the full training set is used to train the Faster R-CNN and the trained model is evaluated by the validation set.

For determination of the translate value \( t \), each image and the corresponding label are processed by one of the following two transformations: 1) randomly translate the image by \( -t \) to \( +t \) on the x-axis; and 2) randomly translate the image by \( -t \) to \( +t \) on the y-axis, where \( t \in [10\%, 15\%, 20\%, 25\%, 30\%] \). Similar to the supervised learning stage in the proposed method, the augmented dataset is combined with the original dataset to form the jointed labeled data \( D_{it} \), while is used for training the Faster R-CNN. Table A.1 illustrates the experimental results of the translate value determination. The Faster R-CNN has achieved the mAP from 90.8% to 91.3% when using different values of \( t \). And the Faster R-CNN has obtained the highest mAP of 91.3% when \( t \) equals 10%. Therefore, the translated value \( t \) is set up to 10% in this research.

The shear value \( s \) and rotate value \( r \) are determined by a similarly experimental process. For the shear value \( s \), each image and the corresponding label are processed by one of the following two transformations: 1) randomly shear the image by \( -s \) degrees to \( +s \) degrees on the x-axis; and 2)
respectively. For the shear value determination, the Faster R-CNN has achieved the highest mAP of 91.5% when the shear value is determined as 15% in this research. For the rotate value determination, the Faster R-CNN has achieved the highest mAP of 91.6% when the rotate value in the weak data augmentation.

As introduced in the discussion section, the Faster R-CNN without any data augmentation has achieved the mAP of 90.8% when training with the full training set. The improvement in mAP is between 0% and 0.5% when adopting different translate values. And the mAP improvement is between 0.4% to 0.7% and 0.5% to 0.8% for shear and rotate transformations, respectively. It is found that finetuning the parameters for each type of weak data augmentation can provide around 1% mAP improvement for Faster R-CNN on the ACID dataset. To be noted, the improvement of 1% mAP is significant in object detection studies when the mAP is larger than 90%. An additional experiment has been conducted to process the images and the corresponding labels by one of translate, shear, and rotate transformations for training, which has obtained the mAP of 91.5%. This result is close to the detection performance of 15% 50%  74.1%  85.6%  95.6%  86.4%  97.9%  94.5%  95.2%  96.3%  91.9%  95.7%  91.3% 10%  75.1%  87.1%  94.8%  86.9%  98.5%  94.3%  94.9%  95.4%  90.9%  96.2%  91.0% 15%  74.2%  86.4%  95.1%  87.0%  98.2%  94.5%  95.7%  96.5%  92.1%  95.5%  91.5% 20%  72.8%  82.1%  95.7%  85.3%  98.1%  95.4%  95.8%  96.5%  93.7%  96.9%  91.1% 25%  74.0%  85.2%  96.3%  83.4%  96.8%  93.9%  96.1%  93.8%  93.1%  95.1%  90.8% 30%  62.7%  85.6%  95.1%  84.7%  98.2%  94.1%  97.1%  96.1%  94.8%  95.2%  90.4%

Table A2 Experimental results of shear value determination.

| Shear values | Translate |  |  |  |  |  |  |  |  |  |  |  |
|--------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| Tower crane  | Mobile crane | Excavator | Dump truck | Grader | Wheel loader | Compactor | Dozer | Concrete mixer truck | Backhoe loader |
| Translate    | mAP      | mAP      | mAP      | mAP      | mAP      | mAP      | mAP      | mAP      | mAP      |
| 10%          | 75.1%    | 87.1%    | 94.8%    | 86.9%    | 98.0%    | 94.3%    | 95.4%    | 96.2%    | 91.7%    | 94.5%    | 91.4%    |
| 15%          | 74.2%    | 86.4%    | 95.1%    | 87.0%    | 98.2%    | 94.5%    | 95.7%    | 96.5%    | 92.1%    | 95.5%    | 91.5%    |
| 20%          | 72.8%    | 82.1%    | 95.7%    | 85.3%    | 98.1%    | 95.4%    | 95.8%    | 96.5%    | 93.7%    | 96.9%    | 91.1%    |
| 25%          | 74.0%    | 85.2%    | 96.3%    | 83.4%    | 96.8%    | 93.9%    | 96.1%    | 93.8%    | 93.1%    | 95.1%    | 90.8%    |
| 30%          | 62.7%    | 85.6%    | 95.1%    | 84.7%    | 98.2%    | 94.1%    | 97.1%    | 96.1%    | 94.8%    | 95.2%    | 90.4%    |

randomly shear the image by $-s$ degrees to $s$ degrees on y-axis. For the rotate value $r$, the transformation is randomly rotating the image by $-r$ degrees to $+r$ degrees. Both the $s$ and $r$ belong to [10%, 15%, 20%, 25%, 30%]. The experimental results are demonstrated in Table A.2 and Table A.3, respectively. For the shear value determination, the Faster R-CNN has achieved the highest mAP of 91.5% when $s$ equals 15%. As such, the shear value is determined as 15% in this research. For the rotate value determination, the Faster R-CNN has achieved the highest mAP of 91.6% when $r$ is 10% and 15%. Actually, the detection performance of 15% $r$ is slightly higher than the 10% $r$ than 0.05% mAP. Therefore, this research selects 15% as the rotate value in the weak data augmentation.

As introduced in the discussion section, the Faster R-CNN without any data augmentation has achieved the mAP of 90.8% when training with the full training set. The improvement in mAP is between 0% and 0.5% when adopting different translate values. And the mAP improvement is between $-0.4\%$ to 0.7% and 0.5% to 0.8% for shear and rotate transformations, respectively. It is found that finetuning the parameters for each type of weak data augmentation can provide around 1% mAP improvement for Faster R-CNN on the ACID dataset. To be noted, the improvement of 1% mAP is significant in object detection studies when the mAP is larger than 90%. An additional experiment has been conducted to process the images and the corresponding labels by one of translate, shear, and rotate transformations for training, which has obtained the mAP of 91.5%. This result is close to that only adopting one type of weak data augmentation, which is reasonable because only one of these transformations has been applied randomly for a certain image.

References

[1] J. Yang, M.-W. Park, P.A. Vela, M. Golparvar-Fard, Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future, Adv. Eng. Informatics. 29 (2015) 211-224, https://doi.org/10.1016/j.aei.2015.01.011.

[2] B. Xiao, Z. Zhu, Two-dimensional visual tracking in construction scenario: A comparative study, J. Comput. Civ. Eng. 32 (2018) 04180006, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000738.

[3] M. Lu, W. Chen, X. Shen, H.-C. Lam, J. Liu, Positioning and tracking construction vehicles in highly dense urban areas and building construction sites, Autom. Constr. 16 (2007) 647-656, https://doi.org/10.1016/j.autcon.2006.11.001.

[4] A. Aidian, B. Quintana, S.A. Prieto, F. Bosch, Scan-to-BIM for ‘secondary’ building components, Adv. Eng. Informatics. 37 (2018) 119-138, https://doi.org/10.1016/j.aei.2018.05.001.

[5] J.S. Bohn, J. Teizer, Benefits and barriers of construction project monitoring using high-resolution automated cameras, J. Constr. Eng. Manag. 136 (2010) 632-640, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000164.

[6] B. Xiao, S. Kang, Vision-based method integrating deep learning detection for tracking multiple construction machines, J. Comput. Civ. Eng. 35 (2021) 04020071, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000957.

[7] H. Son, H. Seong, H. Choi, C. Kim, Real-time vision-based warning system for prevention of collisions between workers and heavy equipment, J. Comput. Civ. Eng. 33 (2019) 1-14, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000845.

[8] Z.-Q. Zhao, P. Zheng, X.-T. Xu, X. Wu, Object detection with deep learning: A review, IEEE Trans. Neural Networks Learn. Syst. PP (2019) 1-21, https://doi.org/10.1109/TNNLS.2018.2876665.

[9] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, L. Zhiwei, Dataset and benchmark for detecting moving objects in construction sites, Autom. Constr. 122 (2021), 103482, https://doi.org/10.1016/j.autcon.2020.103482.

[10] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, K. Murphy. Speed/accuracy trade-offs for modern object detectors, Adv. Eng. Informatics. 50 (2021) 101372
K. Wang, X. Yan, D. Zhang, L. Zhang, L. Lin. Towards human-machine cooperation: C. Doersch, A. Zisserman. Multi-task self-supervised visual learning, IEEE, 2017.

H. Kim, S. Bang, H. Jeong, Y. Ham, H. Kim. Analyzing context and productivity of G. Gualdi, A. Prati, R. Cucchiara. Contextual information and covariance E. Rezazadeh Azar, S. Dickinson, B. McCabe. Server-customer interaction tracker: J. Guo, Q. Wang, Y. Li. Semi-supervised learning based on convolutional neural

H. Tajeen, Z. Zhu. Image dataset development for measuring construction equipment recognition performance, Autom. Constr. 48 (2014) 1-10. https://doi.org/10.1016/j.autcon.2014.07.006.

J. Guo, Q. Wang, Y. Li. Semi-supervised learning based on convolutional neural network and uncertainty filter for façade defects classification. Comput. Civ. Infrastruct. Eng. (2020) 1-16. https://doi.org/10.1011/mice.12632.

M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman. The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2010) 303-338, https://doi.org/10.1007/s11263-009-0275-y.

E. Rezazadeh Azar, S. Dickinson, B. McCabe. Server-customer interaction tracker: Computer vision-based system to estimate dirt-loading cycles, J. Constr. Eng. Manag. 139 (2013) 785-794, https://doi.org/10.1061/(ASCE)JCEM.1943-7626.0000652.

J. Yang, P. Vela, J. Teizer, Z. Shi. Vision-based tower crane tracking for understanding construction activity, J. Constr. Civ. Eng. 28 (2014) 103-112, https://doi.org/10.1061/(ASCE)JCEM.1943-7626.0000242.

S. Chi, G.H. Caldasa. Image-based safety assessment: Automated spatial safety risk identification of earthmoving and surface mining activities, J. Constr. Eng. Manag. 138 (2012) 341-351, https://doi.org/10.1061/(ASCE)JCEM.1943-7626.0000438.

G. Guidali, A. Prati, R. Cucchiara. Contextual information and covariance descriptors for people surveillance: An application for safety of construction workers. EURASIP J. Image Video Process. (2011) 1-16, https://doi.org/10.1186/2050-4829-2011-684819.

A. Heydarian, M. Memarzadeh, M. Golparvar-Fard. Automated benchmarking and monitoring of an earthmoving operation’s carbon footprint using video cameras and a greenhouse gas estimation model, in: Proceedings: Comput. Civ. Eng., American Society of Civil Engineers, Reston, VA, 2012, pp. 509-516, https://doi.org/10.1061/9780784412343.0064.

C. Chen, Z. Zhu, A. Hammad, Automated excavators activity recognition, Autom. Constr. 110 (2020), 103045, https://doi.org/10.1016/j.autcon.2019.103045.

S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137-1149, https://doi.org/10.1109/TPAMI.2016.2577031.

H. Kim, S. Bang, H. Jeong, Y. Ham, H. Kim. Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation, Autom. Constr. 92 (2018) 188-198, https://doi.org/10.1016/j.autcon.2018.04.002.

J. Dai, Y. Li, K. He, J. Sun. R-FCN: Object detection via region-based fully convolutional networks, Adv. Neural Inf. Process. Syst. (2016) 379-387, https://papers.nips.cc/paper/2016/file/5755e1154524da259b6413a7346ae1-Paper.pdf.

B. Xiao, S. Kang. Development of an image data set of construction machines for deep learning object detection, J. Constr. Civ. Eng. 35 (2021) 05020065, https://doi.org/10.1061/(ASCE)JCEM.1943-7626.0000945.

X. Wang, A. Gupta. Unsupervised learning of visual representations using videos, IEEE (2015), 2794-2802, https://doi.org/10.1109/ICCV.2015.220.

C. Doersch, A. Zisserman. Multi-task self-supervised visual learning, IEEE, 2017, pp. 2070-2079, https://doi.org/10.1109/ICCV.2017.226.

K. Wang, X. Yan, D. Zhang, L. Zhang, L. Lin. Towards human-machine cooperation: Self-supervised sample mining for object detection, IEEE, 2018, pp. 1605-1613, https://doi.org/10.1109/CVPR.2018.00173.