Hindawi
Structural Control and Health Monitoring
Volume 2024, Article ID 1032674, 13 pages
https://doi.org/10.1155/2024/1032674

Research Article

Vision-Based Multiscale Construction Object Detection under Limited Supervision

Yapeng Guo,[1] **Yang Xu**,[2] **Hongtao Cui**,[1] **and Shunlong Li**[1]

[1]*School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China*
[2]*School of Civil Engineering, Harbin Institute of Technology, Harbin 150090, China*

Correspondence should be addressed to Shunlong Li; lishunlong@hit.edu.cn

Received 6 December 2023; Revised 17 January 2024; Accepted 5 February 2024; Published 16 February 2024

Academic Editor: Wenai Shen

Contemporary multiscale construction object detection algorithms rely predominantly on fully-supervised deep learning, requiring arduous and time-consuming labeling process. This paper presents a novel semisupervised multiscale construction objects detection (SS-MCOD) by harnessing nearly infinite unlabeled images along with limited labels, achieving more accurate and robust detection results. SS-MCOD uses a deformable convolutional network (DCN)-based teacher-student joint learning framework. DCN uses deformable advantages to extract and fuse multiscale construction object features. The teacher module generates pseudolabels for construction objects in unlabeled images, while the student module learns the location and classification of construction objects in both labeled images and unlabeled images with pseudolabels. Experimental validation using commonly used construction datasets demonstrates the accuracy and generalization performance of SS-MCOD. This research can provide insights for other detection tasks with limited labels in the construction domain.

1. Introduction

Construction site monitoring methods are very important in construction site safety management and productivity analysis. As one of the important basic tasks of intelligent construction, vision-based multiscale construction object detection (MCOD), which aims to accurately localize and recognize various objects with different sizes, can provide important data support for subsequent collision risk warning and construction process optimization [1, 2]. The earlier MCOD algorithms used traditional image processing technology to manually design features and make simple judgments. With the development of machine learning, designing automatic classification algorithms using handcrafted features to conduct MCOD has gradually become the mainstream.

Machine learning-based methods have improved MCOD accuracy to a certain extent, but they are not effective in the case of complex background interference. Deep learning methods have achieved convincing detection accuracy in general object detection field, which greatly improves the detection accuracy of construction objects. Most of these methods have used fully-supervised deep learning approaches and used large numbers of labeled construction image datasets to train detection models [3, 4]. The construction object detection accuracy depends largely on the labeled dataset's quality and quantity.

Building high-quality large-scale MCOD datasets is very challenging and requires lots of time and labor costs. Meanwhile, there are differences in the data distribution between different datasets, and the performance of a MCOD model that performs well on a certain dataset would decrease when tested directly on other datasets [5–7]. Therefore, developing a MCOD model that does not rely heavily on large-scale labeled datasets is of great significance for reducing training costs, expanding data utilization, and improving model generalization capabilities. To achieve this goal, this paper develops a novel semisupervised deep learning-based MCOD approach. As shown in Figure 1, only limited number of labeled images (i.e., small amount) and nearly infinite number of unlabeled images (i.e., large

amount) are needed to achieve better detection accuracy than fully-supervised deep learning (for intradataset and across-dataset).

The research of this paper mainly includes the following two contributions. Firstly, this paper proposes a novel semisupervised MCOD framework, which achieves more accurate and robust detection of construction objects. Secondly, aiming at the multiscale detection problem caused by the large size difference, the presented SS-MCOD uses DCN instead of conventional convolutional network to further improve detection precision. This paper's remaining sections are organized as follows: Section 2 examines the research advancement in vision-based construction object detection; Section 3 introduces the detailed architecture of the proposed SS-MCOD method; Section 4 illustrates the specifics of the implementation; Section 5 exhibits the outcomes of the training and evaluation, as well as the impact of key factors; and Section 6 provides a summary of this paper.

2. Related Studies

Before the widespread adoption of deep learning, most of the vision-based construction object automatic detection methods have used sliding windows to extract image regions of interest and then use hand-crafted features to identify the construction objects contained in the regions. Chi and Caldas [8] used surveillance camera videos to develop a construction object detection algorithm. The background subtraction and morphological operation were used to obtain the area of the construction object, two classifiers were trained to identify the object using shape and texture features. Park et al. [9] conducted a comparative analysis of various manual feature extraction techniques to assess their impact on object recognition accuracy. Azar et al. [10] developed Haar-HOG-based and Blob-HOG-based construction truck detectors, and a part-based excavator recognition method in videos using HOG features. Yuan et al. [11] uses hybrid kinematics shape and key node features to develop an excavator detection algorithm. These approaches, which relied on manually designed features, achieved a partial abstraction of construction objects and led to enhanced processing efficiency and precision in construction object detection.

In the deep learning period, convolutional neural network-based object detection methods have gained wide popularity in construction object detection domain. These deep learning-based methods can be categorized into two-stage anchor-based, single-stage anchor-based, and anchor-free techniques. In the case of two-stage anchor-based techniques, researchers predominantly used Faster R-CNN and its variants for construction object detection. Fang et al. [12] introduced an innovative approach called IFaster R-CNN, specifically designed to detect construction workers and heavy construction equipment. They demonstrated the superiority of their proposed method over the hand-crafted feature-based detection approach. Kim et al. [13] presented the use of transfer learning to address the scarcity problem of training data within the R-CNN series for the construction domain. To detect construction workers in complex backgrounds and changing postures more accurately, Son et al. [14] integrated the deep residual network (152 layers) into Faster R-CNN. Lu et al. [15] implemented fill factor estimation and bucket detection using Faster R-CNN with the feature integration of region proposal network. In the case of single-stage anchor-based techniques, Guo et al. [3] devised an enhanced version of SSD model to ensure accurate detection of dense construction vehicles. Roberts and Golparvar-Fard [16] used RetinaNet with ResNeXt-101 backbone as earthmoving equipment detection module for activity analysis. To accelerate the detection speed, Arabi et al. [17] developed lightweight construction object detection networks using SSD network with MobileNet as backbone. You only look once (YOLO)-v3 is an excellent general object detection framework. Xiao et al. [18] used YOLO-v3 for construction machinery detection in nighttime environment, construction personnel safety device detection, and construction equipment detection in large-scale scenes. In the case of anchor-free techniques, Guo et al. [19] developed an anchor-free method for detecting construction vehicles with arbitrary orientations to realize precise localization of construction vehicles in any orientations.

The construction object detection methods based on deep CNN mentioned above are mostly fully-supervised deep learning, that is, it requires labeled construction datasets for training. In theory, more labeled data with more-parameter models will produce better detection results. To address the challenge of limited labeled data, researchers have dedicated significant efforts to curating benchmark datasets in the field of construction [20], such as, ACID (10,000 images) [5], MOCS (41,668 images) [6], and SODA (19,846 images) [7]. In addition to consuming lots of time and manpower to obtain and annotate datasets, researchers also have tried to use new techniques to automatically generate construction images and corresponding annotations. Soltani et al. [21] developed an automated construction image generation and annotation approach using 3D equipment models. Bang et al. [22] used generative adversarial networks to generate more construction images with various transforms. Hwang et al. [23] investigated to use web crawling technique to acquire construction images and use a segmentation model to automatically label construction objects. The methods using generative approaches to synthesize simulated construction site images have to some extent improved the accuracy of construction object detection. However, these methods still rely on source data from previous limited construction image dataset. The generative models then generate new images that closely match the distribution of this existing construction image set, but they are unable to simulate the diversity of real construction site image distributions. Therefore, the precision and robustness of detecting construction objects using such methods are clearly limited.

The annotation of construction image datasets is time-consuming and laborious, and a large-scale and high-quality annotation is more difficult. Widely-used datasets in general object detection are usually annotated

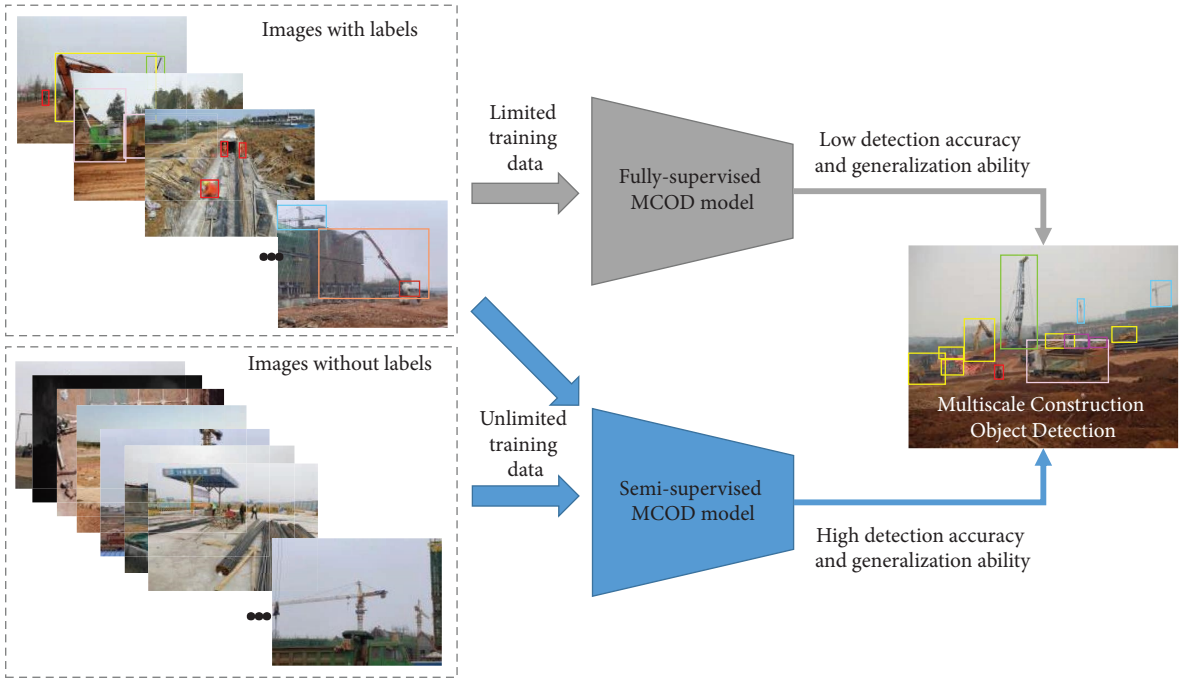

Figure 1: Comparison of MCOD methods with different supervision approaches.

at million-image level. However, it undoubtedly takes a long time to accumulate such numbers of annotations in construction domain. But it is relatively easy to obtain only construction images. Indeed, it is crucial to research more accurate and robust construction object detection algorithms by using limited labeled images and combining with almost infinite unlabeled images. Kim et al. [24] introduced few-shot learning into the construction domain, and successfully realized new construction object category detection using limited labeled samples, and became the pioneer work of construction object detection research under limited labeled samples. Few-shot learning focuses on discovering new categories that are not in the training set, while semisupervised learning is able to effectively use unlabeled data to further enhance detection models [25]. The current application of semisupervised learning in the construction domain focuses on structural damage identification and segmentation. Guo et al. [26] developed a façade defect classification approach with semisupervised learning using a modified mean teacher technique which could train labeled and unlabeled images simultaneously. Wang and Su [27] proposed a surface crack semantic segmentation model and used the semisupervised teacher and student framework with EfficientUNet as the backbone. Zhang et al. [28] presented automatic defect segmentation frameworks integrated with GANs and semisupervised learning to achieve better precision. Unlike object classification and segmentation, semisupervised object detection needs to consider more factors and is more difficult to implement. To fill this gap, this paper proposes the SS-MCOD framework using semisupervised learning technique to realize more precise and robust MCOD.

3. Methodology

As illustrated in Figure 2, the SS-MCOD method proposed in this paper is a teacher-student joint learning framework based on the deformable convolutional network (DCN). The teacher-student joint learning structure is designed to enable semisupervised learning, while the DCN component is leveraged to resolve multiscale issues, thereby enhancing accuracy in construction object detection tasks. During training, labeled construction images are directly input into the student module, and its output is compared with manually labeled data to calculate loss. Unlabeled construction images undergo strong augmentation and are input into the student module to produce pseudolabels (for classification and localization). Additionally, unlabeled construction images undergo weak augmentation and are input into the teacher module, with its output compared with the pseudolabels to calculate loss. The weights of the teacher model are transferred from the weights of the student model using exponential moving average technique.

3.1. Teacher-Student Joint Learning Framework. The proposed SS-MCOD approach introduces a teacher-student joint learning framework for effective semisupervised learning. Specifically, SS-MCOD undergoes training utilizing a combination of labeled and unlabeled data. Both the teacher module and the student module use the identical fully-supervised object detection architecture, which serves as the base construction object detection model. However, the parameters are different between the two models. The teacher module's parameters are transferred through the application of the exponential moving average technique from the student module [29]. This technique enables

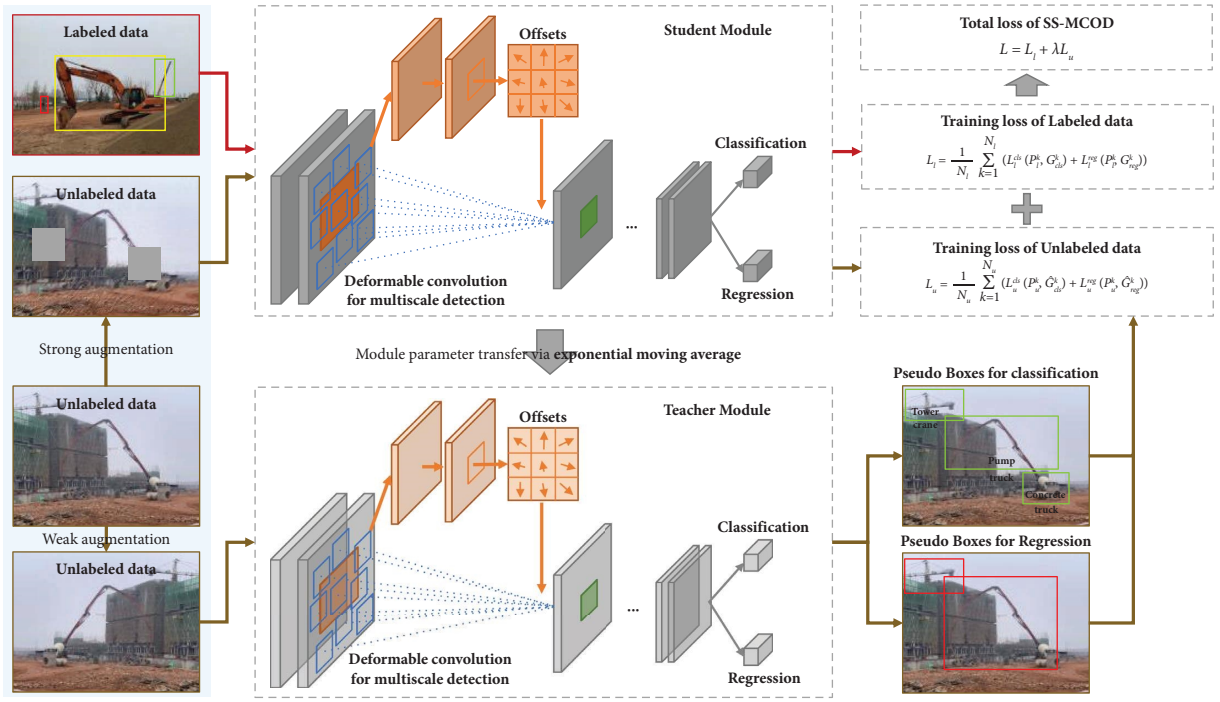

Figure 2: Framework of the proposed SS-MCOD.

a smooth and gradual transfer of knowledge, facilitating enhanced learning and convergence within the semi-supervised framework.

The construction image processing pipeline of the framework is shown in Figure 3. During the training phase of the proposed SS-MCOD method, construction images are divided into labeled and unlabeled data. The labeled data are directly input into the student module, while the unlabeled data undergo strong and weak augmentation, with the former being input into the student module and the latter into the teacher module. All outputs are involved in the calculation of training loss. Additionally, the weights of the teacher module are transferred from the student module. During the inference phase, construction images are directly input into the trained teacher module to obtain bounding boxes representing the position and category of construction objects.

The labeled data uses the same process as the fully-supervised object detection, and is sent to the student module as input, and generates the training loss of the labeled data L_l: labeled classification loss L_l^{cls} and labeled positioning loss L_l^{loc}, as shown in equation (1). N_l represents the labeled construction image number, P_l^k is the corresponding predicted value of the student module, G_{cls}^k represents the labeled construction object category, and G_{loc}^k represents the labeled construction object location.

$$L_l = \frac{1}{N_l}\sum_{k=1}^{N_l}\left(L_l^{cls}\left(P_l^k, G_{cls}^k\right) + L_l^{loc}\left(P_l^k, G_{loc}^k\right)\right). \tag{1}$$

Unlabeled construction images go through two different data augmentation approaches to generate strongly and weakly augmented data. Strongly augmented data is sent into the student module as input, and the predicted construction object detection bounding boxes of unlabeled construction images is output. Weakly augmented data is sent into the teacher module as input with the output of pseudolabels of unlabeled data. The difference between the predicted construction object detection bounding boxes and the pseudolabel is computed, that is, the training loss of unlabeled data L_u: unlabeled classification loss L_u^{cls} and unlabeled location loss L_u^{loc}, as shown in equation (2). N_u is the number of unlabeled construction images, P_u^k is the corresponding predicted construction object detection bounding boxes of the student module, $\widehat{G}_{cls}^k$ is the pseudocategory label, and $\widehat{G}_{loc}^k$ is the pseudolocation label.

$$L_u = \frac{1}{N_u}\sum_{k=1}^{N_u}\left(L_u^{cls}\left(P_u^k, \widehat{G}_{cls}^k\right) + L_u^{loc}\left(P_u^k, \widehat{G}_{loc}^k\right)\right). \tag{2}$$

The quality of construction object pseudolabels is crucial to the training and inference accuracy of SS-MCOD. After the weakly augmented construction images are input into the teacher module, multiple construction object detection bounding boxes will be generated. To eliminate the results of high repetition rate, nonmaximum suppression is used to perform preliminary postprocessing on multiple construction object detection bounding boxes. Referencing to Xu et al. [30], a high threshold is used to filter these bounding boxes after preliminary postprocessing. These construction object detection bounding boxes can be divided into foreground boxes (r_k^{fg}) and background boxes (r_j^{bg}). The foreground boxes are used as the pseudolabel of the classification, and the reliability measure (ω_j) is used to weight the loss of each background box to calculate the unlabeled classification loss L_u^{cls}. As shown in equations (3) and (4),

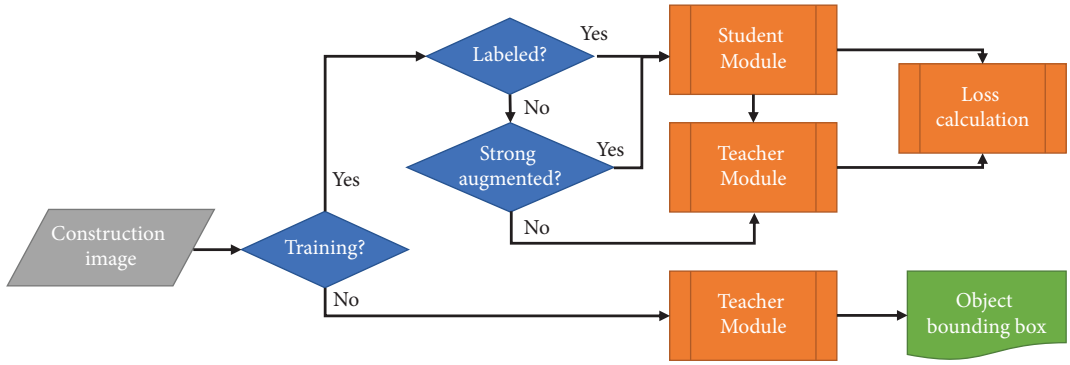

Figure 3: Flowchart of the proposed SS-MCOD.

N_b^{fg} and N_b^{bg} are the number of foreground and background boxes, l_{cls} represents the cross-entropy loss, and m_i is the reliability score for i-th background box.

$$L_u^{cls} = \frac{1}{N_b^{fg}} \sum_{k=1}^{N_b^{fg}} l_{cls}\left(r_k^{fg}, \widehat{G}_{cls}\right) + \sum_{j=1}^{N_b^{bg}} \omega_j l_{cls}\left(r_j^{bg}, \widehat{G}_{cls}\right), \tag{3}$$

$$\omega_j = \frac{m_j}{\sum_{i=1}^{N_b^{bg}} m_i}. \tag{4}$$

Due to the inconsistency of construction object classification and locating tasks, high-quality classification pseudolabels are usually inconsistent with high-quality positioning pseudolabels. This paper uses box jitter method to select the reliable bounding box location coordinates, that is, the variance is calculated after multiple jitters of the foreground box as the reliability measure, and finally the box with high enough reliability (r_k^{fg}) is used as the location pseudolabel, where l_{cls} represents the L1 loss.

$$L_u^{loc} = \frac{1}{N_b^{fg}} \sum_{k=1}^{N_b^{fg}} l_{reg}\left(r_k^{fg}, \widehat{G}_{loc}\right). \tag{5}$$

The total loss L of SS-MCOD is composed of labeled loss and unlabeled loss, where λ is the adjustment coefficient.

$$L = L_l + \lambda L_u. \tag{6}$$

3.2. DCN-Based Student Module. Conventional convolution layers use a consistent convolution operation across various feature maps, with fixed pixel sampling positions (shown in Figure 4(a)). This approach results in the inclusion of numerous background features within the extracted information. Consequently, conventional convolution-based construction object detection networks possess uniform receptive fields for multiscale objects. This limitation hinders the accurate detection of multiscale construction objects.

The deformable convolution is proposed to replace the conventional convolution [31], that is, by adding an offset at the position of the original convolution sampling, as shown in equation (7). X and Y represent input and output convolutional feature maps, ω is the weight function, R is the convolution kernel, u_0 is the location in Y, u_k is the location in R, and Δu_k is the offset.

$$Y(u_0) = \sum_{u_k \in R} \omega(u_k) X(u_0 + u_k + \Delta u_k). \tag{7}$$

The receptive field can be rotated and scaled, which can effectively cover large construction objects and accurately concentrate near small construction objects (as shown in Figure 4(b)). Through the application of deformable convolution operations, disparate features are extracted according to the size and shape of the construction object. This approach minimizes the extraction of extraneous background information, contributing to more accurate object detection.

In the SS-MCOD framework, the base construction object detection model for both the teacher and student modules is derived from the Faster R-CNN-DCN (i.e., FRCD). The backbone of the FRCD architecture is ResNet-50, featuring four convolution stages. Notably, the first stage retains conventional convolutions, while the subsequent stages, from the second to the fourth, incorporate deformable convolutions.

4. Implementation Details

4.1. Dataset. The training dataset used in this paper was sampled from the MOCS dataset [6], which were acquired from 174 different construction sites considering various weather environments using a variety of equipment. The training dataset includes 12 common types of objects in construction sites.

To implement the training of SS-MCOD, the training dataset was segregated into labeled data and unlabeled data. The overall count of the training dataset image is 3000. To explore the influence of different proportions of labeled data and unlabeled data on SS-MCOD, this paper presents to conduct four training cases, and the proportions of labeled data are 2%, 5%, 10%, and 50%, respectively. The numbers of images and objects of different cases are shown in Figure 5. In this paper, objects with a bounding box area (width multiplied by height) smaller than 1024 pixels (32×32) are classified as small objects, while those larger than 9216 pixels (96×96) are categorized as large objects, and the remaining fall under the medium object category.

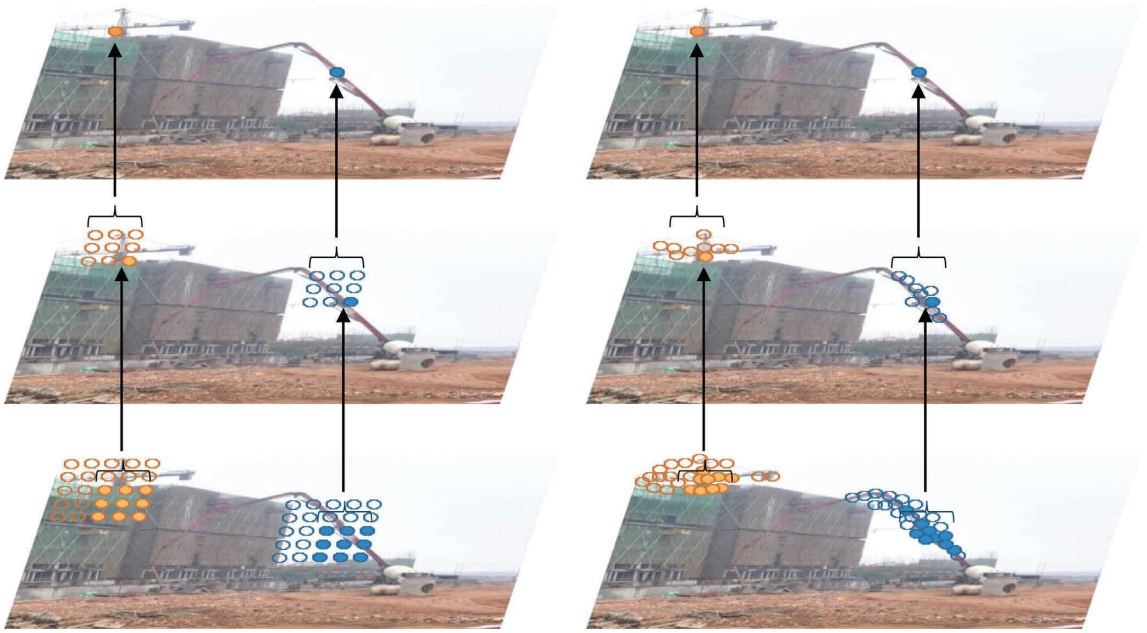

Figure 4: Receptive field comparison of (a) conventional convolution and (b) DCN.

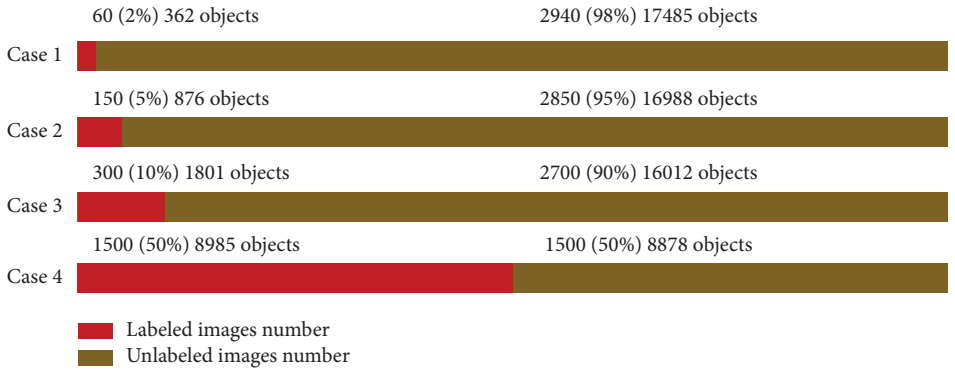

Figure 5: Numbers of images and objects of different training cases in training dataset.

To evaluate the accuracy improvement of SS-MCOD, the MOCS validation dataset (including 4000 images) was used as the validation dataset I (for intradataset evaluation). To evaluate the robustness improvement across different datasets, another dataset called ACID basic (2850 images) [5] was used as the validation dataset II (for across-dataset evaluation), including excavator, truck, and concrete truck.

4.2. Parameter Setting. Batch size, being one of the crucial parameters of deep learning, wields a significant impact on the training procedure. To keep consistent of different experimental cases, the batch size of all cases was both set to 5. In each batch, the ratio of labeled data to unlabeled data was 1 : 4. For each experimental case, the training epochs were set at 45000, while the initial learning rate was configured to 0.005. The learning rate alteration strategy used a multistep approach, involving reductions of 0.3 at the 15000th and 30000th epochs. λ in equation (6) is set as 2.0. The parameter settings for the student model or the teacher model are consistent with those in the original Faster R-CNN [32] and DCN [31].

The hardware platform utilized for training and testing SS-MCOD comprised an Intel Xeon(R) E5-2620 v4 (CPU), an Nvidia RTX 3090 graphics processing unit (GPU), and 128 GB of memory. At the software level, the algorithm is implemented using the MMDetection [33] deep object detection framework based on PyTorch [34]. The main library versions used are PyTorch (1.9.0), Torchvision (0.10.0), Cuda (11.1), OpenCV (4.7.0.72), mmcv (1.3.9), and MMDetection (2.16.0). The overall algorithm implementation system runs on Ubuntu 16.04.

5. Results and Discussions

Training and testing results of the proposed SS-MCOD under four training cases are introduced in this section. Moreover, the influence of pretraining on SS-MCOD is discussed.

5.1. Training Results. The proposed SS-MCOD was trained from scratch by four training cases. Figure 6 illustrates the training process. Training labeled classification loss represents L_l^{cls}; this item is part of the Faster R-CNN loss function. Figure 6(a) shows that L_l^{cls} decreases steadily with the increase of training steps. As the ratio of labeled data increases, so does the value of L_l^{cls} at the same training step, which is due to the lack of fitting. Training labeled classification loss represents L_l^{loc}, Figure 6(b) shows that L_l^{loc} increases first and then decreases, which is due to the characteristics of the two-stage detection method; the coordinates of bounding boxes can be regressed only after the candidate regions are screened. Similarly, as the ratio of labeled data grows, the loss also increases. Training unlabeled classification loss and location loss represent L_u^{cls} and L_u^{loc}, respectively. Figure 6(c) shows that L_u^{cls} increases significantly first and then decreases. Figure 6(d) shows that L_u^{loc} also increases significantly and then decreases slowly. This is because the quality of pseudolabels in the unsupervised learning branch is poor at the beginning of training. As the number of training steps increases, the accuracy of pseudolabel classification and regression improves, leading to a reduction in both two losses. Additionally, with the increase in the ratio of labeled data, L_u^{cls} and L_u^{loc} show a slight decrease, because the reduction of the number of unlabeled data reduces the difficulty of fitting.

Figure 7 is the curve of SS-MCOD total loss L. This loss initially experiences a sharp decrease followed by a gradual increase, and eventually transitions into a slow decrease with advancing training steps. The initial descent corresponds to the rapid data fitting by the supervised training branch of SS-MCOD. The subsequent ascent represents the evolving quality of pseudolabels in the unsupervised training branch. Finally, the subsequent descent reflects the improved data fitting capabilities of both the supervised and unsupervised training branches.

5.2. Intradataset Evaluation Results. To demonstrate the effectiveness of the proposed SS-MCOD, the fully-supervised detection method FRCD and the well-known semisupervised object detection method Soft teacher (Faster R-CNN as the student module) were used for evaluation and comparison. During training, Soft teacher and the proposed SS-MCOD used the same training data, while FRCD used only labeled data for training.

Table 1 presents the testing results of three different methods on validation dataset I, using varying training cases. In Case 1, the mAP achieved by FRCD trained solely with 60 labeled images reached 5.3. Demonstrating the benefits of semisupervised learning, the other two methods exhibit significant enhancements in mAP, highlighting the considerable advantage of the semisupervised detection framework for construction objects when labeled data is scarce. Notably, the proposed SS-MCOD outperforms Soft teacher in terms of evaluation accuracy. Similar trends are observed in Cases 2, 3, and 4. Relative to the fully-supervised FRCD, SS-MCOD yields substantial improvements in evaluation accuracy, with mAP increases of 10.8, 11.4, 10.4, and 13.8 in the respective cases. These improvements represent percentage increases of 204%, 107%, 58%, and 63%, respectively. This underscores the notion that SS-MCOD achieves more pronounced accuracy enhancements as the proportion of labeled data decreases, as well as improvements in recall. In contrast, this relationship is inverted when comparing Soft teacher to SS-MCOD. This phenomenon can be interpreted as follows: Leveraging unlabeled data, the semisupervised COD framework can effectively yield a more precise COD model compared to its fully-supervised counterpart.

The multiscale characteristic of construction objects is an important feature. AP_l, AP_m, and AP_s represent the accuracy of multiscale COD. Compared with FRCD, AP_l, AP_m, and AP_s of SS-MCOD increased by 148%/300%/471%, 85%/117%/116%, 44%/57%/27%, and 47%/58%/56%, respectively. Compared with Soft teacher, AP_l, AP_m, and AP_s of SS-MCOD increased by 10%/14%/11%, 0%/3%/2%, 10%/5%/1%, and 10%/1%/5%, respectively. These results demonstrate that the proposed SS-MCOD achieves significantly higher multiscale detection accuracy compared to FRCD and exhibits improvement in comparison to Soft teacher. This can be explained as follows: SS-MCOD with DCN structure can better extract multilevel features of multiscale construction objects to achieve more accurate detection results.

The ability to identify construction objects across multiple scales is a pivotal feature. In this context, AP_l, AP_m, and AP_s denote the accuracy of MCOD. When juxtaposed with FRCD, the SS-MCOD display increases in AP_l, AP_m, and AP_s by 148%/300%/471%, 85%/117%/116%, 44%/57%/27%, and 47%/58%/56%, in sequential order. In contrast with Soft teacher, the SS-MCOD sees improvements in AP_l, AP_m, and AP_s by 10%/14%/11%, 0%/3%/2%, 10%/5%/1%, and 10%/1%/5%, respectively. This comparison underscores the superior multiscale detection accuracy of the proposed SS-MCOD when set against FRCD and its evident advancement over Soft teacher. The observed enhancement can be rationalized as follows: SS-MCOD, equipped with a DCN structure, adeptly extracts multilevel features from construction objects of varying scales, culminating in more precise detection outcomes.

Figure 8 qualitatively shows the example detection results of SS-MCOD (solid line) and Soft teacher (dotted line) in Case 4. Soft teacher failed to detect the two tower cranes positioned in the middle of the upper left image, the pump truck situated on the left side of the upper middle image, the construction worker located in the lower right of the middle left image, the three construction workers positioned on the right side of the lower left image, as well as the construction vehicle situated in the middle of the lower middle image. In contrast, SS-MCOD successfully detected all of these objects. This indicates that the proposed SS-MCOD can achieve more accurate detection when there are construction objects with large-scale differences in the same image.

Figure 8 provides a qualitative depiction of example detection results for SS-MCOD (represented by a solid line) and Soft teacher (represented by a dotted line) within Case 4. Notably, in the upper left image, two tower cranes

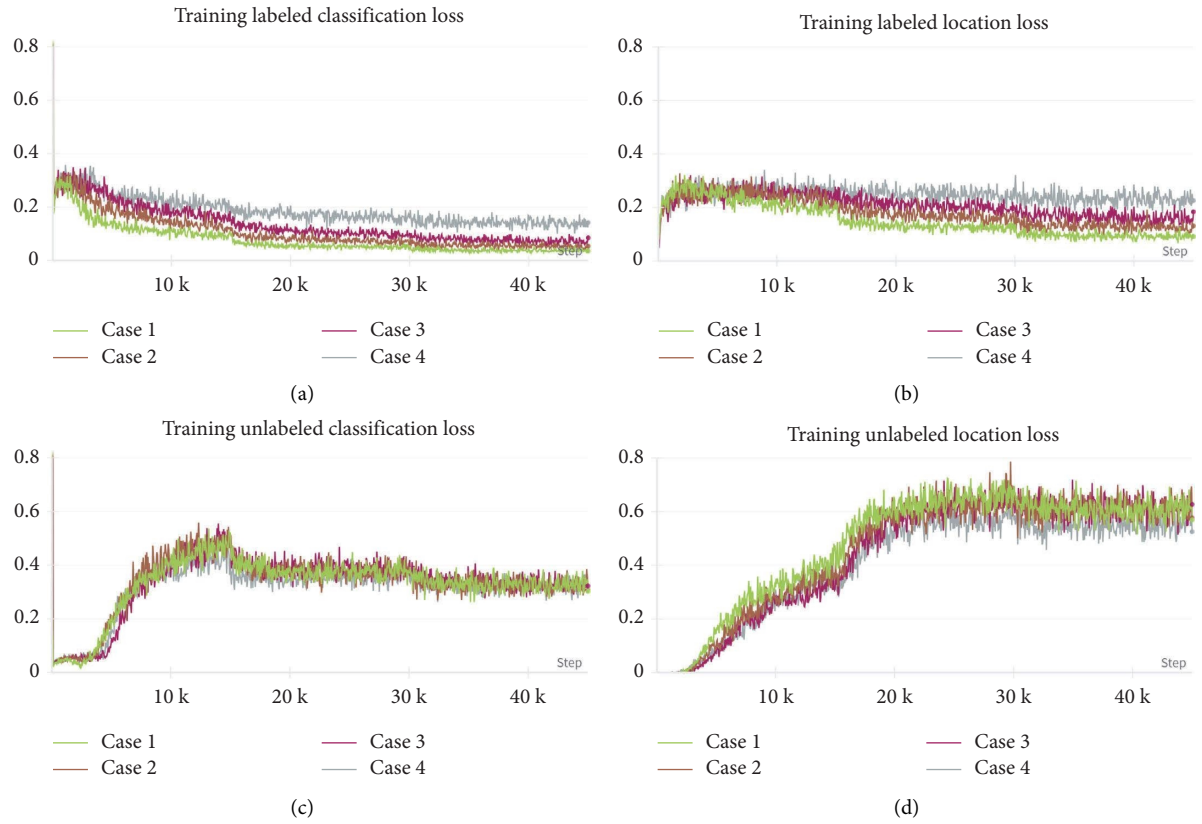

Figure 6: Loss curves of four loss subitems.

Figure 7: Total training loss curve of SS-MCOD.

positioned centrally were successfully detected by SS-MCOD, while eluding detection by Soft Teacher. Similarly, the pump truck situated on the left side of the upper middle image, the construction worker located in the lower right of the middle left image, the three construction workers positioned on the right side of the lower left image, and the construction vehicle at the center of the lower middle image, all went unnoticed by Soft teacher, yet were accurately identified by SS-MCOD. These results underscore the capability of the proposed SS-MCOD to achieve heightened detection accuracy, particularly in scenarios involving construction objects with significant scale variations within a single image.

5.3. Across-Dataset Evaluation Results. To ascertain the generalization capabilities across-datasets, a novel dataset (validation dataset II) was used to evaluate the proposed SS-MCOD. The evaluation results on Validation Dataset II, with various training cases, are illustrated in Table 2.

Similar to the intradataset evaluation, the fully-supervised framework (FRCD) exhibits noticeably lower detection accuracy compared to the semisupervised frameworks. Distinct from the intradataset findings, the performance of the SS-MCOD against Soft Teacher sees substantial enhancement. Specifically, across the four training cases, SS-MCOD demonstrates improvements of 12.6 (49%), 6.4 (17%), 6.2 (16%), and 8.1 (19%) in mAP,

Table 1: Evaluation results on validation dataset I with different training cases.

Case	Method	mAP	$AP_{0.5}$	$AP_{0.75}$	AP_l	AP_m	AP_s	$AR_{0.5:0.95}$
Case 1	FRCD	5.3	16.5	3.3	8.6	3.2	0.7	9.6
	Soft teacher	15.5	34.2	11.7	19.4	11.2	3.6	29.5
	SS-MCOD	**16.1**	**34.7**	**12.9**	**21.4**	**12.8**	**4.0**	**30.1**
Case 2	FRCD	10.6	31.9	8.8	15.5	6.9	2.5	22.7
	Soft teacher	21.4	45.2	16.4	28.7	14.5	5.3	35.2
	SS-MCOD	**22.0**	**46.1**	**19.0**	**28.7**	**15.0**	**5.4**	**36.6**
Case 3	FRCD	17.7	36.8	13.5	25.5	12.7	4.8	32.6
	Soft teacher	25.4	50.9	22.5	33.2	19.0	6.0	39.8
	SS-MCOD	**28.1**	**61.3**	**26.9**	**36.8**	**20.0**	**6.1**	**41.8**
Case 4	FRCD	21.6	45.7	17.1	30.3	16.6	5.7	36.1
	Soft teacher	32.6	59.7	31.7	40.7	24.9	8.5	47.6
	SS-MCOD	**35.4**	**61.3**	**36.2**	**44.7**	**25.1**	**8.9**	**49.1**

Best evaluation results among three methods.

Figure 8: Example detection results of two semisupervised methods in Case 4.

respectively. Furthermore, SS-MCOD's $AP_{0.5}$ and $AP_{0.75}$ outperforms Soft Teacher by 43%/52% and 4%/39%, respectively, alongside a 4%/34% increase in $AP_{0.5}$ and a 3%/40% boost in $AP_{0.75}$. These findings signify that SS-MCOD's accuracy augmentation, when using minimal labeled data (Case 1), arises from enhancements in both low-confidence and high-confidence detection outcomes. In cases with higher proportions of labeled data, accuracy gains primarily stem from improvements in high-confidence detections.

In terms of MCOD accuracy, SS-MCOD displays increased performance compared to Soft teacher. Specifically, SS-MCOD demonstrates growth in AP_l, AP_m, and AP_s by 51%/11%/9%, 18%/3%/2%, 10%/−1%/−12%, and 19%/4%/−2%, respectively. This analysis highlights SS-MCOD's

Table 2: Evaluation results on validation dataset II with different training cases.

Case	Method	mAP	$AP_{0.5}$	$AP_{0.75}$	AP_l	AP_m	AP_s	$AR_{0.5:0.95}$
Case 1	FRCD	10.7	29.7	8.0	13.6	7.3	0.0	19.6
	Soft teacher	25.6	52.2	22.4	26.7	16.5	5.5	41.7
	SS-MCOD	**38.2**	**75.4**	**34.2**	**40.3**	**18.3**	**6.0**	**50.9**
Case 2	FRCD	25.3	51.6	20.5	25.8	16.1	0.0	40.4
	Soft teacher	36.3	75.2	30.1	37.9	21.2	5.5	51.7
	SS-MCOD	**42.7**	**78.4**	**42.1**	**44.7**	**22.5**	**7.1**	**56.0**
Case 3	FRCD	28.7	57.5	24.9	29.1	17.2	3.0	44.0
	Soft teacher	38.7	75.2	35.6	40.3	**25.9**	**9.8**	54.0
	SS-MCOD	**44.9**	**78.3**	**47.7**	**47.1**	25.5	8.6	**58.5**
Case 4	FRCD	31.6	60.2	28.6	32.3	20.8	3.0	47.7
	Soft teacher	42.1	79.8	40.2	43.6	27.7	**12.3**	58.1
	SS-MCOD	**50.2**	**82.7**	**56.5**	**52.2**	**29.0**	12.0	**62.6**

Best evaluation results among three methods.

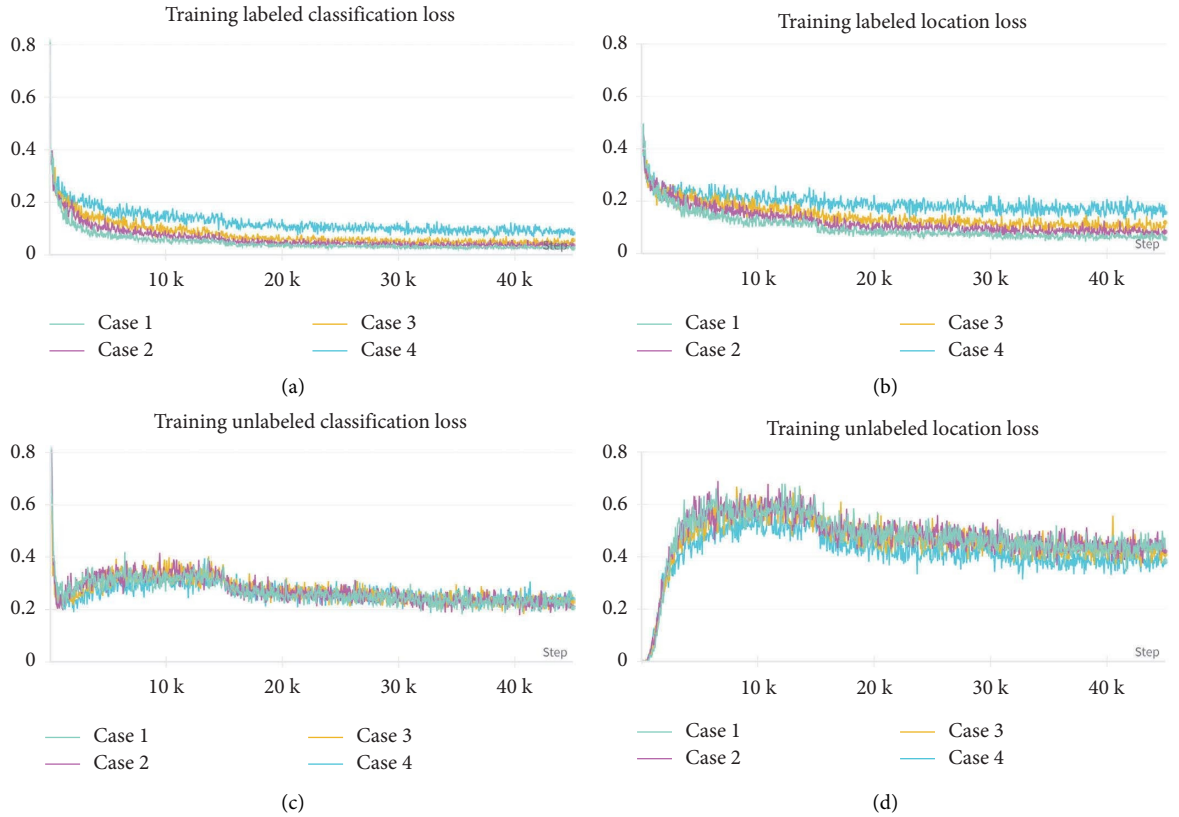

Figure 9: Loss curves of four loss subitems with pretraining.

Figure 10: Total training loss curve of SS-MCOD with pretraining.

Table 3: Evaluation results of SS-MCOD on two validation datasets with pretraining.

Validation dataset	Case	mAP	$AP_{0.5}$	$AP_{0.75}$	AP_l	AP_m	AP_s	$AR_{0.5:0.95}$
Dataset I	Case 1	22.2	35.9	24.2	30.2	14.9	4.3	34.2
	Case 2	33.3	52.2	35.2	44.9	21.0	8.6	46.3
	Case 3	39.7	59.6	43.6	52.5	27.1	8.2	51.5
	Case 4	46.1	68.1	50.8	58.2	33.2	14.9	59.1
Dataset II	Case 1	51.0	77.6	57.8	53.7	24.8	0.1	62.8
	Case 2	56.2	81.0	64.9	59.1	25.1	0.2	66.7
	Case 3	59.1	83.8	68.5	61.8	30.1	0.8	69.5
	Case 4	59.8	84.8	68.7	62.4	31.7	2.4	70.9

primary enhancement in detecting large-scale construction objects and accentuates the noticeable disparity in accuracy among objects of varied scales. This discrepancy can be attributed to the unique proportion of scales in validation dataset II, where it stands at 86%:13%:1%. It is worth noting that this proportion varies to 39% : 36% : 25% in validation dataset I.

5.4. Influence of Pretraining on SS-MCOD. The effectiveness of pretraining in object detection models, which allows for the extraction of more generalized features, has been widely recognized and established. To ensure both representativeness and accessibility of object detection datasets, this study used the training set from the COCO dataset [35] to train the student module of SS-MCOD. The training process spanned 180,000 epochs, after which the trained weight parameters were adopted as the initial weights for SS-MCOD's continued training or fine-tuning.

Illustrated in Figures 9 and 10 are the training loss curves with pretraining. A notable reduction in loss is observed for SS-MCOD with pretraining in comparison to SS-MCOD without pretraining, evident in both partial loss and total loss. This reduction signifies an enhanced capacity of the model to conform to the dataset. Specifically, the trends and patterns of partial losses, L_l^{cls} and L_l^{loc}, within the labeled branches mirror those of SS-MCOD without pretraining, albeit with over 30% reduction in loss values. However, significant changes are noted in the unlabeled training partial losses when compared to SS-MCOD without pretraining. The initial surge in L_u^{cls} is considerably mitigated, and the ascending phase of L_u^{loc} is notably abbreviated, followed by a pronounced reduction. These observations underscore the impact of the unlabeled training branch in expediting pseudolabel generation, thus substantially augmenting the model's fitting capability.

Table 3 presents the evaluation results of SS-MCOD on two validation datasets with pretraining. In the context of intradataset evaluation, SS-MCOD's mAP with pretraining exhibited substantial improvements, registering increments of 37%, 51%, 41%, and 30%, respectively, when compared to SS-MCOD without pretraining. In terms of across-dataset evaluation, SS-MCOD's mAP with pretraining saw noticeable enhancements, with increases of 33%, 31%, 31%, and 19%, respectively, relative to SS-MCOD without pretraining. These findings underscore the significant efficacy of the pretraining strategy in augmenting the performance of SS-MCOD trained on datasets characterized by varying proportions of labeled data.

In addition to pretraining, the choice of backbone for the student or teacher modules in this paper's Faster R-CNN is also a significant factor affecting detection accuracy. The use of a more powerful feature extractor can further improve detection accuracy, but at the same time, the algorithm's processing speed will decrease. Furthermore, by statistically analyzing the size characteristics of construction objects and then determining the size and quantity of anchors in Faster R-CNN, detection accuracy can be further enhanced.

6. Conclusions

In this paper, a novel semisupervised multiscale construction object detection method, SS-MCOD, is introduced. This approach takes advantage of a limited number of labeled samples along with a vast amount of unlabeled construction images for training. As a result, SS-MCOD achieves improved accuracy and robustness in object detection. The following conclusions can be drawn: (1) Superior performance over fully-supervised methods: When contrasted with fully-supervised methods, SS-MCOD achieves substantial improvements in both intradataset and across-dataset evaluations. Notably, for the four cases, the improvements of 204%, 107%, 58%, and 63% in intradataset evaluation and 357%, 168%, 156%, and 158% in across-dataset evaluation have been achieved. These outcomes underscore SS-MCOD's elevated accuracy and its adeptness in generalizing across diverse datasets. (2) Multiscale capability: By harnessing the potent multiscale feature

extraction capabilities inherent in the DCN architecture, SS-MCOD demonstrates pronounced advancements in multiscale COD accuracy compared to the widely recognized semisupervised object detection method, Soft teacher. (3) Impact of pretraining: The incorporation of a pretraining strategy yields a significant enhancement in the accuracy and generalization capabilities of SS-MCOD. Model pretraining using COCO dataset results in an average mAP increase of 40% for intradataset evaluation and 28% for across-dataset evaluation.

The proposed semisupervised framework effectively enhances detection accuracy and robustness, making it applicable for the efficient and cost-effective detection of various valuable objects in civil engineering contexts. However, this study still has the following limitations: the SS-MCOD framework proposed uses Faster R-CNN as the detector, which is a classic two-stage anchor-based detection framework with relatively high detection accuracy but slow running speed. Future research efforts can focus on adopting single-stage anchor-based or anchor-free detectors with higher detection efficiency, ensuring a significant reduction in algorithm runtime while improving detection accuracy.

Data Availability

The data used in this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

Financial support for this study was provided by NSFC (Grant nos. U22A20230 and 52278299) and Fundamental Research Funds for the Central Universities (Grant no. FRFCU5710051018).

References

[1] J. Yang, M.-W. Park, P. A. Vela, and M. Golparvar-Fard, "Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 211–224, 2015.

[2] W. Fang, L. Ding, P. E. D. Love et al., "Computer vision applications in construction safety assurance," *Automation in Construction*, vol. 110, Article ID 103013, 2020.

[3] Y. Guo, Y. Xu, and S. Li, "Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network," *Automation in Construction*, vol. 112, Article ID 103124, 2020.

[4] B. Ekanayake, J. K. W. Wong, A. A. F. Fini, and P. Smith, "Computer vision-based interior construction progress monitoring: a literature review and future research directions," *Automation in Construction*, vol. 127, Article ID 103705, 2021.

[5] B. Xiao and S.-C. Kang, "Development of an image data set of construction machines for deep learning object detection," *Journal of Computing in Civil Engineering*, vol. 35, no. 2, Article ID 05020005, 2021.

[6] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, "Dataset and benchmark for detecting moving objects in construction sites," *Automation in Construction*, vol. 122, Article ID 103482, 2021.

[7] R. Duan, H. Deng, M. Tian, Y. Deng, and J. Lin, "SODA: a large-scale open site object detection dataset for deep learning in construction," *Automation in Construction*, vol. 142, Article ID 104499, 2022.

[8] S. Chi and C. H. Caldas, "Automated object identification using optical video cameras on construction sites," *Computer-Aided Civil and Infrastructure Engineering*, vol. 26, no. 5, pp. 368–380, 2011.

[9] M.-W. Park and I. Brilakis, "Construction worker detection in video frames for initializing vision trackers," *Automation in Construction*, vol. 28, pp. 15–25, 2012.

[10] E. Rezazadeh Azar, B. McCabe, and B. McCabe, "Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos," *Automation in Construction*, vol. 24, pp. 194–202, 2012.

[11] C. Yuan, S. Li, and H. Cai, "Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes," *Journal of Computing in Civil Engineering*, vol. 31, no. 1, Article ID 04016038, 2017.

[12] W. Fang, L. Ding, B. Zhong, P. E. Love, and H. Luo, "Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach," *Advanced Engineering Informatics*, vol. 37, pp. 139–149, 2018.

[13] H. Kim, H. Kim, Y. W. Hong, and H. Byun, "Detecting construction equipment using a region-based fully convolutional network and transfer learning," *Journal of Computing in Civil Engineering*, vol. 32, no. 2, Article ID 04017082, 2018.

[14] H. Son, H. Choi, H. Seong, and C. Kim, "Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks," *Automation in Construction*, vol. 99, pp. 27–38, 2019.

[15] J. Lu, Z. Yao, Q. Bi, and X. Li, "A neural network–based approach for fill factor estimation and bucket detection on construction vehicles," *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 12, pp. 1600–1618, 2021.

[16] D. Roberts and M. Golparvar-Fard, "End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level," *Automation in Construction*, vol. 105, Article ID 102811, 2019.

[17] S. Arabi, A. Haghighat, and A. Sharma, "A deep-learning-based computer vision solution for construction vehicle detection," *Computer-Aided Civil and Infrastructure Engineering*, vol. 35, no. 7, pp. 753–767, 2020.

[18] B. Xiao, Q. Lin, and Y. Chen, "A vision-based method for automatic tracking of construction machines at nighttime based on deep learning illumination enhancement," *Automation in Construction*, vol. 127, Article ID 103721, 2021.

[19] Y. Guo, Y. Xu, J. Niu, and S. Li, "Anchor-free arbitrary-oriented construction vehicle detection with orientation-aware Gaussian heatmap," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 7, pp. 907–919, 2023.

[20] H. Tajeen and Z. Zhu, "Image dataset development for measuring construction equipment recognition performance," *Automation in Construction*, vol. 48, pp. 1–10, 2014.

[21] M. M. Soltani, Z. Zhu, and A. Hammad, "Automated annotation for visual recognition of construction resources using synthetic images," *Automation in Construction*, vol. 62, pp. 14–23, 2016.

[22] S. Bang, F. Baek, S. Park, W. Kim, and H. Kim, "Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques," *Automation in Construction*, vol. 115, Article ID 103198, 2020.

[23] J. Hwang, J. Kim, S. Chi, and J. O. Seo, "Development of training image database using web crawling for vision-based site monitoring," *Automation in Construction*, vol. 135, Article ID 104141, 2022.

[24] J. Kim and S. Chi, "A few-shot learning approach for database-free vision-based monitoring on construction sites," *Automation in Construction*, vol. 124, Article ID 103566, 2021.

[25] E. Karaaslan, U. Bagci, and F. N. Catbas, "Attention-guided analysis of infrastructure damage with semi-supervised deep learning," *Automation in Construction*, vol. 125, Article ID 103634, 2021.

[26] J. Guo, Q. Wang, and Y. Li, "Semi-supervised learning based on convolutional neural network and uncertainty filter for façade defects classification," *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 3, pp. 302–317, 2021.

[27] W. Wang and C. Su, "Semi-supervised semantic segmentation network for surface crack detection," *Automation in Construction*, vol. 128, Article ID 103786, 2021.

[28] G. Zhang, Y. Pan, and L. Zhang, "Semi-supervised learning with GAN for automatic defect detection from images," *Automation in Construction*, vol. 128, Article ID 103764, 2021.

[29] D. Haynes, S. Corns, and G. Kumar Venayagamoorthy, "An exponential moving average algorithm," in *Proceedings of the 2012 IEEE Congress on Evolutionary Computation*, Brisbane, Australia, June 2012.

[30] M. Xu, Z. Zhang, H. Han et al., "End-to-End semi-supervised object detection with Soft teacher," 2021, https://arxiv.org/abs/2106.09018.

[31] J. Dai, H. Qi, Y. Xiong et al., "Deformable convolutional networks," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 22–29, New York, NY, USA, October 2017.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[33] K. Chen, J. Wang, J. Pang et al., "MMDetection: open MMLab detection toolbox and benchmark," 2019, https://arxiv.org/abs/1906.07155.

[34] A. Paszke, S. Gross, F. Massa et al., "PyTorch: an imperative style, high-performance deep learning library," 2019, https://arxiv.org/abs/1912.01703.

[35] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Proceedings of the Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, September 2014.