Automation in Construction 142 (2022) 104499

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

SODA: A large-scale open site object detection dataset for deep learning in construction

Rui Duan [a], Hui Deng [a,b], Mao Tian [c], Yichuan Deng [a,b,*], Jiarui Lin [d]

[a] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
[b] State Key Laboratory of Subtropical Building Science, Guangzhou 510641, China
[c] Sonny Astani Department of Civil and Environmental Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, USA
[d] Department of Civil Engineering, Tsinghua University, Beijing 10084, China

ARTICLE INFO

Keywords:
Dataset
Object detection
Construction site
Deep learning
Computer vision

ABSTRACT

Comprehensive image datasets can benefit the construction industry in terms of serving as the basis for generating deep-learning-based object detection models and testing the performance of object detection algorithms, but building such datasets is complex and requires vast professional knowledge. This paper develops and publicly releases a new large-scale image dataset specifically collected and annotated for the construction site, called Site Object Detection Dataset (SODA), which contains 15 object classes categorized by the worker, material, machine, and layout. >20,000 images were collected from multiple construction sites in different situations, weather conditions, and construction phases, covering different angles and perspectives. Statistical analysis shows that the dataset is well developed in terms of diversity and volume. Further evaluation with two widely-adopted deep learning-based object detection algorithms also illustrates the feasibility of the dataset, achieving a maximum mAP of 81.47%. This research contributes a large-scale open image dataset for the construction industry and sets up a performance benchmark for further evaluation of relevant algorithms.

1. Introduction

The construction industry is still a labor-intensive industry, with most management and interventions of on-site activities relying on manual judgments [1], leading to the difficulty and inefficiency of on-site management. Although the emergence of high-resolution monitoring cameras makes remote and dynamic monitoring of the construction site possible, it still requires a lot of manual intervention [2]. The rapid development of computer vision technology makes it possible to automate tasks that cannot be completed by manpower, improving safety management and production efficiency [3]. The importance of cameras in construction management has become increasingly recognized, and practitioners have begun to adopt automated applications powered by computer vision [4]. For example, video surveillance can detect workers' unsafe behaviors and risks of construction activities [5], where computer vision technology is used to identify workers who do not wear personal protective equipment (PPE) [6–11]. The ability of computer vision technology in construction automation has thus attracted wide attention from academia and industry.

In recent years, deep learning object detection algorithm has developed rapidly, the detection speed and accuracy have been greatly improved. Under the appropriate application scenarios, the recognition accuracy can reach 98% or even higher. The computer vision technology based on deep learning has significant advantages over the traditional image process and recognition methods in terms of detection speed, algorithm robustness and feature extraction without manual design. [12]. Therefore, introducing the deep-learning based object detection in construction site management will be a new direction [13]. However, deep learning algorithms are data-hungry, which means the application of object detection on construction sites requires a customized image dataset in the construction field. The complexity and dynamic nature of construction activities brings challenges for image collection and annotation, which is the reason that well-annotated image datasets designed for the construction industry are hardly found in popular open database such as the ImageNet [14].

For the promotion of the research of object detection in the construction industry, it is necessary to build a large-scale image dataset containing specific objects from the construction site (i.e., worker, material, machine, layout). The existing construction site image datasets are relatively small in scale and have few categories, concentrating on

* Corresponding author at: School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China.
E-mail address: ctycdeng@scut.edu.cn (Y. Deng).

https://doi.org/10.1016/j.autcon.2022.104499
Received 19 February 2022; Received in revised form 17 July 2022; Accepted 23 July 2022
Available online 31 July 2022

people, PPE, and some machines. This is because: (1) The capturing of images of the construction site is more challenging than that of ordinary objects. Due to security concerns, the construction site is generally closed to the public. Moreover, the available online resources of construction site images are less common than daily objects. (2) It is difficult to obtain data from different angles on construction sites by using conventional monocular camera installed on-site, which is easy to cause overfitting of the detection model. (3) The site's environment is usually chaotic and dynamic, which brings difficulty for data collection and increases the cost of annotation. Moreover, professional knowledge is required to correctly annotate the objects in the images taken from the construction site.

A comprehensive image dataset for construction site will benefit the construction industry in terms of serving as the basis for generating deep-learning-based object detection models and testing the performance of object detection algorithms. Considering the professional knowledge required to build such a dataset, it remains a task to be resolved by people in the industry. The purpose of this research is thus to construct a comprehensive object detection image dataset for the construction site as shown in Fig. 1. The research processes include category selection, data acquisition, data cleaning, data annotation, dataset analysis, experimental analysis, and benchmark. A total of 19,846 images were collected, including 286,201 objects with 15 categories of classes. >20,000 images of different construction sites were obtained using various equipment, including monocular cameras, unmanned aerial vehicles (UAVs), and hook visualization equipment. 35 students majoring in Civil Engineering were recruited and trained to annotate the images, and then the statistics of data was analyzed. Finally, the dataset is tested by using the mainstream object detection algorithms, with well performance reached, which also provides the benchmark for selection of algorithms. The images and annotations were then published online. In the future, it will be upgraded regularly, increasing the variety of construction site objects to promote the research related to computer vision in construction.

The remaining of this paper is structured as follows: the second section of this paper introduces the related work of object detection and image datasets. The third section elaborates on the process of building the dataset. In the fourth section, the statistics of the dataset are presented. The fifth section introduces the experimental results of two mainstream one-stage object detection algorithms on the proposed dataset, which provides the benchmark for subsequent research.

2. Related work

2.1. Computer vision and deep learning in construction

The research into computer vision technologies for construction has aroused intense interest in both academia and industry, such as safety monitoring, productivity analysis, and personnel management. Koch et al. [15] reviewed the application of computer vision technology in defect detection and condition assessment of concrete and asphalt civil infrastructure. Xiang et al. [16] proposed an intelligent monitoring method based on deep learning for locating and identifying intrusion engineering vehicles, which can prevent damage to buildings. To prevent construction workers from falling from high, Fang et al. [17] developed an automatic method based on two convolutional neural networks (CNN) models to determine whether workers wear safety belts. Yang et al. [18] used cameras on tower crane to record video data and used MASK R-CNN to identify pixel coordinates of workers and dangerous areas. Chen et al. [19] used construction site surveillance videos to detect, track and identify excavator activities using three different convolutional neural networks. Deng et al. [20] proposed a method combining computer vision with building information modeling (BIM) to realize automatic progress monitoring of tile installation. This method can automatically and accurately measure the construction progress of the construction site. Fang et al. [21] realized the real-time positioning of construction-related entities by using deep learning algorithms combined with semantic information and prior knowledge. Zhang et al. [22] proposed an automatic identification method based on deep learning and ontology. This method can effectively identify the

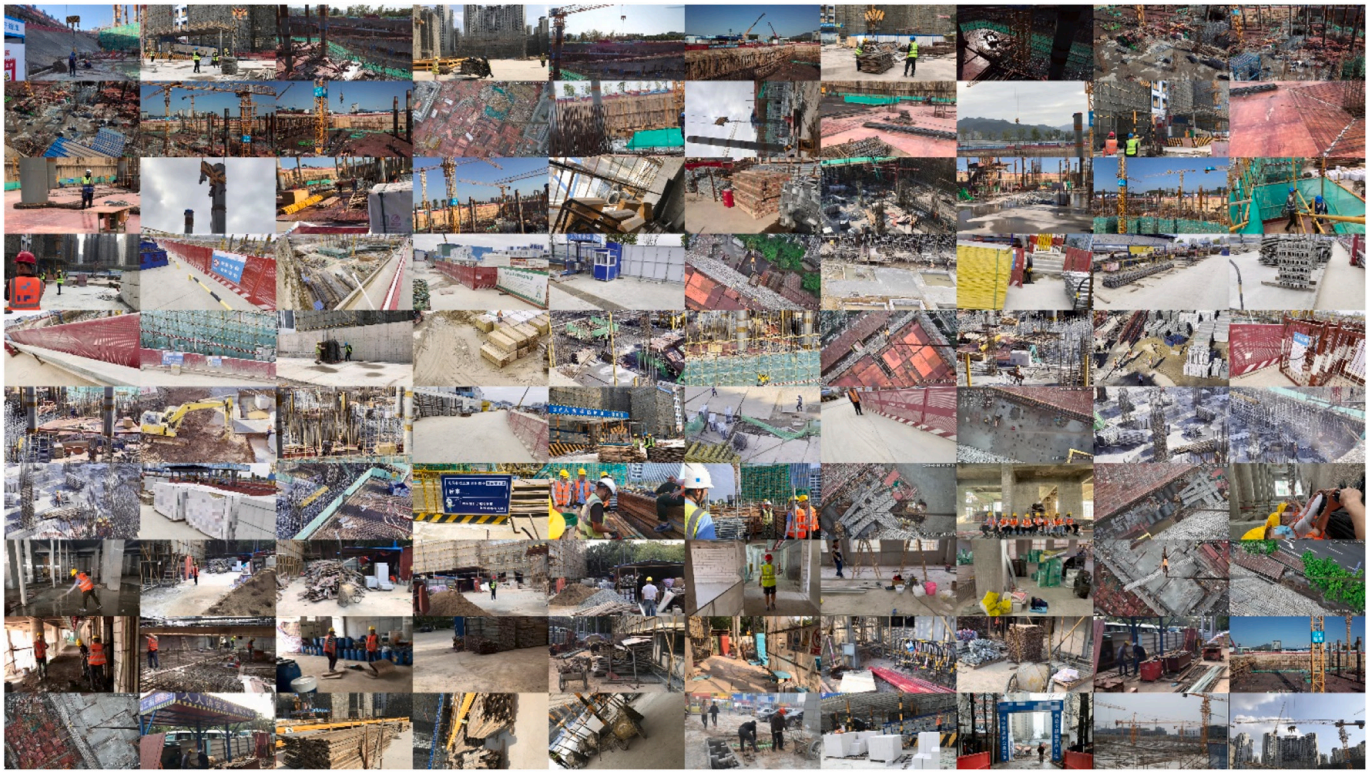

Fig. 1. Example images selected from SODA.

risks in the construction process and prevent the occurrence of construction accidents. Luo et al. [23] used computer vision and deep learning to automatically estimate the posture of different construction equipment in the video taken at the construction site. Pan et al. [24] proposed a novel framework named Video2entities, which combines visual data and common knowledge graph as prior information and uses zero-shot learning (ZSL) to realize the detection of unknown targets and effectively improve the self-learning ability of the object detection algorithm. For the reference of interested readers, this study collected some reviews [25–27] of computer vision in the construction industry. The literature review shows that there is a lack of sufficient scale of image datasets, which will hinder the continuous development of computer vision in the construction field. Moreover, the defect inspection, safety monitoring, and performance analysis of computer vision in the construction field require further investigation.

2.2. Datasets for deep learning based object detection

According to the domain coverage, the existing computer vision datasets can be divided into the general image datasets and the domain image datasets. General datasets include categories in daily life. In contrast, the domain datasets contain the categories of specific fields.

General datasets are mainly composed of natural categories, such as people, animals and vehicles. The Mnist dataset [28] contains 70,000 handwritten digital images, which is an entry-level dataset for deep learning. The PASCAL VOC dataset [29], which is widely used for testing deep learning algorithms, contains 11,000 images of 20 classes. Microsoft COCO [30] is a dataset containing 160,000 images with 91 categories, which has text descriptions of the category and location. Created by Professor Li Feifei's team, ImageNet [14] contains over 14 million images, covering >20,000 categories. Most of the research on image classification, location, and object detection have benefited from this dataset. The capacity and types of general datasets are also constantly increasing and improving, giving birth to a large number of excellent computer vision models (especially deep learning related) that promote the rapid development of computer vision technology.

However, the general image datasets introduced above often lack categories related to the construction field. In recent years, some image datasets have been developed for the construction industry. Tajdeen et al. [31] collected thousands of images of construction machines that cover five kinds of construction equipment (excavators, loaders, bulldozers, rollers, and backhoe diggers). Kim et al. [32] proposed a construction object detection method combining deep convolution network and transfer learning to accurately identify construction equipment. To evaluate the proposed method, a benchmark dataset called AIM was created, containing 2920 construction machine images. Kolar et al. [33] focused on detection of guardrail on construction site. By adding background images to the three-dimensional model of the guardrail, the authors created an enhanced dataset containing 6000 images. Li et al. [6] established and released a dataset containing 3261 images of safety helmet and used the SSD-MobileNet to detect unsafe operation on construction sites. An et al. [34] created the Moving Objects in Construction Site (MOCS) dataset by collecting >40,000 images from 174 construction sites and annotating 13 types of moving objects. They used pixel segmentation to precisely annotate objects and tested them on 15 different deep neural networks. Wang et al. [7] constructed a dedicated image dataset composed of 1330 images for Personal Protective Equipment (PPE) called Color Helmet and Vest (CHV). Xiao et al. [35] developed the Alberta Construction Image Dataset (ACID), an image dataset specially used to identify construction machinery, and manually collected and annotated 10,000 images of 10 kinds of construction machines. Four object detection algorithms, YOLO v3, Inception SSD, R-FCN-ResNet, and Faster-RCNN-ResNet, were used to train the dataset, and satisfactory mAP and average detection speed were obtained.

As mentioned above, computer vision related research in construction has aroused intense interest in both academia and industry. However, there are few related datasets available, and most of them only concentrated on workers, PPE, or machines. According to literature review, there are few studies on the identification and detection of construction materials [36,37] and layouts on the construction site. Therefore, it is necessary to broaden the coverage by developing image datasets containing worker, machine, material, and layout for the research on deep learning in the construction industry.

2.3. Deep learning based object detection algorithms

Object detection is an important application of deep learning, which focuses on determining the category and pixel location of the target [36]. As shown in Fig. 2, object detection algorithms are divided into two-stage object detection and one-stage object detection. The two-stage model is also called the region-based method because of its two-stage processing of images. It extracts a series of checkboxes and then uses convolutional neural networks for classification. Two-stage object detection has higher precision but relatively slow speed. Representative algorithms are R-CNN [38], Fast R-CNN [39], and Faster R-CNN [40]. YOLO [41] is a one-stage object detection framework, which directly obtains the prediction results from the image. The object detection can be achieved only by extracting features once, and the speed is faster than that of the two-stage algorithm. Object detection in the field of construction has a long period of exploration, and many outstanding research results have emerged in defect inspection [15], safety monitoring [26], and performance analysis [27].

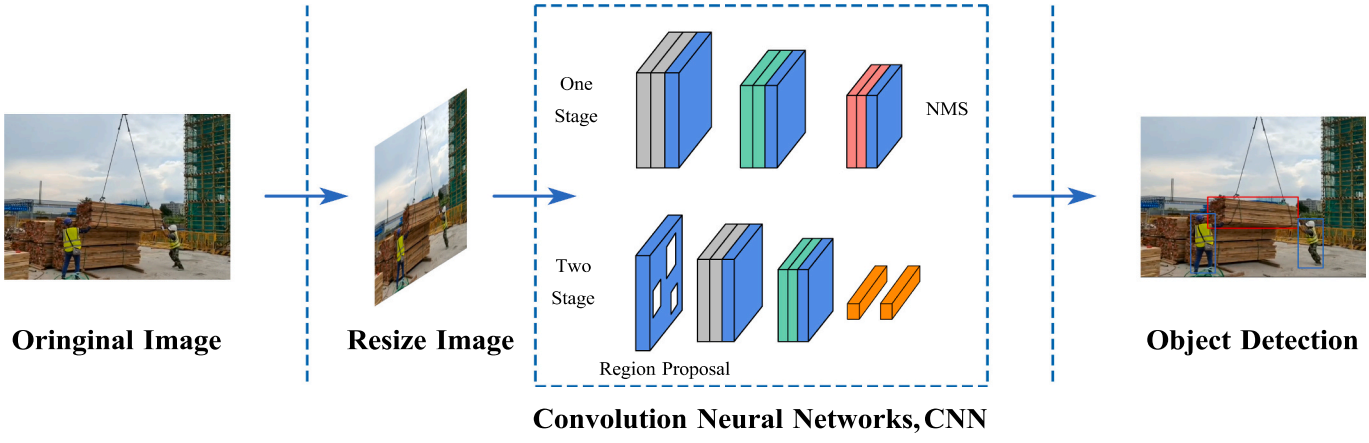

Fig. 2. Flowchart of the object detection algorithm.

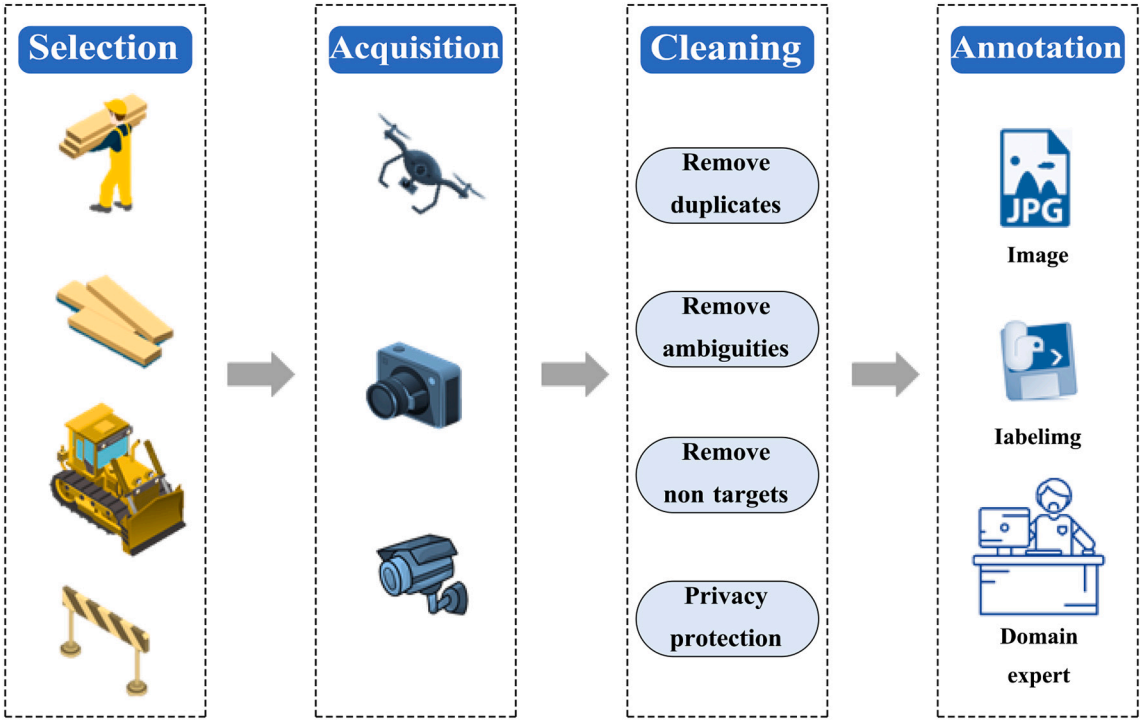

Fig. 3. Building process of SODA.

3. Methodology

This study constructs a new large-scale construction site image dataset, called Site Object Detection Dataset (SODA), which contains 15 classes of objects in four categories. >20,000 images using different equipment at different angles and times of the day were collected on several construction sites in the Greater Bay Area, China. Thirty-five students majoring in civil engineering were trained in image processing and annotation. Each student was responsible for about 600 images, which were then checked by several graduate students and experts in the construction industry. As Fig. 3 shows, the building process of SODA mainly includes four steps: object category selection, image data acquisition, image data cleaning, and image annotation.

3.1. Category selection

Zhou et al. [42] used the principle called 4M1E, which divides the risk factors into (hu)man, material, machine, method, and environment to assess and manage the risks during the construction process. According to the 4M1E principle, the common visible object in the construction site is classified into four categories: worker, material, machine, and layout. These four categories (person, material, machine, and layout) correspond to the physical entities in 4MIE (man, material, machine, and environment). This paper further expands 15 common object detection classes suitable for deep learning object detection from these four categories in the first edition of SODA. The worker category includes person, helmet, and reflective vest; The material category includes board, wood, rebar, brick, and scaffold; The machine category includes handcart, cutter, electric box, hopper, and hook. The layout category includes fence and slogan. Table 1 summarizes the definition of each class, and Fig. 4 shows the example of the corresponding categories. (See Table 2.)

3.2. Data acquisition

The construction of the conventional dataset often collects images from online resources [43]. Therefore, this paper initially applies web crawlers and other methods to obtain online data, but the collected images do not meet the requirements. Since the construction site is often messy, the objects are often mixed with unrelated objects, so images in SODA are all collected on the actual construction site. Data acquisition in construction sites mainly adopts three methods, UAVs, handheld monocular cameras, and construction site monitoring video (hook visualization). All collected images will eventually be converted into JPG format. Ten construction sites covering various construction stages have been visited, from the foundation pit stage to the decoration stage. As a result, a total of 21,863 images were collected.

In the process of data collection, there are some noteworthy problems in data acquisition that may benefit researchers who have the same interest. (1) Compared with worker and material categories, the capacity of machine and layout categories is relatively small, so it is necessary to select a full range of shots of limited targets from multiple angles. (2) Due to the confusion, visual blind spots, and occlusion of construction sites, it is challenging to collect positive samples using a single shooting method. Different shooting methods should be integrated. As shown in Fig. 5, the hook is shot from different angles using different devices to provide comprehensive data. Using UAVs and handheld shooting methods on construction sites are along with security risks, where careful operation is needed. (3) The data obtained from the high-altitude camera can record the panoramic view of the construction site and can also be used to photograph large-scale machines such as tower cranes. However, these video data are usually too vague to annotate for small objects.

3.3. Data cleaning

After the completion of data acquisition, it cannot be annotated directly but should be processed firstly to eliminate invalid data. Four objectives of data cleaning are proposed in this study: removing duplicates, removing ambiguities, removing non-targets, and corresponding privacy protection. The removal criteria and examples are shown in Fig. 6.

Table 1
Related description of selected objects in the construction site.

Construction site object	Description	example
person	Workers working in construction should wear PPE (safety helmet, reflective vest).	Fig. 3(1)
vest		Fig. 3(2)
helmet		Fig. 3(3)
board	Board for construction engineering is used to support the weight and lateral pressure of concrete mixture with plastic flow properties so that it can be solidified according to the design requirements.	Fig. 3(4)
wood	Wood is made into square strips according to the actual processing needs. It is generally used for decoration and door and window materials, template support, and roof truss materials in structural construction.	Fig. 3(5)
rebar	Rebar refers to the steel used for reinforced concrete and prestressed reinforced concrete. Its cross-section is circular, and sometimes it is a square with a round angle.	Fig. 3(6)
brick	Concrete brick is a lightweight porous, thermal insulation, good fire resis-tance, strong plasticity, and seismic capacity of new building materials.	Fig. 3(7)
scaffold	The scaffold is a working platform built to ensure the smooth progress of each construction process.	Fig. 3(8)
handcart	The handcart at the construction site is a two-wheeled, manual push and pull handling vehicle.	Fig. 3(9)
cutter	The cutter is the processing machine used in the material processing at the construction site. Commonly used machines are semi-automatic cutting machines and CNC cutting machines.	Fig. 3 (10)
electric box	All electrical equipment in the construction site must have its own special electric switch box, which is convenient for the switching operation of the circuit and the reasonable distribution of electric energy.	Fig. 3 (11)
hopper	Hopper (tower crane hopper, ash hopper, sand hopper, concrete hopper) are mainly used in building foundations, pouring concrete, piling, and high-rise building construction material transportation.	Fig. 3 (12)
hook	The hook of a tower crane is used to connect objects and ropes.	Fig. 3 (13)
fence	The fence is a protective facility to prevent accidental intrusion in the construction site of building engineering.	Fig. 3 (14)
slogan	The slogan is used to alert workers to civilized and safe construction, spread enterprise culture, etc.	Fig. 3 (15)

Fig. 4. Example of the corresponding classes.

Table 2
Categories and each class of the object label.

category	label				
Person	person	helmet	vest		
Material	board	wood	rebar	brick	scaffold
Machine	handcart	cutter	ebox	hopper	hook
Layout	fence	slogan			

3.3.1. Removing duplicates

The production of the dataset requires that each image must be significantly different from other images in terms of angle, position, and illumination. Therefore, some repetitive images should be manually removed. In the shooting process of the handheld monocular camera, many similar images may be collected accidentally. Moreover, some images of SODA are obtained by intercepting video frames. Although it is set to capture images from video every 30 frames, there are still relatively repetitive images that need to be manually removed due to slow progress in some construction activities.

3.3.2. Removing ambiguities

In the shooting process, some videos and images are blurred due to weather, lighting, human, and other reasons. These images are not only challenging to annotate but also affect the training effect of the deep learning model, which are supposed to be deleted before or during annotation.

3.3.3. Removing non-targets

In the process of on-site shooting, it is inevitable to take some images that meet the conditions of non-repetition and non-ambiguity but do not contain the target of our dataset. This kind of image is of no significance for the research and needs to be manually removed.

3.3.4. Privacy protection

Privacy is an essential issue for publishing public datasets. This study spends a lot of time making effort on privacy processing (200 working hours for all annotators, accounting for two-thirds of the data cleaning). Considering the engineering ethics, the proposed privacy protection method is divided into two parts, the company information (LOGO) processing and human characteristics processing. To avoid infringing on the relevant company's commercial secrets or triggering related property rights issues, all the company LOGOs have been blurred. Meanwhile, to avoid ethical issues, the face of the on-site worker is fuzzified accordingly. The above-specific operations are performed manually by the recruited students.

The data cleaning process took about 300 h, accounting for 16.16% of the total dataset construction time. In the removal process, 2017 invalid images were deleted. In data privacy processing, after a round of privacy processing and a joint inspection by authors and experts, students still have 1307 omissions (37 omissions per person).

3.4. Data annotation

The Visual Object Classes (VOC) format dataset is the standard dataset of the world-class Computer Vision Challenge (PASCAL VOC Challenge) [29], which is widely used in the field of computer vision and is an industry-recognized standard dataset format. In order to ensure the high quality of annotation, three standards are strictly followed when annotating an image: (1) The annotation box must frame the target and not intersect the target. (2) Reducing the frame selection of irrelevant background. (3) For similar targets with close mutual distance, avoid using a frame to select multiple targets. (4) When the target has partial occlusion or is impossible to annotate the whole size of it, it should be omitted. As shown in Fig. 7, the annotation of the blue box is accepted instead of the red box.

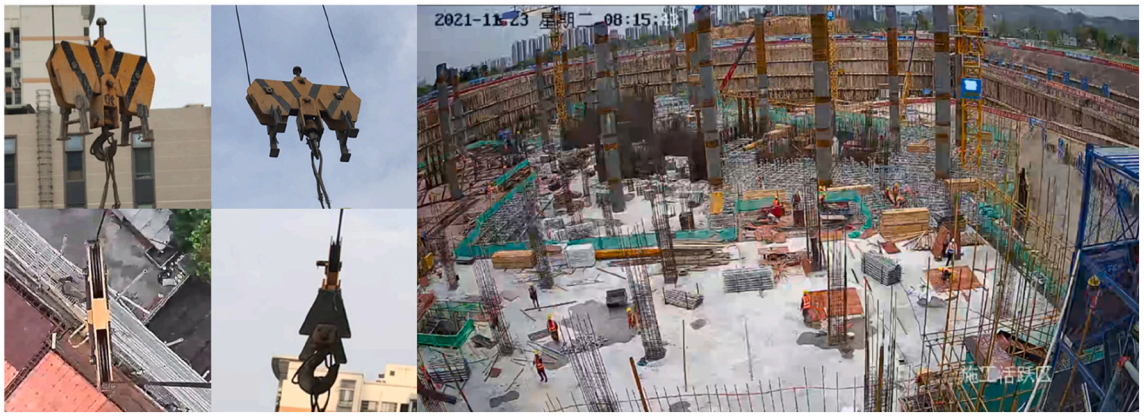

Fig. 5. Example of different angles using different devices.

Fig. 6. Example of images that need to be removed and processed.

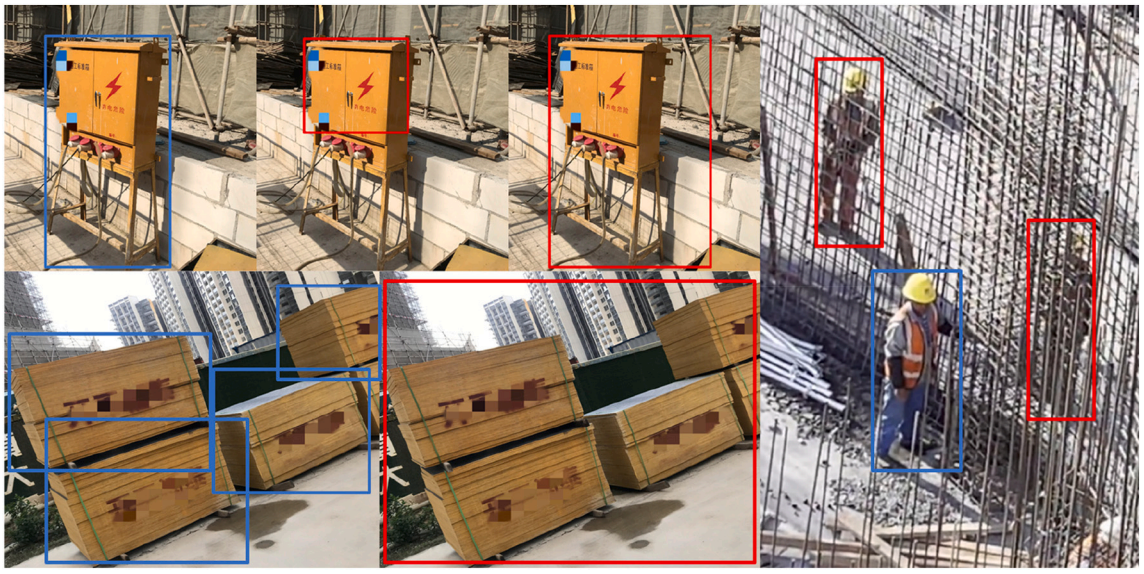

Fig. 7. Example of annotation standard.

Fig. 8. Interface of labelImg.

As shown in the following Fig. 8, the software labelImg is used by students to annotate 15 categories of images. LabelImg is an open-source image annotation tool written in Python [44]. The labeled image data are saved as XML files in PASCAL VOC format. The XML format is shown in Fig. 9, which contains information such as storage path, image name, and coordinate. 35 students are trained as annotators after receiving tutorials that explained how to annotate and frame target. To ensure the accuracy of the annotation, the label completed by the students are regularly examined by a graduate student and the authors.

After obtaining the annotation documents, a round of inspection with another expert is conducted. Although students have been well trained, there are still some unexpected errors. 1) The spelling error of label words, 2) The plural error of the label, and 3) Unknown labels. The label errors have been introduced as follows.

As Table 3 shows, all the word spelling and plural errors were listed. In addition, there is another kind of error that 48 'w' and 2 'wwww' labels appear. The reason for the 'w' error is that the shortcut key of the tagging software is w, and the annotator typed the label name incorrectly. All of these errors have been corrected in the officially released dataset.

After completing the above four processes, the annotations of 19,846 images and the corresponding XML files are obtained. A statistic of the resources (time, manpower, tools, sites) used in the whole process is presented. First of all, in terms of manpower and time, three graduate students took pictures using various equipment on the construction sites in the data acquisition process, which took about 24 working hours. Data cleaning and data annotation process are carried out by 35 undergraduates, and all students' data cleaning and annotation time are counted. As Fig. 10 shows, data cleaning costs 300 working hours (account for 16.16%), and data annotation costs 1246 working hours (account for 67.13%). During the inspection process, two graduate students and two construction industry experts conducted two rounds of sampling inspection, taking a total of 286 h (account for 15.41%). UAVs, surveillance cameras, mobile phones, single-lens reflex cameras (SLRs),

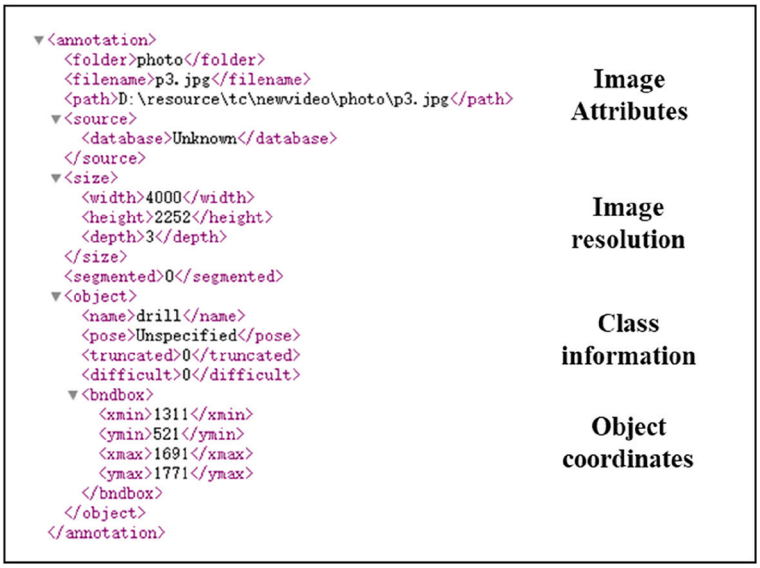

Fig. 9. XML format files.

Table 3
Example of the spelling mistake.

Sample							
Right	scaffold	hopper	helmet	board	person	vest	brick
Error	seaffold	hooper	helemt	borad/boadrd	preson	vests	bricks
Number	14	8	8	4/1	1	1	345

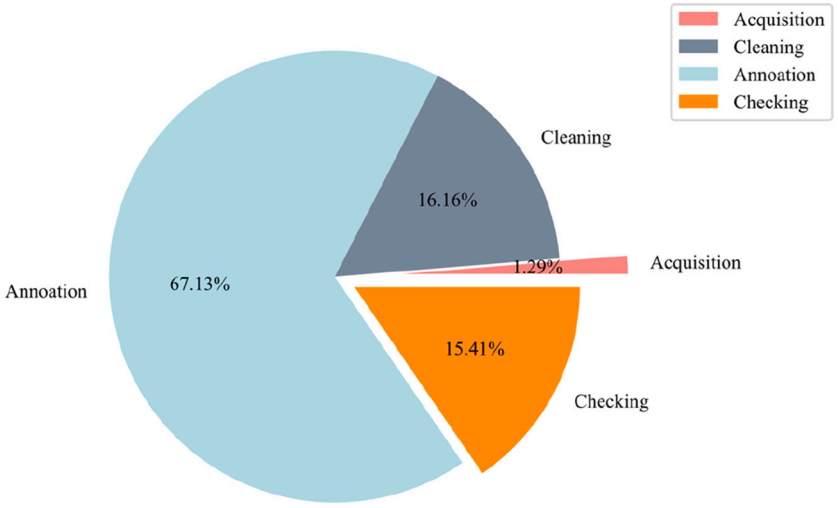

Fig. 10. Time distribution for each process of building the dataset.

and hook visualization equipment are utilized as image capturing tools. A total of 10 construction sites in Guangzhou, Shenzhen, and Dongguan city in the Greater Bay Area, China have been visited.

4. Statistics of the dataset

In this part, the proposed dataset is analyzed and compared with the current open object detection image dataset in the construction industry. The proposed SODA contains a total of 19,846 images, and the size of dominated images in the dataset is 1920 * 1080, accounting for 86%. A total of 286,201 objects are annotated, and the object distribution is shown in Fig. 11 and Fig. 12, providing a quantitative understanding of the dataset. Among them, the number of labels of worker is the largest, and the labels of the machine and layout are less, which is more consistent with the situation of the construction sites. Each class has over 1000 targets, and each category has >20,000 objects.

Some object detection models generally specify the length-width ratio and range of the detection object, there is a statistical sample length-width clustering requirement for the dataset. The k-means clustering algorithm [45] is thus used to analyze the sample data by visualizing the length-width ratio and range (shown in Fig. 13).

To reflect the characteristics of the original bounding box data distribution of the SODA and compare the data distribution characteristics

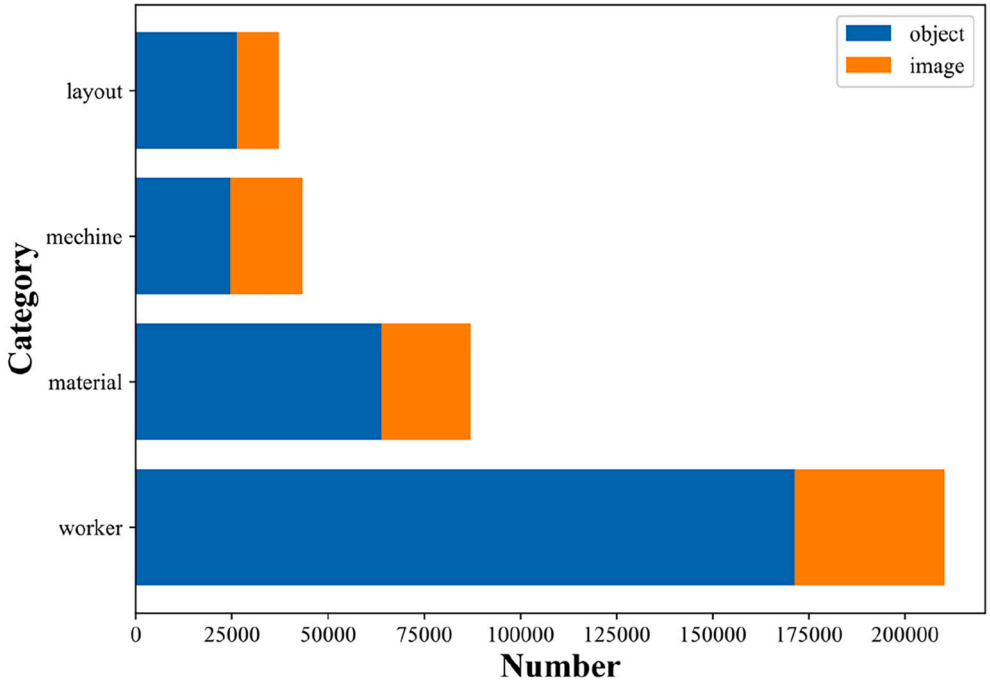

Fig. 11. Number of objects and images in each category.

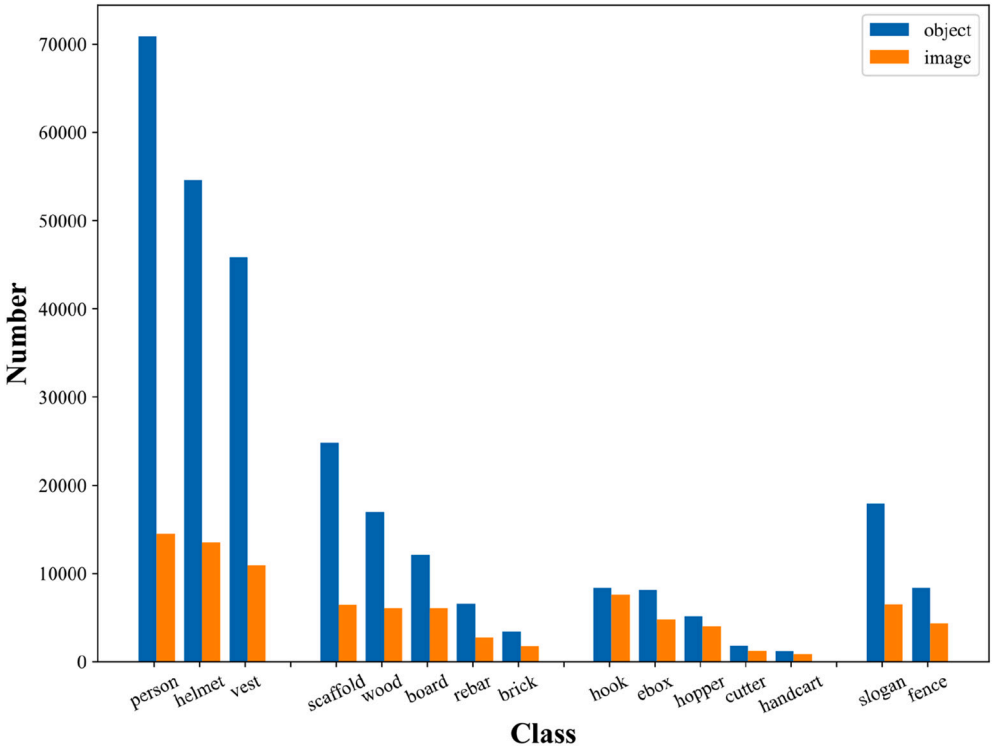

Fig. 12. Number of objects and images in each class.

of each class, the Box-plot of the length-width distribution of each class of bounding boxes without the outliers is shown in Fig. 14 and Fig. 15.

It is also worth noting that the images are taken from different angles from hand-held camera short-range shooting perspective, hand-held camera long-range shooting perspective, UAV perspective, tower crane hook visual system perspective, to achieve full coverage. The distribution is shown in Fig. 16.

Table 4 shows a comparison of the proposed dataset with the current popular open object detection dataset in the construction industry. The result shows that the proposed SODA not only contains the largest number of objects and categories but also is the first time to realize the full coverage of the four categories of worker, material, machine, and layout. Moreover, this study firstly uses the image data gathered from hook visualization equipment, which is also a neglected part of previous studies.

In summary, the following points distinguish SODA from other

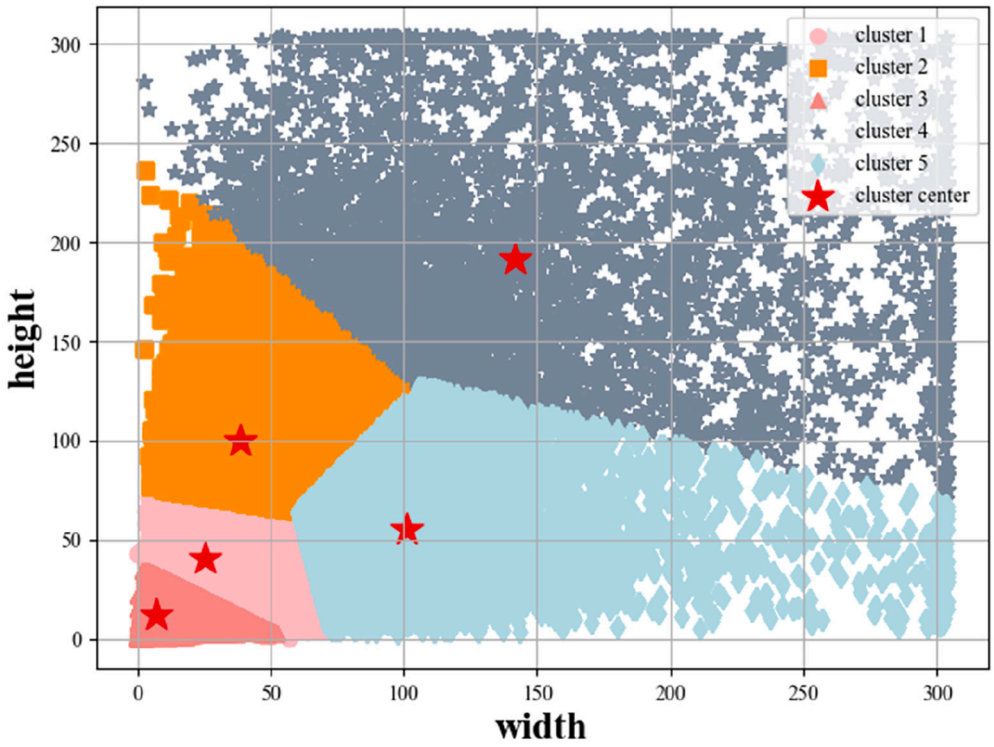

Fig. 13. Anchors statistical results of the k-means clustering algorithm.

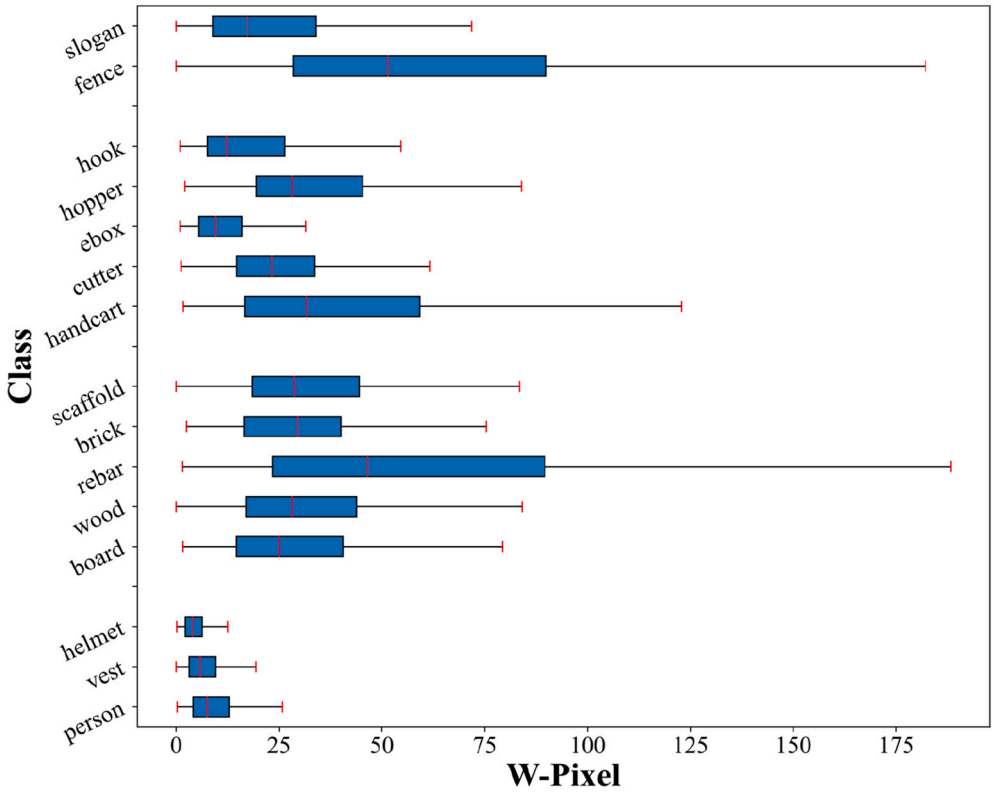

Fig. 14. Box-plot of the width distribution of each class.

published datasets. (1) SODA not only contains the largest number of objects and categories but also is the first time to realize the full coverage of the four categories of worker, material, machine, and layout. SODA initially avoids selecting most categories covered by existing mature published datasets and is committed to complementing existing datasets rather than completely replacing them. (2) SODA also achieves authenticity and diversity by collecting images using a variety of equipment from different angles at the construction sites. Although

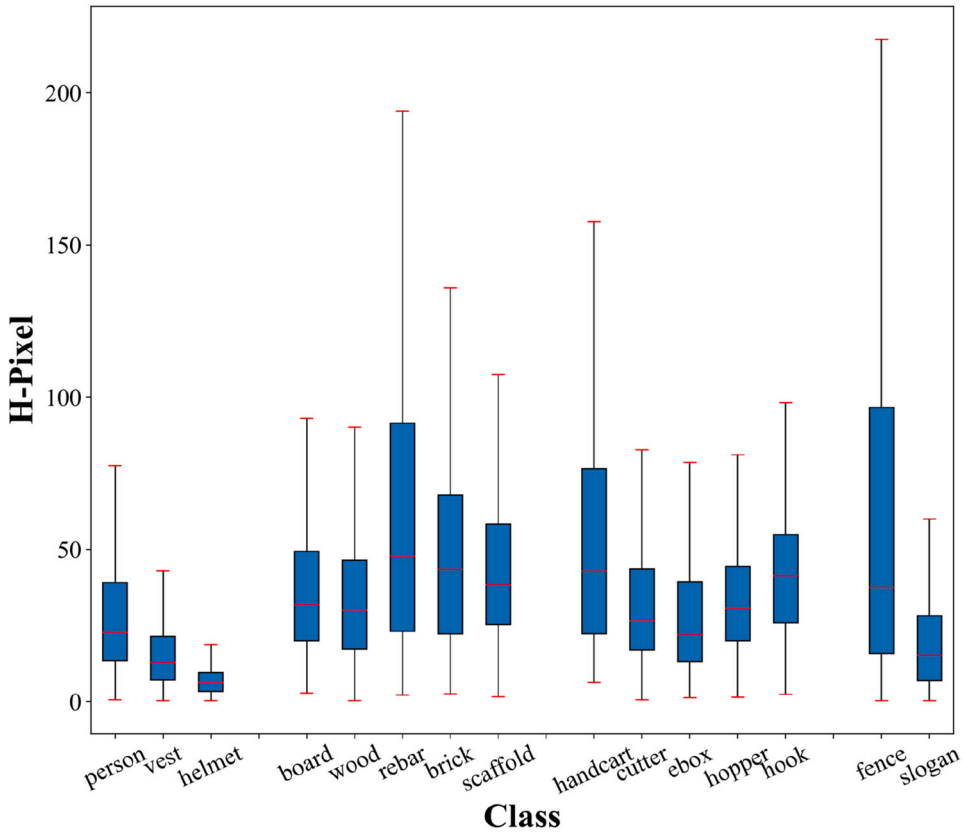

Fig. 15. Box-plot of the length distribution of each class.

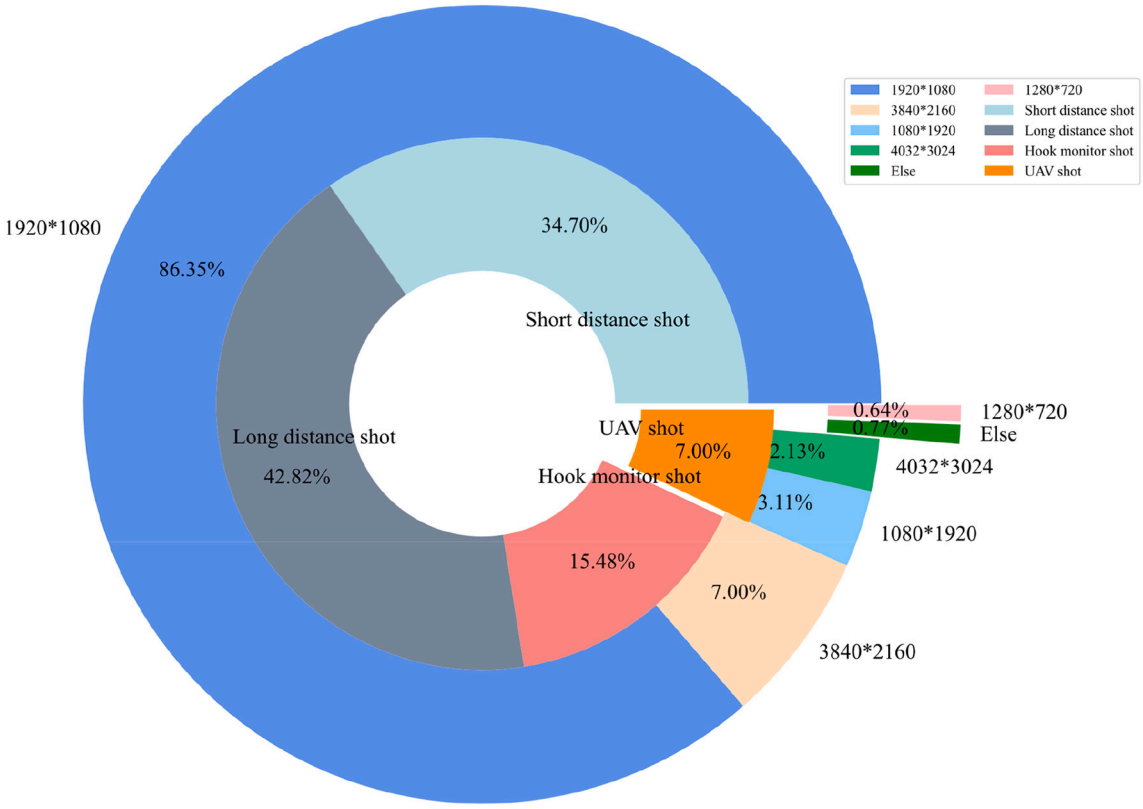

Fig. 16. Distribution of image resolution and shooting perspective.

Table 4
Comparison of SODA dataset with other datasets in the construction industry.

Dataset	Image	Object	Category				Class	Image size	Year
			Person	Material	Machine	Layout			
SODA	**19,846**	**286,201**	**3**	**5**	**5**	**2**	**15**	**1920*1080 and higher**	**2022**
CHV	1330	9209	6	–	–	–	6	608*608	2021
MOCS	41,668	222,861	1	–	12	–	13	1200*–	2020
ACID	10,000	15,767	–	–	10	–	10	>608*608	2020
AIM	2920	2920	–	–	5	–	5	500*-	2017
Tajeen's	2000	–	–	–	5	–	5	–	2014

some datasets have achieved multi-view image data collection, this study firstly uses the image data gathered from hook visualization equipment, which is also a neglected part of previous studies. (3) SODA elaborates on the construction process of the dataset about the problems encountered and the solutions, which provides convenience for subsequent research in establishing such an image dataset and the training of deep learning models. (4) SODA is continuously updated, with new categories and objects enriched in subsequent updates. Moreover, capability expansion including instance segmentation and image caption will be uploaded in the near future.

5. Experiments on the dataset

This study also aims to provide benchmarks for researchers to select appropriate algorithms for relevant studies. The authors also welcome other researchers to use different deep learning algorithms to verify SODA and compare the performances with the results of this study. The experiment proves the feasibility of using the deep learning object detection algorithm to detect the construction-related worker, material, machine, and layout in images and videos.

5.1. Algorithm selection

Xiao et al. [34] concluded that the one-stage detection algorithm is more suitable for the identification of construction engineering than the two-stage detection algorithm. Although the two-stage algorithm is able to reach a higher accuracy than the single-stage algorithm, the difference is insignificant, and the one-stage algorithm is much faster. It can also avoid background errors and learn the generalization characteristics of objects, which is more suitable for complex scene recognition on construction sites. Considering the YOLO is developed with 'flexible' properties and its popularity, YOLO v3 and YOLO v4 are selected to show their performance on SODA.

As a one-stage detector, YOLO does not generate the proposal region but directly divides the image into S × S grid cells, and each grid detects the object falling into the center. The YOLO v3 network is shown in Fig. 17, which is mainly composed of the backbone (Darknet53), Neck (FPN), and Head (YOLO Head). The residual network in Darknet53 is continuously used for convolution to extract image features. Each convolution layer is a unique convolution structure, where l2 regularizations are performed at each convolution, after which BatchNormalization and LeakyReLU are performed. The neck is responsible for

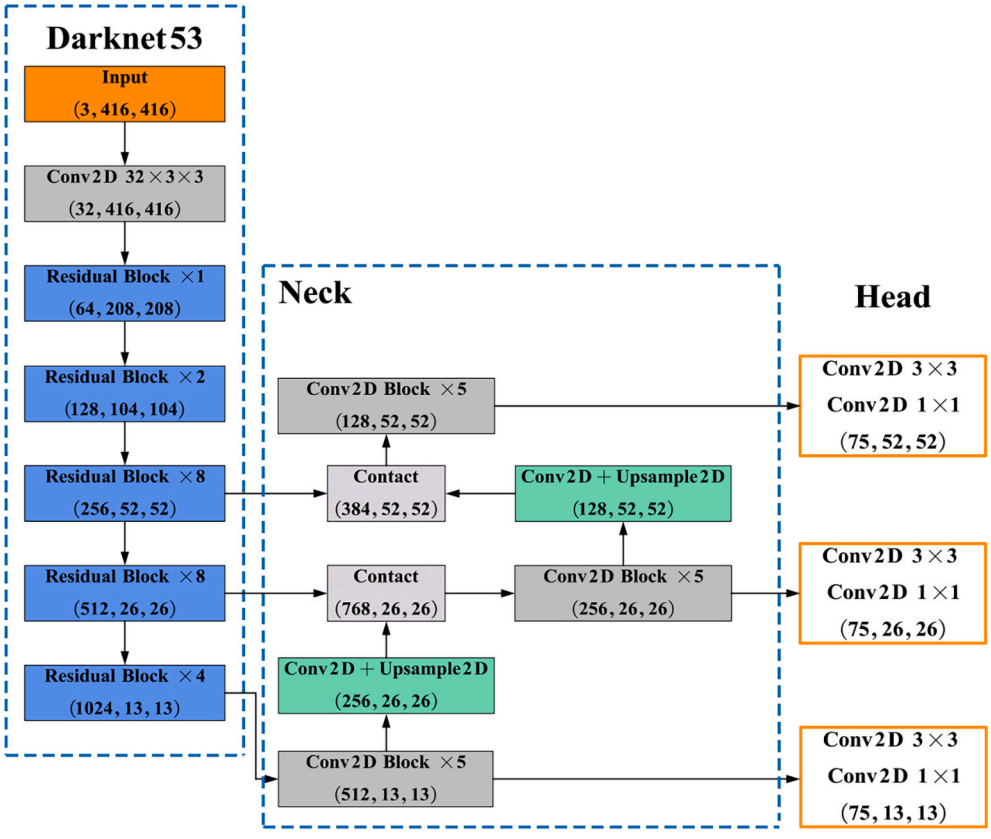

Fig. 17. Network structure of YOLO v3.

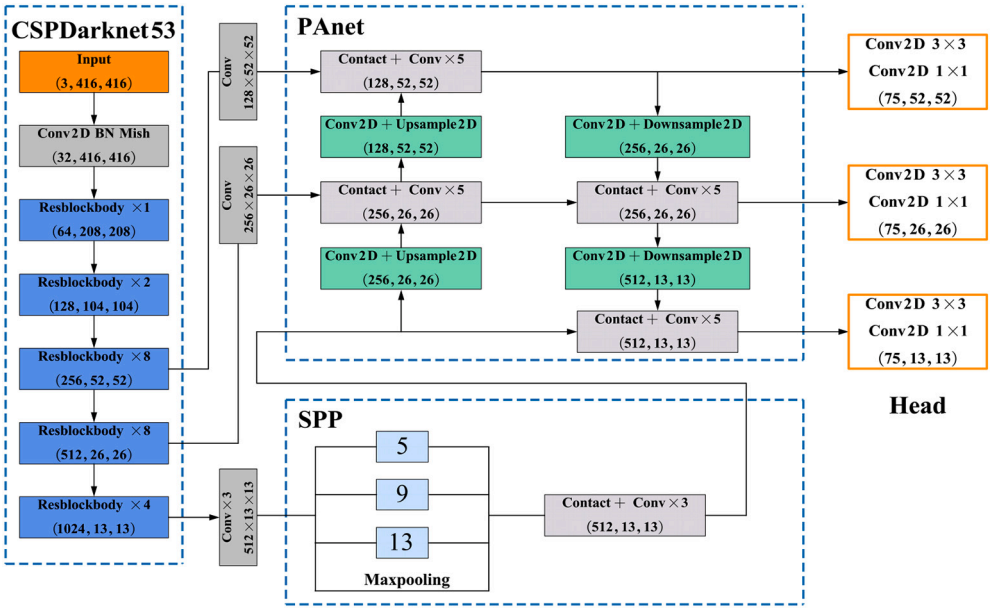

Fig. 18. Network structure of YOLO v4.

constructing the Feature Pyramid Network (FPN), using image features for up-sampling and feature fusion. The processed enhanced features are output to three YOLO Heads for the prediction the results. Three YOLO Heads are designed for large, medium, and small-scale object prediction. YOLO Head uses 3 × 3 convolution to integrate features and then uses 1 × 1 convolution to adjust the number of output channels to predict the results. YOLO v4 can be seen as an improved version of YOLO v3. YOLO v4 has added a series of tips based on the YOLO v3 network, making the training and prediction process better. The YOLO v4 network is shown in Fig. 18. In terms of backbone, DarkNet53 is replaced by CSPDarkNet53. The activation function in the convolution layer is replaced by the Mish function from the LeakyRelu in YOLO v3, and normal Resblock is changed into a CSPnet structure. SPP and PAN network structures are used to form newly FPN on Neck. The SPP structure is added to the last convolution layer of the backbone, and the parallel maximum pooling of the convolution kernels of 13 × 13, 9 × 9, 5 × 5, and 1 × 1 is carried out so that the network can increase the receptive field and then separate the salient features. The PANet structure i adds down-sampling feature extraction from top to bottom on the basis of the traditional up-sampling feature extraction steps of FPN. The PANet structure strengthens the feature extraction by combining up-sampling and down-sampling.

5.2. Evaluation metrics

In this study, the mean average precision (mAP) [46] is applied to evaluate the performance of the model. The use of mAP can eliminate the limitation of using a single evaluation index by combining Precision and Recall. Intersection over Union (IOU) is a basic indicator for evaluating performance of object detection algorithm, which is used to measure the coincidence degree between the detection box and the ground truth box. The calculation of IOU is shown in Fig. 19. In the molecular part, the value is the overlap area between the detection box and the ground truth box. In the denominator part, the value is the total area occupied by the detection box and the ground truth box. After obtaining the IOU index, the TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) need to be calculated. Respectively, T or F represents whether the sample is correctly predicted. P or N represents whether the sample is predicted to be positive or negative. TP represents the positive sample is predicted correctly, TN represents the negative sample is predicted correctly, FP represents the negative sample is predicted wrongly, and FN represents the positive sample is predicted wrongly. After obtaining the four indicators, the Precision and Recall can be calculated (shown in formulas (1) and (2)).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \int_0^1 p(r)dr \tag{3}$$

$$mAP = \frac{1}{N}\sum_{n}^{i=1} AP \tag{4}$$

Average precision (AP) considers the combination of different Precision and Recall points, it can be calculated by measuring the area under the curve after integration (shown in formulas (3)). By calculating the AP of all classes, the Mean Average Precision (mAP) of the overall performance of the model can be estimated when detecting (shown in formulas (4)).

5.3. Result

The experiment is performed on a computer with the following configuration: NVIDIA Geforce RTX 2060, Intel (R) Core (TM) i7-10750H CPU @ 2.60GHz 2.60GHz 16.0GB. All algorithms are operated using the PyTorch framework. SODA is randomly divided into a training set (17,861 images) and a test set (1985 images) according to

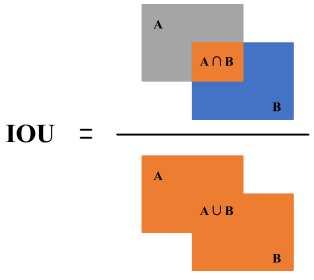

Fig. 19. Calculation process of the IOU.

9:1. While training, 100 epochs are adopted for training, which are divided into two stages: the freezing and thawing stages. The first 50 epochs freeze the main parameters of the model, which will increase the learning rate and help the model training jump out of the local optimal solution. In the second 50 epochs, by thawing model backbone parameters and reducing the learning rate, model backbone parameters are greatly changed in this stage.

The following Fig. 20 is the loss curve (training and verification loss) of two deep learning algorithms on SODA, in which the red line is the training loss and the blue line is the verification loss. It can be seen that the two deep learning algorithms are well fitted to the SODA dataset. Moreover, it should be noted that the loss values of different algorithms are not necessarily comparable because the test algorithm implements different loss functions. The training loss and verification loss keep decreasing and finally reach a stable value with the increase of epoch, which shows that the model has been convergent. Meanwhile, they decrease after the rise of the 50th thaw epoch, and the verification loss is higher than the training loss. The learning curve shows the robustness and universality of the dataset.

After the training, the detector is evaluated on the test dataset. The overall performance index analysis results of the test dataset in YOLO v3 and YOLO v4 are shown in Table 5 and Figs. 21-22. The results show well detection performance with 71.22% mAP on YOLO v3 and 81.47% mAP on YOLO v4. Due to the change in network structure, each category has a different performance, but the training results of the two models show that the mAP of the material is higher, and the results of the workers are lower. The highest AP is the hook (92.81% in YOLO v3) and the hopper (95.18% in YOLO v4). The lowest AP is the fence (50.99% in YOLO v3) and the helmet (55.62% in YOLO v4). In terms of detection speed, YOLO v4 performed better at 31.94 FPS than the 25.06 FPS of YOLO v3. Table 6 shows the comparison of the object detection model trained in SODA with the other image datasets in the construction industry. In the case of the same detector and backbone, the mAP of SODA reached 81.47, lower than the highest ACID score of 87. As shown in Fig. 14-15, the SODA dataset contains numerous objects with different pixel sizes. Therefore, it is reasonable to believe that the detection process of SODA is more complex than detecting a single category of objects. The proposed SODA has the highest recognition speed of 31.94 fps, the fastest in the current construction industry dataset. Moreover, SODA and CHV made a similar conclusion that the performance (mAP, speed) of YOLO v4 is better and more comprehensive than YOLO v3.

Fig. 23 shows the examples of detection results. The images are selected from the validation set, which is never seen in the training process. The four images in the first row are taken by cameras and mobile phones. The second row of images is captured by hook visualization devices, UAVs, and surveillance. From the comparisons in Fig. 23, it is found that the identified objects are correctly classified, and there is less error detection in both YOLO v3 and YOLO v4. The detection effects of YOLO v3 and YOLO v4 are not significantly different with the simple background and fewer objects. In the comparison of Fig. 23(2), YOLO v4 has a better detection effect in some remote and small objects and partially occluded objects. And in large-scale images such as Fig. 23(7), YOLO v4 can identify more small objects. In the complex and crowded scenes, YOLO v3 has some missing detection compared with YOLO v4. In summary, YOLO v3 and YOLO v4 have similar recognition performance in simple scenes, but YOLO v4 has higher recognition ability in complex scenes. In the comparison of Fig. 23(9), neither YOLO v3 nor YOLO v4 identified people sheltered by steel cages, but that's reasonable because this study did not annotate obscured workers in the annotation criteria. The reason why YOLO v4 performs better than YOLO v3 is that YOLO v4 makes some improvements on the basis of the YOLO v3 network, which broadens the receptive field and strengthens the feature extraction so that its detection effect in small remote targets and partial occlusion is better. Some other results also met the expectations. For example, in the category of worker, the accuracy ranking is person, vest, and helmet, which is consistent with the size of the target. And the highest AP scores of YOLO v3 and YOLO v4 all appear in the machine category which is less challenging in detection. As for the small target such as helmet, the AP of YOLO v4 is nearly 3% higher than YOLO v3. Choosing a more advanced backbone feature extraction network may get better detection results in practical engineering. Owing to space limitations and the theme of publishing the dataset, further analysis of the algorithm is not discussed in this paper.

6. Conclusions

In this research, a dataset called SODA is developed for object detection on the construction site. SODA is an image dataset in VOC format containing 19,846 images and annotation information of 286,201 objects. SODA has 15 categories, which cover most of the common objects on the construction site. SODA is tested on two mainstream one-stage object detection algorithms, after which a series of training results and benchmarks are obtained.

In summary, the contributions of SODA are as follows: (1) SODA dataset is specifically built for the construction industry with high-quality image data. All the images from SODA are collected from actual construction sites at different construction stages, angles, and time. The training on the proposed dataset will better improve the practicability and effect of object detection in the field of the construction industry than general datasets. In addition, SODA has more images and 4M1E categories than the existing datasets in the construction industry, which broadens the detection range of common construction workers, PPE, and machines. (2) We share relevant experience in building such a dataset, according to which the researchers can increase the number of categories and images to update the proposed dataset iteratively. A detailed data analysis of the image data is carried out, including the statistical analysis of the number of images from different categories and the number of identified objects, the statistical analysis of image resolution size and image shooting perspective, the clustering analysis of all bounding boxes, and the box-plot analysis of pixel length and width of each class. A comparison of the SODA dataset with other open datasets in the construction industry is presented and analyzed. Researchers can not only choose some specific categories of images for further in-depth research but also expand data on the basis of SODA. (3) To explore the applicability of the deep learning-based object detection on the construction site using SODA, two one-stage object detection algorithms, YOLO v3 and YOLO v4 have been selected to conduct the benchmark tests. Experimental results show that both YOLO v3 and YOLO v4 spend less time to train the model and reach a relatively high accuracy rate, which prove their well performance in terms of speed and accuracy. The experiment conducted in this study is also the first case of detecting multiple 4M1E categories in a single object detection model. This work can provide insights for research on integrating multiple tasks regarding construction site safety monitoring, construction progress analysis, emergency response, and civilized construction.

This study is the first open dataset to cover the largest scale of object categories of the construction site, and elaborates on the construction process of the dataset about the problems encountered and the solutions. Meanwhile, it also establishes the benchmark to provide insights for other researchers to select detectors. Notably, SODA is continuously updated, with new categories and objects enriched in subsequent updates. Moreover, capability expansion including instance segmentation and image caption will be uploaded in the near future.

Several limitations of research should be mentioned. (1) It is recommended to add more categories and more data to enrich the SODA dataset. Although the category and capacity of SODA are larger than other datasets in the construction industry, they are still relatively smaller than other datasets in the general computer vision community. (2) Further research on annotation tasks is needed. In this study, the annotation of the SODA dataset is object-level, and only the boundary frame of the object is annotated rather than pixel-level. This indicates that SODA can only be trained on deep learning object detection

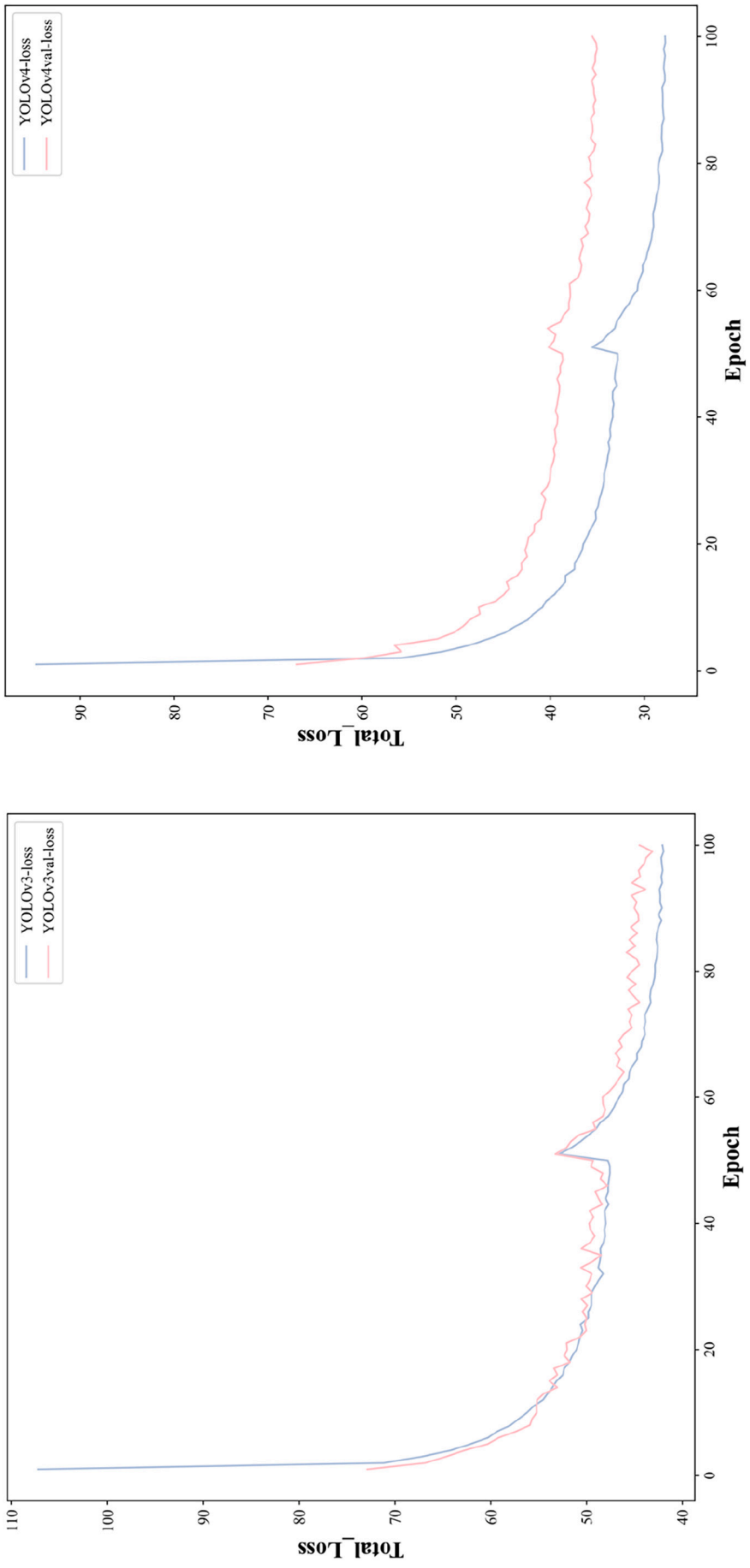

Fig. 20. Training loss and validation loss of YOLO (v3, v4).

Table 5
Performance comparison of SODA dataset in YOLO v3 and YOLO v4*.

Algorithm	Worker (%)			Material (%)					Machine (%)					Layout (%)		mAP
	person	helmet	vest	board	wood	rebar	brick	scaffold	handcart	cutter	ebox	hopper	hook	fence	slogan	
YOLO v3	73.71	52.88	57.67	83.34	83.32	54.73	76.02	86.43	74.34	64.61	69.05	89.42	**91.76**	**50.99**	60.07	71.22
YOLO v4	72.42	**55.62**	66.09	92.79	91.41	73.21	90.89	92.92	89.45	87.96	76.11	**95.18**	92.81	74.11	71.08	81.47

* The best performance and the worst performance of the model are roughened with red and green in the chart.

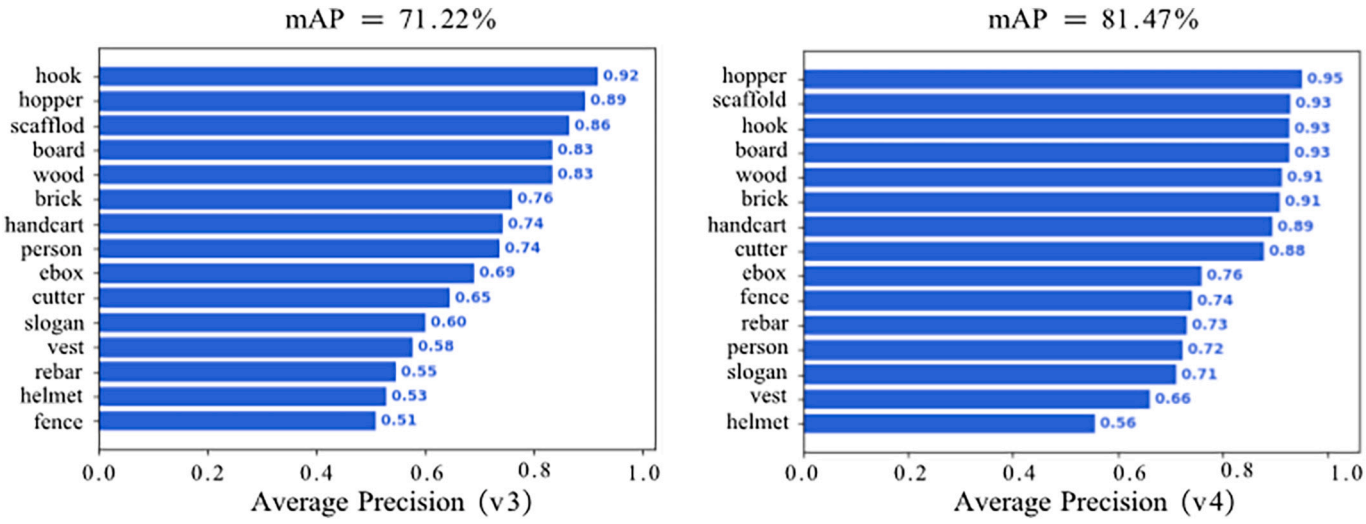

Fig. 21. mAP of YOLO (v3, v4).

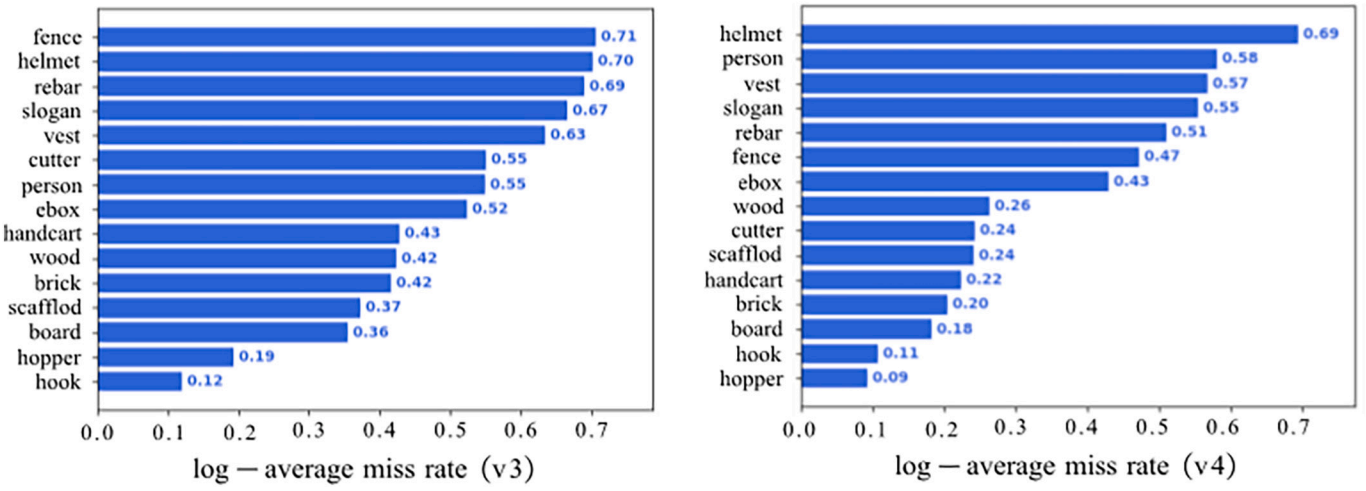

Fig. 22. Log-average miss rate of YOLO (v3, v4).

Table 6
Comparison of object detection results by different algorithms between SODA and others.

Dataset	Detector	Backbone	Input Size	mAP	AP (highest)	AP (lowest)	Speed(fps)
SODA	YOLOv3	Darknet53	416 × 416	71.22	91.76 (hook)	50.99 (fence)	25.06
	YOLOv4	CSPDarknet53		81.47	95.18 (hopper)	55.62 (helmet)	31.94
MOCS	YOLOv3	Darknet53	608 × 608	33	56.809 (worker)	16.782 (hook)	30
ACID	YOLOv3	Darknet53	608 × 608	87.8	94.9 (truck)	62.0 (tower crane)	26.3
CHV	YOLOv3	Darknet53	416 × 416	82.65	88.19 (white helmet)	77.51 (blue helmet)	27.15
	YOLOv4	CSPDarknet53	(608 × 608)	84.16	90.57 (white helmet)	78.14 (blue helmet)	30.18
AIM	R-FCN	Resnet50	400–700	96.33	99.20 (mixer truck)	94.43 (dump truck)	14
Tajeen's	Torralba et al. [46]		375 × 250	96.8	98.0 (dozer)	90.0 (excavator)	18.6 s/per picture
	Felzenszwalb et al. [47]				99.0 (roller)	95.0 (excavator)	1.71 s/per picture

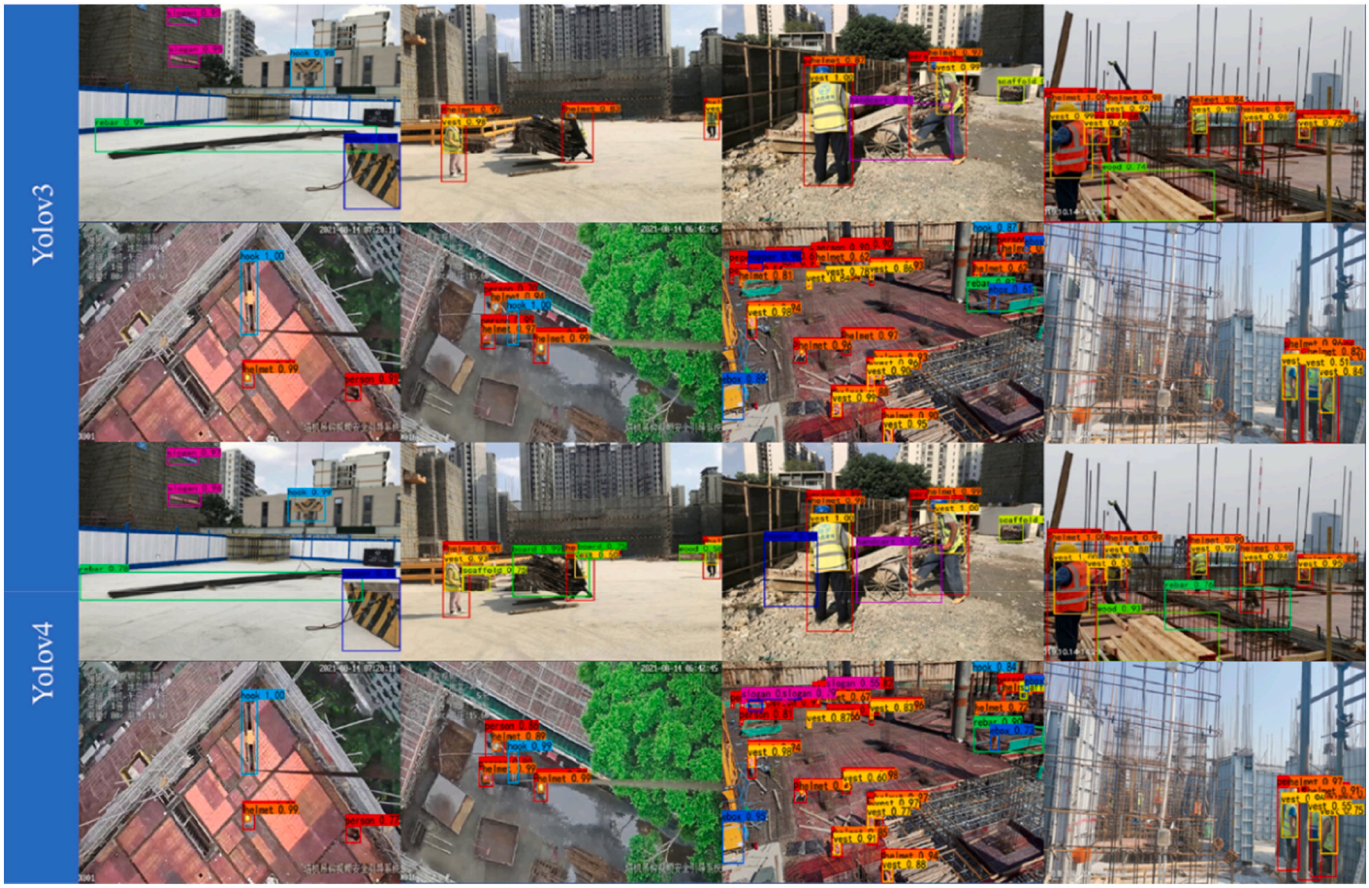

Fig. 23. Identification result of YOLO v3 and YOLO v4.

algorithms instead of the object segmentation algorithms. Regarding this, SODA will not only continue to increase new categories and capacity in subsequent updates but also make some update iterations regarding instance segmentation and image caption. More annotation methods such as crowdsourcing [49] and automatic annotation should also be explored. (3) Images of SODA are collected in the Greater Bay Area of China, and model trained on SODA may not be directly applicable to other countries. The application and feasibility of the SODA should be further investigated on construction sites of different countries and regions considering the variety of construction practices. Several previous published datasets [32,34] have also encountered this problem and they all agree that transfer learning could be used to solve the problem of application across different countries and regions. What's more, cooperation with more scholars and companies in different countries and regions is needed.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to acknowledge the support by Guangdong Science Foundation (Grant No. 2022A1515010174); the support by the State Key Lab of Subtropical Building Science, South China University of Technology (No. 2022ZB19); the support by the Guangzhou Science and Technology Program (No. 202201010338); and support by the National Natural Science Foundation of China (No. 51908323, No. 72091512).

The author would also like to pay special tribute to students who contribute to the data cleaning and annotation process of SODA at the South China University of Technology. Their names are Junxiong Zhang, Fengning Chen, Hongfeng Chen, Jianhe Chen, Jingjun Chen, Zhentao Chen, Yina E, Jie Fan, Xingyu Gao, Jiaxuan He, Jiayi Huang, Jingyuan Huang, Ying Huang, Yuefan Huang, Jiaxi Jiang, Liki Lei, Jufang Lin, Rui Liu, Junjie Ma, Yinchao Qiu, Wanxi Su, Ying Sun, Jiaquan Wang, Xinyuan Wang, Jide Wu, Haopeng Yan, Yuqi Zeng, Aiwaner Zeng, Xiaolan Jan, Yang Zhang, Honglong Zheng, Yuxian Zhu, Junze Zheng, Zhu Chao, and Yelin Ru.

References

[1] J. Teizer, Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites, Adv. Eng. Inform. 29 (2) (2015) 225–238, https://doi.org/10.1016/j.aei.2015.03.006.
[2] A. Assadzadeh, M. Arashpour, A. Bab-Hadiashar, T. Ngo, H. Li, Automatic far-field camera calibration for construction scene analysis, Comp.-Aided Civ. Infrastruct. Eng. 36 (8) (2021) 1073–1090, https://doi.org/10.1111/mice.12660.
[3] B.A.S. Oliveira, A.P.D.F. Neto, R.M.A. Fernandino, R.F. Carvalho, A.L. Fernandes, F.G. Guimaraes, Automated monitoring of construction sites of electric power substations using deep learning, IEEE Access. 9 (2021) 19195–19207, https://doi.org/10.1109/access.2021.3054468.
[4] S.V.T. Tran, T.L. Nguyen, H.L. Chi, D. Lee, C. Park, Generative planning for construction safety surveillance camera installation in 4D BIM environment, Autom. Constr. 134 (2022), 104103, https://doi.org/10.1016/j.autcon.2021.104103.
[5] R. Xiong, Y. Song, H. Li, Y. Wang, Onsite video mining for construction hazards identification with visual relationships, Adv. Eng. Inform. 42 (Oct.) (2019), https://doi.org/10.1016/j.aei.2019.100966, 100966.1-100966.10.

[6] Y. Li, H. Wei, Z. Han, J. Huang, W. Wang, Deep learning-based safety helmet detection in engineering management based on convolutional neural networks, Adv. Civ. Eng. 6 (2020) 1–10, https://doi.org/10.1155/2020/9703560.
[7] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, Y. Zhao, Fast personal protective equipment detection for real construction sites using deep learning approaches, Sensors. 21 (10) (2021) 3478, https://doi.org/10.3390/s21103478.
[8] S. Kumar, H. Gupta, D. Yadav, I.A. Ansari, O.P. Verma, YOLO v4 algorithm for the real-time detection of fire and personal protective equipments at construction sites, Multimed. Tools Appl. (2021) 1–21, https://doi.org/10.1007/s11042-021-11280-6.
[9] R. Cheng, X. He, Z. Zheng, Z. Wang, Multi-scale safety helmet detection based on SAS-YOLO v3-tiny, Appl. Sci. 11 (8) (2021) 3652, https://doi.org/10.3390/app11083652.
[10] D. Benyang, X.C. Luo, Y. Miao, Safety helmet detection method based on YOLO v4, in: 2020 IEEE Conference on Computational Intelligence and Security (CIS), IEEE, 2020, pp. 155–158, https://doi.org/10.1109/CIS52066.2020.00041.
[11] H. Wang, Z. Hu, Y. Guo, Z. Yang, F. Zhou, P. Xu, A real-time safety helmet wearing detection approach based on csyolov3, Appl. Sci. 10 (19) (2020) 6732, https://doi.org/10.3390/app10196732.
[12] P. Wang, E. Fan, P. Wang, Comparative analysis of image classification algorithms based on traditional machine learning and deep learning, Pattern Recogn. Lett. 141 (2021) 61–67, https://doi.org/10.1016/j.patrec.2020.07.042.
[13] Y. Zhang, Safety Management of Civil Engineering Construction Based on artificial intelligence and machine vision technology, Adv. Civ. Eng. (2021) 1–14, https://doi.org/10.1155/2021/3769634.
[14] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.
[15] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, P. Fieguth, A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure, Adv. Eng. Inform. 29 (2) (2015) 196–210, https://doi.org/10.1016/j.aei.2015.01.008.
[16] X. Xiang, N. Lv, X. Guo, S. Wang, A. El Saddik, Engineering vehicles detection based on modified faster R-CNN for power grid surveillance, Sensors. 18 (7) (2018) 2258, https://doi.org/10.3390/s18072258.
[17] W. Fang, L. Ding, H. Luo, P.E. Love, Falls from heights: a computer vision-based approach for safety harness detection, Autom. Constr. 91 (JUL.) (2018) 53–61, https://doi.org/10.1016/j.autcon.2018.02.018.
[18] Z. Yang, Y. Yuan, M. Zhang, X. Zhao, Y. Zhang, B. Tian, Safety distance identification for crane drivers based on mask R-CNN, Sensors. 19 (12) (2019) 2789, https://doi.org/10.3390/s19122789.
[19] C. Chen, Z. Zhu, A. Hammad, Automated excavators activity recognition and productivity analysis from construction site surveillance videos, Autom. Constr. 110 (2020), 103045, https://doi.org/10.1016/j.autcon.2019.103045.
[20] H. Deng, H. Hao, D. Luo, Y. Deng, J.C. Cheng, Automatic indoor construction process monitoring for tiles based on BIM and computer vision, J. Constr. Eng. Manag. 146 (1) (2020), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001744.
[21] Q. Fang, H. Li, X. Luo, C. Li, W. An, A sematic and prior-knowledge-aided monocular localization method for construction-related entities, Comp.-Aided Civ. Infrastruct. Eng. 35(9 (2020) 979–996, https://doi.org/10.1111/mice.12541.
[22] M. Zhang, M. Zhu, X. Zhao, Recognition of high-risk scenarios in building construction based on image semantics, J. Comput. Civ. Eng. 34 (4) (2020), https://doi.org/10.1061/(ASCE)cp.1943-5487.0000900, 04020019.
[23] H. Luo, M. Wang, P.K.Y. Wong, J.C. Cheng, Full body pose estimation of construction equipment using computer vision and deep learning techniques, Autom. Constr. 110 (2020), 103016, https://doi.org/10.1016/j.autcon.2019.103016.
[24] Z. Pan, C. Su, Y. Deng, J.C. Cheng, Video2entities: a computer vision-based entity extraction framework for updating the architecture, engineering and construction industry knowledge graphs, Autom. Constr. 125 (2021), 103617, https://doi.org/10.1016/j.autcon.2021.103617.
[25] M. Zhang, R. Shi, Z. Yang, A critical review of vision-based occupational health and safety monitoring of construction site workers, Saf. Sci. 126 (2020), 104658, https://doi.org/10.1016/j.ssci.2020.104658.
[26] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Inform. 29 (2) (2015) 239–251, https://doi.org/10.1016/j.aei.2015.02.001.
[27] B. Zhong, H. Wu, L. Ding, P.E. Li, H. Li Love, H. Luo, L. Jiao, Mapping computer vision research in construction: Developments, knowledge gaps and implications for research, Autom. Constr. 107 (2019) 102919, https://doi.org/10.1016/j.autcon.2019.102919.
[28] L. Deng, The mnist database of handwritten digit images for machine learning research, IEEE Signal Process. Mag. 29 (6) (2012) 141–142, https://doi.org/10.1109/msp.2012.2211477.
[29] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, https://doi.org/10.1007/s11263-009-0275-4.
[30] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, Microsoft Coco: Common Objects in Context, in: European Conference on Computer Vision, Springer, Cham, 2014, September, pp. 740–755, https://doi.org/10.1007/978-3-319-10602-1_48.
[31] 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.
[32] H. Kim, H. Kim, Y.W. Hong, H. Byun, Detecting construction equipment using a region-based fully convolutional network and transfer learning, J. Comput. Civ. Eng. 32 (2) (2018) 04017082, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000731.
[33] Z. Kolar, H. Chen, X. Luo, Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images, Autom. Constr. 89 (2018) 58–70, https://doi.org/10.1016/j.autcon.2018.01.003.
[34] X.H. An, L. Zhou, C.Z. Wang, P.F. Li, Z.W. Li, Dataset and benchmark for detecting moving objects in construction sites, Autom. Constr. 122 (2021) 103482, https://doi.org/10.1016/j.autcon.2020.103482.
[35] B. Xiao, S.C. Kang, Development of an image data set of construction machines for deep learning object detection, J. Comput. Civ. Eng. 35 (2) (2021) 05020005, https://doi.org/10.1061/(ASCE)cp.1943-5487.0000945.
[36] J. Song, C.T. Haas, C.H. Caldas, Tracking the location of materials on construction job sites, J. Constr. Eng. Manag. 132 (9) (2006) 911–918, https://doi.org/10.1061/(ASCE)0733-9364(2006)132:9(911).
[37] A. Dimitrov, M. Golparvar-Fard, Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections, Adv. Eng. Inform. 28 (1) (2014) 37–49, https://doi.org/10.1016/j.aei.2013.11.002.
[38] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 580–587. https://arxiv-1311.2524.
[39] R. Girshick, R.-C.N.N. Fast, Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2015, pp. 1440–1448, https://doi.org/10.1109/ICCV.2015.169.
[40] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in, Adv. Neural Inf. Proces. Syst. (2015) 91–99, https://doi.org/10.1109/TPAMI.2016.2577031.
[41] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788, arxiv-1506.02640.
[42] H. Zhou, Y. Zhao, Q. Shen, L. Yang, H. Cai, Risk assessment and management via multi-source information fusion for undersea tunnel construction, Autom. Constr. 111 (2020), 103050, https://doi.org/10.1016/j.autcon.2019.103050.
[43] X.C. Luo, H. Li, Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks, J. Comput. Civ. Eng. 32 (3) (2018) 4018012.1, https://doi.org/10.1061/(asce)cp.1943-5487.0000756.
[44] T. Lin, LabelImg. https://github.com/tzutalin/labelImg, 2015.
[45] J.A. Hartigan, M.A. Wong, Algorithm AS 136: a K-means clustering algorithm, J. R. Stat. Soc. 28 (1) (1979) 100–108, https://doi.org/10.2307/2346830.
[46] Z. Zou, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A Survey, arXiv (2019) preprint arXiv:1905.05055, https://arxiv.org/abs/1905.05055v2.
[47] A. Torralba, K. Murphy, W.T. Freeman, Sharing features: efficient boosting procedures for multiclass object detection, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition vol. 2 II-II, CVPR, 2004, https://doi.org/10.1109/CVPR.2004.1315241.
[49] S. Standing, C. Standing, The ethical use of crowdsourcing, Bus. Ethics: A Eur. Rev. Bus. Ethics. 27 (1) (2018) 72–80, https://doi.org/10.1111/beer.12173.