Combining Self-Supervised Learning and Yolo v4 Network for Construction Vehicle Detection

Ying Zhang, Xuyang Hou, and Xuhang Hou

School of Electrical and Control Engineering, Shenyang Jianzhu University, Shenyang 110000, China

Correspondence should be addressed to Xuyang Hou; 1633098217@qq.com

Received 18 May 2022; Revised 9 August 2022; Accepted 27 August 2022; Published 20 September 2022

Academic Editor: Salvatore Carta

Copyright © 2022 Ying Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, there are many application fields of target detection, but it is very difficult to apply intelligent traffic target detection in the construction site because of the complex environment and many kinds of engineering vehicles. A method based on self-supervised learning combined with the Yolo (you only look once) v4 network defined as "SSL-Yolo v4" (self-supervised learning-Yolo v4) is proposed for the detection of construction vehicles. Based on the combination of self-supervised learning network and Yolo v4 algorithm network, a self-supervised learning method based on context rotation is introduced. By using this method, the problem that a large number of manual data annotations are needed in the training of existing deep learning algorithms is solved. Furthermore, the self-supervised learning network after training is combined with Yolo v4 network to improve the prediction ability, robustness, and detection accuracy of the model. The performance of the proposed model is optimized by performing five-fold cross validation on the self-built dataset, and the effectiveness of the algorithm is verified. The simulation results show that the average detection accuracy of the SSL-Yolo v4 method combined with self-supervised learning is 92.91%, 4.83% detection speed is improved, 7–8 fps detection speed is improved, and 8–9% recall rate is improved. The results show that the method has higher precision and speed and improves the ability of target prediction and the robustness of engineering vehicle detection.

1. Introduction

The technology of target detection has made great achievements in many fields. For example, target detection techniques are used in the field of medicine for cell identification and segmentation [1]. In the field of manufacturing, the detection network is used to determine whether the target is defective [2]. In the field of traffic, target detection technology is used to identify license plates for integrated traffic control and to identify autonomous driving targets in bad weather or at night [3–6]. Vehicle detection is one of the applications of the computer vision technology. At present, the research on construction vehicle detection methods can be roughly divided into two categories: the traditional image processing machine learning method and the deep learning based method [7]. In the traditional methods, the vehicle speed, vehicle color, number, and other information are mostly used for detection. For example, the visual-based virtual detection line method meets the requirements of vehicle supervision at large sites [8]. Aiming at the vehicle detection method based on sensors, this method is simple to operate and does not need complex procedures, but the environment adaptability is poor [9]. The combination of HOG features and support vector machine provides a new idea for construction vehicle identification: first, the extracted image is preprocessed, and then the target area is extracted according to the shape, color, and other characteristics of the construction vehicle, which reduces the target detection range effectively [10]. The CNN (convolutional neural network) is improved and applied to the intelligent monitoring of the detection of intrusion engineering vehicles, but there are problems such as installation difficulties, serious occlusion, and low large area inspection efficiency [11–13]. The researchers came up with the method of combining depth learning features and edge feature and proposed the FCOS algorithm, which has good tracking effect but not good classification effect [14].
At present, the popular deep learning algorithm is improved by the Yolo algorithm. Yolo is a real-time object detection system based on the CNN proposed in 2015. It has been widely used in medical, industrial, production, and other aspects. In recent years, to improve the detection effect of the convolutional network, researchers have continuously improved the residual network structure, deepening the network layer number, and other operations [15–17]. For example, the improved Yolo v3 detection algorithm uses context features for fusion and multiscale training, which greatly improves the detection accuracy [18]. A method by using freely acquired multimodal content for training computer vision algorithms was proposed by Lanaro et al. [19]. Through the idea of self-supervised learning of visual features, to mine the large-scale multimodal (text and image) document corpus, using the text corpus found in the hidden semantic structure and a topic modeling technology (TextTopicNet) to do the multimodal [20, 21]. Wu et al. promote self-supervised learning through knowledge transfer, proposing to reduce pseudo-label transfer knowledge on unlabeled datasets [22].

In life, due to the limitation of actual conditions, the open datasets of construction vehicles are often small samples, and the accuracy of supervised training on the basis of deep learning is not high enough because the types and number of samples collected by the datasets are small, and the feature extraction process cannot be effectively trained. While the supervised training will be affected by other factors, the manual labeling label is missing, errors, and other situations, and the labeling process is also very difficult. Due to the complexity of the construction environment, there is still a problem of poor small object detection accuracy by using the deep learning algorithm. Mainly because the pixel will change after multiple convolution training, the coefficient will appear with the improvement of convolution accuracy, which will affect the detection process. To solve the above problems, the design training process collects the corresponding dataset by itself and introduces the context-based self-supervised learning method, and the self-supervised network through the auxiliary task training, combined with the later deep learning algorithm. While ensuring the pixels of the datasets, 3–4 times data enhancement can improve the model robustness.

2. Correlation Methods

2.1. Self-Supervised Learning. Supervised learning requires a large number of manual operations during the generation of manual labeling and labels and a large number of data samples in the training of deep learning [23]. Label labeling of a large number of samples is still a bottleneck for supervised, as the amount of training data is crucial in data-driven models [24]. To reduce the burden of data collection, unsupervised or semisupervised learning strategies can be adopted. Unsupervised learning only does not require manual intervention and operational training, and self-supervised learning belongs to a set of unsupervised learning [25]. In self-supervised learning, auxiliary supervised tasks are set by entering certain properties of the data to achieve the training purpose, without manually marking the data. For example, divide the picture into different sizes, restore the picture, extract the main features of the picture, and predict the location of the picture.

2.2. Yolo v4 Algorithm. Yolo (you only look once) is a real-time object detection system based on the CNN, in 2015 [26]. The Yolo algorithm treats object detection as a regression problem and predicts bounding box coordinates and class probabilities directly from the full image. In recent years, the Yolo v2 algorithm has improved in the prediction accuracy, identifying more objects, and speed [27]. The Yolo v3 algorithm has changed the size of the model structure to measure the speed and accuracy of detection, and improved the detection range through multiple downsampling layers, and then improved the detection accuracy. The Yolo v4 algorithm through the CSPDarknet53 network features extract the image in S × S grid, target detection through the target center in the grid, using the residual network sampling and sampling features, the maximum pooling of different scales after stacking, finally after the size of the target category and position [28].

\[
\text{Mish} = x \ast \tanh \left( \ln \left( 1 + e^x \right) \right).
\]  

Yolo v4 network front-end innovation introduces mosaic data enhancement, SAT (self-adversarial training), its backbone network is CSP Darknet53 network, and adopts Mish activation function. The anchor frame mechanism of the output layer of the Yolo v4 algorithm is the same as the Yolo v3, and the main improvement is the loss function during the training [29]. The loss function of Yolo v3 consists of frame loss, confidence loss, and category loss, and the Yolo v4 algorithm innovates in the surrounding frame loss. As there will be an overlap in the detection process, the frame loss mode adopts CIOU, mainly considering three factors: aspect ratio, overlapping area, and distance to the central point.

\[
L = L_{\text{CIOU}} + L_{\text{CONF}} + L_{\text{CLASS}}
\]

\[
L_{\text{CIOU}} = 1 - \text{IOU}(A, B) + a \ast v + \frac{d^2}{c^2}.
\]

In the type, \( \text{IOU} \) is the union ratio between the prediction box and the real box, \( a \) is the weight coefficient, \( v \) is the similarity ratio of length to width, \( d \) is Euclidean distance between the center point of the prediction box and the real box, \( c \) is the diagonal distance between the minimum closure region of the prediction box and the real box. \( w^\text{ref} \) and \( h^\text{ref} \) are the width and height of the real box, \( w \) and \( h \) predicted the width and height.
\[ v = \frac{4}{\pi^2} \left( \frac{\tan \omega_{in}}{\mu_{in}} - \frac{\tan \omega_i}{\mu_i} \right)^2, \]

\[ a = \frac{\nu}{1 - IOU} + \nu, \]

\[ L_{CONF} = \sum_{i=0}^{x} \sum_{j=0}^{y} k[-\log p + (BCE(\tilde{n}, n))], \]

\[ L_{CLASS} = \sum_{i=0}^{x} \sum_{j=0}^{y} \log[y(n_{obj})[-\log(1 - p_c)]. \]

In the type, \( S \) is the number of grids, \( B \) is the number of prior boxes in each grid, \( K \) is the weight, determine whether the \( i \) prior box of the \( j \) grid is responsible for the object. If it is, the value is 1, otherwise, it is 0, and there is a probability that the current prior box has objects. The Yolo v4 algorithm requires that the output size image should be fixed. When the input image size is greater than or less than the specified output image size, the input image will be compressed or stretched, and this process will lead to distortion of the image. When there are small targets in the picture, it is easy to be blurred or even lost. To solve this problem, this paper proposes the SSL-Yolo v4 algorithm to improve the original data enhancement method of Yolo v4 by contrast enhancement, to improve the accuracy of network identification, positioning, and detection.

### 3. Research Methods

#### 3.1. Data Augmentation

In the process of data set construction, due to the complexity of construction vehicles and environment, there are few complete data sets available. In terms of data collection, to ensure the authenticity of the data and contact with the construction site, various construction vehicles including cranes and excavators around the transmission lines under different backgrounds, such as trees and houses, were collected. The obtained vehicle datasets are put in the network model for training, and the data is enhanced through random rotation, denoising, and other operations. To improve the accuracy of the detector, the dataset used in this design is to independently complete the construction vehicle dataset in the MATLAB environment.

1. The collected video of the engineering vehicle is divided into 500 frames, and the original image is distributed according to the ratio of 6:3:1.60% of the dataset is randomly rotated, 30% of the dataset is self-supervised detection, and 10% of the dataset is detection.

2. Perform a \( 0^\circ \sim 180^\circ \) random rotation operation on the image, which can increase the diversity of the sample. Several images are randomly generated with no position type and saved as JPG pictures with transparency information.

#### 3.2. Context-Based Self-Supervised Learning Methods

A context-based self-supervised learning strategy is adopted to generate and input unlabeled data into the training network, and model the unlabeled data together with the precollected labeled data. Context-based self-supervised learning can construct a large number of task information.
such as image mosaic, repair, coloring, rotation, and so on. With the rotation image as input and the predicted rotation angle of the image as output, the images with the building background were rotated 90°, 180°, and 270°, combined with the dataset of the network training front-end, the problem of blurred rotation angle of the input image is avoided. Because this study cannot fully simulate the complexity of the building background, the image of the building background is spliced with the image after rotation to simulate the complex building background. Using the untrained Resnet50 network as the training network, the validity of the Resnet50 network training and the accuracy of the classification were proved by the previous experiments. In this study, we changed the number of nodes in the full connection layer to 4 because we needed to predict 4 different classifications. After each convolution and before the activation of the normalized operation to improve the ability of feature extraction. In the residual error block of deep convolution, the input and output are controlled by setting convolution-related parameters to increase processing and avoid the loss of gradient of the deep network. The self-supervised learning process not only increases the number of images, but also improves the pixel quality. Previous studies have used self-supervised learning networks to increase the number of images. In this study, we removed mosaic data enhancement and proposed cutout and mix-up based on self-supervision. The self-supervised folders classified by rotation angles, with four different overlapping images and add noise on the images, are jointly introduced into the self-adversarial training network at the front end of the Yolo v4 network to train the enhancement results to improve the robustness of the model. The bottom right shows the Yolo v4 network structure diagram in Figure 4, blue represent the highly convolutional module such as CSP, and the output is the 3 required output dimensions. The CNN is a self-adversarial training network (SAT network), which uses the calculation process loss of the CNN, and then backpropagation to the image to modify the image information. It is worth noting that this operation does not need to change the network weight and directly put the modified picture into the training network [30].

When there are many targets in the picture, the accuracy of the model should be improved, while the self-supervised model only achieves the local optimization in the training process and fails the global optimization. To solve this problem, we combine self-supervised learning with the Yolo v4 network front-end to improve the data enhancement algorithm of the Yolo v4 network, and then use the self-adversarial network to backpropagate the information to modify the original picture. The original Yolo v4 algorithm adopts the mosaic data enhancement method, which combines 4 pictures into one training picture with the cut-mix method. The cut-mix method is to randomly cut pictures of different shapes and sizes and replace them with pictures of the same size and different kinds, to predict the occurrence probability of different kinds of targets. This method can improve the positioning ability and training efficiency, but because of the similar background pictures are forced splicing but not the area of the target, the background confusion will increase the difficulty of detection.

3.3. Building the SSL-Yolo v4 Algorithm Network. This training uses MATLAB to complete the comparative training and research of a variety of advanced target detection networks. In view of the complexity of the
Figure 2: Self-supervised learning network structure based on rotation.

Figure 3: Some results of online generation of labels in self-supervised learning network.

Figure 4: SSL-Yolo v4 network structure.
construction site, the similarity, occlusion problems and multiscale changes, and other complex engineering problems between the construction vehicles, the detection speed, and accuracy are suitable for the detection network of the construction site. By collecting the actual video of the construction site, the label data generated after self-supervised learning is input into the data enhancement network, and the pictures after the noise adding cutout operation and the random picture overlapping mix-up operation are first experienced to the front-end self-confrontation network of the SSL-Yolo v4 network.

1. Preprocessing the enhanced picture preparation after pretraining to adjust the image size, scale the pixel size, and batch process the input pictures.

2. When the input picture size and the specified network output picture size are inconsistent, according to the feature extraction network input size, adjust the input frame and anchor frame and adjust the input dataset size to the appropriate size of the feature extraction network.

3. Reset the parameters of the SSL-Yolo v4 network, set the number of anchor boxes to 8, and pass the anchor boxes data to the configure yolo v4 function, for the correct network arrangement, the configure yolo v4 function can improve the running rate of the network.

4. Create the Yolo v4 target detection network and set network training parameters; Yolo v4 network training optimization method adopts stochastic gradient descent momentum (SGDM), the initial learning rate is 0.001, Yolo v4 is divided into 16 subsets, the maximum training number of 100. The anchor box was estimated with the prediction anchor box from the size of the target in the training data, considering that the image size is adjusted before training, the size of the training data used to estimate the anchor box is also adjusted to set the “CheckpointPath” to a temporary position. This saves the partially trained detector during the training process. If the training is interrupted due to a power failure or a system failure, you can continue the training from the saved checkpoint. For detection, the pretrained network is downloaded, the yolov4 network, and the test image is read. Set the anchor frame and introduce the target type category, detect the target image in the figure, and visualize the detection results. The display results include the target position, size category, and detection accuracy.

4. Results and Discussion

To accurately evaluate the detection performance of the proposed SSL-Yolo v4 algorithm, the detection accuracy (average precision), detection speed (detection speed), and regression rate (recall) are selected. Set the correct number detected as TP, false positive calls the number of errors detected as FP, and false negative calls the number not identified as FN. IOU (intersection union) is a standard to measure the accuracy of detecting the corresponding object in a specific dataset. There are multiple bounding boxes to predict together, and then the network will choose the well-predicted bounding box (that is, IOU large) online to predict [31]. The intersection ratio (IOU) is the two regions divided by the set of the two regions.

\[
\text{AveragePrecision} = \frac{\text{TP}}{\text{TP} + \text{FP}},
\]

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

\[
\text{IOU} = \frac{\text{AreaofOverlap}}{\text{AreaofUnion}}.
\]

Previous experiments divided the data into training set and test set, the test set is independent of the training data, completely not involved in training, for the evaluation of the final model. But in the training process, the problem of fitting is that the model can match the training data well, but cannot predict the data outside the training set well. In order to optimize the model effect and verify the network generalization performance, the experiment adopts five-fold cross-validation method to get 5 models.

At first, five-fold cross-validation is adopted, and then three different algorithms are used to illustrate the comparison diagram. The dataset used in this experiment is a self-built dataset, split different construction site video to get 10,000 pictures, including 15 different construction vehicle targets, on an average, there are 1.2 goals in a picture. Divide the dataset into five small datasets, data 1, data 2, data 3, data 4, and data 5, each containing 2000 images. Using data 1, data 2, data 3, and data 4, four datasets as the training set, data 5 as the detection dataset, the precision of the first round of experiments and the regression rate were obtained. In the second experiment, data 1, data 2, data 3, and data 5 were used as the training set, and data 4 was used as the detection dataset. The precision and regression rate of the second experiment were obtained. By analogy, we carried out five rounds of experiments and got the regression rates of the five models, taking the average value based on the precision value. Table 1 shows the results of five-fold cross-validation, and after five trainings we can see that the third experiment had the best detection accuracy and regression rate, with the average detection accuracy of the model reaching 0.933.

To verify the validity of the context-based self-supervised learning model classification, two public datasets were selected: Pascal VOC and CIFAR-10. The Pascal VOC dataset contained 11530 images for training and testing, calibrating 27450 regions of interest. The dataset grew from four categories to the last 20 in eight years: human, animal, airplane, automobile, motorcycle, train, dining table, sofa, television, and so on. The CIFAR-10 dataset is divided into 5 training sets and 1 test set, each containing 10000 images. Each RGB image contain 32 * 32 in size. Planes, cars, birds, cats, deer,
dogs, frogs, horses, boats, and trucks fall into ten broad categories. In this experiment, 50, 100, 150, 200, 250, and 300 images were randomly selected as different test sets. The self-supervised method is to use the self-supervised learning method to build the training model, and the supervised method is to directly use the label data information to build the training model.

Table 2 is the IOU of supervised detection, Yolo v4 algorithm detection and SSL-Yolo v4 algorithm proposed in this paper. The three algorithms have different datasets (including 50, 100, 150, 200, 250, and 300 detection images). It can be seen that the present algorithm and Yolo v4 algorithm have a high detection speed, while the accuracy of IOU has not been greatly reduced. When the number of detection images gradually increases, both the detection speed and the recall rate increase. However, as shown in Figure 5, compared with self-supervised learning, the results of supervised learning detection are lower, and the SSL-Yolo V4 algorithm proposed in this paper has higher detection accuracy and recall rate, and faster detection speed.

Using the same datasets and different training and detection methods, different results are obtained. Figure 6 shows the supervised detection results, Figure 7 shows the

<table>
<thead>
<tr>
<th>The number of images tested</th>
<th>Supervision and testing IOU (%)–speed (fps)</th>
<th>Yolo v4 algorithm detection IOU (%)–speed (fps)</th>
<th>The algorithm of this paper IOU (%)–speed (fps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>85.3–22</td>
<td>85.2–22</td>
<td>86.9–20</td>
</tr>
<tr>
<td>100</td>
<td>85.6–20</td>
<td>88.0–19</td>
<td>87.4–20</td>
</tr>
<tr>
<td>150</td>
<td>85.6–20</td>
<td>88.6–17</td>
<td>88.1–14</td>
</tr>
<tr>
<td>200</td>
<td>88.0–18</td>
<td>89.1–12</td>
<td>90.7–10</td>
</tr>
<tr>
<td>250</td>
<td>88.2–17</td>
<td>90.9–12</td>
<td>92.3–8</td>
</tr>
<tr>
<td>300</td>
<td>87.2–17</td>
<td>90.7–10</td>
<td>94.0–8</td>
</tr>
</tbody>
</table>

**Figure 5:** Comparison diagram of detection accuracy and recall rate under different number of detection image types. (a) Recall curve. (b) Detection accuracy curve.
detection results after introducing self-supervised learning in the Yolo v4 network, and Figure 8 introduces the detection results of self-supervised learning after improving the Yolo v4 data enhancement. According to the detection accuracy under different circumstances, it can be seen that the loss of the detection box in Figure 6 is serious, while Figure 7 diagram introducing self-supervised learning can detect small targets, but, because the helmet covers the face, it is not completely detected. Figure 8 is the detection results after improving the data enhancement method and introducing the contrast enhancement of different targets, which can clearly see that the detection coverage rate and detection accuracy have been improved. The algorithm proposed can simulate different external environments and mark the vehicle position more accurately when the vehicle features are not obvious. By comparison, it shows that the proposed SSL-Yolo v4 algorithm has higher detection accuracy and more accurate detection type when the camera is above and blocked.

5. Conclusions

Due to the complexity of the construction detection environment, there are many uncertainties in the target detection process, which will more or less have a certain impact on the results. As an effective means of security, the video surveillance system requires high requirements on attention, vigilance, and especially the ability to respond to abnormal situations. This paper proposes the SSL-Yolo v4 algorithm, which introduces a self-supervised learning method, turns the manual annotation detection box problem into automatic or semiautomatic annotation, and saves artificial methods while realizing data enhancement. At the same time, improving the Yolo v4 data enhancement method, adding contrast training, also achieves the data enhancement and improves the model robustness, and improves the detection accuracy and speed. Pretraining and training on images containing 2000 images on three different datasets yielded the SSL-
Yolo v4 network. The comparison of the simulation results shows the detection accuracy and recall of the detection accuracy and speed. However, the algorithm proposed still has some disadvantages. When the input picture pixels are not high enough, the detection accuracy will decline or even appear as classification errors, which will be further made in future research.

Data Availability
The data used to support the findings of this study are included within the article.

Conflicts of Interest
The authors declare that there are no conflicts of interest.

Acknowledgments
The work described in this article was supported by the funds from the Basic Scientific Research Projects of the Educational Department of Liaoning Province (grant no. LJKZ0585) and the project of Ministry of Housing and Urban-Rural Construction of Foundation (grant no. 2019-K-168), thirty thousand RMB.

References


