Improved Pose Estimation Network Based on Spatial Registration Model

Zexiang Liu[1,2], Ziao Dong[1,2], Xin Yin[1,2],Yanbing Liang[1,2*]
[1]Hebei Key Laboratory of Data Science and Application, North China University of Science and Technology, Tangshan, Hebei, China
[2]College of Science, North China University of Science and Technology, Tangshan, Hebei, China
**Corresponding Author*

Abstract: Position and attitude estimation refer to estimating the distance and attitude between the object to be measured and the sensor device from the input information captured by the sensor device. The traditional computer vision processing method to optimize the feature vector extracted by the feature extraction algorithm often requires a large number of resources to optimize the model. The introduction of deep learning provides a new solution. In this paper, we use the depth learning Mask-R-CNN framework to recognize and segment the object, and obtain the size and contour features of the object; Then VGG-16 algorithm is used to extract RGB image features; Then, the mask information extracted by Mask R-CNN is fused for convolution and pooling; Finally, the convolved feature map is collected into two fully connected layers to predict the translation matrix and rotation matrix respectively. At the same time, after the fusion, the feature map is sampled up and convolved, and finally the feature map with the same size as the original input image is obtained. The output is expected to be consistent with the input mask. By establishing the sensor space registration model, the least square method and the generalized least square method are used to estimate the error parameters and compensate them to the sensor system, so as to reduce the error impact caused by the equipment observation data and obtain more accurate position and attitude information. The experimental comparison shows that the average proportion of the traditional Pose CNN to predict the object's position and attitude error is 64.6 within 8 cm. The average accuracy of the position and attitude estimation network studied in this paper is 81.5, and the position and attitude estimation network have better generalization ability.

Keywords: Position and Attitude Estimation; System Error; Spatial Registration Model; Target Detection and Segmentation; Deep Learning

1. Introduction

The subdivision of manufacturing disciplines and more com-plex manufacturing processes require industrial robots to work more intelligently and finely [1]. In modern machine production, machine assembly automation has become the key link to improve the production efficiency of the entire manufacturing system, reduce costs and stabilize product quality. Among them, the automatic assembly of visual robots is a key point of development. The automatic assembly of visual robots focuses on how to determine the position and orientation of the assembly workpiece [2]. The position and orientation estimation is a key link in the automatic assembly of visual robots. The accuracy of the position and orientation estimation directly deter-mines the correctness of the assembly of visual robots. In this paper, the pose estimation of objects will be studied in depth. Malik J et al. proposed a fully supervised depth network [3], which learns to jointly estimate the complete 3D hand mesh representation and pose from a single depth image. By using a joint training strategy with real and synthetic data, 3D hand mesh and pose can be recovered from the real image within 30ms. Liu Y, Xing C and others proposed to extract edges by fusing Canny edge detection method and depth learning method using double threshold method [4], and then using nonparametric statistical method to fit edge pixel residuals to form an error

function, so as to obtain the pose estimation of the target object. Lu H et al. used the ASIS method to segment and preprocess the three-dimensional point cloud [5], and mapped the three-dimensional point cloud to the two-dimensional plane, thus generating the depth feature map and extracting the pose features, which greatly improved the accuracy of pose estimation. Yu C et al. built a dense point cloud image by adding the module of building a dense point cloud in the pure vision SLAM system [6], showing some information more favorable to the surrounding environment, providing a solution to the problem of position and pose estimation of sparse point cloud image. Other researchers used the Scale Invariant Feature Transform (SIFT) algorithm or the Normal Aligned Radial Feature (NARF) to extract the required point cloud features [7]. Then, in order to eliminate the wrong feature pairs, they used the Random Sample Consistency RANSAC algorithm to obtain an initial transformation matrix, Then the final transformation matrix is estimated by the ICP algorithm [8].

In this paper, the depth learning algorithm is introduced into the vision positioning and guidance of measurement sensor equipment, and then the spatial registration model is established to compensate the error parameters of the sensor system. Chapter 1 discusses and summarizes the existing pose estimation methods; In chapter 2, the spatial registration method is analyzed theoretically; In chapter 3, the pose estimation model based on depth learning is proposed, including the target detection module based on Mask R-CNN and the pose estimation module combined with VGG network fusion; Chapter 4 verifies by selecting the data set CamVid of urban road scenes published by Cambridge University; Chapter 5 summarizes and analyzes the experimental results, and the pose estimation network has certain practical value.

2. Sensor Data Space Registration Algorithm

Because of their strong execution and high target recognition rate, robots have demonstrated incomparable advantages in various fields of industry and life [9]. From the angle of the sensor, the measurement of the target position and coordinates is always aimed at a certain coordinate system. The selection of different coordinate systems determines the diversity of the expression of the object motion state model. By comparing and verifying the motion state under different co-ordinate systems, selecting a reasonable coordinate system as the benchmark for the sensor to describe the object's motion state can improve measurement accuracy, re-duce errors, and reduce the amount of calculation. The space matching criterion is a process of estimating and compensating the system deviation of the sensor when the sensor detects the space target. It can improve the fusion accuracy [10].

2.1 Three-Dimensional Coordinate Conversion

The distance between the camera and the object can be measured by using the triangle positioning method of the camera [11]. The object can be observed from different positions. The three-dimensional reconstruction of the object can be completed by synthesizing the observed data and results. We call this measurement method stereo vision measurement.

According to the principle of the stereo vision measurement method, the number of different cameras can be divided into the monocular vision method, binocular vision method and multi-vision method. The measurement method using single-camera equipment is called the monocular vision method. The system structure is simple, the recognition method is simple, the recognition time is short, and the field of view of a single camera is limited. Therefore, this measurement method is limited to the field of view of a single camera. The binocular vision method refers to two cameras and introduces geometric constraints between objects, which increases the matching accuracy of objects, but the improvement of accuracy also increases the measurement error and reduces the measurement efficiency.

Similar to the way people use their eyes to capture objects, cameras in binocular vision capture the same object in the same space from different perspectives. The captured information is calculated and detected by the computer to determine the position information of the object.

2.2 Theoretical Analysis of Measurement Scheme

There are some defects and disadvantages in

the use of stereo vision measurement technology [12], because it is implemented in a two-dimensional plane. Even if the high-order approximation is used to ensure the accuracy of coordinate transformation, because the geodetic coordinate system is an ellipsoid rather than a strict sphere, new errors are still introduced in the approximate calculation, making the system error change rather than a fixed constant. In the two-dimensional coordinate system, the slant distance error and azimuth error can be estimated by mathematical modeling, but the pitch angle error cannot be estimated. For the new error and the need to estimate the systematic error of the pitch angle, the 3D space registration model based on the 3D space coordinate system is adopted.

From the perspective of sensors, the types of sensors used in the 3D spatial registration model are essentially different from those used in the 2D spatial registration model; The expression dimension of the 3D spatial registration model has an additional layer. Three-dimensional space can express more in-formation, but the amount of computing increases very much, which puts forward higher requirements for computer performance.

3.3 Multi-Sensor Spatial Registration Model

The measurement space registration model equation of multi-sensor is obtained. Assuming that a group of multi-sensor measurement results are available, the common methods for system error solution are the least square method, generalized least squares method [13], etc.

3. A Pose Estimation Model Based On The Mask R-CNN Framework

For pose estimation, depth learning method is used to explore, aiming at estimating the transformation matrix from the target object coordinate system to the camera coordinate system, including translation matrix T and rotation matrix R, when a color image is given. Convolution network can extract more effective features from images and make better use of various information in images.

3.1 Target Detection Network

Target detection [14] [15] refers to finding out the position of an object in a given input image and giving the category of the object through model classification or pre-diction. In traditional computer vision methods, it is necessary to extract and select image features from each input image, and then put them into the model for training to obtain results[16]. The appearance of deep learning expands the boundary of digital image processing, provides a new idea and method, and is widely used in image classification, semantic segmentation, object detection and other fields.

R-CNN (Region CNN) is the first to successfully introduce the depth learning method into the field of object detection, which realizes the object detection of the input image [17]. The R-CNN algorithm idea is summarized as four steps: extracting candidate regions, feature extraction, image classification, and non-maximum suppression. The selected search algorithm is used to extract candidate regions from the input digital image to extract about 2000 to 3000 candidate regions that may contain objects. The above candidate regions are uniformly converted into 227 * 227 dimensions and transferred to AlexNet CNN to extract feature vectors, The output feature vectors are classified and predicted using the trained SVM model [18]. Fast R-CNN [19] is improved based on R-CNN. It uses convolutional network to extract features of input digital images, and temporarily stores all features in video storage to obtain complete image features. The pooling operation of regions of interest unifies the feature maps of different sizes, optimizes the SVM classification process in the R-CNN framework, and uses the joint training strategy to output the classification results and boundary boxes of regions at the same time [20]. Faster R-CNN [21] introduces the RNP network based on Fast-CNN. The use of the RNP network simplifies the training model and significantly improves the running speed.

Mask R-CNN is a target detection algorithm proposed in 2017, which is an extension of Faster R-CNN [22]. It mainly uses Softmax classifier, LDA (Latent Dirichlet Allocation) and SVM (Support Vector Machine) classifier. It can output the mask of the image by adding a branch of prediction segmentation on the bounding box recognition in parallel, and train the parameters through the loss function to achieve better image detection and segmentation effect. Mask R-CNN has also achieved good detection results in other fields,

such as human key point recognition and detection. The algorithm flow of Mask R-CNN is divided into two stages. In the first stage, all possible frames including the target object are obtained mainly through the regional recommendation network (RNP network). In the second stage, through the detection and recognition of each object, the two outputs of the object's frame information in the image and the predicted category are obtained [23]. On this basis, a third output branch is added to mask the output object.

In feature extraction, Mask R-CNN selects ResNet50 residual network and feature pyramid network (FPN) [24]. After training the FPN and ResNet50 networks, a feature map containing strong semantic information and strong spatial information will be obtained. According to the feature map, the region candidate network (RPN) can obtain candidate regions through non maximum suppression. At the same time, in order to solve the problem of misalignment between the feature map and the original image, Mask R-CNN uses the Rol Align method to improve. Rol Align is a pooling layer. By mapping the feature map and the original image pixels, the approximate spatial position is reserved, so that the results are mapped to the original image, and the detection accuracy of the algorithm is improved. Finally, the full junction layer and convolution layer receive ROL for position regression, mask prediction, category prediction, etc.

3.2 Position and Attitude Estimation Model

VGG-16 is a convolutional neural network model with strong generalization ability and deep structure. Among various combinations of the stacked network layer, the VGG-16 network performs better. It processes the data provided by the upper layer in the lower layer, and there is no shared feature information between the two layers. It has 16 hidden layers, composed of 13 convolution layers and 3 full connections. The salient features of the image are extracted mainly by using 3 × Convolution Kernel of 3 and 2 × 2. Therefore, the VGG-16 network has a good target object detection effect, can well extract the characteristics of target objects, and can maximize the research of attitude estimation networks. Its network architecture is shown in **Figure 1.** below.

The expression form of the predicted translation matrix is $T=(x_1,x_2,x_3)$, which respectively represents the three outputs of the translation matrix T, and the expression form of the true value of the corresponding translation matrix is:

$$\tilde{T}=\left(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3\right) \tag{1}$$

Similarly, R and $\tilde{R}$ are used to represent the predicted rotation matrix and its true value respectively. Considering the characteristics of the rotation matrix, there will be redundancy, so the predicted rotation matrix is expressed by a quaternion. The dimensions of the two full connection layers in the network are 1024 and 512 respectively, and the dimensions of the last full connection layer are 3 and 4 respectively, representing the translation matrix T and the rotation quaternion R.

The prediction rotation matrix and translation matrix belong to the regression problem. The mean absolute error function (MAE) is often used as the loss function in regression problems, and its formula is as follows:

$$MAE=\sum_{i=1}^{n}|\ y_{\mathrm{i}}-y_i^p\ | \tag{2}$$

However, in practice, simply using MAE as the loss function of network training often fails to achieve the desired results. Considering that the changes of the target object in the three-dimensional space are not consistent with the changes caused by the two-dimensional image, specifically, in the three-dimensional space, when the target object moves in the direction away from the optical axis of the camera and the target object moves in the direction parallel to the pixel plane, the changes caused by the target object in the two-dimensional image are more intense. Considering this situation, and in order to directly estimate the position and orientation of the target object from the network, a weighted loss function is proposed:

$$loss=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{Q}\left\|\tilde{\omega}_j\left(x_{ji}-\tilde{x}_{ji}\right)\right\|^2 \tag{3}$$

Wherein, x_{ji} represents the jth component of the ith training sample $\tilde{x}_{ji}$ represents the jth component of the ith training sample predicted by the network. Q represents the dimension of the prediction vector, N represents the number of samples, and $\tilde{\omega}_j$ represents the weight of

each dimension. The weight is calculated as follows:

$$\omega_j = \left(\frac{1}{N} \sum_{k=1}^{N} \left\| x_k \right\|_1 \right)^{-1} \tag{4}$$

$$\tilde{\omega}_j = \frac{\omega_j}{\sum_{i=1}^{M} \omega_j} \tag{5}$$

Where, M represents the number of samples, and k represents the kth value of the sample.

The evaluation of the pose estimation network adopted this time adopts the method of calculating the average distance. The average distance is calculated by calculating the real rotation matrix R and translation T of the target object center point projected to the cloud point in space and the rotation matrix R and translation T predicted by the pose estimation calculation method mapped to the cloud point in space. The specific calculation formula is as follows:

$$\mathrm{AD} = \frac{1}{m} \sum_{x \in A} \sqrt{\{(\mathrm{Rp} + T) - (\tilde{R}\mathrm{p} + \tilde{T})\}^2} \tag{6}$$

Where R represents the rotation matrix, T represents the translation matrix, $\tilde{R}$ represents the predicted rotation matrix, $\tilde{T}$ represents the predicted translation matrix, A represents the set of all points in the point cloud model and $p = (x, y, z)$ represents a point coordinate in the three-dimensional point cloud.

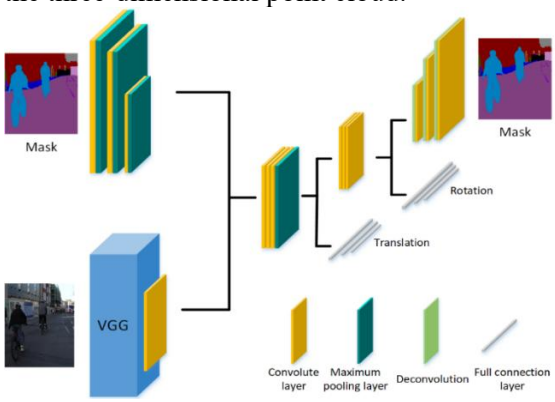

Figure 1. Schematic Diagram of Pose Estimation Network

This research combines the depth learning method and the common target detection framework to try and explore, to achieve the conversion relationship of estimating the object to be measured from the target coordinate system to the specified coordinate system of the sensor equipment when a given color image is input, including systematic error and random error. Convolution neural network is widely used in image feature extraction, and the effect of image feature extraction is remarkable, so we choose to use a convolution neural network combined with pose estimation to establish the model.

4. Analysis of Experimental Results

4.1 Data Set

This paper selects CamVid, a dataset of urban road scenes published by Cam-bridge University. The data set includes about 700 images, and the segmentation ac-curacy is evaluated for 11 commonly used categories, namely Symbol, Car, and Pedestrian. The ratio of randomly dividing training set, test set and verification set according to different scenes is 4:2:1. The training set is used to estimate the model, the verification set is used to determine the parameters of the network model, and the test set is used to test the performance of the finally selected optimal model.

4.2 Results of Motion Model Parameter Estimation Algorithm

It is assumed that sensor equipment A and sensor equipment B are erected at (0,0) and (400, 0) of the coordinate system. In this hypothetical system, it is assumed that the observed deviation of the sensor equipment is 1m, the deviation of the azimuth angle is 0.0085, and the measured object moves along the parallel line direction of the connecting line of the sensor equipment.

Table 1. Estimation of System Error Parameters by Least Squares Method

σ_n	Δr_A	$\Delta\theta_A$	Δr_B	$\Delta\theta_B$
0	1.0183	0.0087	1.0077	0.0087
0.5	1.0466	0.0086	1.0156	0.0088
1.0	1.0521	0.0086	1.0765	0.0090
1.5	1.0774	0.0084	1.0865	0.0090
2.0	1.4332	1.0066	1.1432	0.0112

Under different random errors, the approximate estimation of the 3D spatial registration system is obtained by computer calculation after repeated sampling 150 times. The results are shown in **Table 1.** and **Table 2.**

Among them, σ_n represents the noise covariance, Δr_A represents the difference between the estimated distance of equipment A and the actual value, $\Delta\theta_A$ represents the

difference between the estimated azimuth of equipment A and the actual value, Δr_B represents the difference between the estimated distance of equipment B and the actual value, and $\Delta\theta_B$ represents the difference between the estimated azimuth of equipment B and the actual value. It can be seen from **Table 1.** and **Table 2.** that due to the increase in noise covariance, the difference between the estimated value of the least squares method and the actual value becomes larger, and the error is relatively high. The generalized least squares method considers the noise error, and the result is better than the least squares method.

Table 2. System Error Parameter Estimates of Generalized Least Squares Method

σ_n	Δr_A	$\Delta\theta_A$	Δr_B	$\Delta\theta_B$
0	1.0031	0.0087	1.0042	0.0086
0.5	1.0112	0.0086	1.0088	0.0089
1.0	1.0256	0.0085	1.0386	0.0089
1.5	1.0455	0.0084	1.0122	0.0092
2.0	1.0350	0.0077	1.0422	0.0096

Table 3. Summary of Comparison Results

Target object	Pose CNN	Research Network
001_Bicyclist	52.8	76.6
002_Pedestrian	62.0	83.2
003_Car	68.1	85.2
004_Fence	61.6	67.1
005_SignSymbol	72.1	88.5
006_Tree	70.3	77.3
007_Pavement	61.8	75.4
008_Road	78.3	79.9
009_Pole	60.3	86.5
010_Building	71.6	80.3
011_Sky	51.7	96.5
Average	**64.6**	**81.5**

According to the above research conclusions, in order to better calculate the system error, relevant issues should be studied as carefully and comprehensively as possible in combination with the specific conditions of the system. See the following section for the specific research conclusions and analysis.

4.3 Experimental Results

In order to improve the accuracy of the network, when training the network, the samples are randomly sampled with return. The size of each batch of samples is set to 20 groups, the learning rate is set to 0.002, and the number of iterations is set to 5. The initial parameters of the first 13 convolution layers in the network are obtained through VGG-16 network training on ImageNet. Set the confidence level of the target detection object of Mask R-CNN to 0.6, and set the threshold value of the average distance AD to 8cm. The visual results of model prediction are shown in **Figure 2.**

In the experiment, in order to verify the effect of the network in this section, an experimental comparison is made with Pose CNN, an object pose estimation network. The following table summarizes the proportion of estimated pose errors of 11 objects within 8 cm. The experimental results are shown in **Table 3.**

It can be seen from the experimental results that, in the end of the training using the CamVid dataset, the error of the network recognition results and pose estimation in this paper is better than that of Pose CNN network, because the introduction of a better target recognition and object segmentation mask makes the pose estimation network proposed in this paper obtain a better result, and the network has a better generalization ability than Pose CNN network. Because there are some objects to be predicted with small volumes or seriously occluded by other objects in the data set, the extracted features are very few. Objects with smooth surfaces and single shapes have the same problem, so the prediction effect of the pose estimation network is poor.

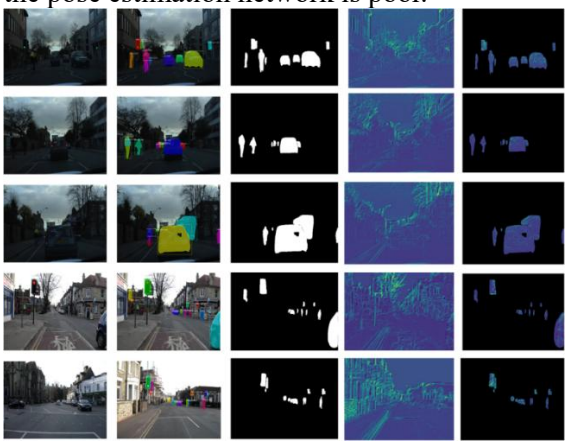

Figure 2. Visualization Results of Some Model Forecasts in Different Scenarios

5. Conclusion

In this paper, multi-sensor spatial registration model and attitude estimation model are established. The pose estimation model is composed of two modules: target detection and pose estimation. The target detection selects

the Mask RCNN framework to detect the target and obtain the mask information; In the attitude estimation part, the feature map extracted from VGG16 network and the Mask information extracted from Mask R-CNN are fused to predict the translation matrix and rotation matrix respectively. By improving the loss function and adjusting the parameters, the model can predict the pose of the target object from the image to be measured. Through experimental comparison, the prediction effect of the network model in this paper is better than that of Pose CNN. According to the experimental results, the improved pose estimation network based on multi-sensor spatial registration model established in this paper has a certain degree of application value. In terms of accuracy, further research will be conducted in the future.

REFERENCES

[1] Togias, T., Gkournelos, C., Angelakis, P., Michalos, G., & Makris, S. (2021). Virtual reality environment for industrial robot control and path design. Procedia CIRP, 100, 133-138.

[2] Song, R., Li, F., Fu, T., & Zhao, J. (2020). A robotic automatic assembly system based on vision. Applied Sciences, 10(3), 1157.

[3] Malik, J., Elhayek, A., Nunnari, F., Varanasi, K., Tamaddon, K., Heloir, A., & Stricker, D. (2018, September). Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In 2018 International Conference on 3D Vision (3DV) (pp. 110-119). IEEE.

[4] Liu, Y., & Lew, M. S. (2016). Learning relaxed deep supervision for better edge detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 231-240).

[5] Lu, H., & Shi, H. (2020). Deep learning for 3d point cloud understanding: a survey. arXiv preprint arXiv:2009.08920.

[6] Yu, C., Liu, Z., Liu, X. J., Xie, F., Yang, Y., Wei, Q., & Fei, Q. (2018, October). DS-SLAM: A semantic visual SLAM towards dynamic environments. In 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 1168-1174). IEEE.

[7] Li, C., Xia, Y., Yang, M., & Wu, X. (2022). Study on TLS point cloud registration algorithm for large-scale outdoor weak geometric features. Sensors, 22(14), 5072.

[8] Zhao, B., Chen, X., Le, X., Xi, J., & Jia, Z. (2021). A comprehensive performance evaluation of 3-D transformation estimation techniques in point cloud registration. IEEE Transactions on Instrumentation and Measurement, 70, 1-14.

[9] Sun, Y. (2021, April). Design and Research on Distributed Control System of Humanoid Robot Based on Automation Technology. In 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) (pp. 1063-1066). IEEE.

[10] Wang, J., Zeng, Y., Wei, S., Wei, Z., Wu, Q., & Savaria, Y. (2021). Multi-sensor track-to-track association and spatial registration algorithm under incomplete measurements. IEEE Transactions on Signal Processing, 69, 3337-3350.

[11] Yang, A., Yang, X., Liu, W., Han, Y., & Zhang, H. (2019). Research on 3D positioning of handheld terminal based on particle swarm optimization. Journal of Internet Technology, 20(2), 563-572.

[12] Li, H., & Zhang, B. (2021). Application of integrated binocular stereo vision measurement and wireless sensor system in athlete displacement test. Alexandria Engineering Journal, 60(5), 4325-4335.

[13] Brown, J. D., & Brown, J. D. (2018). Generalized Least Squares Estimation. Advanced Statistics for the Behavioral Sciences: A Computational Approach with R, 189-217.

[14] Brownlee, J. (2019). Deep learning for computer vision: image classification, object detection, and face recognition in python. Machine Learning Mastery.

[15] Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257-276.

[16] Gao, H., Zhang, Y., Chen, Z., Xu, S., Hong, D., & Zhang, B. (2023). A multidepth and multibranch network for hyperspectral target detection based on band selection. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-18.

[17] Zou, X. (2019, August). A review of object detection techniques. In 2019 International conference on smart grid and

electrical automation (ICSGEA) (pp. 251-254). IEEE.

[18] Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2023). A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 10(1), 183-208.

[19] Xu, X., Zhao, M., Shi, P., Ren, R., He, X., Wei, X., & Yang, H. (2022). Crack detection and comparison study based on faster R-CNN and mask R-CNN. Sensors, 22(3), 1215.

[20] Josifovski, J., Kerzel, M., Pregizer, C., Posniak, L., & Wermter, S. (2018, October). Object detection and pose estimation based on convolutional neural networks trained with synthetic data. In 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS) (pp. 6269-6276). IEEE.

[21] Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., & Zhang, C. (2021). Defrcn: Decoupled faster r-cnn for few-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8681-8690).

[22] Bharati, P., & Pramanik, A. (2020). Deep learning techniques—R-CNN to mask R-CNN: a survey. Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019, 657-668.

[23] Vukicevic, A. M., Macuzic, I., Mijailovic, N., Peulic, A., & Radovic, M. (2021). Assessment of the handcart pushing and pulling safety by using deep learning 3D pose estimation and IoT force sensors. Expert Systems with Applications, 183, 115371.

[24] Kirillov, A., Girshick, R., He, K., & Dollár, P. (2019). Panoptic feature pyramid networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6399-6408).

