ISSN: 2664-9640

Dangerous behavior recognition method based on Pose-HRnet Neural Network

Ming Zhang [1,*], Pengli Hu [2]

[1] Guangdong Vocational College of Post and Telecom, Guangzhou 510006, China;

[2] School of Intelligent Engineering, Sun Yat-sen University, Guangzhou 510006, China.

Abstract

In order to effectively predict whether the high-altitude workers are in a dangerous state, this paper studies the identification method of dangerous behaviors that lead to safety accidents of high-altitude workers. With the arrival of the wave of deep learning research, convolutional neural network as the representative of deep learning algorithm can extract image features more accurately and effectively. Therefore, this paper applies deep learning to the field of human behavior recognition. Based on the HRnet Neural Network proposed in reference [1], by introducing the Squeeze and Exception module, and using the improved activation function L-swish, this paper proposes the improved network structure Pose-HRnet, and uses it to judge whether the workers' behaviors are dangerous when working at heights. The experimental results show that compared with HRnet, the improved network structure Pose-HRnet reduces the size of the model while maintaining the accuracy, and the average accuracy of the classification experiment reaches 98.5%.

Keywords

Neural Network, Pose-HRnet, Feature Extraction, Behavior Recognition.

1. Introduction

In recent years, Chinese construction industry has developed rapidly, but the probability of construction safety accidents is also rising. The research shows that there are two main factors for safety accidents in high-altitude operation: one is the subjective factor[2-4], the dangerous behavior of operators, such as fatigue operation, nonstandard construction action, not wearing safety belt, etc.; the other is objective factor, there are safety hidden dangers in the construction environment, such as the loosening of scaffold pole and the falling off of safety rope.

The existing safety measures to prevent High-altitude Falling mainly include strengthening the physical protection measures , such as wearing safety rope, using scaffold, safety net and toe board[5] . These measures can only give a certain degree of protection to the operators after the occurrence of the danger, and can not play a role in the risk prediction. In addition, the safety monitoring and early warning system in the process of personnel operation also belongs to a kind of safety measures. For example, Melo[6] has proposed to use unmanned aerial vehicles to monitor personnel operation behavior in construction sites. However, this method has high equipment cost and requires real-time human monitoring, so its application is limited. Jia Guangshe[7] studied the influencing factors of construction workers' safety behavior and proposed a safety behavior early warning system based on a cross-level perspective, which can pre control safety behavior, but the real-time performance of this system is not high and can not send early warning signals in time.

With the rise of deep learning, neural network[8] has shown remarkable performance in feature extraction and behavior recognition. It has been proved that the neural network is feasible to detect the human behavior characteristics. Before 2015, the human detection model

ISSN: 2664-9640

used regression method[9] to get the coordinates of key points of bone, but the experimental effect is not ideal. Therefore, the commonly used transition processing method is to take human behavior recognition as a detection problem, so as to obtain the hot spot map. The CPM method proposed by Wei[10] has strong robustness, in 2016. The contribution of CPM lies in the use of sequential convolution architecture to express spatial information and texture information. The model can solve the occlusion problem effectively, but in order to reduce the computational complexity, the image accuracy is lost and some high-resolution features are lost. Ma Z [11] used CA-CenterNet for supervising the dangerous driving behavior .In order to solve the problem of low image resolution, Sun K[1] proposed a high-resolution Network named HRnet, which can maintain high resolution in the whole network link of data flow, greatly improving the accuracy of image recognition, but at the same time, it also brings the problems of parameter increase and slow running speed.

In order to solve the above problems, on the basis of HRnet, this paper proposes an improved network structure named Pose- HRnet, by introducing the squeeze and exception module and using the improved activation function L-swish. The improved network structure is applied to the identification of dangerous behaviors of high-altitude workers, and is used to judge whether the operators have dangerous behaviors, and then send warning signals to prevent dangers happen.

2. Network architecture improvement

Most of the existing 2D human pose estimation algorithms are based on the hourglass model structure proposed in reference[12]. The basic flow of the algorithm is as follows:

Firstly, convolution is performed on the input image, at the same time, the down sampling operation is carried out to obtain the features with lower resolution, thus reducing the computational complexity. Then, the low-resolution features are upsampling to increase the resolution of the image, so as to realize multi-scale feature extraction. The process of feature extraction is carried out in serial mode, so the efficiency and accuracy of the results are not high.

The network model proposed in this paper is further improved on the basis of HRnet network structure. HRnet can maintain the high resolution of feature map in the whole process of feature extraction,by gradually adding sub network of low resolution feature map into the main network of high-resolution feature map, and carry out multi-scale fusion and feature extraction between different networks[13]. In HRnet network, high resolution network and low resolution network are connected in parallel. Therefore, the method can maintain high resolution and the predicted feature map is more accurate in space. The joint team of China University of science & technology and Microsoft Asia Research Institute applied HRnet network to key point detection, object pose estimation and other fields, and achieved good accuracy. Specifically, the HRnet network model is shown in Figure 1.

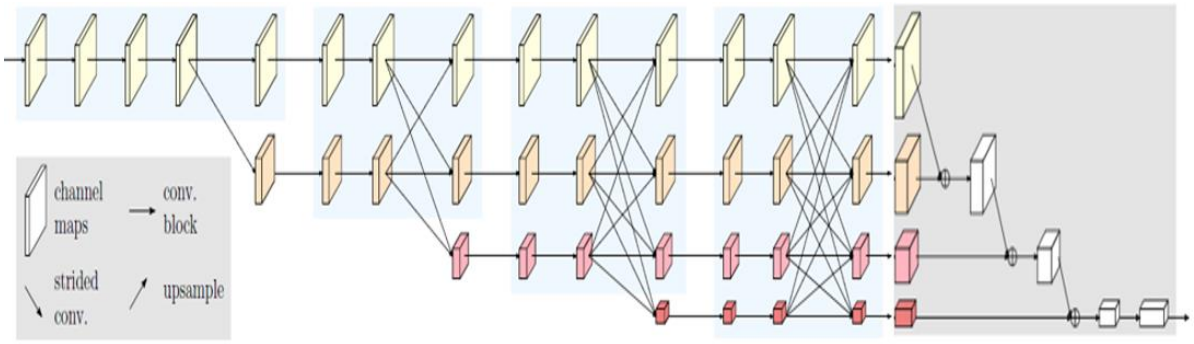

Figure 1: HRnet network model diagram

In order to meet the application of the model in the embedded equipment in the field of aerial work (such as deploying and running in the intelligent camera), on the premise of maintaining

ISSN: 2664-9640

the accuracy of the HRnet network model, the convolution layer will adopt the depth separable convolution[14] , which greatly reduces the network parameters, reduces the model volume and improves the running speed. The improved activation function L-swish is used, and the Squeeze and Exception (SE) module[15] is introduced to further improve the accuracy of the model.

2.1. Depth separable convolution

Depth separable convolution mainly decomposes standard convolution into depth convolution and pointwise convolution, which can greatly reduce the amount of parameters and calculation without much loss of accuracy. The decomposition process is shown in Figure 2.

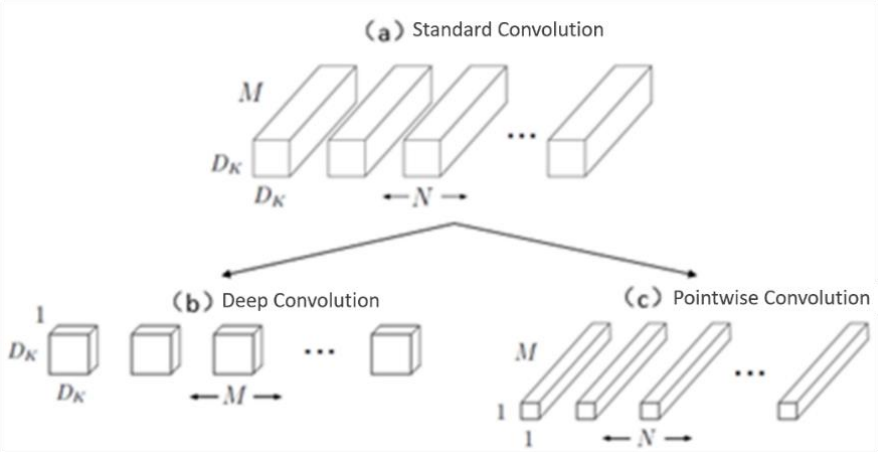

Figure 2: Standard convolution decomposition process

As shown in Figure 2, it is assumed that the size of input feature map f of neural network is represented by (D_F, D_F,M).

If the standard convolution method is used, that is, as shown in part (a) of Fig. 2, assuming that the size of standard convolution K is (D_K, D_K,M,N) and the size of the output feature map G is (D_G, D_G,N), then the calculation formula of standard convolution can be expressed by formula (1):

$$G_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m} \quad (1)$$

According to the above formula, the input channel number is M, the output channel number is N, and the corresponding calculation amount is: $D_k \cdot D_k \cdot M \cdot N \cdot D_F \cdot D_F$.

If the depth separable convolution is used, the depth convolution of part (b) and the pointwise convolution of part (c) are separated. The former is responsible for filtering and the latter is responsible for conversion channel. If the depth convolution size is set to (D_K, D_K,1,M) and the powise convolution size is set to (1, 1,M,N), then the depth separable convolution can be expressed by formula (2):

$$\hat{G}_{k,l,n} = \sum_{i,j} \hat{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m} \quad (2)$$

According to formula (2), the computational complexity of deep separable convolution is obtained as follows: $D\_K \cdot D\_K \cdot M \cdot D\_F \cdot D\_F + M \cdot N \cdot D\_F \cdot D\_F$.

Compared with the traditional standard convolution, the computational complexity of the two convolution methods is reduced:

$$\frac{D_k \cdot D_k \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_k \cdot D_k \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2} \quad (3)$$

It can be seen that the use of depth separable convolution in HRnet will greatly reduce the number of parameters in the network, thus reducing the volume of network model and providing calculation speed, which has great application value in the construction scene with shortage of resources and complex environment.

ISSN: 2664-9640

2.2. L-Swish activation function

As a common activation function in neural network, ReLU function can map the input end of neuron to the output end, which has the advantage of fast convergence. However, when the ReLU gradient is 0 when x is less than 0, the negative gradient will be set to zero under the action of the activation function ReLU, and the neuron may never be activated by any data again, resulting in neuron "Necrosis". This forced sparse processing will reduce the effective capacity of the network model and shield too many features, resulting in the model unable to learn effective features[16] . In order to avoid the problem of ReLU activation function, Swish function is proposed. Experiments show that Swish[17] activation function is a better nonlinear activation function than ReLU, as shown in formula (4):

$$Swish[x] = x \cdot sigmoid(\beta x) \tag{4}$$

Among them, β is a constant or trainable parameter. Swish has the characteristics of no upper bound, next generation, smooth and non monotonic. But compared with ReLU, it is more complex because it contains sigmoid function. In order to further improve the accuracy of the model and reduce its computational cost, the piecewise function L-Sigmoid (as shown in formula (5)) is used to simulate the sigmoid function. The comparison effect is shown in Fig. 3.

$$l - sigmoid = \frac{\min(\max(\alpha x, x+3), 6)}{6} \tag{5}$$

Where α = 0.01. The improved Swish function is shown in formula (5):

$$l - swish[x] = x\frac{\min(\max(\alpha x, x+3), 6)}{6} \tag{6}$$

Figure 3 shows the comparison of sigmoid and l-sigmoid activation functions. Figure 4 shows a comparison of swish and L-swish activation functions.

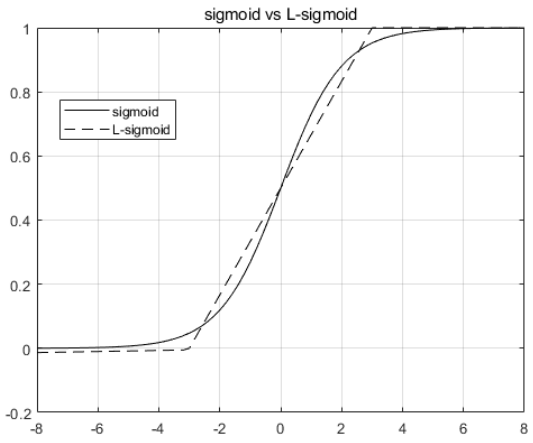

Figure 3: sigmoid vs L-sigmoid

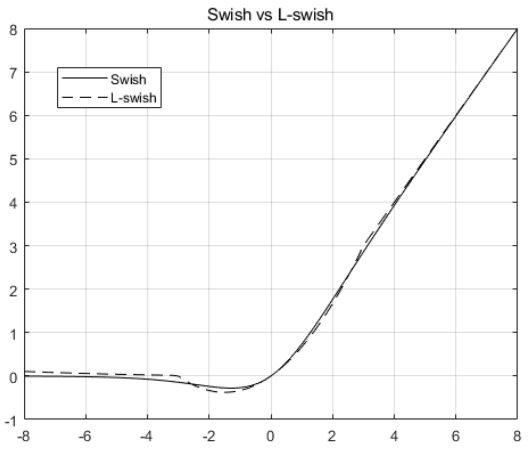

Figure 4: Swish vs L-swish

Considering the cost of applying nonlinear activation function, we apply L-swish to the parallel low-resolution subnet in the model design.

2.3. Squeeze and Exception module

The SE module can learn and construct the feature weight according to the target loss function loss. Finally, the effectiveness of the feature map is positively correlated with the weight value. This training method further improves the accuracy of the model.

ISSN: 2664-9640

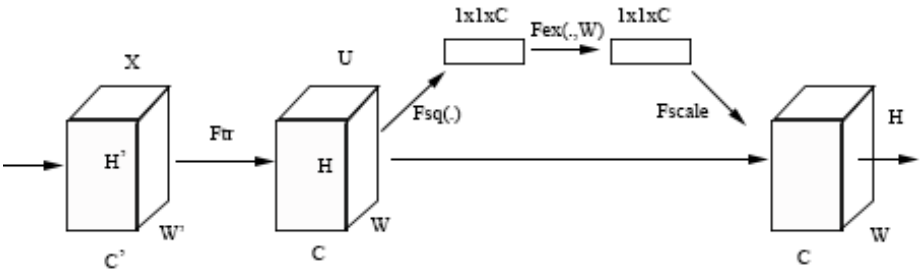

Figure 5: SE model structure diagram

Figure 5 shows the structure of SEnet. In the figure, F_{tr} is a traditional convolution structure, and X and U are the input ($C'_xH'_xW'$) and the output (C_xH_xW) of F_{tr}, respectively. Different from the general convolution neural network, SEnet recalibrates the previously acquired features by Squeeze, Exception and Reweigh, as described below.

First, Squeeze operation (Fsq (·)) is used to compress features along the spatial dimension. Each two-dimensional feature channel is transformed into a real number, which has a global receptive field to some extent, and the output dimension matches the number of input feature channels. It represents the global distribution of the response on the feature channel, and makes the layer near the input obtain the global receptive field, which makes the extracted features more accurate. The second is the Exception operation (Fex (·) in the figure), which is a mechanism similar to the gate in cyclic neural networks. The weight of each feature channel is generated by the parameter W, where the parameter W is learned to explicitly model the correlation between feature channels. Finally, reweight operation is performed. The weight of the output of exception is regarded as the importance of each feature channel after feature selection, and then the weight is weighted to the previous feature channel by channel through multiplication to complete the recalibration of the original feature in the channel dimension [18].

2.4. SE module algorithm

As shown in Figure 5, first of all, F_{tr} is a conversion operation. The definition of input and output is as follows:

$$F_{tr}: X \to U, X \in R^{H' \times W' \times C'}, U \in R^{H \times W \times C} \tag{7}$$

Then the expression of Ftr is shown in formula (8):

$$u_c = V_c * X = \sum_{s=1}^{C'} V_c^s * X^z \tag{8}$$

The U obtained by F_{tr} is the second three-dimensional matrix on the left in Fig. 5, which is called C characteristic graphs with the size of H * W, also known as vectors. U_C is the C-th two-dimensional matrix in U, and the subscript c is the channel.

Next is the Squeeze operation. The formula is essentially the global average pooling operation:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{9}$$

Therefore, formula (8) converts the input of H * W * C into the output of 1 * 1 * C, corresponding to the Fsq operation in Fig. 5. The results of this step show the numerical distribution of the C characteristic graphs in this layer, also known as global information.

Finally, the Exception operation is implemented with two full connection layers, as shown in formula (9). The result of the squeeze operation in front of the last equal sign is z. here we multiply W_1 by z, which is a full connection layer operation. The dimension of W_1 is C / (r * C), and the parameter r represents the compression ratio. The purpose of r is to reduce the number of channels and thus reduce the amount of calculation. Because the dimension of Z is 1 * 1 * C, the result of W_{1z} is 1 * 1 * C / r; then, through a relu layer, the output dimension remains

ISSN: 2664-9640

unchanged; then, the output result is fully connected with W2, that is, multiplication operation is performed. Because the dimension of W_2 is C * C / r, the dimension of output result is 1 * 1 * C; finally, through the sigmoid function, s is obtained.

$$s = F_{ex}(z, w) = \sigma\big(g(z, w)\big) = \sigma(W_2 \delta(W_1 z)) \tag{10}$$

In other words, the dimension of s is 1 * 1 * C, and C is the number of channels, which is used to describe the weight of C characteristic graphs in vector U. Moreover, this weight is obtained by learning the full connection layer and nonlinear layer, so the end-to-end training can be carried out. The function of these two full connection layers is to fuse the feature map information of each channel, because the previous squeeze is operated in the characteristic graph of a certain channel. Finally, channel multiplication is performed on the initial vector U by formula (9), which is the fscale process in Fig. 5:

$$\widetilde{X_c} = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{11}$$

Where u_c is the two-dimensional matrix and s_c is the weight value. In this paper, SE module is introduced into HRnet network to further improve the accuracy of the algorithm.

3. Improved Pose-HRnet

Through the analysis and optimization of deep separable convolution, L-swish activation function and SE module, this paper proposes an improved network structure, Pose-HRnet, based on the original network structure HRnet. In addition, the Pose-HRnet is used to separate the volume resolution of the network. The network reduces the computational complexity and improves the accuracy of the model. The improved module is shown in Figure 6. For each data channel, a weight is used to indicate the importance of the channel in the next stage.

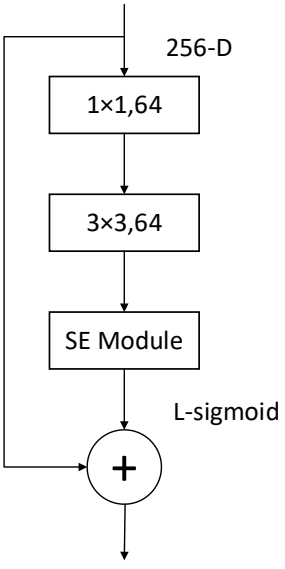

Figure 6: SE-Bottleneck module

In order to effectively understand the performance of the proposed model, we use our own data set PHA to train the Pose-HRnet model. The data set PHA obtains the operation behavior diagram related to high-altitude operation through web crawler, and classifies all kinds of behaviors into dangerous behavior and safety behavior according to the actual needs. The input image size is set as 256x256x3, then the high-resolution subnet is taken as the first stage, and the high-resolution subnet is gradually added to form a new stage. The multi-resolution subnet is connected in parallel, and the switching unit across the parallel subnet is introduced, so that each subnet can repeatedly receive information from other parallel subnets Finally, the images are divided into two categories by softmax classifier, and the recognition results of human actions are output. Table 1 shows the Pose-HRnet structure diagram, where input and output

ISSN: 2664-9640

represent the number of input and output channels, operate represents network operation, SE represents whether to add SE module, NL represents activation function adopted by network, SE refers to ReLU activation function, Le represents improved L-swish activation function, B represents number of blocks, fuse indicates whether the module is integrated.

Table 1: Improved Pose-HRnet network structure diagram

Input	Operate	Out	SE	NL	B	Fuse
3	Conv2d	64	-	-	-	False
64	Conv2d	64	-	RE	-	False
64	Bottleneck	256	√	LS	4	False
256	Conv2d	32 64	-	RE	-	False
32 64	BasicBlock	32 64	-	LS RE	4 4	True
32 64 128	Conv2d	32 64 128	-	LS RE RE	-	True
32 64 128 256	BasicBlock	17	-	LS LS RE RE	4	True
17	Conv2d	2	-	-	-	-

4. Experimental results and analysis

This experiment adopts pytorch-1.8, Intel (R) Xeon (R) CPU e5-2630 V4 @ 2.20GHz architecture, and NVIDIA Titan x Pascal is used for GPU. The personnel behavior classification experiment was conducted on the data set PHA which was made by ourselves. The data source of PHA was mainly obtained from Google pictures and Baidu image websites through web crawler technology, whether the personnel were wearing safety ropes and whether they were in a state of fatigue, as shown in pytorch Fig. 7 and Fig. 8.

ISSN: 2664-9640

Figure 7: Example diagram of PHA security behavior of data set

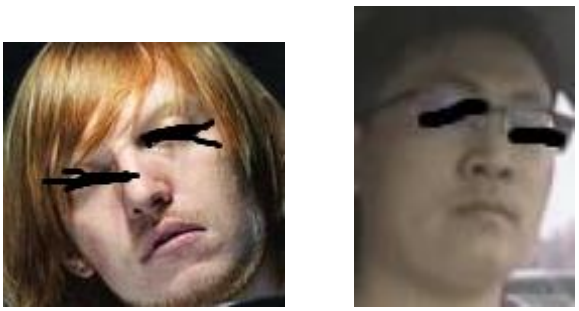

Figure 8: Example of PHA dangerous behavior in data set

In order to further verify the improved network, the improved Pose-HRnet network model and HRnet network model are used to test the human behavior classification on the data set PHA. The identification accuracy and parameter quantity of the two network models are tested and compared. The test results are shown in Table 2. The experimental results show that the improved Pose-HRnet network model can reduce the parameters without damaging the accuracy, which is of great significance to the deployment of network model in some embedded devices with limited computing resources.

Table 2: Comparison results of two network models

Model	Input image size	Recognition accuracy	Parameter quantity（M）
HRnet	128x128x3	95.4%	81.7
Pose_HRnet	128x128x3	98.5%	54.6

5. Conclusion

Aiming at the task of high-altitude workers' dangerous behavior recognition, this paper proposes an improved deep convolution neural network model based on HRnet, and uses the improved network model to test on the data set PHA. The test results show that the accuracy of the improved network is improved, the number of parameters is greatly reduced, and the model has a smaller volume, which makes the model embedded in It is possible to apply it in the portable equipment. At present, deep convolution neural network algorithm has been successfully applied in image recognition, image segmentation and other fields. However, if it is to be deployed in the construction industry, the acquisition of dangerous action data set of construction workers, deployment of cameras and consumption of computing resources for model operation need to be further solved. In the application scenarios of construction industry, the accuracy and real-time performance of the model are required. Although the proposed scheme has improved the operation speed, there is still a long way to go for practical application.

ISSN: 2664-9640

Therefore, further research on model compression and accuracy improvement can be done in the future.

References

[1] Sun K, Xiao B, Liu D, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[J]. arXiv preprint arXiv:1902.09212, 2019.

[2] Liu xinlai. Analysis of the causes of safety accidents in construction and preventive measures [J]. Science and technology outlook, 2014 (21): 66-67.

[3] Rao LAN, Zhang Xia. Cause analysis of construction safety accidents in China [J]. Pearl River modern construction, 2010 (2): 32-34

[4] Heinrich, H. W. Industrial accident prevention: a scientific approach[J]. 2011, 4(4):609-609.

[5] Wu Junxian. Safety management measures for fall prevention [J]. China new technology and new products, 2009 (18): 168-168

[6] Melo, Roseneia Rodrigues Santos etc. "Applicability of unmanned aerial system (UAS) for safety inspection on construction sites." Safety Science 98:174-185.

[7] Jia Guangshe, he Changquan, Chen Yuting, et al. Early warning of safety behavior of construction workers from a cross level perspective [J]. Journal of Tongji University (NATURAL SCIENCE EDITION), 2018, 47 (04)

[8] Chen Y , Jiang H , Li C , et al. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(10):6232-6251.

[9] Yin Jianjie. Review and application of logistic regression model analysis [D]. Heilongjiang University, 2011

[10] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[J]. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4724-4732.

[11] Ma Z , Yang X , Zhang H . Dangerous Driving Behavior Recognition using CA-CenterNet[C]// 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, 2021.

[12] Newell A, Yang K, Deng J. Stacked Hourglass Networks for Human Pose Estimation[J]. Computer Vision and Pattern Recognition, 2016, pp 483-499.

[13] Gai Li, Jing Guodong. Behavior recognition based on multiscale method combined with convolutional neural network [J]. Computer engineering and applications, 2019, 55 (2): 100-103

[14] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. CoRR, abs/1704.04861 (2017).

[15] Jie H, Li S, Albanie S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99):1-1.

[16] Tian Juan, Li Yingxiang, Li Tongyan. Comparative study of activation function in convolutional neural networks [J]. Computer system applications, 2018, v.27 (07): 45-51

[17] Mi Shuo, Tian Fengshou, sun Ruibin, et al. Performance of swish activation function on small and medium-sized data sets [J]. Science and technology innovation and application, 2018 (1): 4-5

[18] Shao Jiaqi, Qu Changwen, Li Jianwei. Performance analysis of convolutional neural network for SAR target recognition [J]. Radar science and technology, 2018, 16 (05): 65-72.