

AttentionNet: An Efficient Scheme for Human Activity Recognition

Wei Yang¹ , Xiaojun Jing², Hai Huang², Chao Li² and Botao Feng³

¹College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China

²School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

³College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China

yangwei@sztu.edu.cn

Abstract. To solve the problem of noise interference in two-dimensional radar signals, an efficient human behavior recognition scheme based on attention mechanism and convolutional neural network (CNN) is proposed. This algorithm combines attention mechanism and spatial pyramid pooling (SPP) layer with CNN, and fuses hierarchical feature maps generated by the network to reduce noise interference. By comparing the performance of proposed method with that of common CNNs, the experimental results show that the proposed scheme acts more effectively under different noises. Especially, when the signal-to-noise ratio (SNR) is higher than -10dB, an accuracy rate of more than 90% could be achieved.

Keywords: Human activity recognition; Radar signal; Attention mechanism; CNN

1 Introduction

Human behavior recognition primarily integrates human movement data through a variety of information media, and the generated model can be applied to circumstances in which human behavior recognition may be required. Human behavior recognition is a vital technology with many applications in human-computer interface and human behavior analysis. Researchers' attention has been focused to human behavior recognition technologies in recent years due to advancements in technology and growing application requirements.

Human behavior recognition technology can be categorized into three groups depending on various data gathering mediums, that is, visual images, sensors, and radar signals. Human behavior recognition techniques based on visual images are extremely sensitive to changes in the environment and have strict criteria for application circumstances and light photography angles because of the influence of optical properties. Besides, sensor-based human behavior detection solutions necessitate wearing equivalent sensors due to limits in signal gathering methods, which significantly reduces convenience. Therefore, the convenience of deployment and robustness under different

scenarios in radar-based human behavior recognition technology bring substantial benefits in reality.

Radar-based human behavior recognition algorithms mainly include traditional machine learning methods and deep learning methods. Most of these two methods are based on feature extraction based on micro-Doppler diagrams converted from radar images. At present, domestic and foreign scholars have proposed different solutions to improve the classification accuracy and efficiency of micro-Doppler spectrograms, including manual feature extraction methods and end-to-end automatic classification methods. In the early stages of research, there are also some algorithms trying to extract features from radar echo signals [1], such as spectral analysis method [2, 3], independent component analysis method [4], empirical mode decomposition method, support vector machine [5], etc. However, these algorithms heavily rely on the prior knowledge of the micro-Doppler effect, the accuracy and efficiency of their classification are also related to the feature selection as well as extraction methods. For the problem of human behavior recognition, some scholars have also proposed different solutions, including using cameras to obtain high-resolution natural image sequences [6]. This vision-based method is more sensitive to the surrounding environment and strongly depends on lighting conditions. Compared with optical systems, radar images are more robust and can overcome imaging difficulties caused by weather [7], they can also measure target distances and detect small frequency changes corresponding to target speeds [8].

Typically, recognition algorithms based on human activity radar data are noise-sensitive and perform poorly in real time. This research proposes a human behavior detection scheme based on attention and a convolutional neural network to address the problem of noise interference in two-dimensional radar signals. To reduce noise interference, this algorithm incorporates attention mechanism with spatial pyramid pooling layer into network topology, as well as fuses the network's hierarchical feature maps.

2 Deep Learning

2.1 CNN

CNN is a one-way neural network. As the number of network layers increases, more detailed features can be recovered. It has a deep network structure and performs convolution calculations. The proposal and application of convolutional neural networks are mostly focused on visual systems. The convolutional layer continuously filters the information in the image to generate a receptive field, which is then processed by the neural network for classification.

The convolution kernel, a parameter matrix with a typical size of 3×3 or 5×5 , is the most crucial component of the convolutional layer. The image is actually a computerized matrix. The matrix could be a three-dimensional RGB image or a two-dimensional grayscale image. The process of convolution involves scanning the image with a specific step size. Every step ahead will produce a result of the convolution process, which will then be mapped to the appropriate location in the feature map. The feature map of the image is the output result of the convolution operation performed by this convolution kernel when it has scanned every place in the image.

The convolution operation process also reveals that the convolution kernel primarily extracts the portion of the image that most closely resembles its own texture. The

resulting feature map is the corresponding feature map, which can represent the portion of the original image that most closely resembles the part distribution of the convolution kernel.

2.2 Attention Mechanism

Deep learning researchers have paid close attention to the attention mechanism in recent years, and it is crucial to many deep learning applications. This paper applies the attention mechanism to human behavior recognition based on convolutional neural networks to increase the robustness to noise. The attention mechanism, as its name implies, is a mechanism that affects neural networks' attention. The principle is that the brain naturally assigns different attention to distinct sections or paragraphs depending on experience when a person observes a certain image or reads a particular passage of text. The area that receives the majority of attention typically contains a lot of information, which conserves resources. The attention mechanism efficiently distributes the neural network's computational resources by imitating the way human brain assigns to various data locations and converting them into weights of different sizes. Usually, for a neural network with attention mechanism could be able to discern the importance of data, these weights need to be learnt during network training. As a result, input with a higher influence on the outcome will be given more weight, and vice versa. The Self-Attention structure in Fig. 1 illustrates how the neural network can extract more significant information from a large amount of input data by adjusting weights in this manner.

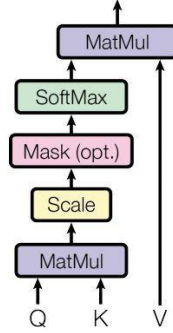


Fig. 1. Self-Attention structure.

As shown in Fig. 1, the inputs Q , K , and V of the attention mechanism correspond to different matrices, and the matrices will be updated along with the network. The weight calculation of the attention mechanism is showed in Eq. (1). Firstly, matrix multiplication is performed on Q and K . Then, the results will be controlled by a coefficient. Finally, the weight distribution is obtained by converting it into a probability value through the *Softmax* function and multiplying it with V .

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

It can be seen from Eq. (1) that attention mechanism will finally obtain a normalized probability distribution, and its output can be used as a method for neural network to re-adjust the data weight. In the convolutional neural network, we will use attention mechanism to adjust the weight of the feature map to improve the model's robustness to noise.

3 Model

Two-dimensional radar emissions can be used to recognize human behavior with a high degree of accuracy. The micro-Doppler time-frequency diagrams are often processed as images in the human behavior recognition process. Therefore, the noise produced during movement will unavoidably result in interference. By using short-time Fourier transform (STFT) and further processing, the high resolution range profile (HRRP) produces the micro-Doppler time-frequency diagram. The principle is that micro-Doppler will be responsible for the echo that is returned to the radar when a person moves relative to the radar. On micro-Doppler time-frequency diagrams, the micro Doppler effect results in frequency shifts that can be utilized as characteristics to categorize action behaviors. Since CNN has a good effect in image visual feature extraction, this article will introduce an attention mechanism based on convolutional neural network to reduce noise interference in radar images, thereby boosting the robustness of human behavior identification algorithm.

3.1 Network Structure Design

Table 1 gives each layer's detailed parameters of the network in proposed algorithm, which mainly include five sections.

Table 1. CNN network structure based on attention mechanism.

Conv	1×1 , C = 32
DB 1	3×3 conv, P = 1, C = 32, InsNorm, LeakyReLU(0.2)
	1×1 conv, C = 32, InsNorm, LeakyReLU(0.2)
CA 1	1×1 conv, C = 4, ReLU
	1×1 conv, C = 32, Sigmoid
Pool 1	2×2 pool, max pooling
	Output = 7×7 , spatial pyramid pooling
DB 2	3×3 Conv, P = 1, C = 64, InsNorm, LeakyReLU(0.2)
	1×1 Conv, C = 64, InsNorm, LeakyReLU(0.2)
CA 2	1×1 Conv, C = 8, ReLU
	1×1 conv, C = 64, Sigmoid
Pool 2	2×2 Max pooling
	Output = 7×7 Spatial pyramid pooling
DB 3	3×3 Conv, P = 1, C = 128, InsNorm, LeakyReLU(0.2)
	1×1 Conv, C = 128, InsNorm, LeakyReLU(0.2)
CA 3	1×1 Conv, C = 16, ReLU
	1×1 Conv, C = 128, Sigmoid
Pool 3	2×2 Max pooling
	Output = 7×7 Spatial pyramid pooling
DB 4	3×3 Conv, P = 1, C = 256, InsNorm, LeakyReLU(0.2)
	1×1 Conv, C = 256, InsNorm, LeakyReLU(0.2)
CA 4	1×1 Conv, C = 32, ReLU
	1×1 Conv, C = 256, Sigmoid
Pool 4	2×2 Max pooling
	Output = 7×7 Spatial pyramid pooling
DB 5	3×3 Conv, P = 1, C = 512, InsNorm, LeakyReLU(0.2)
	1×1 Conv, C = 512, InsNorm, LeakyReLU(0.2)
CA 5	1×1 Conv, C = 64, ReLU
	1×1 Conv, C = 512, Sigmoid
Pool 5	2×2 Max pooling
	Output = 7×7 Spatial pyramid pooling
DB 6	3×3 Conv, P = 1, C = 1024, InsNorm, LeakyReLU(0.2)
	1×1 Conv, C = 1024, InsNorm, LeakyReLU(0.2)
CA 6	1×1 Conv, C = 128, ReLU
	1×1 Conv, C = 1024, Sigmoid
Pool 6	2×2 Max pooling
	Output = 7×7 Spatial pyramid pooling
CA	1×1 Conv, C = 504, ReLU
	1×1 Conv, C = 4032, Sigmoid
Conv	1×1 Conv, C = 4032, InsNorm, LeakyReLU(0.2)
Pool	7×7 Max pooling
Conv	1×1 Conv, C = 6

- Convolution Layer: The 3×3 convolution kernel is used here, which functions as the convolution kernel's operating unit. Comparable to a filter, the convolution kernel moves with a step size of 3 on the time-frequency diagram. Every covered block is a potential location for feature extraction. LeakyReLU is the activation function among them, and its coefficient is 0.2.
- Pooling Layer: This article uses the maximum down sampling layer. Maximum pooling can retain as many key features in the image as possible, thereby effectively reducing the calculation amount of subsequent network layer convolutions.
- Spatial Pyramid Pooling (SPP) Layer: Since attention mechanism is added to the proposed solution, the size of the output feature maps of each layer is inconsistent.

The SPP layer [9] will be used for the fusion of hierarchical feature maps to solve inconsistency in size of the image.

- Dense network Layer: Dense blocks, which are mostly utilized for convolutional layer connections, develop from residual blocks. Studies [10, 11] has demonstrated that substituting cross-connection blocks for convolutional layers can lessen overfitting and enhance neural network performance.
- Attention Mechanism: The attention mechanism is used to model feature maps in different channels [12] and learn through other layers for recalibration. We additionally substitute convolutional layers for the linked layers in original attention module in order to lessen overfitting.

3.2 Network Training

Firstly, the data is preprocessed to convert the HRRP matrix into a time distance image, as shown in Fig. 2. Behavioral radar returns distance information since the HRRP matrix's columns contain time information. Consequently, a temporal distance graph represents the corresponding image.

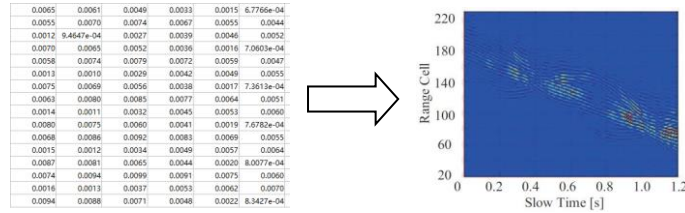


Fig. 2. HRRP matrix converted into time distance graph.

The time distance picture is then turned into a Doppler time-frequency image using STFT in Fig. 3, which can handle the Fourier transform problem of non-stationary signals, and the next computation method will use window function. Assuming that the signal received by window function is stationary, the power spectrum at each different moment could be determined by performing a rapid Fourier transform and sliding the window with a given step size. The Doppler time-frequency graphic is then blended using a sliding window sequence.

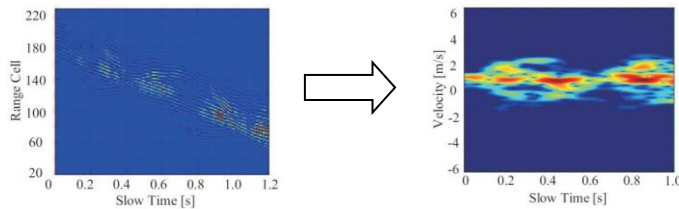


Fig. 3. Time distance diagram to Doppler time-frequency diagram.

We increase the data by improving the sliding window, for the reason that the variety and volume of data impact the network's performance. To create time-frequency graphs, sliding windows with varying widths are slid across the HRRP matrix. The

short-time Fourier transform is carried out concurrently. Ultimately, for each action, 2000 Doppler time-frequency maps were generated. Among them, the training dataset and test dataset is 3:1, which indicates that there are 1500 data in training set and 500 in test set.

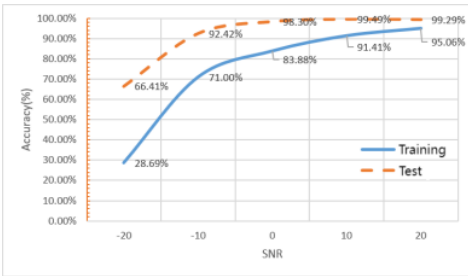
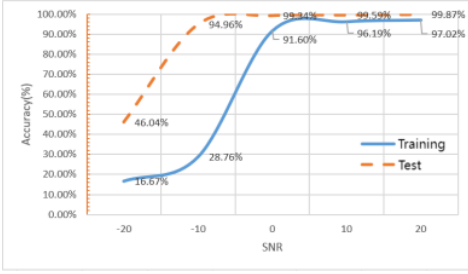
Usually, the convolutional neural network's computing speed is influenced by the size of the images. Therefore, each Doppler time-frequency image needs to be resized to 100*100 before addicted to the network. The training set is split up into a group of 256 at the same time and submitted into the network to be trained. The training hardware is built on NVIDIA 1080Ti GPU and CUDNN acceleration framework, while the network structure uses PyTorch.

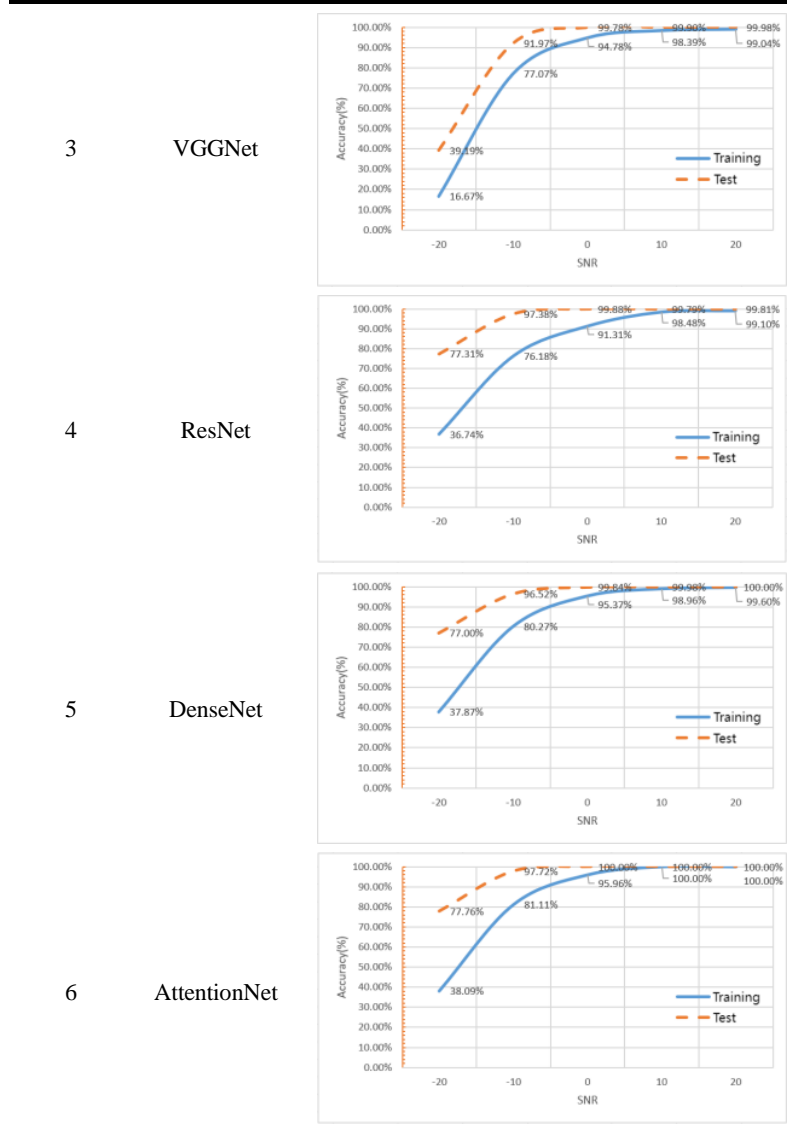
4 Experiments

The experiment is carried out using Matlab platform, in which the radar center frequency is 4GHz, pulse repetition frequency is 600Hz, fast sampling time is 65ns, and the fast sampling frequency is 3GHz. By adding different level of Gaussian noises to training set, the models are trained separately. Experiment contains four different levels of noise, forming a total of five types of SNR, namely, -20dbm, -10dbm, 0, 10dbm, and 20dbm. In addition, five convolutional neural networks, that is, KimNet [13], AlexNet [14], VGGNet [15], ResNet [16] and DenseNet [17], were selected for comparative experiments.

Table 2 compares the accuracy of six convolutional neural networks mentioned above in training and test datasets under various SNR.

Table 2. Comparison of recognition accuracy rates with different CNN and different SNR.

Number	Type	Accuracy Comparison
1	KimNet	
2	AlexNet	



It can be seen that, in Table 2, the recognition accuracy of six convolutional neural networks improves with the increase in SNR. This is because the SNR has a direct impact on network's efficiency and quality while extracting feature maps. In the case of low SNR, noise will influence the retrieved features, causing interference with final recognition effect. Compared with other five networks, the proposed convolutional neural network based on attention mechanism is more resilient to noise.

Under different SNR conditions, the accuracy performance comparison of six convolutional neural networks in training dataset and test dataset could be illustrated in Table 3.

Table 3. Comparison of recognition accuracy rates with different CNN and same SNR.

Number	SNR (dbm)	Accuracy Comparison																					
1	-20	<table border="1"> <caption>Accuracy Comparison at SNR -20 dbm</caption> <thead> <tr> <th>Model</th> <th>Training Accuracy (%)</th> <th>Test Accuracy (%)</th> </tr> </thead> <tbody> <tr> <td>KimNet</td> <td>28.69%</td> <td>66.41%</td> </tr> <tr> <td>AlexNet</td> <td>16.67%</td> <td>46.04%</td> </tr> <tr> <td>VGGNet</td> <td>16.67%</td> <td>39.19%</td> </tr> <tr> <td>ResNet</td> <td>36.74%</td> <td>77.31%</td> </tr> <tr> <td>DenseNet</td> <td>37.87%</td> <td>77.70%</td> </tr> <tr> <td>AttentionNet</td> <td>38.09%</td> <td>77.76%</td> </tr> </tbody> </table>	Model	Training Accuracy (%)	Test Accuracy (%)	KimNet	28.69%	66.41%	AlexNet	16.67%	46.04%	VGGNet	16.67%	39.19%	ResNet	36.74%	77.31%	DenseNet	37.87%	77.70%	AttentionNet	38.09%	77.76%
Model	Training Accuracy (%)	Test Accuracy (%)																					
KimNet	28.69%	66.41%																					
AlexNet	16.67%	46.04%																					
VGGNet	16.67%	39.19%																					
ResNet	36.74%	77.31%																					
DenseNet	37.87%	77.70%																					
AttentionNet	38.09%	77.76%																					
2	-10	<table border="1"> <caption>Accuracy Comparison at SNR -10 dbm</caption> <thead> <tr> <th>Model</th> <th>Training Accuracy (%)</th> <th>Test Accuracy (%)</th> </tr> </thead> <tbody> <tr> <td>KimNet</td> <td>71.00%</td> <td>92.42%</td> </tr> <tr> <td>AlexNet</td> <td>28.67%</td> <td>94.96%</td> </tr> <tr> <td>VGGNet</td> <td>77.07%</td> <td>96.97%</td> </tr> <tr> <td>ResNet</td> <td>76.18%</td> <td>97.38%</td> </tr> <tr> <td>DenseNet</td> <td>80.27%</td> <td>96.52%</td> </tr> <tr> <td>AttentionNet</td> <td>81.11%</td> <td>97.72%</td> </tr> </tbody> </table>	Model	Training Accuracy (%)	Test Accuracy (%)	KimNet	71.00%	92.42%	AlexNet	28.67%	94.96%	VGGNet	77.07%	96.97%	ResNet	76.18%	97.38%	DenseNet	80.27%	96.52%	AttentionNet	81.11%	97.72%
Model	Training Accuracy (%)	Test Accuracy (%)																					
KimNet	71.00%	92.42%																					
AlexNet	28.67%	94.96%																					
VGGNet	77.07%	96.97%																					
ResNet	76.18%	97.38%																					
DenseNet	80.27%	96.52%																					
AttentionNet	81.11%	97.72%																					
3	0	<table border="1"> <caption>Accuracy Comparison at SNR 0 dbm</caption> <thead> <tr> <th>Model</th> <th>Training Accuracy (%)</th> <th>Test Accuracy (%)</th> </tr> </thead> <tbody> <tr> <td>KimNet</td> <td>83.88%</td> <td>98.30%</td> </tr> <tr> <td>AlexNet</td> <td>91.60%</td> <td>99.34%</td> </tr> <tr> <td>VGGNet</td> <td>94.78%</td> <td>99.78%</td> </tr> <tr> <td>ResNet</td> <td>91.31%</td> <td>99.88%</td> </tr> <tr> <td>DenseNet</td> <td>95.37%</td> <td>99.84%</td> </tr> <tr> <td>AttentionNet</td> <td>95.96%</td> <td>100.00%</td> </tr> </tbody> </table>	Model	Training Accuracy (%)	Test Accuracy (%)	KimNet	83.88%	98.30%	AlexNet	91.60%	99.34%	VGGNet	94.78%	99.78%	ResNet	91.31%	99.88%	DenseNet	95.37%	99.84%	AttentionNet	95.96%	100.00%
Model	Training Accuracy (%)	Test Accuracy (%)																					
KimNet	83.88%	98.30%																					
AlexNet	91.60%	99.34%																					
VGGNet	94.78%	99.78%																					
ResNet	91.31%	99.88%																					
DenseNet	95.37%	99.84%																					
AttentionNet	95.96%	100.00%																					
4	10	<table border="1"> <caption>Accuracy Comparison at SNR 10 dbm</caption> <thead> <tr> <th>Model</th> <th>Training Accuracy (%)</th> <th>Test Accuracy (%)</th> </tr> </thead> <tbody> <tr> <td>KimNet</td> <td>91.41%</td> <td>99.49%</td> </tr> <tr> <td>AlexNet</td> <td>96.19%</td> <td>99.59%</td> </tr> <tr> <td>VGGNet</td> <td>96.39%</td> <td>99.90%</td> </tr> <tr> <td>ResNet</td> <td>98.48%</td> <td>99.79%</td> </tr> <tr> <td>DenseNet</td> <td>98.80%</td> <td>99.98%</td> </tr> <tr> <td>AttentionNet</td> <td>99.98%</td> <td>100.00%</td> </tr> </tbody> </table>	Model	Training Accuracy (%)	Test Accuracy (%)	KimNet	91.41%	99.49%	AlexNet	96.19%	99.59%	VGGNet	96.39%	99.90%	ResNet	98.48%	99.79%	DenseNet	98.80%	99.98%	AttentionNet	99.98%	100.00%
Model	Training Accuracy (%)	Test Accuracy (%)																					
KimNet	91.41%	99.49%																					
AlexNet	96.19%	99.59%																					
VGGNet	96.39%	99.90%																					
ResNet	98.48%	99.79%																					
DenseNet	98.80%	99.98%																					
AttentionNet	99.98%	100.00%																					
5	20	<table border="1"> <caption>Accuracy Comparison at SNR 20 dbm</caption> <thead> <tr> <th>Model</th> <th>Training Accuracy (%)</th> <th>Test Accuracy (%)</th> </tr> </thead> <tbody> <tr> <td>KimNet</td> <td>95.06%</td> <td>99.29%</td> </tr> <tr> <td>AlexNet</td> <td>97.02%</td> <td>99.87%</td> </tr> <tr> <td>VGGNet</td> <td>99.04%</td> <td>99.98%</td> </tr> <tr> <td>ResNet</td> <td>99.10%</td> <td>99.81%</td> </tr> <tr> <td>DenseNet</td> <td>99.60%</td> <td>100.00%</td> </tr> <tr> <td>AttentionNet</td> <td>100.00%</td> <td>100.00%</td> </tr> </tbody> </table>	Model	Training Accuracy (%)	Test Accuracy (%)	KimNet	95.06%	99.29%	AlexNet	97.02%	99.87%	VGGNet	99.04%	99.98%	ResNet	99.10%	99.81%	DenseNet	99.60%	100.00%	AttentionNet	100.00%	100.00%
Model	Training Accuracy (%)	Test Accuracy (%)																					
KimNet	95.06%	99.29%																					
AlexNet	97.02%	99.87%																					
VGGNet	99.04%	99.98%																					
ResNet	99.10%	99.81%																					
DenseNet	99.60%	100.00%																					
AttentionNet	100.00%	100.00%																					

The line chart in Table 2 and the bar chart in Table 3 above represent the experimental findings from training and test datasets, respectively. It is clear that the identifi-

cation accuracy of six networks will be higher when the SNR is high. Among them, when the SNR is -10dbm, the proposed technique may achieve more than 90% accuracy in test dataset. AttentionNet outperforms the other convolutional neural networks in terms of recognition accuracy. This is because the attention mechanism has potential to improve CNN's adaptive feature learning capabilities, as well as enhance the network's overall robustness and generalization.

5 Conclusion

Human behavior recognition has emerged as a research hotspot in recent years, which brings promising application potential and research value. Due to restrictions in application scenarios and algorithm performance, the development of new algorithms is urgently required. This paper provides an enhanced detection method for human activity radar signals that incorporates attention mechanism into convolutional neural network. Specifically, the convolutional neural network extract features from two-dimensional images better, while attention mechanism benefits the CNN to learn features adaptively. The results of comparative experiments demonstrate that the proposed scheme could reduce the impact of noise on recognition efficiency.

Acknowledgments. This work is supported by the General Program of Continuous Support Foundation of Shenzhen City (No. 20220715114600001), the Scientific Research Capacity Improvement Project from Guangdong Province (No. 2021ZDJS109), and the Shenzhen Post-doctoral Science Foundation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmed, S. and Cho, S.H.: Machine Learning For Healthcare Radars: Recent Progresses In Human Vital Sign Measurement and Activity Recognition. *IEEE Communications Surveys & Tutorials* (2023)
2. Xu, G., Zhang, B., Yu, H., Chen, J., Xing, M. and Hong, W.: Sparse Synthetic Aperture Radar Imaging From Compressed Sensing and Machine Learning: Theories, Applications, and Trends. *IEEE Geoscience and Remote Sensing Magazine*, 10(4), pp.32-69 (2022)
3. Abdu, F.J., Zhang, Y., Fu, M., Li, Y. and Deng, Z.: Application of Deep Learning On Millimeter-Wave Radar Signals: A Review. *Sensors*, 21(6), p.1951 (2021)
4. Bhavanasi, G., Werthen-Brabants, L., Dhaene, T. and Couckuyt, I.: Patient Activity Recognition Using Radar Sensors and Machine Learning. *Neural Computing and Applications*, 34(18), pp.16033-16048 (2022)
5. Ahmed, S., Kallu, K.D., Ahmed, S. and Cho, S.H.: Hand Gestures Recognition Using Radar Sensors For Human-Computer-Interaction: A Review. *Remote Sensing*, 13(3), p.527 (2021)

6. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G. and Liu, J.: Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), pp.3200-3225 (2022)
7. Farley, A. and Ham, H.: Real Time IP Camera Parking Occupancy Detection Using Deep Learning. *Procedia Computer Science*, 179, pp.606-614 (2021)
8. Rahman, M.M., Gurbuz, S.Z. and Amin, M.G.: Physics-Aware Generative Adversarial Networks For Radar-Based Human Activity Recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 59(3), pp.2994-3008 (2022)
9. He, K., Zhang, X., Ren, S. and Sun, J.: Spatial Pyramid Pooling In Deep Convolutional Networks For Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), pp.1904-1916 (2015)
10. He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning For Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778 (2016)
11. Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q.: Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700-4708 (2017)
12. Hu, J., Shen, L. and Sun, G.: Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132-7141 (2018)
13. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G. and Liu, J.: Human Action Recognition From Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), pp.3200-3225 (2022)
14. Krizhevsky, A., Sutskever, I. and Hinton, G.E.: Imagenet Classification With Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25 (2012)
15. Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks For Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014)
16. Girshick, R., Donahue, J., Darrell, T. and Malik, J.: Rich Feature Hierarchies For Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587 (2014)
17. Taigman, Y., Yang, M., Ranzato, M.A. and Wolf, L.: Deepface: Closing The Gap To Human-Level Performance In Face Verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701-1708 (2014)