IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 August 2022, accepted 18 August 2022, date of publication 25 August 2022, date of current version 31 August 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3201496*

RESEARCH ARTICLE

# Sleep Behavior Detection Based on Pseudo-3D Convolutional Neural Network and Attention Mechanism

**RUI GUO, CHAO ZHAI, (Member, IEEE), LINA ZHENG, AND LUYU ZHANG**
School of Information Science and Engineering, Shandong University, Qingdao 266237, China

Corresponding author: Lina Zheng (zhenglina@sdu.edu.cn)

This work was supported in part by the Shandong Provincial Natural Science Foundation under Grant ZR2020QF002.

**ABSTRACT** Good sleep quality is very important for everyone to protect physical and mental health. People's sleep behavior at night reflects their sleep status. In this paper, we propose a method to detect people's sleep behavior at night by adopting Pseudo-3D (P3D) convolution neural network with attention mechanism. In particular, we propose a new structure, which integrates Squeeze-and-Excitation (SE) blocks into P3D blocks, named P3D-Attention. For the input video, we use P3D blocks to extract spatial-temporal features, and use SE blocks to pay more attentions to the important channel features. The proposed network is tested on the Sleep Action (SA) dataset, which consists of five different actions, namely turn over, get up, fall off bed, play mobile phone, and normal sleep. Experimental results show that the proposed network achieves reasonably good detection results, and the accuracy rate on the test set can reach 90.67%. Compared with 3D convolutional neural networks (C3D), our proposed network can increase the accuracy by about 6% with only 1/6 model parameter size, and achieves an average prediction speed about 1.75 item/s. Compared with the residual spatiotemporal convolution network (R(2+1)D), our proposed network can increase the accuracy rate by about 1.5% with less than 1/2 model parameter size.

**INDEX TERMS** Pseudo-3D convolution, attention mechanism, spatial-temporal feature extraction, sleep behavior detection.

## I. INTRODUCTION

People's physical and mental health is closely related to sleep. Having sufficient sleep time and high-quality sleep is a necessary condition for everybody to keep a good state. According to the annual sleep report of China (2022), the overall sleep status of Chinese residents is not optimistic. In 2021, 67.5% respondents slept less than eight hours per day on average [1]. For some particular application scenarios, such as nursing home, elderly home endowment, child care, hospital, home disease rehabilitation, depression treatment, and prisons, etc., monitoring the personnel sleep state can reveal the body condition of specific people, so timely medical treatment and psychological counseling can be provided, which can effectively reduce the damage to health and avoid the occurrence of risk event.

Traditionally, the sleep status detection is evaluated by using the polysomnography (PSD) [2], [3]. However, only specialized technicians or doctors can draw the PSD, and testers need to wear a large number of heavy and complex equipments, which severely impairs the sleep comfort. Subsequently, the activity recorder is invented, which uses sensors to measure and record the wrist acceleration [4], [5]. Although, sensors can measure the number of turn over and judge the effective sleep time, testers still need to wear equipments during sleep. The high experimental cost and user-unfriendly wearing experience restricts its application. In recent years, benefiting from the great development of computer technology and the emergence of large-scale datasets, such as ImageNet [6], convolutional neural network (CNN) has made remarkable progress in image classification. Many deep learning models have emerged, such as VGGNet [7], GoogLeNet [8], Residual Network [9], are widely used in pattern recognition, artificial intelligence,

The associate editor coordinating the review of this manuscript and approving it for publication was Okyay Kaynak.

communications, bioinformatics and natural language processing, etc. In bioinformatics and computational biology, deep learning models have been implemented to identify DNA 6mA sites from the cross-species genome [10], and identify Flavin mono-nucleotide (FMN) biding sites in electron transport chains [11]. It is natural to use CNNs for the video recognition. Video is a series of continuous image frames playing at a certain speed. Different from images, video contains not only spatial information of behaviors and scenes, but also temporal information of dynamic evolutions of behaviors. Video behavior recognition based on 3D convolutions has been widely studied, as it can improve the accuracy of human action recognition. 3D convolution is an extension of 2D convolution, which extends the spatial convolution to the spatial-temporal convolution, to simultaneously extract both spatial and temporal features. However, it faces some serious issues, such as large amount of computation, high memory consumption, and high requirements on hardware performance, etc.

In this work, we propose a novel neural network architecture based on Pseudo-3D (P3D) block and attention mechanism to detect sleep behaviors of people. The network is composed by a dual-channel structure, which uses P3D blocks instead of the residual structure. We also add a time dimension in the squeeze-and-excitation (SE) block, which is palced after the $1\times3\times3$ spatial convolution filter to extract the interdependent relationship between channel features. The main contributions of this work are summarized as follows:

1) We design a dual-channel network structure (DCS). For shallow networks, this structure can effectively reduce the difference of output feature maps, which is caused by the different arrangements of spatial and temporal convolutions in P3D blocks. We stack these structures and propose a lightweight and efficient deep learning network model.
2) We add a time dimension into the SE block and apply it after the $1 \times 3 \times 3$ convolution filter. For a group of successive frames, attention is paid to each channel of frame. This approach can increase the relevance of important channel features, amplify important features, and extract temporal features between successive frames.
3) We recruit volunteers to record their sleep videos in real scenes and build a sleep behavior dataset named sleep action (SA) dataset. This dataset contains a total of 2393 video clips captured from the upper left and upper right angles above the bed. There are five basic sleep behaviors: turn over, get up at night, fall off bed, play mobile phones, and normal sleep. We use the backward propagation algorithm to train the network.
4) Experimental results show that our proposed network can achieve an accuracy rate of 90.67% for the sleep behavior recognition over the test set. The model parameter size of our proposed network is only 14.47M, which is about 1/6 of the C3D size. Furthermore, our network has an average prediction speed of 1.75 item/s. Thus, our proposed network can increase the accuracy of sleep behavior detection, and meanwhile decrease the requirement of computation resources.

The remainder of this paper is organized as follows. Section 2 reviews the relevant works on video behavior recognition. Section 3 illustrates our proposed network structure. In Section 4, we introduce the dataset and experiment. Finally, conclusion and future work are outlined in Section 5.

## II. RELATED WORKS

### A. VIDEO BEHAVIOR RECOGNITION

Before the advent of 3D convolution, there are two main methods to extract temporal features for the video based behavior recognition. One method is using optical flow information and 2D convolution neural network (CNN) for the behavior recognition. Wang and Schmid used the optical flow to obtain the track in the video sequence, and extracted features along the track [12]. Simonvan and Zisserman designed a two-stream network in [13], which trained the CNN model for spatial and temporal features, respectively, and finally integrated the results of these two networks. Wang *et al.* proposed the temporal segment network (TSN) based on the long-range temporal structure modeling [14]. It combines a sparse temporal sampling strategy and video-level supervision to train the whole video. Another method is to use the recurrent neural network (RNN) to extract temporal features of image frames and the 2D CNN to extract spatial features of image frames. In [15], the long short-term memory (LSTM) cell is connected to the output of the underlying CNN.

Ji *et al.* exploited 3D convolution for the behavior recognition [16]. Tran *et al.* proposed the classical C3D network and revealed that the optimal size of convolution kernel is $3 \times 3 \times 3$ [17]. Li *et al.* proposed a hierarchical structure of video with multiple granularities and use a multi-stream deep learning architecture to model each granularity [18]. The Two-Stream Inflated 3D Convolution Network (I3D) is proposed in [19], where the initial convolution kernel in the 2D CNN Inception-V1 is expanded as 3D. However, due to the surge of computing volume, the high computing cost, and the limitation of computing resources, 3D convolution network can't reach a high level of depth. In order to solve this problem, Qiu *et al.* proposed the P3D block, which splits 3D convolution into $1 \times 3 \times 3$ spatial and $3 \times 1 \times 1$ temporal convolutions separately to extract temporal and spatial features, respectively [20]. Three basic structures were proposed with different arrangements of spatial and temporal convolutions, which replace 2D convolution in Resnet Network and achieve much better performance on public datasets. In order to speed up the video behavior recognition process without degrading accuracy, the CNN layer basically uses P3D blocks to replace 3D convolution, as shown in [21], [22], [23], and [24], and in [21] the authors decomposed CBAM into spatial and temporal attention blocks to form a dual channel attention mechanism, which is embedded into the P3D structure and can be regarded as an improvement

of the attention mechanism. In our work, in order to achieve the lightweight, we use convolution instead of the original residual structure after embedding the improved SE module to form a two-way structure.

### B. ATTENTION MECHANISM

The attention mechanism is used to imitate human visual, which pays more attention to the useful information in images and videos, so the model can learn to focus on key information and ignore irrelevant information. Attention mechanism is widely used in the areas of machine translation, image processing, and natural language processing, etc. For the classification task, Chen *et al.* used the attention mechanism to softly weigh the multi-scale features [25]. Wang *et al.* constructed a Residual Attention Network by stacking multi-layer Attention modules [26]. Girdhar and Ramanan proposed a method for the fine-grained video classification, which adopts attentional pooling instead of average pooling or max pooling before the last full-connection layer [27]. Kar *et al.* implemented adaptive scan pooling to predict the importance of each image frame of the video [28]. Zhang *et al.* introduced the Context Encoding Module to capture the global context information and highlight the category information associated with the scene [29]. Zhao *et al.* proposed the self-attention mechanism, which can make each position in the feature map connect with all the other positions through the adaptive predictive attention map [30]. In [31], authors adopted SE blocks to model the interdependence between channels, adaptively iterates the characteristic responses of channels, and their network won the first place in the ILSVRC classification contest in 2017. By using the self-attention GAN (SAGAN), Kim *et al.* proposed a module that can extract four features of video information: spatial information, time information, slow motion information, and fast motion information [32]. In [33], authors proposed an interaction-aware self-attention model, which can learn attention maps by extracting the information of interactions between feature maps. Vaswani *et al.* proposed the Transformer structure, where the whole structure is composed of Self-Attention and Feed Forward Neural Network [34]. In addition, the authors built a trainable neural network by stacking Transformers, which has achieved great success in machine translation [34]. Compared with other attention mechanisms, the SE block is a substructure, which can be embedded into any layer of classification detection models with less computation required. Therefore, we integrate SE blocks into P3D blocks to enhance attentions on key frames.

## III. SLEEP BEHAVIOR DETECTION MODEL

### A. ATTENTION MECHANISM

The attention mechanism can imitate the human visual mechanism. When people watch pictures or videos, they always focus on key areas and changes of prominent features, and pay less attention to background factors and non-key areas. Attention mechanism can increase the focus to the most salient features in images. It's often used for image feature extraction. The SE block can perform Squeeze and Excitation operations to assign different weights for different channels [31]. The structure of the SE block is shown in Fig. 1, where $F_{sq}$ denotes the Squeeze operation, $F_{ex}$ denotes the Excitation operation, and $F_{scale}$ represents that the weighted feature vector is multiplied by the original feature map.

The Squeeze operation is mainly performed at a Global Average Pooling layer to compress the spatial dimension of input images, and compress the global spatial information into channels, as shown in formula (1) [31]:

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

where $x_c$ represents the input image, $H$ and $W$ represent height and width of the input image, respectively.

The Excitation operation uses the global information obtained in the Squeeze operation, which forms a bottleneck structure consisting of two full-connection layers. This bottleneck structure can model the correlation between feature channels, and parameterize the gating mechanism. Finally, the Sigmoid function is used for the activation, as shown in formula (2) [31]:

$$s = F_e x(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \tag{2}$$

where $\delta$ and $\sigma$ represent ReLU and Sigmoid activation functions, respectively, $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$ represent weight matrices of the first and the second full-connection layers, respectively.

In order to integrate SE blocks into P3D blocks, we add a temporal dimension into the SE block to form the 3D-SE block. For the input successive frames, the 3D-SE block pays attention to the channel of each frame separately. This approach focuses on the relationship of continuity between frames, and makes the change of action between frames more obvious. Take the input $X = [x_1, x_2, \ldots, x_c]$ as an example with $x_c \in R^{(C \times F \times H \times W)}$, where $C$ represents the number of channels per frame, $F$ represents the number of input frames, $H$ and $W$ represent height and width of each frame, respectively. The feature map of $(C, F, H, W)$ will be transposed as $(F, C, H, W)$, which is input into 3D-SE blocks. After the Squeeze operation, the feature map becomes $x'_c \in R^{F \times C \times 1 \times 1}$. Later, we let $x'_c$ learn the weights of $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$ to determine the importance of each channel. Finally, $x'_c$ is multiplied with the original feature. The 3D-SE operations is shown in Fig. 2.

### B. P3D WITH ATTENTION MECHANISM

Qiu *et al.* replaced 3D convolution with P3D blocks ($1 \times 3 \times 3$ spatial convolution filter $S$ is used in the spatial domain, and $3 \times 1 \times 1$ temporal convolution filter $T$ is used in the temporal domain), and designed three bottleneck building blocks, namely P3D-A, P3D-B, P3D-C, by considering two points: (1) whether spatial and temporal convolution filters directly or indirectly influence each other; (2) whether the two

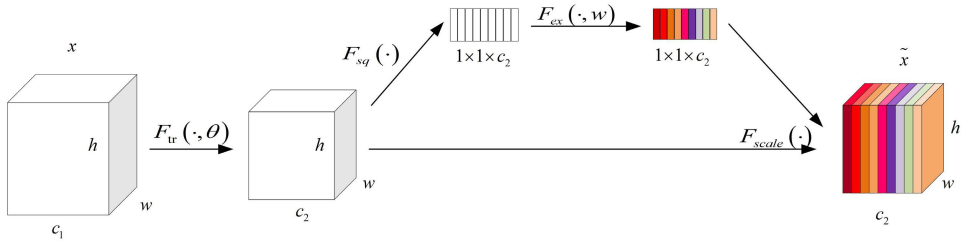


**FIGURE 1.** Schematic diagram of the SE block module structure.

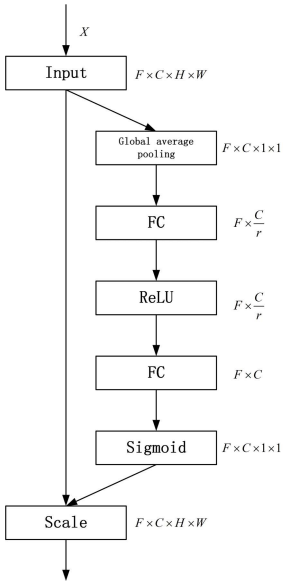


**FIGURE 2.** 3D-SE operations.

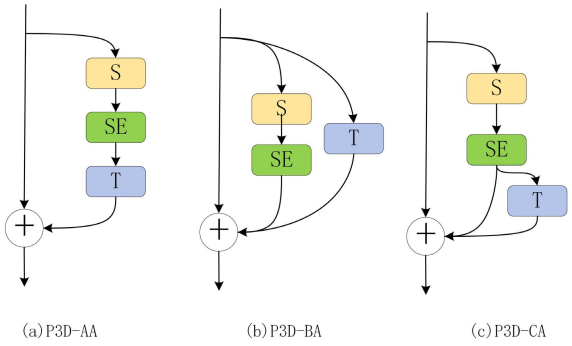


**FIGURE 3.** P3D-Attention structure diagram.

convolution filters directly affect the final output [20]. On the basis of P3D blocks, we introduce the attention mechanism to form the P3D-Attention structure as shown in Fig. 3.

P3D-AA: The attention mechanism is added after the spatial convolution $S$, and the temporal convolution $T$ directly affects the final output. The structure of P3D-AA can be expressed as:

$$(I + T \cdot AT \cdot S) \cdot x_t := x_t + T(AT(S(x_t))) = x_{t+1}, \quad (3)$$

where $x_t$ and $x_{t+1}$ represent the input feature map and the output of the P3D-Attention structure, respectively. They have the same dimension. $T$ represents the temporal 1D convolution, $S$ represents the spatial 2D convolution, $I$ represents doing nothing, and $AT$ represents the 3D-SE block.

P3D-BA: The attention mechanism is added after the spatial convolution $S$, and the structure allows the two filters to process the input data separately. Although there is no mutual influence between the two convolutional filters, they both affect the final output directly. The structure of P3D-BA can be expressed as:

$$(I + T + AT \cdot S) \cdot x_t := x_t + AT(S(x_t)) + T(x_t) = x_{t+1}. \quad (4)$$

P3D-CA: The attention mechanism is added after the spatial convolution, which establishes direct influence between $S$ and $T$. And they both affect the final output directly. Specifically, P3D-CA implants the direct connection between $S$ and the final output based on P3D-AA structure. The structure of P3D-CA can be expressed as:

$$(I + AT \cdot S + T \cdot AT \cdot S) \cdot x_t := x_t + AT(S(x_t)) + T(AT(S(x_t))) = x_{t+1}. \quad (5)$$

### C. DUAL-CHANNEL STRUCTURE

We propose a DCS based on the P3D-Attention. This structure integrates the attention mechanism to make the key frame play a greater role in the classification, and it can also narrow the gap between the outputs of different P3D-Attention blocks thanks to the different arrangements of spatial and temporal convolutions. Compared with the single-channel structure, our DCS is more helpful to excavate the deep semantic features of video and improve the detection accuracy.

The P3D-Attention decouples 3D convolution into 1D and 2D convolutions on the basis of residual structure. However, P3D-AA adopts a series of spatial and temporal convolutions, while only the temporal convolution connects directly to the final output. P3D-BA adopts spatial and temporal convolutions in parallel, and both convolutions can connect to the final output directly. P3D-CA adopts the series-parallel compromise mode to influence the final output jointly by spatial and temporal convolutions. Considering the depth of the proposed network model, there will be no serious network performance degradation, so we replace the residual structure with P3D-Attention and form DCS. Fig. 4 shows three

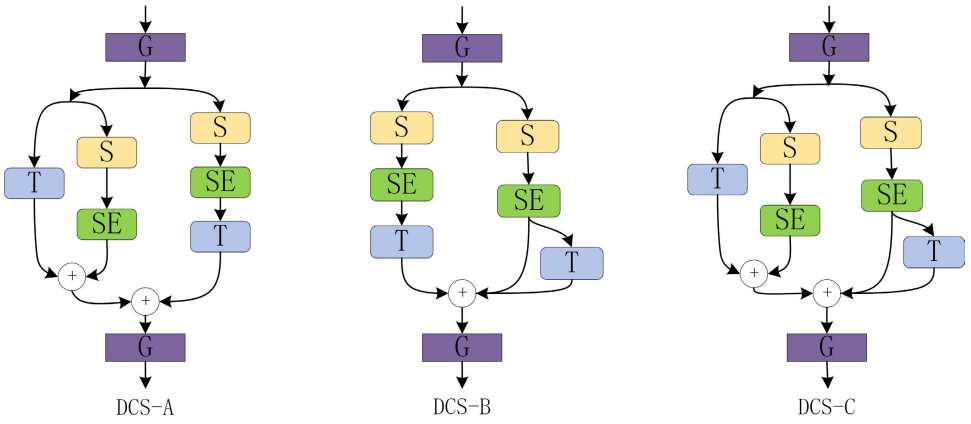


**FIGURE 4.** Schematic diagram of dual-channel structures.

DCS, which consist three dual-channel schematic diagrams denoted as DCS-A, DCS-B, and DCS-C, respectively.

DCS-A is composed of both P3D-AA and P3D-BA, which can be expressed as:

$$\begin{aligned}&(T \cdot AT \cdot S) \cdot x_t + (AT \cdot S + T) \cdot x_t \\ &\quad := T(AT(S(x_t))) + [AT(S(x_t)) + T(x_t)] = x_{t+1}, \end{aligned} \tag{6}$$

where $x_t$ and $x_{t+1}$ represent the input feature map and the output of a DCS structure, respectively. They have the same dimension. $G$ represents the $1 \times 1 \times 1$ convolution, $T$ represents the temporal 1D convolution, $S$ represents the spatial 2D convolution. $AT$ represents the 3D-SE block structure.

DCS-B is composed of both P3D-AA and P3D-CA, which can be expressed as:

$$\begin{aligned}&(T \cdot AT \cdot S) \cdot x_t + (AT \cdot S + T \cdot AT \cdot S) \cdot x_t \\ &\quad := T(AT(S(x_t))) + [AT(S(x_t)) + T(AT(S(x_t)))] = x_{t+1}. \end{aligned} \tag{7}$$

DCS-C is composed of both P3D-BA and P3D-CA, which can be expressed as:

$$\begin{aligned}&(AT \cdot S + T) \cdot x_t + (AT \cdot S + T \cdot AT \cdot S) \cdot x_t \\ &\quad := [AT(S(x_t)) + T(x_t)] + [AT(S(x_t)) + T(AT(S(x_t)))] = x_{t+1}. \end{aligned} \tag{8}$$

We preserve the $1 \times 1 \times 1$ convolution layer in the proposed DCS. The $1 \times 1 \times 1$ convolution layer can flexibly change the matrix dimensions without changing the properties of matrix. For example, the $1 \times 1 \times 1$ convolution layer at the beginning of the structure is used to reduce the number of channels, which can reduce the matrix dimension and hence the computation. The $1 \times 1 \times 1$ convolution layer is used at the end to restore the reduced dimensions. In addition, the $1 \times 1 \times 1$ convolution layer can also increase the network complexity at a low cost and make it closer to the accurate target model.

### D. NETWORK MODEL

Fig. 5 shows our proposed sleep behavior detection model based on P3D blocks and attention mechanism. The model consists of four DCS-B structures, five pooling layers, one batch normalization (BN) layer, one $1 \times 7 \times 7$ convolution layer, and one full-connection layer. Among them, the main function of the first $1 \times 7 \times 7$ convolution layer is to downsample the input image. In the premise of not increasing the number of channels, downsampling can reduce the computation cost and preserve as much information as possible about the original images. The BN processing method is proposed by Ioffe and Szegedy in 2015 [35], which is used to speed up the training and convergence of the network, and prevent over-fitting to a certain extent. As shown in Fig. 4, the DCS adopts DCS-B, which can fully extract the spatial-temporal features by using the P3D-Attention, and improve the correlation of important channel features by using 3D-SE blocks. In order to prevent the occurrence of over-fitting, the dropout layer is added to make the network model more robust. At the end of the network, a full-connection layer is used as a classifier, followed by a softmax activation function to output the detection result. The detailed parameters of the network model are shown in Table 1 given on top of the next page.

We evaluate the uncertainty of the model by predicting a video on the same model $3 \sim 5$ times to obtain the predicted value and calculate the prediction entropy. The prediction entropy calculation formula is as follows:

$$\begin{aligned}&H[y|x, D_{train}] \\ &\quad := -\sum_c p(y = c|x, D_{train}) \log p(y = c|x, D_{train})]. \end{aligned} \tag{9}$$

where $x$, $y$, $D_{train}$, $p$ are labels, input videos, training data, and probabilities, respectively. If the value is low, the model is sure about its predictions, and if the result is high, the model doesn't know what's in the video.

## IV. EXPERIMENT

### A. DATASET

#### 1) SLEEP ACTION (SA) DATASET

So far as we know, there are no publicly available datasets of sleep behavior on the Internet. We recruit four volunteers with different characteristics and record their night sleep videos to build a dataset of sleep action. The volunteers are all graduate

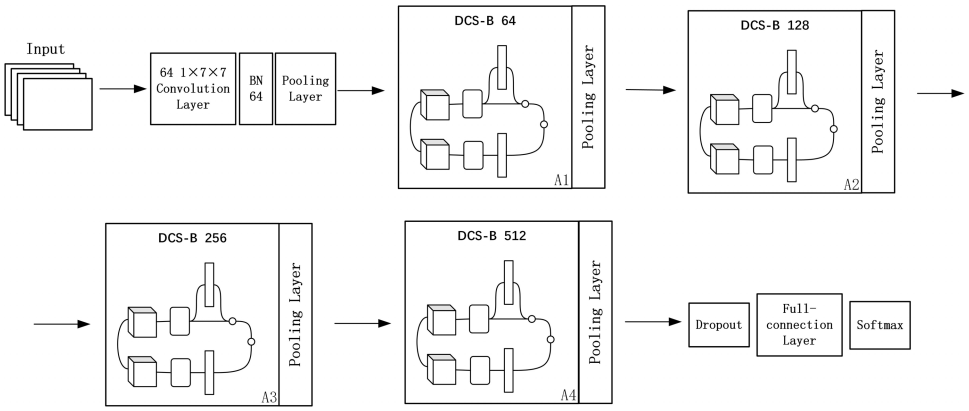


**FIGURE 5.** Schematic diagram of network model structure.

**TABLE 1.** Example of a network model. We show the structure and output size with 16 frames of pictures input.

| Layer | Architecture | Output |
|---|---|---|
| 3D-input | stack | $3 \times 160 \times 160$ |
| convolution 1 | $64, 1 \times 7 \times 7, stride = (1, 2, 2)$ | $64 \times 16 \times 80 \times 80$ |
| maxpool 1 | $2 \times 3 \times 3, stride = 2, padding = (0, 1, 1)$ | $64 \times 8 \times 40 \times 40$ |
| A1 | $\begin{bmatrix} 256, 1 \times 3 \times 3, stride = 1, padding = (0, 1, 1) \\ 256, 3 \times 1 \times 1, stride = 1, padding = (1, 0, 0) \end{bmatrix}$ | $256 \times 8 \times 40 \times 40$ |
| maxpool 2 | $256, 2 \times 1 \times 1, stride = (2, 1, 1)$ | $256 \times 4 \times 40 \times 40$ |
| A2 | $\begin{bmatrix} 512, 1 \times 3 \times 3, stride = 1, padding = (0, 1, 1) \\ 512, 3 \times 1 \times 1, stride = 1, padding = (1, 0, 0) \end{bmatrix}$ | $512 \times 4 \times 20 \times 20$ |
| maxpool 2 | $512, 2 \times 1 \times 1, stride = (2, 1, 1)$ | $512 \times 2 \times 20 \times 20$ |
| A3 | $\begin{bmatrix} 1024, 1 \times 3 \times 3, stride = 1, padding = (0, 1, 1) \\ 1024, 3 \times 1 \times 1, stride = 1, padding = (1, 0, 0) \end{bmatrix}$ | $1024 \times 2 \times 10 \times 10$ |
| maxpool 2 | $1024, 2 \times 1 \times 1, stride = (2, 1, 1)$ | $1024 \times 1 \times 10 \times 10$ |
| A4 | $1024, 3 \times 3, stride = 1, padding = 1$ | $2048 \times 5 \times 5$ |
| avgpool | $2048, 5 \times 5, stride = 1$ | $2048 \times 1 \times 1$ |

students in the school and sleep videos are recorded in the student dormitory. The angle of the camera with respect to the bed is set as upper left corner and upper right corner, respectively. A high-definition infrared camera is used to record the volunteers' sleep states for seven consecutive days from 00:00 a.m. to 8:00 a.m. Since the light is bright from 6:00 a.m. to 8:00 a.m. like daytime, the dataset is a mix of black and white and color. A total of 145 videos were recorded with each lasting for 2-3 hours. Among them, volunteers are required to simulate the fall off bed behavior in a more natural way for part of the time. In the rest of time, the camera records actions such as turn over, get up at night, play mobile phones, and normal sleep under normal conditions.

In order to meet the experiment requirements, we manually clip 145 videos and reserve the required action parts. Some sleep behaviors can be completed in a relatively short period of time, for example, turn over, get up at night, and fall off the bed. Such behaviors can be completed within about 4 seconds, while playing mobile phones and normal sleep is a continuous behavior, which can last from ten to dozens of minutes. Therefore, each video clip lasts for 4 seconds after cutting. After clipping, each video frame pixel is adjusted from 1920 × 1080 to 320 × 240. As shown in Table 2, after cutting there are 555 turn over actions, 211 fall off actions, 270 get up at night actions, 528 play mobile phone actions, and 829 normal sleep. There are totally 2393 videos, we named it SA dataset.

**TABLE 2.** Number of action clips after clipping.

| | Turn over | Fall off bed | Get up | Play mobile phone | Normal sleep |
|---|---|---|---|---|---|
| Number | 555 | 211 | 270 | 528 | 829 |

### 2) DATASET PROCESSING

The length of each video in the SA dataset is 4s, the frame width is 320, the frame height is 240, and the frame rate is 25 frames/s. Since continuous image frames are input into our network, we cut each video in the dataset into image frames at

**FIGURE 6.** Example of modified image.

**TABLE 3.** Number of dataset.

| | Training | Validation | Test |
|---|---|---|---|
| Number | 1523 | 384 | 479 |

**TABLE 4.** Accuracy of different DCS structures.

| | DCS-A | DCS-B | DCS-C |
|---|---|---|---|
| Accuracy | 86.43% | 90.89% | 88.10% |
| Recall | 86.54% | 93.13% | 87.23% |
| Precious | 88.63% | 88.78% | 88.69% |

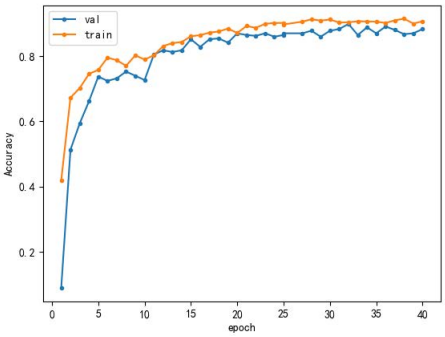


**FIGURE 7.** Accuracy of training and validation.

the rate of 25 frames/s. In order to improve the computation efficiency, we reduce the number of reference frames, and remove the frames with little changes of actions between continuous frames. Through intercepting one frame in every four frames, we finally obtain 25 reference frames for each video clip. The size of the video frame is still relatively large for the network model. When such pictures are input into the network, there is too much useless information, which will cause interference to the network data fitting and increase the computation overhead. Therefore, we use the resize method in PyTorch to adjust the size of each picture when cutting the reference frame. Fig. 6 shows an example of a modified RGB image (fall off bed action). After processing, each video will be decomposed into 25 pictures of size.

### B. EVALUATION

We evaluate the proposed network on the SA dataset. The experiment was conducted on a desktop workstation, which is equipped with 64 bit Ubuntu 20.4 operating system, CUDA 11.4, CUDNN 8.2.4. The CPU is Intel Core I9-10900X, and GPU is GeForce RTX 3090 with 24GB display memory. The deep learning framework is PyTorch, version 1.10.0, and Python version 3.7.11.

When the dataset is loaded, it will be randomly divided into training set, verification set, and test set with a ratio of 6:2:2. Table 3 shows the number of samples in training set, validation set, and test set. When the training set is used for training, the data of the validation set and test set are not used. After each training, the validation set is used to evaluate the generalization ability of the model, and the test set is used to verify the recognition ability of the model. During the model training, the network randomly selects 16 consecutive frames from the 25 frames, randomly intercepts each frame with a size of $160 \times 160$, and input them into the network. The batch size is 64, and the initial learning rate is 0.001. The learning rate attenuates gradually with the increase of training rounds, and the attenuation coefficient is 0.1, which attenuates once every ten rounds. We adopt the cross entropy loss, and use the Adam optimizer to update network parameters through following the back-propagated gradient information of the network. Regularization built into the optimizer is used to prevent overfitting, and the value is $8 \times 10^{-4}$.

In 40 rounds of training, after eath round, the validation set will be used to observe the accuracy of the model and the loss function timely, so as to facilitate the mediation of hyperparameters. Finally, the generalization ability of the model is tested with the test set every five rounds of training.

Fig. 7 shows the line graph of the result of training and validation set on the SA dataset. It can be seen that the fitting of training and validation is very good. Fig. 8 shows the accuracy over the test set. In addition, we draw the ROC (receiver operating characteristic) curve with TPR (true posive rate) as the ordinate and FPR (false postive rate) as the abscissa, and the area enclosed by it is the AUC (Area Under Curve) value. Fig. 9 shows the ROC of each category and calculate the AUC, The larger the AUC value, the better the classification effect of the classifier.

### C. COMPARISON

#### 1) COMPARISON OF DUAL-CHANNEL STRUCTURES

In order to verify the efficiency of DCS, we compare the three DCS proposed in subsection III.C on the SA dataset,

**TABLE 5.** Comparison of different network models in SA dataset.

| Network structure | Accuracy | Recall | Precision | Model parameter size /MB | Average prediction speed item/s |
|---|---|---|---|---|---|
| C3D | 84.13% | 86.19% | 88.76% | 78.02 | 2.18 |
| R(2+1)D | 89.32% | 86.16% | 88.37% | 33.18 | 1.26 |
| Our network without SE | 88.30% | 90.43% | 89.57% | 14.47 | 0.85 |
| Our network with SE | 90.89% | 93.13% | 88.78% | 14.47 | 0.89 |

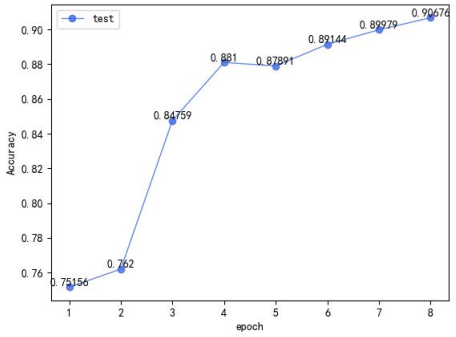


**FIGURE 8.** Accuracy of test.

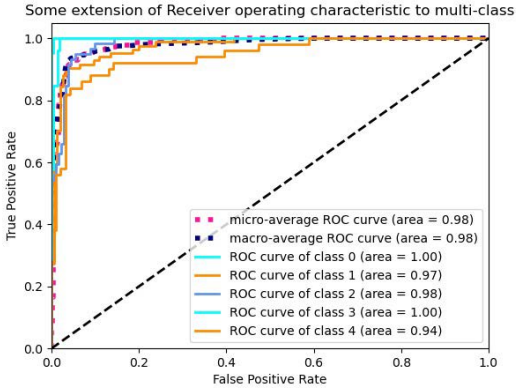


**FIGURE 9.** Receiver operating characteristic.

the results are shown in Table 4. DCS-B structure combines P3D-A and P3D-C with the SE structure. In [20], the authors have verified that the accuracies of P3D-A and P3D-C are higher than P3D-B, so the dual-channel structure DCS-B composed of P3D-A and P3D-C is better than the other two structures.

### 2) COMPARISON OF NETWORK MODELS

In order to verify the efficiency of the proposed network, this experiment compares the recognition effects of C3D network and R(2+1)D network on the SA dataset in terms of accuracy, recall, precision, model parameter size, and average prediction speed (1 item equivalent to 1 batch size), respectively. Since the whole data samples are divided into three subsets according to the ratio of 6:2:2, we perform 5-flod cross-validation. We perform 5 times of experiments to train, validate, and test our proposed model separately. In each experiment, we build a training set, a validation set, and a test set. For the 5 experiments, the test sets have no intersections. Finally, the test results of these 5 experiments are averaged.

C3D network is the first applies 3D-CNN for the behavior detection, it has representative significance in convolutional neural networks. But its network structure is very simple, including only 8 convolution layers, 5 pooling layers, and 2 full-connection layers, so it has the worst effect and takes the longest time on the SA dataset. Through comparing with C3D, we can show the efficiency of splitting 3D convolution into 2D spatial convolution and 1D temporal convolution in sleep behavior recognition. R(2+1)D network is jointly created by Tran, founder of C3D [17], and Lecun, inventor of CNN [36]. R(2+1)D uses a similar method as the P3D block, adopts 2D spatial convolution and 1D temporal convolution, and learns through Residual Network. For the behavior recognition, R(2+1)D network has reasonably good accuracy. However, for the sleep behavior recognition, our proposed dual-channel structure can achieve much better performance than R(2+1)D.

As show in Table 5, R(2+1)D and our proposed network have highter accuracy compared with C3D, so it is a reliable method to split 3D convolution into 2D spatial convolution and 1D temporal convolution. Compared with C3D, the accuracy of our proposed network improves by about 6% with only 1/6 model parameter size. The average prediction speed of the proposed method is about 1.75 item/s. Compared with the R(2+1)D network, the accuracy is improved by 1.5% with only 1/2 model parameter size.

In order to prove the effectiveness of our experiments, we performed T-tests for different models, and the test methods are as follows: If two algorithms perform the same, they should have the same test error rate in the same training/testing set, that is $\hat{\xi}_{iA} = \hat{\xi}_{iB}$. Difference between the error rates for each set of tests, we get $\Delta_i = \hat{\xi}_A - \hat{\xi}_B$. If the performance of the two models is the same, the difference is 0. If they are different, the T-test is performed according to $\Delta_1, \Delta_2, \Delta_3, \ldots \Delta_k$ to calculate the mean $|\bar{\Delta}|$ and variance $S^2$ of the difference, then:

$$\frac{|(\bar{\Delta} - \Delta_i)|}{S/\sqrt{k}} \sim t(k-1) \tag{10}$$

If postulate $H_0 : \xi_A = \xi_B$, $H_1 : \xi_A \neq \xi_B$, and: $\frac{|\bar{\Delta}|}{S/\sqrt{k}} \sim t_{\delta/2}$. Since we use 5-flod cross validation, we get five $\Delta_i$ for validation. In this paper, we verified C3D and our proposed network, R(2+1)D and our proposed network, respectively. After verification, the performance of our proposed network is not statistically the same as others.

## V. CONCLUSION AND FUTURE WORK

We propose a network structure by integrating P3D blocks and attention mechanism. In order to improve the network performance, we extend the SE block by adding a time dimension after the $1 \times 3 \times 3$ spatial convolution. To verify the effectiveness of the proposed model, we build a SA dataset and conduct experiments on it. We compare the proposed model with several classical models in terms of accuracy, model size, and computational cost. The results show that our proposed model can achieve highly competitive performance with smaller model size and less computational cost.

In our model, the attention mechanism is only added to the channel at the spatial level, which improves the extraction ability of the underlying spatial features. The extraction of temporal features is not enough, which is a promising direction to study. In addition, inspired by [16], [22], and [34], it is meaningful to build a network by combining CNN and RNN, and introduce the Transformer attention mechanism to extract the spatiotemporal feature more sufficiently.

## REFERENCES

[1] J. Wang, Y. Zhang, and Y. Liu, *Annual Sleep Report of China 2022*. Peking, China: Social Sci. Academic Press, Mar. 2022.

[2] B. A. Chaudhary, "Introduction to polysomnography," in *Primary Care Sleep Medicine*. Totowa, NJ, USA: Humana Press, 2007, pp. 295–304.

[3] N. Butkov and S. A. Keenan, "An overview of polysomnographic technique," in *Sleep Disorders Medicine*. New York, NY, USA: Springer, 2017, pp. 264–279.

[4] H. Miwa, S.-I. Sasahara, and T. Matsui, "Roll-over detection and sleep quality measurement using a wearable sensor," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2007, pp. 1507–1510.

[5] Y.-C. Zeng and W.-T. Chang, "Estimation of sleep status based on wearable free device for elderly care," in *Proc. IEEE 5th Global Conf. Consum. Electron.*, Kyoto, Japan, Oct. 2016, pp. 604–607.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jul. 2015, pp. 1–9.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[10] N. Q. K. Le and Q.-T. Ho, "Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes," *Methods*, vol. 204, pp. 199–206, Aug. 2022.

[11] N.-Q.-K. Le and B. P. Nguyen, "Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2189–2197, Nov. 2021.

[12] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.

[13] K. Simonvan and A. Zisserman, "Two-stream convolutional network for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 568–576.

[14] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 20–36.

[15] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[18] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo, "Learning hierarchical video representation for action recognition," *Int. J. Multimedia Inf. Retr.*, vol. 6, no. 1, pp. 85–98, Feb. 2017.

[19] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[20] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Beijing, China, Oct. 2017, pp. 5534–5542.

[21] Y. Zhuang and Y. Qi, "Driving fatigue detection based on pseudo 3D convolutional neural network and attention mechanisms," *J. Image Graph.*, vol. 26, no. 1, pp. 143–153, 2021.

[22] J. Wei, H. Wang, Y. Yi, Q. Li, and D. Huang, "P3D-CTN: Pseudo-3D convolutional tube network for spatio-temporal action detection in videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 300–304.

[23] T. Chen, X. Liu, and G. Li, "Attention based on pseudo 3D convolution residual network for action recognition of earth-moving machinery," in *Proc. Int. Conf. Comput. Inf. Sci. Artif. Intell. (CISAI)*, Sep. 2021, pp. 93–98.

[24] B. Lu, Z. Lv, and S. Zhu, "Pseudo-3D residual networks based anomaly detection in surveillance videos," in *Proc. Chin. Autom. Congr. (CAC)*, Hangzhou, China, Nov. 2019, pp. 3769–3773.

[25] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3640–3649.

[26] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.

[27] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. 31th Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 34–45.

[28] A. Kar, N. Rai, K. Sikka, and G. Sharma, "AdaScan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5699–5708.

[29] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7151–7160.

[30] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Loy, D. Lin, and J. Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 270–286.

[31] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2017.

[32] M. Kim, T. Kim, and D. Kim, "Spatio-temporal slowfast self-attention network for action recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2206–2210.

[33] W. Hu, H. Liu, Y. Du, C. Yuan, B. Li, and S. J. Maybank, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 373–389.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and L. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37. Lille, France, Jul. 2015, pp. 448–456.
[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

**RUI GUO** received the B.Sc. degree in communication engineering from Qingdao Agricultural University, Qingdao, China, in 2021. He is currently pursuing the M.Sc. degree with the School of Information Science and Engineering, Shandong University. His research interests include deep learning, neural networks, and behavior recognition.

**CHAO ZHAI** (Member, IEEE) received the B.Sc. degree in communication engineering and the M.Sc. degree in communication and information systems from Shandong University, Jinan, China, in 2007 and 2010, respectively, and the Ph.D. degree in electrical engineering from The University of New South Wales, Sydney, Australia, in 2013.

From 2011 to 2012, he was a visiting Ph.D. student with The Chinese University of Hong Kong. In 2013, he was a visiting Ph.D. student with The Hong Kong University of Science and Technology. From 2014 to 2016, he held a postdoctoral position with Shandong University. He joined the School of Information Science and Engineering, Shandong University, where he is currently an Associate Professor. His research interests include artificial intelligence, cognitive spectrum sharing, cooperative relaying, massive MIMO, and energy harvesting wireless communications.

**LINA ZHENG** received the B.Sc. degree in communication engineering and the M.Sc. and Ph.D. degrees in communication and information system from Shandong University, Jinan, China, in 2001, 2004, and 2011, respectively.

Since 2004, she has been with the School of Information Science and Engineering, Shandong University, where she is currently an Associate Professor. Her research interests include artificial intelligence, cooperative communications, energy harvesting wireless communications, and cross-layer design of routing and MAC protocols in ad-hoc networks.

**LUYU ZHANG** received the B.Sc. degree in communication engineering from Shandong University, Weihai, China, in 2021, where she is currently pursuing the M.Sc. degree with the School of Information Science and Engineering. Her research interests include deep learning, big data analysis, simultaneous wireless information and power transfer, and cooperative communications.