IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 3, 2020, accepted December 14, 2020, date of publication December 25, 2020, date of current version June 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047455

Study on the Identification Method of Human Upper Limb Flag Movements Based on Inception-ResNet Double Stream Network

ZHONG YUE, JIQING LUO, FANG HUSHENG, FAMING SHAO, AND ZHOU RANZHI
Department of Mechanical Engineering, College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China

Corresponding author: Jiqing Luo (ljqlgd@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 2016YFC0802904, and in part by the Key Research and Development Program of China under Grant 61671470.

ABSTRACT To investigate the recognition effect of flag motions based on 9 axis sensors, starting with deep learning methods, this paper proposes a framework for feature extraction and recognition of flag movements recognition by employing an improved Inception-ResNet dual-stream network. In traditional signal recognition studies, Support Vector Machine (SVM), Random Forest (RF) and One Dimensional Convolutional Neural Network (CNN) are usually used to extract signals' features. In the meanwhile, the time series data sets such as flag movements are usually the standard data set after processing. Therefore, there are usually some limitations in traditional systems. First, in actual environment, there exists a lack of effective segmentation detection method for the samples of long time series, resulting in the deviation of the data set in the recognition process. Second, the One-Dimensional CNN framework used and the machine learning frameworks used in previous studies are difficult to process large quantities of data with too high computational memory. Based on these problems, this study proposes a signal change point detection model based on the diversity factory function in the signal segmentation and detection stage, miniaturizes the convolution kernel in the original CNN by using the Inception-ResNet(I-R) dual-flow network separable convolution method, and proposes a CrossEntropy-Logistic(C-L) joint classification loss function. Through conducting comparative experiments, it is found that the average calculation parameter of the CNN framework based on the Inception-ResNet model is 2.7×10^7, which is approximately 37% lower than the number of 3.7×10^7 in the original CNN model. Finally, the recognition rate between C-L joint loss function and other models such as Multi-Layer Perceptron and Ensemble Learning in recent years are compared. Compared with Ensemble Learning-CrossEntropy (ELC) model, the C-L joint loss function can improve the recognition rate by nearly 5% according to the results of flag movements identification measured by several classification models.

INDEX TERMS CrossEntropy-logistic, deeply separable convolution, difference threshold search, diversity factor function, inception-ResNet.

I. INTRODUCTION

Flag movement technology is a command technology with wide application in traffic, ship navigation and engineering fields [1]. Among them, the research on motion data acquisition method and recognition network framework is currently the key content. The traditional methods of flag gesture acquisition include the first generation of human flag gesture acquisition technology based on optical fiber equipment and the second generation of flag gesture image acquisition and recognition technology based on computer vision. However, the traditional signal-sign movement recognition method requires high requirements on data acquisition and equipment. In practical application, the signal-sign commander cannot move around the field or change the body position. The equipment deployment is not flexible enough, and the recognition effect needs to be improved [2]. With the development of sensor technology, human motion recognition based

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

on various sensors gradually shows a new development trend. At present, sensor-based motion recognition has been widely used in medical treatment and human health state detection.

Wearable Technology is an information acquisition technology that employs technologies such as multimedia, sensors and wireless communication embedded in clothing or on the body to capture real-time signals [3]. It is of importance to select the appropriate sensor data type for feature extraction and recognition. Under normal circumstances, the direction, angle and moving speed of the motion can be obtained by collecting the acceleration data of the sensor. For instance, Wu *et al.* [4] and Kwon *et al.* [5]. used acceleration sensor and gyroscope sensor to collect the combined data of acceleration and magnetic deviation Angle, finding that the combination of sensor data can enhance the identification accuracy. Matsui *et al.* [6] used the 9-axis attitude sensor to collect the data of acceleration, magnetic deflection and rotation angular velocity of the human body in daily behavior and movement. Experimental results show that the accuracy of behavior recognition can be improved by collecting different types of data. Based on various researches, the acceleration and Angle fusion data collected by accelerometer and gyroscope have achieved satisfying results in the application of motion recognition.

A. ORIGINAL CNN

In recent years, the application of deep learning to human motion recognition has become a research hotspot [7]. Compared with the traditional machine learning method, the advantage of deep learning lies in the self-learning of features through a large amount of data and the gradual extraction of higher level abstract semantic features from a large number of data sets, thus avoiding the limitation of manual extraction of features. Among them, CNN is an important method in deep learning. Meanwhile. On the basis of CNN convolutional neural network, several commonly used classical convolutional neural network models are born, such as LeNet-5, AlexNet, VGGNet, Inception-v3 and ResNet. At present, the above methods have been applied in various different fields. For example, Spanhol *et al.* [8] sampled BreaKHis data set by random and fixed sliding Windows and trained the network with AlexNet blocks of different scales. The recognition rate could reach 89.6%, so as to satisfy the requirements of clinical detection. Deniz *et al.* [9]. pretrained the AlexNet and VGG16 network models by means of migration learning method, and then optimized the weights and bias values in the convolutional neural network by employing the loss back-propagation method of classification function. Thus, automatic classification of benign and malignant breast cancer pathologic data was realized, and the final recognition rate reached 93.78%. Kim and Cho [10]. collected real-time acceleration data from construction workers using 17 sensors worn all over the body. Long and Short Time Memory (LSTM) deep learning network is established to realize self-learning of the acceleration data characteristics of human body during construction with the aim to detect the body's working condition and predict the risk coefficient during the whole body movement.

B. LIMITATIONS OF THE EXISTING NETWORK FRAMEWORK

Based on the above motion recognition and classification techniques, it can be concluded that in the field of deep learning, the processing of standard motion data set by using CNN or Recurrent Neural Network (RNN) model is currently the main method for relevant researches. In the past research, a one-dimensional CNN network framework was also used to extract features from the fused 9-axis sensor data, 3 axis of input data fusion signal directly for processing, the input dimension of 9×100, by setting the corresponding convolution layer and pooling of input processing, feature extraction, as well as the experiment results show that the original one-dimensional CNN framework has increased a lot of technical parameters, and for a non-standard semaphore with limited data set is lack of signal divided work, and thus lead to appeared in the process of identifying the extremely obvious over fitting phenomenon. At the same time, according to the rest of the methodologies summarized, the existing researches lack non-standard long time series signal processing, and therefore difficult to be applied to real-time training environment, and the classification algorithm types in the existing network framework are relatively single. The adaptability to all kinds of data sets is not enough, and thus there exists large space to improve identification precision.

C. MAIN WORKS AND SECTIONS ARRANGEMENT

This paper designs an improved Inception-ResNet dual-stream network system for the collection, pre-processing analysis, feature extraction and classification recognition of sensor data worn on human hands. Its main contributions are presented as follows:

i. The preprocessing work of flag signal is improved. An algorithm based on long time series variation point segmentation is proposed. Principal Component Analysis (PCA) is employed to establish the differentiation function model of three main features pulse index I, peak factor C, and waveform factor W in the 9-axis long time series. In addition, the hyperplane search principle in SVM is used to search the threshold value of the difference function, and the threshold value λ is obtained to determine whether there are signal change points between adjacent sample points.

ii. An Inception-ResNet dual-flow network model based on separable convolution idea is proposed. The original one-dimensional signal dimension is transformed from 18×100 to $30 \times 30 \times 2$ two-dimensional data by matrix reconstruction method, which is subsequently used as the input of the network. The original one-dimensional CNN large-scale convolution kernel is decomposed into multiple small-scale convolution kerns by means of decomposition. Thus, the computational parameters in the network framework are greatly reduced.

iii. A CrossEntropy-Logistic joint loss function based on the fusion of dichotomous task and multi-categorical task is proposed. First, CrossEntropy loss function is used to judge the initial category of the signal, and then the low-dimensional plane feature is mapped to the high-dimensional space feature by feature mapping. Then, dichotomous Logistic loss function is used to distinguish the Angle in the signal class. Finally, the category of the sample is determined. The experimental results show that the overfitting phenomenon can be greatly reduced and the recognition accuracy can be greatly improved by task fusion.

The sections of the whole flag movements recognition system construction task are arranged as follows:

Section II: Flag Motion Capture Experiment Based On 9-axis Attitude Sensor. In the present section, the flag movements acquisition and processing equipment are mainly introduced, and the flag movements data set collected is introduced.

Section III: Time Series Detection and Segmentation Algorithm Based on Difference Threshold Search. This section mainly introduces the long time series signal segmentation and detection algorithm in the pre-processing stage of flag movements, establishes the difference function between the sample points before and after the signal through three types of feature fusion methods, and determines the hyperplane search algorithm to find the difference function threshold λ.

Section IV. Feature Extraction Algorithm Based on Inception-ResNet. In the current section, the structure of the dual-flow network framework is constructed in the phase of the signal feature extraction phase, and introduces the operating principle and method improvement of the Inception- ResNet double flow network structure. The original one-dimensional CNN convolution kernel can be separated and improved, significantly reducing the network calculation parameter.

Section V. C-L Joint Loss Function Classification Model. This section mainly improves the algorithm of classification recognition task, explains the limitations of the traditional single Cross Entropy loss function, puts forward the algorithm model of C-L joint loss function, and introduces the core ideas of binary classification and multi-classification task fusion method, aiming to establish the form of C-L joint loss function.

Section VI. Implementation and result. This section is a practical part of the above algorithm content. In the current experiment, the final signal difference threshold was obtained by comparing the standard deviation of various, signal lengths and the average proportion of key information under different λ values. The joint loss function coefficients were trained by Adagrad gradient descent method, and the final C-L combined loss function parameters were determined by comparing the average loss values of various actions after 1000 rounds of iteration. After establishing the semaphore gesture recognition system frame structure, the collected 6 class standard classified semaphore movement training and testing, comparing the Inception-ResNet network framework and SVM, RF, SVM, 6 classes action within the framework of the one-dimensional CNN a total of 3000 samples independent sample average recognition rate. By comparing the convergences of the identification accuracy with that of the training curves, it can be concluded that the Inception-ResNet network converges and exerts the best identification effect during the identification process.

II. FLAG MOTION CAPTURE EXPERIMENT BASED ON 9-AXIS ATTITUDE SENSOR

This paper adopts a 9-axis attitude sensor worn at the wrist of both hands to collect six common upper limb flag movements. Figure 1 shows the sensor structure. The WT901C digital attitude sensor used in the experiment is integrated with a high-precision 3-axis gyroscope, 3-axis accelerometer and 3-axis geomagnetic field sensor [11]. The real-time motion attitude of the module can be quickly solved using a high-performance microprocessor and advanced dynamic solution. The attitude measurement accuracy of the sensor is static 0.05 degree and dynamic 0.1 degree. The range is acceleration: $\pm$2/4/8/16g, angular velocity: $\pm$250/500/1000/2000°/s, angle: $\pm$180° and data output frequency: 0.1Hz ~200Hz.

FIGURE 1. WT901C digital attitude sensor.

In the experimental stage, the sensor was worn on the experimenter's hands and wrists and connected to the acquisition device (computer) through USB interface. Various signals collected were transmitted to the computer in text forms in real time. The sensor motion acquisition and receiving platform and wearing mode are shown in Figure 2. The data monitoring terminal is presented in Figure 3.

(a) (b)

FIGURE 2. Experiment platform and sensor wearing mode Sensor wear mode; (b) Motion acquisition platform.

The data of 18 channels of acceleration, angular velocity and magnetic deflection (roll Angle (X axis), pitch Angle (Y axis) and course Angle (Z axis) of the three axes of the dual sensor were collected in the experiment. There are six

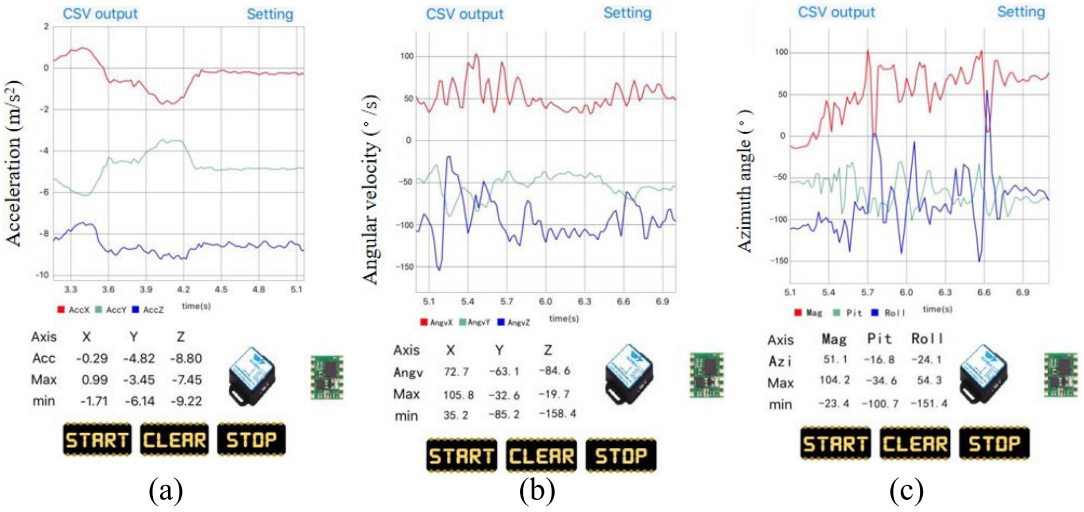


FIGURE 3. Data transmission monitoring terminal(built by Shenzhen Wit-Motion company) (a) Acceleration acquisition interface; (b) Angular velocity acquisition interface; (c) Azimuth acquisition interface.

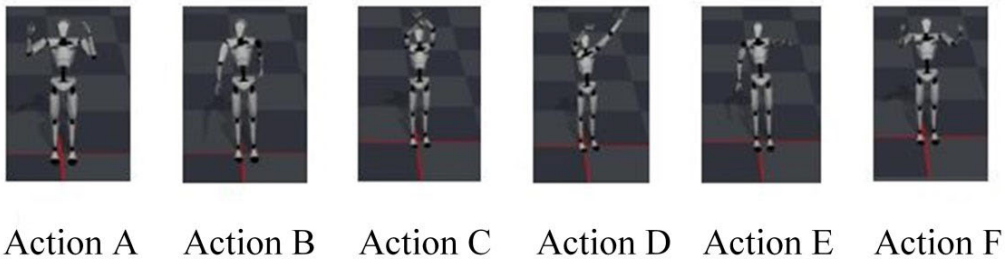


FIGURE 4. Diagram of six flag action to be tested.

TABLE 1. Movement essentials of flag to be tested.

Upper limb flag movement number	Essential of flag actions
A	The upper arm is perpendicular to the body 90, the elbow is curved, the forearm is placed on the side of the head and the cross swings back and forth.
B	Drop arms naturally, stick to sides and swing back and forth.
C	Raise arms vertically above head, repeatedly crossing and separating arms.
D	Raise arms straight over head and swing them parallel to each other.
E	Straighten left arm and make a counterclockwise circle in front of body.
F	Stretch arms straight up and down the body.

different types of flag gestures, numbered A,B,C,D,E and F respectively. Their demonstration and key points of actions are shown in Figure 4 and Table 1 respectively.

III. TIME SERIES DETECTION AND SEGMENTATION ALGORITHM BASED ON DIFFERENCE THRESHOLD SEARCH

A. SIGNAL DETECTION AND SEGMENTATION ALGORITHM

After completing the establishment of the experimental platform and the collection of samples, the flag movement data obtained is a time series whose length is proportional to the length at the time of collection. In the data preprocessing stage, the detection and segmentation and denoising of key motion signals are mainly completed. The present study mainly investigates the time series segmentation and detection algorithm. According to the previous research [12], sliding window is used to segment the one-dimensional time series. By designing the time window and sliding at any step length on the one-dimensional time series, the data fragments contained in each sliding window are intercepted to obtain the required data fragments. However, when the sliding window is used to segment the data, it is mainly divided in accordance with the peak and valley value of the timing signal. Therefore, when the size of the sliding window is selected, information

redundancy and loss caused by too large or too small window distance will appear, bringing considerable difficulties to the recognition process. Similarly, improper setting of the slide step can also cause the above problems.

According to the existing research results, the detection methods based on signal can be divided into shard unit alignment method and boundary detection method [13]. The difference between the two methods lies in that, based on the segmentation unit alignment method, the signal detection should be combined with other methods such as ANN(artificial neural network) or clustering based on the characteristics of the signal [14]. Thus, the signal detection based on the segmented unit alignment method has a high demand for prior knowledge, especially in motion recognition, requiring a high number of samples of the detected classified signals. However, the method based on boundary detection is different. Its basic idea is to detect the existing unit boundary according to the various characteristics of the signal and its mutation. Consequently, this kind of method does not need prior signal characteristics and is easier to construct the signal segmentation system.

According to the introduction of time domain waveform and action essentials of flag signal, the research object is a one-dimensional flag signal time series. It is found that in a piece of signal containing all kinds of flag signal, acceleration abrupt point and periodic component exist between different kinds of signals. Therefore, it is easier to segment and process flag signal than mechanical fault signal and continuous motion signal. Nevertheless, the flag signal contains a large number of low-frequency components, and thus it is of necessity to increase the sliding window scale as much as possible while learning the low-frequency characteristics of the signal in accordance with the actual research [15]. When the sliding window increases, due to the signal acquisition actions do not match the length of time, it will eventually lead to the segmentation of the signals in two situations: 1. Some samples contained a single signal, two or more flags action, and thus a single sample has multiple classification results appearing in the process of recognition. 2. Some samples become negative samples due to the absence of some signals, leading to the problem of missing features in subsequent feature extraction and learning due to the absence of signal information. Therefore, the present study proposes a time series change point detection algorithm based on differential degree threshold search, in which the detection of change points is mainly targeted at the moment when the change occurs from one state (action) to another state (action) in the time series, which is equivalent to the starting point and end point of a single sample.

The collected data can be classified into linear time series according to the signal characteristics. Suppose a flag movement signal is F, as shown in formula (1), where f_i is the value of signal at time i. Since the signals collected in the experiment include acceleration, angular velocity and Angle data, in the strict sense, f_i should be regarded as the 3D vector information on the time series, in contrast [16]. However, three kinds of data with the variation of wave amplitude, acceleration data has significant changes in signal change characteristic. Therefore, time sequence change point detection based on acceleration of the 9 shaft signal can be seen as change point detection, thus simplifying the problem to one dimensional linear time series change point detection problem.

$$F = \{f_1, f_2, \ldots, f_n\} \tag{1}$$

B. ALGORITHM RESEARCH AND MODELING ANALYSIS

Figure 5 presents a time series containing multiple flag actions. It was assumed that the sequence at the moment τ was divided into two different action sample sequences, which are denoted as $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_m\}$ where m and n represent the signal length. However, the signal detection problem in this paper can be transformed into the calculation of the amplitude of signal characteristic change at any time point τ of time series, aiming to determine whether time τ is the change point of two signals before and after. Therefore, the next task is how to find the change point between different signals by establishing discriminant function or characteristic difference model.

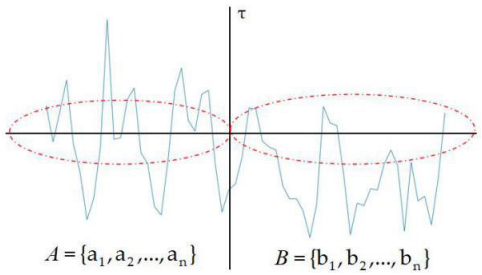


FIGURE 5. Schematic diagram of signal to be segmented in linear time series.

According to the research findings, SVM performs well in signal definition detection. Chui *et al.* [17]. proposed a threshold-searching method for sleepiness of based on SVM genetic algorithm to determine the human fatigue degree during driving using human electrocardiogram signal. Symmetric and antisymmetric information in different kinds of ECG signals are captured by cross-correlation and convolution kernel respectively. The final threshold distribution is determined by multi-objective genetic algorithm. Similarly, Mingjing Wang and Chen [18] proposed a CMWOAFS-SVM model for the study of disease classification based on Dermatology Database. Through integrating chaotic and multi-swarm strategies, they adopted whale optimization algorithm to solve the problem of parameter setting and feature selection of SVM, obtaining the best classification results in the comparison experiment with other algorithms. On this basis, this paper proposes a discriminant function model based on sliding search for change points on time series. Through the characteristic changes caused by the time t change in the time series, the hyperplane of

the two types of samples before and after the partition is found by SVM. In the proposed method, the principal component analysis method is combined with the feature fusion method to directly establish the statistical feature model for the original acceleration a, angular velocity ω and azimuth θ three-dimensional scatter information, so as to obtain the optimal solution of the time points of the action samples before and after the change [19]. Specific ideas and methods are presented as follows. Through data acquisition and preprocessing, the sequential relationship between acceleration a, angular velocity ω, azimuth θ and time t was obtained.

1) A time series sample set T is established as shown in formula (2).

$$T = \{(a_1, \omega_1, \theta_1), (a_2, \omega_2, \theta_2), \cdots, (a_n, \omega_n, \theta_n)\} \quad (2)$$

At the same time, each sampling point in the time series is mapped to three-dimensional space to form a series of three-dimensional scatter plots, as presented in Figure 6:

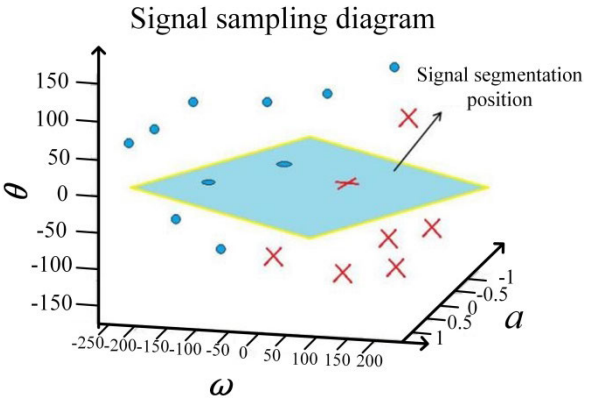


FIGURE 6. Three-dimensional scatterplot of time series signals.

2) The category definition of each signal scatter point is given. Let x_i be the sample feature and y_i be the sample category, where $x_i = (a_i, \omega_i, \theta_i), y \in \{1, -1\}$ and i is the serial number of the sampling point in the time series. The original signal data is converted into feature set by principal component analysis (PCA) [20] and features are fused by feature pyramid model. The following presents the specific details: 1. Calculate and centralize the sample mean. 2. Calculate the covariance matrix and its eigenvalues, and arrange them according to the size. 3. By setting the cumulative variance contribution rate matrix and selecting the corresponding eigenvalues, the projection matrix is formed. 4. Low-order mapping of high-dimensional features to obtain linear sets. 5. Through the feature pyramid model, the feature fusion bit value is expressed as x_{i}.

3) An interclass discriminant function for time series point I is established. In the present study, the discriminant function is defined by sliding anchor points, which are defined as. The sliding anchor points are progressively moved in the time series sampling points, and the variable A and B anomaly degree of the two categories before and after is defined, where $I = (\mathrm{A} \cup \mathrm{B} \mid \varphi)$ represents the difference degree function of A and B when the sliding anchor is located at any sampling point.

4) By setting threshold λ, the value of difference degree at time τ is measured, and the following threshold test discriminant is given. If $I = \{A \cup B \mid \varphi\} < \lambda$, then the time series does not change at this moment. If $I = \{A \cup B \mid \varphi\} < \lambda$, the time series subsequently changes at this moment.

5) The sliding anchor point is combined with the hyperplane search theory in SVM [21] to set the sliding point function:

$$F_i = U(a_i, \omega_i, \theta_i) \quad (3)$$

where $F(i)$ represents the sample quantization characteristics of sliding anchor point at sequence point i obtained by acceleration a_i, angular velocity ω_i and azimuth θ_i, and the difference degree function expressed by sliding anchor point can be obtained:

$$(A \cup B \mid \varphi) = F(i+1) - F(i) \quad (4)$$

Thus, $F(i) = x(i)$. As a result, the hyperplane function can be defined as $\kappa F(i) + \delta = 0$. The geometric interval of the sampling points slid by the sliding point can be set as:

$$\gamma = \|F(i), F(i+1)\| = y_i \left(\frac{\kappa}{\|\kappa\|} \times F(i) + \frac{\delta}{\|\kappa\|} \right) \quad (5)$$

The maximum value of the sample interval $\max_{i=1,2,\ldots,n}$ obtained is the difference threshold λ, which measures whether the sample points constitute the change of category. Therefore, the problem of solving threshold λ can be transformed into solving the maximum hyperplane problem, namely, constrained optimization problem [22]. Using SVM model, the solution:

$$\min_{\kappa,\delta} \frac{1}{2}\|\kappa\|^2, s.t. y_i (\kappa x_i + \delta) \geqslant 1, \quad i = 1, 2, \ldots, n \quad (6)$$

By transforming the objective function into the newly constructed Lagrangian objective function L(ω, δ, α) and then combining the objective functions of the two conditions inside and outside the constraint range of formula (6), the new objective function is obtained where $L(\kappa, \delta, \alpha)$ is shown in formula (7), and the new objective function is presented in formula (8):

$$L(\kappa, \delta, \alpha) = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{n} \alpha_i (y_i (\kappa x_i + \delta) - 1) \quad (7)$$

$$\zeta(\kappa) = \begin{cases} \frac{1}{2}\|\kappa\|, y_i (\kappa x_i + \delta) \geqslant 1 \\ +\infty, y_i (\kappa x_i + \delta) < 1 \end{cases} \quad (8)$$

where α_i is the Lagrange multiplier and $\alpha_i 0$. where $\zeta\kappa$ is expressed as formula (9). Thus, the solution to the original problem can be expressed as formula (10).

$$\zeta(\kappa) = \max_{\alpha_i \geq 0} L(\kappa, \delta, \alpha) \quad (9)$$

$$\min_{\kappa,\delta} \zeta(\kappa) = \min_{\kappa,\delta} \max_{\alpha_i \geq 0} L(\kappa, \delta, \alpha) = p^* \quad (10)$$

When solving the biextremum problem, the partial derivative of the objective function is obtained by taking the parameter of the objective function and setting it as 0. The function of x_i, x_j, y_i and y_j is obtained. Then, the maximum value is converted to the minimum value by the conversion of positive and negative relations. Meanwhile, the following objective function and constraint conditions (11) are obtained:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(x_i \cdot x_j\right) - \sum_{i=1}^{n}\alpha_i$$
$$s.t.\ \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geqslant 0,\ i = 1, 2, \ldots, n \qquad (11)$$

For the nonlinear optimization of such mulivariate functions, penalty parameter P could be introduced and hinge function [23] could be used to constrain it. As a result, the constraint condition of Equation (11) could be rewritten as $0 \leq \alpha_i \leq P, i = 1, 2, \ldots, n$. After we obtain the optimal $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)^T$, calculate $\kappa^* = \sum_{i=1}^{n}\alpha_i^* y_i x_i$, by selecting the value in α^* that satisfies the (11) constraint and calculate formula (12). The separation hyperplane after sliding anchor point search is obtained as presented in $\kappa x + \delta = 0$.

$$\delta^* = y_j - \sum_{i=1}^{n}\alpha_i^* y_i\left(x_i \cdot x_j\right) \qquad (12)$$

Then, the nearest sampling point φ of the distance separation overfrequency surface refers to the signal change point in the desired time series, so as to complete the category detection and segmentation task of the time series. Its implementation process can be found in Figure 7:

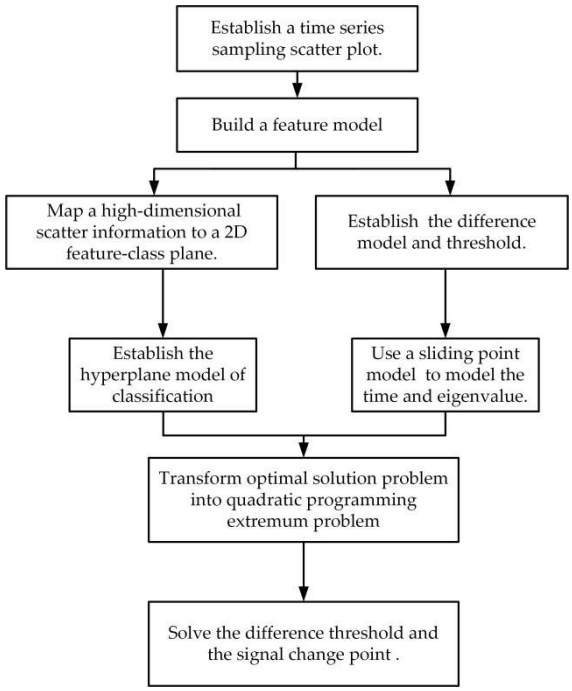


FIGURE 7. The process diagram of sliding segmentation detection signal algorithm based on SVM.

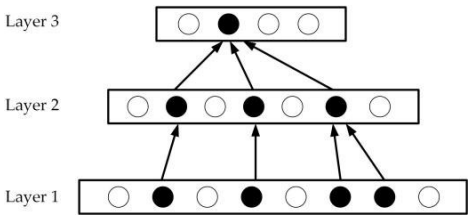


FIGURE 8. Based on Inception Net, the unit connection mode of network is constructed layer by layer.

IV. FEATURE EXTRACTION ALGORITHM BASED ON INCEPTION-ResNet

A. RESEARCH STATUS AND PROGRESS

Inception and ResNet networks are two types of convolutional neural network models that have been extrensively used in recent years. The Inception model was first proposed and used for handwritten number-recognition in 2014 [24]. Inception network in depth increases model at the same time, so as to avoid the excessive growth of network parameters, avoiding the phenomenon of over fitting the training sample. Its main thought method is to connect all and even the general convolution neural network into a sparse connection. Performing the analysis of the statistical properties of the activation values and the output of highly correlated clustering can construct an optimal network step by step, guaranteeing the network information and performance under the premise of not loss as far as possible to simplify the network connection [25]. Its structure is shown in Figure 8.

Residual Neural Network (ResNet) model was first proposed by He and Jegelka [26] in 2015. ResNet, also known as residual network, is a deep learning network that utilizes residual network structure to solve gradient dispersion and precision degradation problems in deep networks. Its main feature is the introduction of an idea of "Identity Shortcut Connection" by establishing a residue learning unit (a residue block). Specifically, congruent mapping layer is added to the deep learning to make the original input x directly serve as the initial result of the next stage after it is propagated, aiming to transform the learning of the original output $O(x)$ into the learning of the input-output difference $D(x) = O(x) - x$ in the training target [27]. Its structure is shown in Figure 9.

According to the characteristics of two kinds of network models: Inception networks are primarily the use of feature dimensions, whereas ResNet networks are more concerned with the convergence of loss functions as the result of enhanced network depth [28]. The combination of the two neural networks has become a new research hotspot. As early as 2014, Simonyon *et al.* [29]. first proposed the application of dual-stream network to solve the frame recognition problem in action video. The recognition accuracy of UCF101 and HMDB51 is 88.0% and 59.4%, respectively, Feichtenhofer *et al.* [30]. made further improvement on the original Inception-ResNet network, and realized information

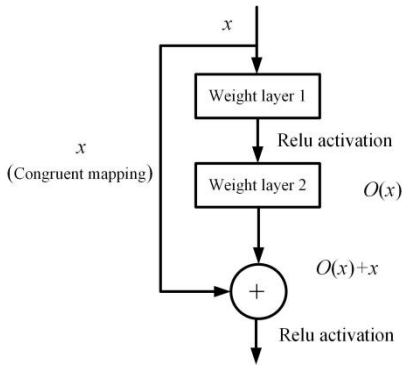


FIGURE 9. ResNet congruent mapping layer network structure.

interaction between airspace and time domain by adding residual connection between dual-flow networks. The experimental results demonstrate that the extension of the residual network from space to space-time network can increase the time residual connection, thus increasing the network's ability to understand temporal signals.

As far as current research is concerned, the Inception-ResNet network is mostly used for image or time-frequency data recognition tasks [31], while the Inception-ResNet network is rarely used for time series data set recognition tasks. The main reasons are as follows: 1. Compared with public data sets such as handwritten Numbers and face data, the number of motion data sets based on attitude sensors is relatively small; 2.The sensor data set mostly 3 axis or 9 axis data, so data format usually d $3 \times n$ or $9 \times n$ matrix format, and often lack depth information. Therefore, the Inception-ResNet model will appear due to the lack of depth information Inception-ResNet network structure so deep into the depth of feature extraction in the process of feature dimension of different imbalances leading to the final results show overfitting problem. Therefore, this paper will focus on the above two types of problems when processing 9-axis sensor data by employing the Inception-ResNet network.

B. INCEPTION-ResNet MODEL CONSTRUCTION BASED ON 9 AXIS SENSOR DATA

According to a variety of Inception-ResNet structures designed by Szegedy *et al.* [32], it was found that by adding stem layer (bottleneck layer) to the overall network structure, the depth information can be increased while the height and width information can be compressed. Two 3×3 convolution kernels are replaced by the original 5×5 convolution kernel by the convolution kernel decomposition idea, and the 3×3 convolution kernel is replaced by 1×3 and 3×1 convolution kernels. Based on the parameter calculation formula of convolutional neural network, the calculated parameters can be reduced by 33% on the original basis. In addition, more nonlinear transformations are established to increase the network's ability to learn features.

This experimental model refers to the Inception-ResNet-V3 network bottleneck layer structure and convolutional decomposition method proposed by Lu *et al.* [33]. Using the traditional structure of convolutional neural networks, a deeply separable model of convolutional neural networks is proposed and the overall structure is shown in Figure 10.

Table 2 shows the overall Inception-ResNet network parameters.

In the experiment, since the data collected is double-9-axis sensor data and the sampling points of each type of sample are set as 100 (that is, the 9-axis data at 100 consecutive moments are recorded), each single sample divided can be regarded as a two-dimensional matrix of 100×18. In the input vector setting, the 100×18 matrix is first converted into the $30 \times 30 \times 2$ input vector, because the data size is small. The lower sampling process can be skipped during the data preprocessing. In a general Inception-V3 model, there are usually nine blocks, and the standard input data dimension is at least 139×139, and 299×299 is the standard input dimension, whereas in the current experiment, only the first four blocks are used because the input dimension is small. First, the bottleneck layer of Block 1is set up. Its structure is shown in Figure 11.

After passing through the Stem layer, the original acceleration signal is transformed from $30 \times 30 \times 2$ dimensions to $18 \times 18 \times 192$ dimensions, and then used as input for the Inception-ResNet layer. As the core layer of Two-Dimensional CNN, Inception-ResNet layer is composed of residual structure and block convolution structure where the left residual structure retains the depth information in the original input, while in the right partitioned convolution structure, the 5×5 convolution kernel is decomposed into two 3×3 convolution kernels and one 1×1 convolution kernel. The convolution operation is performed on the activation results of Relu function in the previous layer. Then, the result obtained by convolution is added to Linear convolution layer, so as to transform the nonlinear features in the original input into Linear features. At the same time, the left and right convolution results are added and activated by Relu function to obtain the output vector with dimensions of $18 \times 18 \times 384$ where the Inception-ResNet structure is shown in Figure 12.

In order to further enhance Inception-ResNet to increase network width and reduce network computing parameters, this paper proposes to replace the original Reduction-A convolution block in the Inception network structure with deep separable convolution. Depth separable convolution is a form of width expansion of ordinary convolution. A depth separable convolution block is composed of a depth convolution and a 1×1 point convolution kernel in which depth convolution is used to deepen the network depth, and the point convolution kernel is used to adjust the number of output channels. The traditional convolution method is shown in formula (13). in which I represents the input eigenvector, whose magnitude is $(k + i\text{-}1) \times (i + j\text{-}1) \times m$, O represents the output eigenvector, whose size is $k \times l \times n$; K denotes the convolution kernel, whose size is $i \times j \times m$ and number is n. Formula (13) shows the operational relationship between the degree of output eigenvectors and the dimensions of input eigenvectors and

TABLE 2. Inception-ResNet network parameter settings.

Name	Layer		Output shape	Parameter		Connect to
Block 1 (Stem)	Conv1		18×18×192	2320		Input
	Conv2			2320		
	Conv3			2320		
	Maxpooling1			160		
	Conv4			2320		
	Conv5			2320		
	Maxpooling2			160		
	Activation1			0		
Block 2 (Inception-ResNet)	Conv1		18×18×384	4640		Block 1
	Conv2			9248		
	Conv3			9248		
	Conv4			544		
	Activation2			0		
Block 3(Reduction-A)	Traditional Convolution	Separable Convolution	18×18×256			Block 2
	Averagepooling1	Averagepooling1		256	256	
	Conv1	Conv1		18496	18496	
		Conv2			36928	
	Conv2	Conv3		36928	2112	
	Conv3	Conv4		2112	36928	
		Conv5			36928	
	Conv4	Conv6		180593	36928	
	Add	Conv7		0	36928	
		Add			0	
Block 4(Reduction-B)	Maxpooling		9×9×736	0		Block 4
	Conv1			18432		
	Conv2			76800		
	Conv3			82944		
	Conv4			18432		

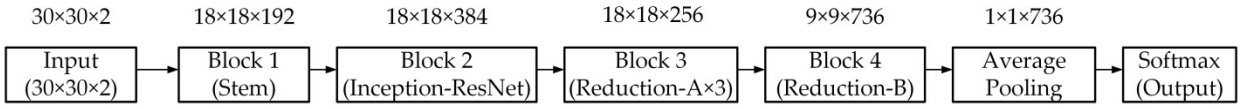


FIGURE 10. Inception-ResNet network architecture based on dual speed sensors.

convolution kernel.

$$O_{k,l,n} = \sum_{i,j} K_{i,j,n} I_{k+i-1,i+j-1,m} \tag{13}$$

$$O'_{k,l,m} = \sum_{i,j} K'_{i,j,m} F'_{k+i-1,i+j-1,m} \tag{14}$$

$$O_{k,l,n} = \sum_{\mathrm{m}} P_n O'_{k,l,m} \tag{15}$$

Equations(14) and (15) present the relationship between the output feature vector, input and convolution kernel after deep separable convolution, where K' represents the convolution kernel in deep convolution operation with the size of $i \times j \times m$, P is the convolution kernel in point convolution operation with the size of $1 \times 1 \times m$ and the number of convolution kernels is n.

Through Equations (14) and (15), it can be found that in the deeply separable convolution, the output eigenvector is obtained by the input eigenvector after two convolution transformations. In the first convolution operation process, the depth separable convolution operation is the same as the ordinary convolution operation. The difference refers to that in the process of deep separable convolution, the input vector is transformed to the same size as the output vector with the convolution kernel at a relatively shallow level.

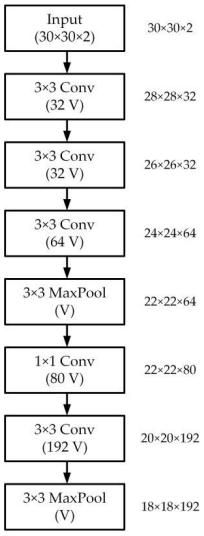


FIGURE 11. Block 1 Stem layer structure.

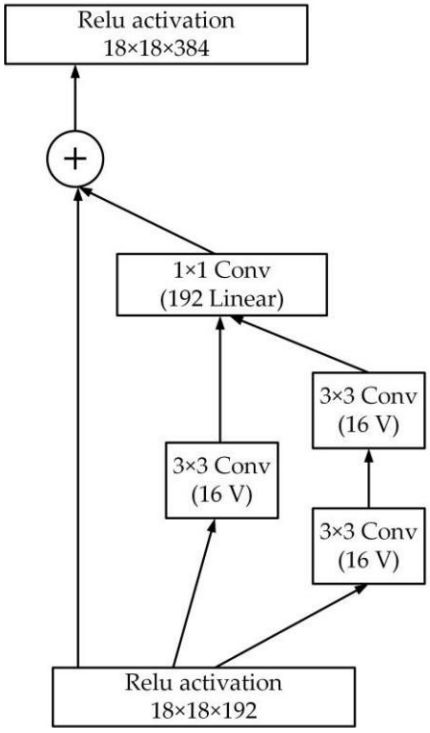


FIGURE 12. Block 2 Inception-ResNet structure.

In the second convolution process, a 1-Dimensional deep convolution kernel is used to carry out convolution operation on the intermediate output vector O', and the depth is adjusted to the required depth of output feature vector under the condition that the vector size remains unchanged. Figure 13 shows the deeply separable convolution structure. Figure 14 and 15 respectively present the traditional convolution block and the depth-separable convolution block in the dimension Reduction layer of Reduction-A.

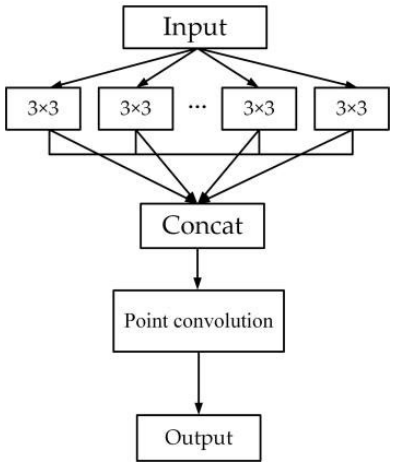


FIGURE 13. Depth separable convolution network structure.

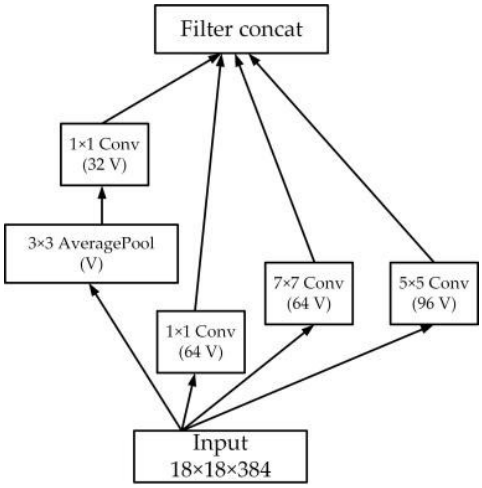


FIGURE 14. Reduction-A Traditional convolutional layer.

After obtaining the output vector activated by Relu function, convolution and pooling are carried out with it as the input vector of block 3 Reduction layer. Block 3 consists of three Reduction (dimension Reduction layer), and the original vector is input into the traditional convolutional layer and the deep separable convolutional layer respectively. In the Reduction-A depth separable convolution layer, the original 5×5 and 7×7 convolution kernel was decomposed into five 1×1 convolution kernel, two 3×3 convolution kernel and one 3×3 pooling kernel. Its structure can be found in Figure 15.

According to the results of convolution operation, both the traditional convolutional layer and the deeply separable convolutional layer can obtain $18 \times 18 \times 256$ output eigenvectors. However, by comparing the calculated parameters, it can be found that the two types of convolutional layers differ greatly in structure. Formula (16) and (17) respectively represent the total amount of parameters calculated by ordinary convolution and depth separable convolution, where I, K and O respectively represent input, convolution kernel and

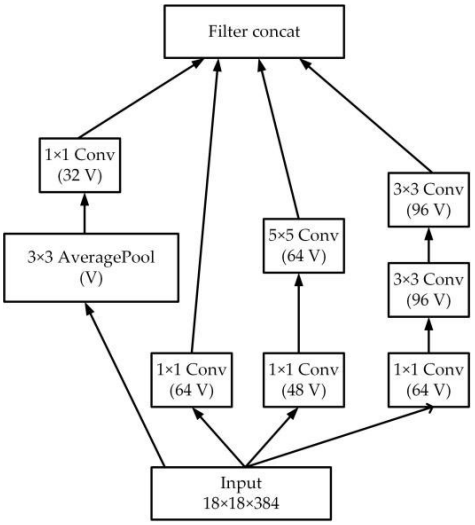


FIGURE 15. Block 3 Reduction-A depth separable convolution network structur.

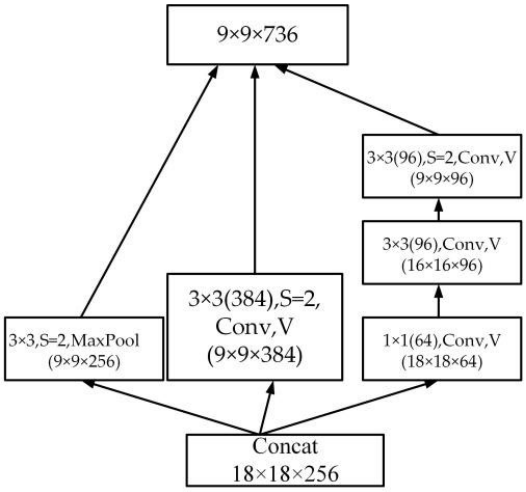


FIGURE 16. Block 4 Reduction-B layer structure.

output, and subscript w and h. i and o respectively represent the height and width of input and output; The height width of the convolution kernel.

$$N_o = I_w \times I_h \times K_i \times K_o \times O_w \times O_h \tag{16}$$

$$N_D = I_w \times I_h \times K_o \times O_w \times O_h + K_i \times K_o \times O_w \times O_h \tag{17}$$

According to the formula, the same left average pooling layer and the middle 1×1 convolutional layer can be calculated and removed. In the ordinary convolution process, the total number of calculated parameters is $N_o = 18 \times 18 \times 64 \times 18 \times 18 + 18 \times 18 \times 2 \times 96 \times 18 \times 18 = 2.7 \times 10^7$. In the process of deep separable convolution, the total number of calculated parameters is $N_D = (18 \times 18 \times 64 + 64 + 18 \times 18 \times 96 + 96 \times 2) \times 18 \times 18 = 1.7 \times 10^7$. Through comparing the total amount of calculated parameters, it can be found that adding a depth separable convolutional layer to the dimension Reduction layer of Reduction-A can reduce the calculated parameters by 37%.

It was further input into the block 4 Reduction-B layer, where the input was $18 \times 18 \times 256$. Three 3×3 and one 1×1 convolution kernel were employed for dimensionality Reduction decomposition. Meanwhile, the maximum pooling kernel of 3×3 was used to extract local key features in the original input. Finally, the output feature vector of $9 \times 9 \times 736$ was obtained, Its structure is shown in Figure 16.

The output vector was input with dimension $9 \times 9 \times 736$ into the average pooling layer of 9×9, and finally the output result of $1 \times 1 \times 736$ was obtained. Through Softmax layer, the output vector in each category was mapped to the probability distribution between (0, 1). By comparing the probability distribution size of each category, the final classification result was obtained.

V. C-L JOINT LOSS FUNCTION CLASSIFICATION MODEL

A. THEORY OF LOSS FUNCTION

In section IV, the final output feature vectors of each flag action sample are obtained through Inception-ResNet Two-Dimensional convolutional neural network. The final stage is the selection of classifier. The Inception-ResNet network is used to extract the feature vectors of the flag movement signals, and the 1×736 dimensional feature vectors are finally obtained. In the process of classification and recognition, Softmax classifier was used to take the feature vector as input and set the dimension of the output vector as the same as the number of categories [34]. Regarding each input flag action sample $x^{(i)}$, mark it as $y^{(i)} \in \{1, 2, \ldots, k\}$, a total of k class. The probability $p(y^{(i)} = j \mid x^{(i)})$ of each input sample belonging to each category j is expressed by setting the sample probability function. Then, the sample probability function hi can be expressed as formula (18).

$$f_k^i = \begin{cases} \|F(x_i)\| \cos p\theta_{(k,i)}, & k = i \\ \|F(x_i)\| \cos \theta_{(k,i)}, & k \neq i \end{cases}$$

$$h_i = \begin{pmatrix} p(y^{(i)} = 1 \mid x^{(i)}; \theta) \\ p(y^{(i)} = 2 \mid x^{(i)}; \theta) \\ \vdots \\ p(y^{(i)} = k \mid x^{(i)}; \theta) \end{pmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{pmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{pmatrix} \tag{18}$$

where,

$$\theta = \begin{pmatrix} \theta_{11} & \theta_{21} & \cdots & \theta_{k1} \\ \theta_{12} & \theta_{22} & \cdots & \theta_{11} \\ \cdots & \cdots & \cdots & \cdots \\ \theta_{1n} & \theta_{2n} & \cdots & \theta_{kn} \end{pmatrix}$$

is the parameter matrix of $736 \times k$, n = 736; $\frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}$ is the normalization of the probability distribution. In matrix θ, each column parameter represents the prediction of any sample in each action category. $LossFunction(\theta)$ is established

and the parameters in the convolutional neural network are trained to minimize the LossFunction and obtain the final weight parameter [35].

In CNN model, cross entropy is usually used as the loss function of training network. However, when the traditional CrossEntropy loss function is used to classify small data sets, overfitting is easy to occur [36]. To solve this problem, Xin-Yu *et al.* [37]. adopted a centerloss-softmax (CS) and A-Softmax (AS) joint optimization of loss function to identify and classify small face data sets of freshmen, with an average recognition rate of over 98%. In addition, the application effect based on joint loss function in small-scale data set is verified.

B. FLAG SIGNAL CLASSIFICATION LOSS FUNCTION MODEL BASED ON THE FUSION OF DICHOTOMY AND MULTI-CLASSIFICATION TASK

In the current experiment, the traditional CrossEntropy loss function is improved, and a loss function based on the combination of CrossEntropy and logistic weighting is established. CrossEntropy loss function with wide application in multi-classification task represents the difference between the predicted output and the real value in various dimensions. Bosman *et al.* [38]. proposed a gradient based random sampling method to empirically study the loss surfaces generated by two different error measures including quadratic loss and entropy loss. It is proved that entropy loss has stronger gradient and fewer fixed points compared with the secondary loss. Besides, the entropy loss function has better global search. At the same time, in the application of neural network, the global minimum value of loss surface is successfully captured. Logistic regression loss function, widely used in dichotomous tasks, represents the attribute of (0,1) belonging to a certain category for a single sample. The researchers applied Logistic loss function in the classical pattern recognition data set Iris to conduct modeling and classification prediction based on partial features and all features, proving that Logistic loss function can achieve higher classification accuracy in the modeling of higher weighted feature combination [39].The loss functions of logistic and CrossEntropy are shown in formula (19) and (20) respectively:

$$loss\left(h_\theta(x), y\right) = \sum_{i=1}^{m}\left[-y_i \lg\left(h_\theta(x)\right) - (1-y_i)\lg\left(1-h_\theta(x)\right)\right] \tag{19}$$

$$loss\left(x_i, y\right) = \sum_{i=1}^{m}\left(-x_i + \log\sum_j e^{(x_j - y_j)}\right) \tag{20}$$

where $h_\theta(x)$ is the observed value of each sample of the dichotomy task; y denotes the actual value of each sample; y_i indicates the observed value of each type of sample; x_j is the predicted probability of each category in the observed value of a sample, m is the total number of samples, j is the number of categories, the observed value of each sample in the multi-classification task, x_i is between (0,1), y_j is the actual classification of each sample under different categories, is the one-hot value. When the data is directly linearly separable, Logistic regression can find a linear decision boundary through the original linear charactristics of the data. However, for acceleration, angular velocity and azimuth fusion data, direct linear separable can not be realized. Consequently, the feature map needs to be linearly segmented in the high-dimensional data space. The Cross Entropy loss function is suitable for the classification mode of multi-class data fusion while it is prone to ''dimensional disaster'' caused by excessively high dimension of feature mapping, leading to the phenomenon of over-fitting of classification results [40]. Therefore, in the stage of experimental classification recognition, a combined loss function of linear addition of Softamx and Logistic function is adopted, as presented in formula (21), where A and B are weight coefficients of dichotomous and multi-classification tasks respectively, $a + b = 1$.

$$loss\left(h_\theta(x), x_i, y\right) = aloss\left(h_\theta(x), y\right) + bloss\left(x_i, y\right) \tag{21}$$

C. C-L JOINT LOSS FUNCTION MODEL PROCESS

The classification steps and thinking methods based on CL combined loss function algorithm are as follows:

(1) Two intersecting linear functions were used to classify the entire data set into five initial categories, respectively, C_1, C_2, C_3, C_4 and C_5.

(2) The inter-class distance problem is transformed into an Angle problem, and the Angle interval parameter is adjusted by initializing and updating the weight and bias value of the full connection layer of the convolutional neural network, aiming to continuously adjust the included Angle between classes. The specific formula is as follows, where $F(x_i)$ is all the characteristics of the ith sample, and 2 norm is taken for it in the formula; f_k^i is the characteristic of the ith sample belonging to class k; $\theta_{(k,i)}$ is the included Angle between the kth column of weight ω of the full connection layer and the ith sample, p is the Angle interval parameter. Each parameter update normalizes the kth column of weight ω, and sets the bias value to 0.

$$f_k^i = \begin{cases} \|F\left(x_i\right)\| \cos p\theta_{(k,i)}, k = i \\ \|F\left(x_i\right)\| \cos \theta_{(k,i)}, k \neq i \end{cases} \tag{22}$$

(3) Carry out high-dimensional feature mapping for logistic initial classification data, and convert planar data into multidimensional data.

(4) CrossEntropy multi-classification function is used to carry out the linear classification task of category C_6 as well as C_1, C_2, C_3, C_4 and C_5 in the high-dimensional data space for the five types of initial classification data after mapping.

VI. IMPLEMENTATION AND RESULT

A. DATA IMPORT AND ESTABLISHMENT OF 9-AXIS FLAG SIGNAL FEATURE FUNCTION

In the recognition of flag movement based on 9-axis attitude sensor data, it is common to directly extract signal features such as signal mean value, variance and skewness. However,

TABLE 3. Detection and segmentation of 7 types of original signals under different λ values.

No. Of signals	The number of sample categories divied			Standard deviation of signal length(S.D.)			The average proportion of key information in the divided sample(%)		
	λ=3	λ=4.5	λ=6	λ=3	λ=4.5	λ=6	λ=3	λ=4.5	λ=6
S1	22	21	18	0.72	0.36	0.94	84.9	82.3	54.3
S2	21	20	19	0.83	0.41	1.03	80.4	79.4	67.2
S3	22	20	20	0.86	0.33	1.06	85.6	84.5	83.2
S4	23	20	17	0.54	0.24	0.83	78.4	76.3	63.2
S5	22	20	20	0.26	0.08	0.61	85.3	84.6	81.3
S6	21	20	18	0.48	0.13	0.83	82.6	81.4	73.0
S7	20	19	19	0.79	0.22	0.55	79.2	77.5	75.6

in the experimental process, it was found that the information overlap existed in the data of velocity, angular velocity and magnetic deviation Angle in the 9-axis data. Therefore, excessive computation and poor visibility of data would occur when common signal feature extraction methods were used. Therefore, in principal component analysis of original data, ''feature redundancy'' needs to be considered due to information overlap. In the process of flag signal detection, the original 9-axis signal is extracted to form feature set to form feature set $F = [f_1, f_2, \ldots, f_m]$, and then the principal component $C = [C_1, C_2, \ldots, C_r]$(where , r is the final selected fusion feature) is established and all features are represented as linear combinations, as shown in formula (23) where a_{ij} is determined by the original data eigenvalue, and C_i is the one with the largest variance of linear combination.

$$C_i = a_{i1}f_1 + a_{i2}f_2 + \ldots + a_{im}f_m \quad \forall i \in \{1, \ldots, r\} \quad (23)$$

By comparing the contribution value of various features to the cumulative variance, three main features based on the data of acceleration a, angular velocity ω and magnetic declexion θ were selected, respectively, signal mean μ, peak P and energy E. Since the subsequent feature fusion process requires linear fusion of the three types of features, the three types of features are transformed into dimensionless indicators, namely pulse index I, peak factor C, and waveform factor W. The calculation formula is as follows, where Z_{rms} is the root mean square value of the signal.

$$I = \frac{P}{\mu} \quad C = \frac{P}{Z_{rms}} W = \frac{Z_{rms}}{\mu} \quad (24)$$

Thus, the representation of any sampling point in the flag signal is in the characteristic form, that is, $F_i = (I_i, C_i, W_i)$. Meanwhile, using the idea of multi-scale feature fusion in the feature pyramid [43], the parameter x_i is expressed as the vector sum of the three main feature components in the class-feature plane, that is:

$$x_i = \sqrt{I_i^2 + C_i^2 + W_i^2} \quad (25)$$

After establishing the function model of original sample and fusion feature, the hyperplane was searched by importing 9-axis signal data and using SVM to determine the threshold λ of discriminant degree of sample difference before and after. Table 3 shows the different detection segmentation results for the 7 original signals under the condition of different threshold. A total of 4096 sampling points are known in each original sample, including 20 types of basic flag actions. Through observing the data in Table 3, it can be found that with the continuous increase of the value of λ, the length of the divided single action sample increases, and thus the divided sample category decreases. At the same time, as the length of a single signal increases, the average proportion of key information in the sample tends to decrease. Figure 17 shows the visual difference between the signal segmentation with different λ values and the real segmentation.

According to figure 17, when $\lambda = 4.5$, the visual segmentation of the signal is closest to the real process. When λ is less than 4.5, the size of the sample is too large to cause some action films to be missing. When λ is greater than 4.5, it will also occur for the reason that the signal segmentation is too dense to cause some false samples to appear.

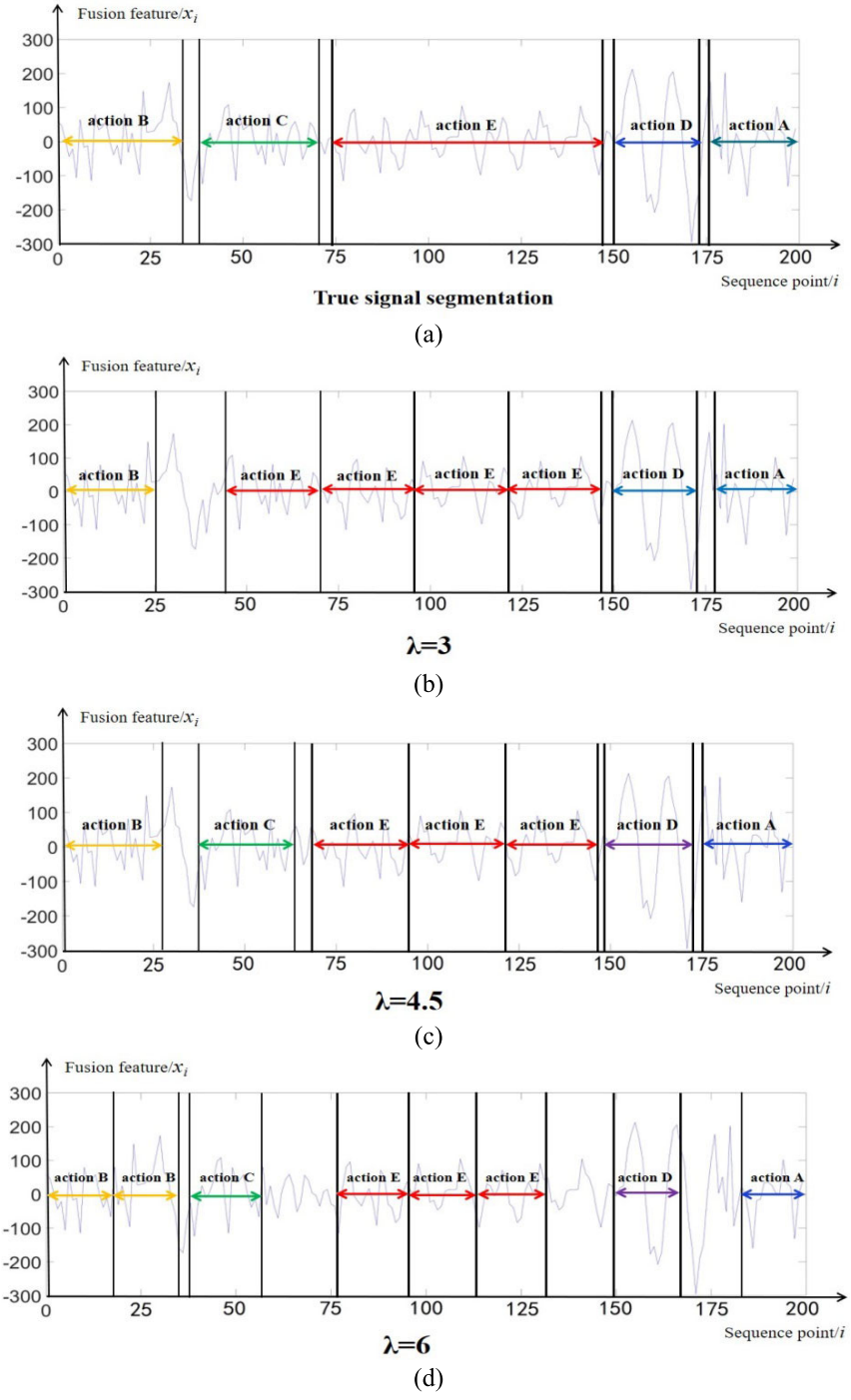


FIGURE 17. Visual segmentation results of each λ value (a) Standard condition; (b) $\lambda = 3$; (c) $\lambda = 4.5$; (d) $\lambda = 6$.

According to the standard deviation comparison, it is found that when $\lambda = 4.5$, the mean standard deviation of the 7 types of signals is small, indicating that the length difference between samples is small. Based on the comparison results of seven kinds of signal segmentation detection, it is concluded that when the threshold of signal difference fluctuates around 4.5, the signal segmentation detection effect is the best, so as to determine the final threshold of signal difference between the classes.

B. COMPARISON OF RECOGNITION ACCURACY OF VARIOUS ACTIONS BEFORE AND AFTER USING JOINT LOSS FUNCTION

This study aims to avoid the oscillation of the loss function in iteration and the problem of searching the optimal value locally [41]. After determining the form of loss function, the coefficient of loss function was trained by Adagrad gradient descent method. Table 4 shows the average loss values of various data movements with different precision.

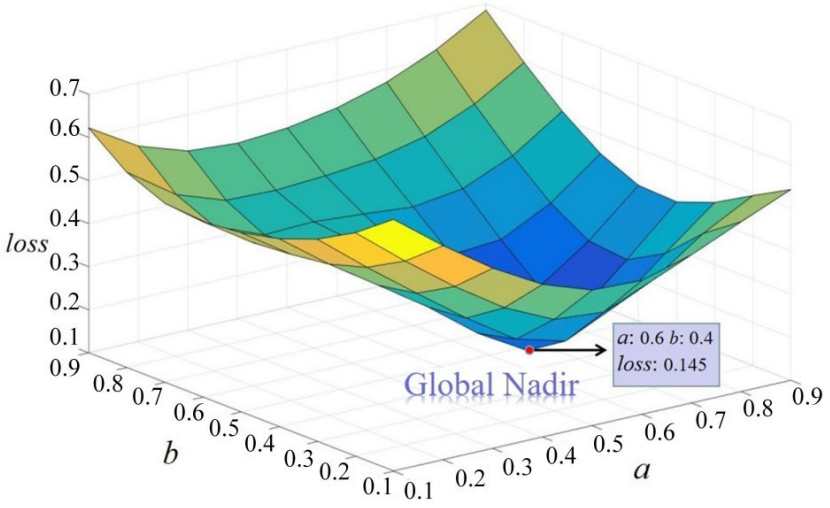


FIGURE 18. Quadric fitting surface of average Loss value varying with coefficient *a* and *b*.

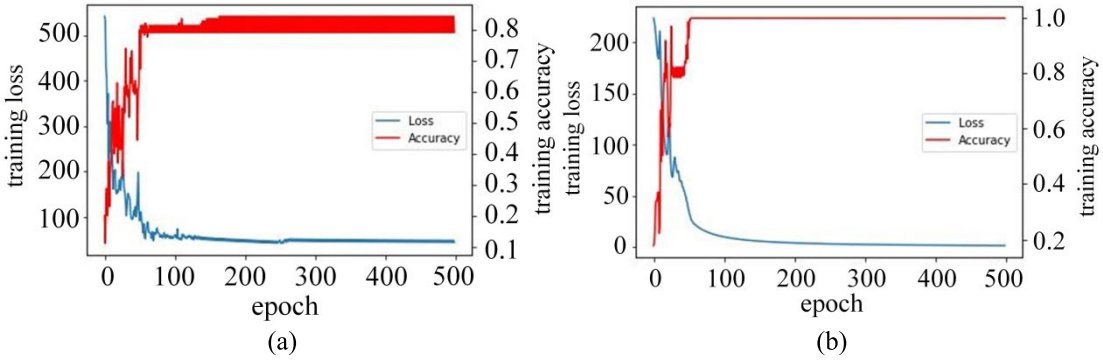


FIGURE 19. Figure 18: Before and after using C-L combined loss function, (a) classification results before using C-L loss functions, (b) classification results after using C-L loss functions.

TABLE 4. Loss values under different weights of *a* and *b*.

Coefficient values	Loss value
a=0.9 b=0.1	0.543
a=0.8 b=0.2	0.433
a=0.7 b=0.3	0.256
a=0.6 b=0.4	0.145
a=0.5 b=0.5	0.273
a=0.4 b=0.6	0.385

According to the data in the table, it can be found that with the continuous decrease of a value, the average loss value as a whole shows a downward trend, reaching the lowest point when $a = 0.6$ and $b = 0.4$. Subsequently, the average loss value rose again. As shown in Figure 18, quadric surface fitting of a, b and loss values can obtain the approximate variation trend of average loss value of joint loss function.

In order to compare the effects of C-L joint loss function and single CrossEntropy in the training set, two kinds of loss functions were used to classify all kinds of pre-processed data. Among them, the highest recognition rates of two kinds of loss functions in different data were selected for comparison. Figure 19 presents the effect changes before and after the use of CrossEntropy-Logistic loss function. Before using CrossEntropy-Logistic function, the loss values are distributed in the range of 0-500 and the oscillation amplitude is severe. The recognition rate only fluctuates around 80%. With CrossEntropy-Logistic combined loss function, the binary classification and multi-classification task weighted fusion of loss value can significantly reduce the predicted loss brought by training and keep loss value stable within 50, greatly improving the training accuracy to 99.4%.

After determining that better training effect can be achieved by using the Inception-ResNet network, in next stage, the main task is to compare the recognition effect of C-L joint loss function and traditional CrossEntropy loss function in various actions. Additionally, the same data set as the previous section is also selected for experimental comparison.

Before network operation, network parameters need to be initialized. Consequently, in the experiment, 0 initialization, random initialization and HE initialization were selected respectively. Similarly, the learning rate was set to 0.0001. Figure 20 shows the convergence of the predicted loss values under the three initialization methods.

According to Figure 20, it can be found that the predicted results of HE parameter initialization method are significantly better than the other two methods. As a result, in the subsequent classification function comparison test, HE parameter initialization method will be used to conduct initial weight training on the Inception-ResNet network.

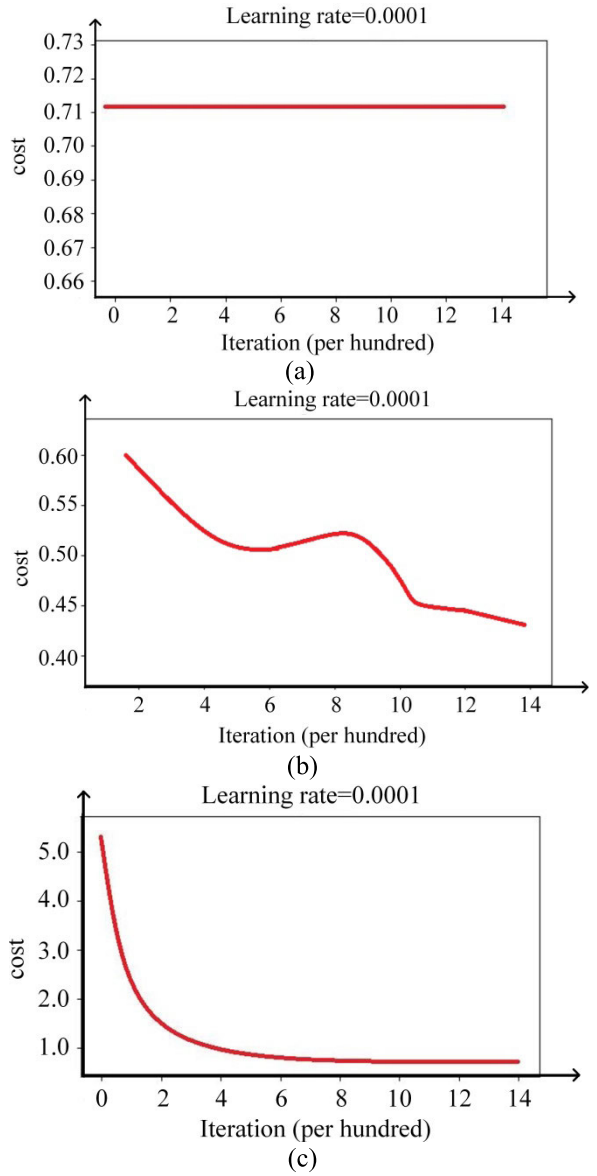


FIGURE 20. The prediction results of three types of parameter initialization methods (a) 0 initialization; (b) random initialization (c) HE initialization.

After obtaining the best method for Inception-ResNet network parameter initialization, it is of necessity to discuss the number of convolution kernels in the four convolutional layers of its core block Block2. Although, within a certain range, more feature maps can be obtained by adding the number of convolution kernels, when the number of convolution kernels increases gradually, introducing too many feature Numbers will lead to the overfitting of the final recognition, leading to the gradual decline of the recognition rate. Therefore, in the current experiment, the optimal number group of convolution kernels in the four convolutional layers will be determined by changing the number of convolution kernels layer by layer. In order to make the results more reliable, while changing the number of convolution kernels, the final combination of the number of convolution kernels is determined by averaging multiple experiments. The specific steps are presented as follows: first, the number of convolution kernels in the second, third and fourth convolutional layers is fixed. Then, the number of convolution kernels in the first convolutional layer is gradually changed from 10. After conducting the experiments, the number of convolution kernels in the first layer is obtained when the average recognition rate is the highest. Next, the number of kernel of layer 1 is fixed, and the number of convolution kernel of layer 2 is changed, aiming to obtain the number of convolution kernel of layer 2 when the average recognition rate is highest. By analogy, the number of convolution kernels in the four convolutional layers with the highest recognition rate is obtained. Figure 21 shows the change of the recognition rate with the number of convolution kernels in the four convolutional layers.

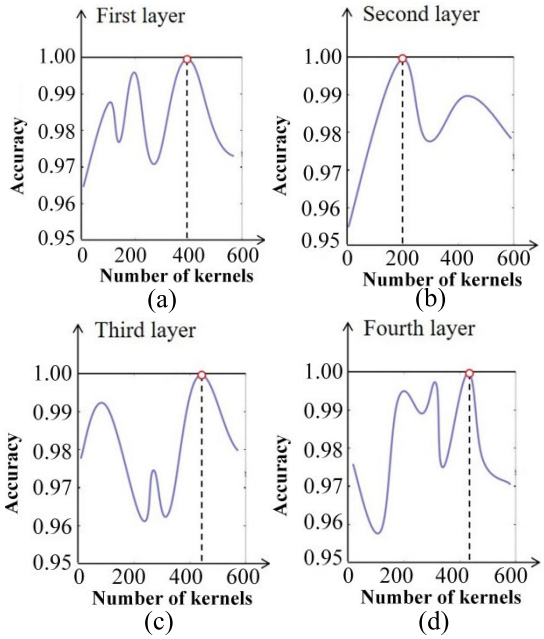


FIGURE 21. The change curve of each layer's kernels – accuracy (a) First layer; (b) Second layer; (c) Third layer; (d) Fourth layer.

According to the curve of recognition rate variation in the figure, it can be concluded that when the number of convolution kernels of layer 1 is 400, that of layer 2 is 220, that of layer 3 is 440, and that of layer 4 is 420. The parameters

TABLE 5. Ten-fold cross validation result.

Validation epoches	Method (Training time(s)/accuracy(%))					
	SVM	RF	KNN	CNN	I-R	I-R+ convolution separation
ULP1	18.33 / 81.2	14.18 / 77.3	9.84 / 63.5	22.84 / 91.5	38.36 / 96.4	27.64 / 96.7
ULP2	19.24 / 83.2	12.47 / 74.3	8.69 / 72.1	23.52 / 87.6	42.33 / 97.5	31.22 / 98.4
ULP3	17.58 / 82.6	15.11 / 80.6	10.14 / 68.4	27.68 / 89.3	44.27 / 96.4	32.62 / 97.6
ULP4	20.02 / 81.3	13.27 / 81.4	9.68 / 63.1	28.63 / 88.6	41.58 / 99.2	30.71 / 99.4
ULP5	18.50 / 85.5	12.53 / 80.8	10.32 / 65.2	26.33 / 86.4	39.66 / 95.4	28.64 / 94.6
ULP6	19.57 / 79.4	13.88 / 73.4	8.63 / 70.4	25.08 / 87.5	40.42 / 96.7	30.58 / 98.6
ULP7	18.54 / 78.9	11.52 / 77.6	9.04 / 58.6	27.40 / 89.5	39.17 / 97.4	28.56 / 97.6
ULP8	16.27/ 81.6	15.17 / 76.8	9.00 / 58.3	24.57 / 86.4	40.87 / 98.6	29.04 / 99.1
ULP9	15.49 / 82.4	14.16 / 79.4	8.91 / 55.4	23.02 / 90.4	46.14 / 96.7	33.27 / 96.7
ULP10	18.81 / 86.9	12.55 / 78.3	7.62 / 68.2	25.18 / 86.4	44.08 / 97.8	31.09 / 98.3

setting at this time satisfies the requirement of the highest accuracy of recognition rate.

C. THE EFFECT OF SEGMENTATION SIGNAL IS COMPARED IN VARIOUS NETWORK FRAMES

The Inception-ResNet-V2 motion recognition model was developed using Tensorflow 2.0, which is an artificial intelligence library using data flow graphs to build models [42]. The whole experiment is operated on a computer equipped with i5 CPU and 8GB RAM. Integrated development environment for Spyder editor. The learning rate of convolutional neural network was set as 0.0001, and the number of test iterations was set as 500.

After the completion of the whole neural network feature extraction and recognition framework, the flag movement data after segmentation and detection is input. In order to compare the classification results of various actions in the experimental model, a total of 500 samples of each of six types of actions A, B, C, D, E and F were selected for the experiment. The action samples were divided into 10 data sets on average, denoted as ULP1, ULP2. . .ULP10. The average data set contains 50 samples for each of six types of actions. The data sets from ULP1 to ULP10 were trained by 10-fold cross-validation. ULP1 to ULP10 are taken as test sets and the remaining 9 sets are regarded as training sets respectively. Table 5 shows the comparison of networks running time and training accuracy after 10 Cross-Validation results under several model structures respectively.

According to the table data, the following conclusions can be drawn. Judging from the network running training time, the average running time of the three types of deep learning networks exceeds that of the other three types of machine learning methods. Among them, KNN learning mode network has the shortest running time, also proving its characteristics of less arithmetic process and simple network framework. The Inception-ResNet network has the longest running time due to its deep structure and a lot of computational parameters. Through observing the comparison of the running time of I-R network before and after separable convolution, it can be found that the addition of separable convolution layer can not only greatly reduce the network calculation parameters, but also reduce the running time by about 10 seconds on average. Based on the comparison of recognition rates, although KNN method has a short running time, due to its limited computational parameters, its low recognition accuracy still appears in the classification of complex signals. The Inception-ResNet network based on separable convolution has obviously achieved high identification accuracy. In the meanwhile, it can be concluded that under the condition of not changing the network framework, the network running time of any algorithm is usually inversely proportional to the recognition accuracy. Therefore, considering the operation time and recognition rate of the network, the Inception-ResNet network based on separable convolution has the best comprehensive performance.

At the same time, the Inception-ResNet-V2 network, traditional CNN, Random Forest, SVM and K-Nearest Neighbors were selected as experimental comparison respectively. Totally 20 random experiments were conducted for each recognition framework, and the most general Cross Entropy loss function was selected from the selection of classification function of convolutional neural network. According to the classification results, the recognition results are compared as shown in Figure 22. As can be seen from the data in the figure, Inception-ResNet-V2 model achieves the best effect during flag movement recognition. In 20 random training experiments, the average recognition accuracy was 96%, the K-nearest Neighbors framework achieves the worst recognition effect in recognition and the average accuracy was 66%. In addition, it can also be found that the two kinds of deep learning methods have obvious advantages over the traditional three representative machine learning methods in

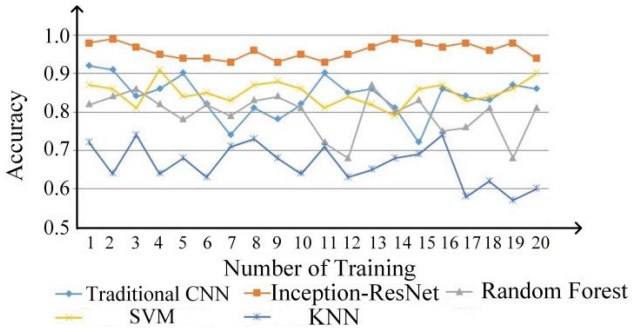


FIGURE 22. The accuracy of five classification models.

the process of flag signal recognition. Through comparing the two types of deep learning methods, it can be found that adding depth separable convolution and residual structure on the basis of traditional CNN can increase the recognition rate to a certain extent while reducing the calculation parameters.

Figure 23 shows the identification result confusion matrix of loss function in 6 kinds of actions by using traditional cross entropy loss function and C-L joint loss function respectively.

According to the confusion matrix in Figure 23, it can be found that in the comparison of recognition effects of six types of actions, action E achieves the best recognition effect, while action D has poor recognition effect in both types of loss functions. Its action characteristics show that the main components of action E are circular acceleration and angular velocity signals, and thus there will be more of the same components in the angular velocity characteristics. As a result, it is easy to distinguish it from other actions. In the operation process of action D, the swinging motion of the arm is similar to that of Action C and Action A. Besides, the action of one hand coincides with that of action C and D. Therefore, the size and change rule of some data values of the single sensor are similar. Thus, part of the data in action D will appear in the recognition process. Since the data in the first several axes are similar to action C, the local optimal solution similar to action C will be obtained directly when the loss function is employed to calculate the iterative loss value of action D. Thus, C and D action obfuscation in the obfuscation matrix appears. Although the two methods have this phenomenon, the horizontal comparison of the two types of data shows that the mixing efficiency of C-L joint loss function is significantly lower than the traditional cross entropy loss function, even if the C and D actions are confused. The reason refers to hat the C-L joint loss function adds logistic regression dichotomy on the basis of the traditional cross entropy loss function, more than in the initial classification task, on the basis of analysis on the characteristics of the sample mapping to high dimension space, again using logistic regression in high-dimensional space mission comparing differences between similar signal, to a certain extent, to avoid the traditional cross entropy loss function in the classification task of ''access to local optimal solution'', thus raising the similar action between recognition rate.

Figure 24 shows the overall recognition rate curves of the two types of loss functions respectively. According to the data shown in the table, compared with the traditional cross entropy loss function, the Inception-ResNet-V2 model based on C-L joint loss function has achieved obvious advantages in the process of classification and identification of six basic flag actions.

According to the recognition rate curve in Figure 24, it can be found that the C-L joint loss function tends to converge around 15 rounds of iteration. Besides, the average recognition rate can reach more than 99%. However, the traditional CrossEntropy loss function oscillates significantly in the course of 50 rounds of iterative training. Although the initial recognition rate is slightly higher, the highest recognition rate still remains lower than the C-L joint loss function. Meanwhile, it oscillates greatly between 97% and 98%. It is also proved that the traditional cross entropy loss function appears the problem of ''interval oscillation'' in the process of gradient descent, causing the misclassification of the local optimal solution to the global optimal solution, which can thus lead to the difference in recognition.

D. THE RECOGNITION RESULTS OF FLAG MOVEMENT UNDER DIFFERENT RECOGNITION FUNCTIONS

Through the comparison experiments of feature extraction framework and recognition function, the network framework for the optimal result of flag signals classification task is obtained, that is, feature extraction is carried out through separate Inception-ResNet network, and classification task is completed by CrossEntropy-Logistic joint loss function. Simultaneously, in order to compare the recognition effect difference between this model and several mainstream network models at present, in the practical application, the original signals are collected again, and the signals are classified and recognized by employing three network frameworks, respectively, C4.5, Multi-Layer Perceptron (MLP) and Ensemble Learning.

The MLP method is proposed by Catal *et al.* [44], and the method of classification is used by the Logical regression algorithm, which has achieved a satisfying recognition effect compared with the method of J48, which is used by Wenchao *et al.* [45]. Kwapisz *et al.* [46]. extracted the fusion features of actions by using C4.5, and also employed logistic regression function to classify them, obtaining a high recognition rate of actions. Alsheikh *et al.* [47]. improved the Ensemble Learning algorithm, extracted the depth features of the action samples, and used a single CrossEntropy loss function to identify and classify them, achieving good results in the case of a single sample with a small feature dimension. In Table 6, the effect comparison between the network frame used and the other three frames in flag movement recognition can be found.

Table 7 presents the classification and recognition rates of four types of methods in A, B, C, D, E and F six actions respectively, so as to obtain the difference in recognition rates of several types of recognition frameworks in specific actions.

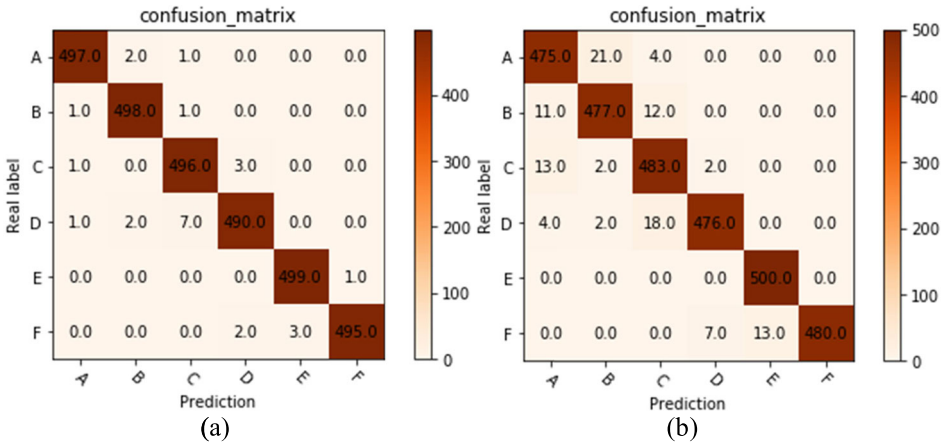


FIGURE 23. Comparison of recognition effect of two kinds of loss functions, (a) C-L joint loss function recognition effect, (b) Cross Entropy loss function recognition effect.

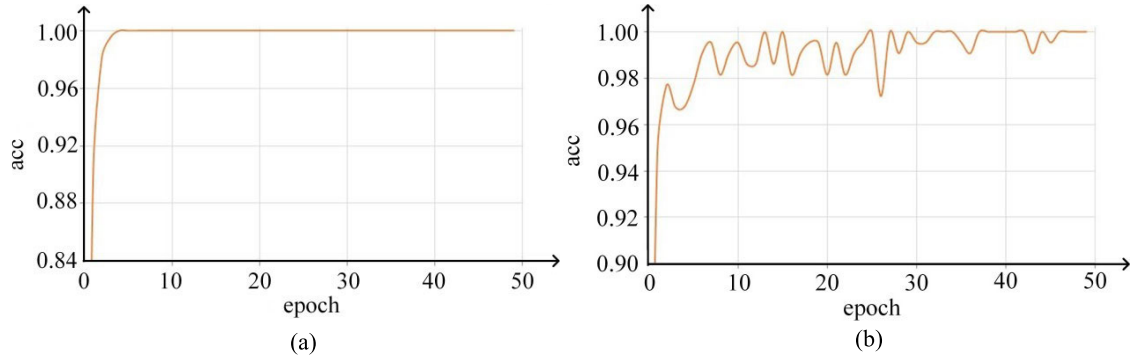


FIGURE 24. Comparison of recognition rate curve of two kinds of loss functions, (a) C-L joint loss function recognition, (b) Cross Entropy loss function recognition.

TABLE 6. A comparison between the proposed method and existing methods.

Reference	Method	Accuracy (%)
Catcal et al [44].	MLP+Logistic	91.7
Kwapisz et al [46].	C4.6+Logistic	85.1
Alsheikh et al [47].	Ensemble Learning+CrossEntropy	94.3
Ours	I-R (with Separable convolution)+C-L	99.4

TABLE 7. Comparison of recognition rates of six types of flag movements recognition (%).

	MLP+Logistic	C4.6+Logistic	Ensemble Learning+Cross Entropy	I-R (with Separable convolution)+CL
A	88.4	81.6	91.3	98.7
B	89.1	83.7	93.4	99.3
C	88.7	86.3	94.6	99.2
D	90.1	84.2	95.1	99.1
E	92.4	88.9	96.2	99.8
F	87.6	87.0	92.4	98.9

It is evident from the data in the table that the Inception-ResNet network framework based on the C-L joint loss function significantly improves the recognition rate of actions in comparison with the existing methods. The average recognition rate can still be improved by 4.9% compared with the Ensemble Learning method with the highest recognition rate. Thus, it is proved that the joint loss function has an inestimable advantage in the recognition of more complex upper limb movement signals. At the same time, according to the specific recognition rate of each action in Table 7, it can be observed that there exists little difference between

the recognition results of several methods in the six types of flag movement, among which the recognition rate of action E is the highest, which is not different from the experimental results in the I-R network training stage. Therefore, it can be concluded that when the data difference between flag gestures is small, the results are usually better under various recognition frameworks.

VII. CONCLUSION

To conclude, in present studys, the flag signal acquisition and classification learning systems are improved from three aspects of signal segmentation detection, feature extraction framework and classification recognition. The sliding point search signal mutation point model was used to obtain relatively complete samples to be classified. Meanwhile, the Inception-ResNet model based on a deep separable convolution block was used to reduce the calculation parameters of feature extraction from the network by 37%, consequently reducing the response time of feature classification task. Finally, on the basis of the original cross entropy loss function, the recognition rate was increased from 94.5% to over 99% by adding logistic regression dichotomous task, and the convergence effect of classification loss value was obtained in the number of iterations with few rounds. Through the improvement of the whole framework of flag movement acquisition and classification, the recognition task requirements for totally 6 kinds of basic movements are fulfilled. Finally, compared with the other three classification methods, the separable I-R network based on C-L joint loss function has achieved a significant improvement in the recognition rate of six flag actions.

REFERENCES

[1] J. H. Richardson and C. J. Witkowski, "Remote communication system and method using modified semaphore flags," U.S. Patent 12 551 025, Dec. 24, 2009.

[2] G. G. Demisse, "Pose encoding for robust skeleton-based action recognition," in *Proc. Vis. Understand. Hum. Crowd Scene (CVPRW)*, 2018, pp. 188–194.

[3] Z. Enyang *et al.*, "Design of application software for wearable medical monitor system," *China Med. Devices*, vol. 33, no. 1, pp. 53–56, 2018.

[4] W. Wu, "Classification accuracies of physical activities using smartphone motion sensors," *J. Med. Internet Res.*, vol. 14, no. 5, p. 130, 2012.

[5] Y. Kwon, K. Kang, and C. Bae, "Unsupervised learning for human activity recognition using smartphone sensors," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6067–6074, Oct. 2014.

[6] S. Matsui, N. Inoue, Y. Akagi, G. Nagino, and K. Shinoda, "User adaptation of convolutional neural network for human activity recognition," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 753–757.

[7] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015

[8] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1868–1873.

[9] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, "Transfer learning based histopathologic image classification for breast cancer detection," *Health Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–7, Dec. 2018.

[10] K. Kim and Y. K. Cho, "Effective inertial sensor quantity and locations on a body for deep learning-based worker's motion recognition," *Autom. Construct.*, vol. 113, May 2020, Art. no. 103126.

[11] O. Gronz, P. H. Hiller, S. Wirtz, K. Becker, T. Iserloh, M. Seeger, C. Brings, J. Aberle, M. C. Casper, and J. B. Ries, "Smartstones: A small 9-axis sensor implanted in stones to track their movements," *Catena*, vol. 142, pp. 245–251, Jul. 2016.

[12] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, Apr. 2014.

[13] G. Hu, K. Wang, Y. Peng, M. Qiu, J. Shi, and L. Liu, "Deep learning methods for underwater target feature extraction and recognition," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–10, Oct. 2018.

[14] A. Sun, T. Zhao, J. Chen, and J. Chang, "Comparative study: Common ANN and LS-SVM exchange rate performance prediction," *Chin. J. Electron.*, vol. 27, no. 3, pp. 561–564, May 2018.

[15] K. Ma, F. Dong, and B. Yang, "Large-scale schema-free data deduplication approach with adaptive sliding window using MapReduce," *Comput. J.*, vol. 58, no. 11, pp. 3187–3201, Nov. 2015.

[16] L. Huang, H. Jiang, and H. Wang, "A novel partial-linear single-index model for time series data," *Comput. Statist. Data Anal.*, vol. 134, pp. 110–122, Jun. 2019.

[17] K. T. Chui, K. F. Tsang, H. R. Chi, B. W. K. Ling, and C. K. Wu, "An accurate ECG-based transportation safety drowsiness detection scheme," *IEEE Trans. Ind. Informat.*, vol. 12, no. 4, pp. 1438–1452, Aug. 2016.

[18] M. Wang and H. Chen, "Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis," *Appl. Soft Comput.*, vol. 88, Mar. 2020, Art. no. 105946.

[19] L. Bo and N. Qian, "Mechanical bearing fault diagnosis based on feature fusion and KPCA_GA-SVM," *Mod. Comput.*, vol. 11, pp. 32–38, 2019.

[20] M. Imaizumi and K. Kato, "PCA-based estimation for functional linear regression with functional responses," *J. Multivariate Anal.*, vol. 163, pp. 15–36, Jan. 2018.

[21] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: A review," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 803–855, Aug. 2019.

[22] 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.* Cham, Switzerland: Springer, 2016.

[23] H.-J. Xing and M. Ji, "Robust one-class support vector machine with rescaled hinge loss function," *Pattern Recognit.*, vol. 84, pp. 152–164, Dec. 2018.

[24] P. Chudzik, S. Majumdar, F. Caliva, B. Al-Diri, and A. Hunter, "Exudate segmentation using fully convolutional neural networks and inception modules," *Image Process.*, vol. 10574, Mar. 2018, Art. no. 1057430.

[25] W. Zhang, J. Cen, and H. Zheng, "Temporal inception architecture for action recognition with convolutional neural networks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3216–3221.

[26] H. Lin and S. Jegelka, "Resnet with one-neuron hidden layers is a universal approximator," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6169–6178.

[27] J. Duan *et al.*, "A novel ResNet-based model structure and its applications in machine health monitoring," *J. Vib. Control*, pp. 1–15, 2020.

[28] M. Habibzadeh Motlagh, M. Jannesari, Z. Rezaei, M. Totonchi, and H. Baharvand, "Automatic white blood cell classification using pre-trained deep learning models: ResNet and inception," in *Proc. 10th Int. Conf. Mach. Vis. (ICMV)*, Apr. 2018, Art. no. 1069612.

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

[30] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition processing systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3468–3476.

[31] F. Baldassarre, D. G. Morín, and L. Rodés-Guirao, "Deep koalarization: Image colorization using CNNs and inception-resnet-v2," 2017, *arXiv:1712.03400*. [Online]. Available: https://arxiv.org/abs/1712.03400

[32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: https://arxiv.org/abs/1602.07261

[33] W. Lu, H. Sun, J. Chu, X. Huang, and J. Yu, "A novel approach for video text detection and recognition based on a corner response feature map and transferred deep convolutional neural network," *IEEE Access*, vol. 6, pp. 40198–40211, 2018.

[34] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Inf. Sci.*, vol. 428, pp. 49–61, Feb. 2018.

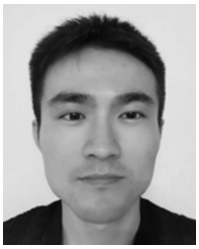
[35] K. Agbodah "The determination of three-way decisions with decision-theoretic rough sets considering the loss function evaluated by multiple experts," *Granular Comput.*, vol. 4, pp. 285–297, May 2018.
[36] Z. Li, K. Kamnitsas, and B. Glocker, "Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 402–410.
[37] Z. Xin-yu, Y. Ming-yu, Z. Jiang, and H. Xuan, "Face recognition of small-scale dataset based on joint loss functions," *Trans. Beijing Inst. Technol.*, vol. 40, no. 2, pp. 163–168, 2020.
[38] A. S. Bosman, A. Engelbrecht, and M. Helbig, "Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions," *Neurocomputing*, vol. 400, pp. 113–136, Aug. 2020.
[39] Q.-T. Nguyen-Vuong, Y. Ghamri-Doudane, and N. Agoulmine, "On utility models for access network selection in wireless heterogeneous networks," in *Proc. IEEE Netw. Oper. Manage. Symp. (NOMS)*, Brazil, Salvador, Apr. 2008, pp. 144–151.
[40] Z. Xiaohui, "Data classification based on Logistic regression," *Intell. Comput. Appl.*, vol. 6, no. 6, pp. 139–140 and 143, 2016.
[41] Y. Li and Y. Li, "Feature extraction of underwater acoustic signal using mode decomposition and measuring complexity," in *Proc. 15th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2018, pp. 757–763.
[42] M. Abadi, "TensorFlow: Learning functions at scale," *ACM SIGPLAN Notices*, vol. 51, no. 9, p. 1, Dec. 2016.
[43] K. Xiao and D. Mingzhi, "Detection of maize seeds based on multi-scale feature fusion and extreme learning machine," *J. Image Graph.*, vol. 21, no. 1, pp. 24–38, 2016.
[44] C. Catal, S. Tufekci, E. Pirmit, and G. Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition," *Appl. Soft Comput.*, vol. 37, pp. 1018–1022, Dec. 2015.
[45] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1307–1310.
[46] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newslett.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.
[47] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H. P. Tan, "Deep activity recognition models with triaxial accelerometers," in *Proc. Workshops 13th AAAI Conf. Artif. Intell.* 2015, pp. 1–7.

ZHONG YUE was born in 1996. He received the bachelor's degree in engineering from the East China University of Science and Technology, Shanghai, China, in 2018. He is currently pursuing the master's degree with the Army Engineering University, People's Liberation Army. His research interests include pattern recognition and digital signal processing.

JIQING LUO was born in 1997. He received the bachelor's degree in engineering from the Taiyuan University of Technology, Taiyuan, China, in 2019. He is currently pursuing the master's degree with the Army Engineering University, People's Liberation Army. His research interests include pattern recognition and target detection.

FANG HUSHENG was born in 1979. He is currently pursuing the Ph.D. degree with the Army Engineering University, People's Liberation Army. He worked with the Army Engineering University, People's Liberation Army. He is also an Associate Professor. His research interests include pattern recognition and embedded technology.

FAMING SHAO was born in 1978. He received the Ph.D. degree from the Army Engineering University of PLA, China. He is currently an Associate Professor with the Army Engineering University of PLA. His research interests include signal processing, deep learning, and software engineering.



ZHOU RANZHI was born in 1995. He received the bachelor's degree in engineering from the Inner Mongolia University of Science and Technology, Inner Mongolia, China, in 2018. He is currently pursuing the master's degree with the Army Engineering University, People's Liberation Army. His research interests include computer vision and deep learning.