Abstract: Multimodal research is an emerging field of artificial intelligence, and the analysis of dangerous driving behavior is one of the main application scenarios in the field of multimodal fusion. Aiming at the problem of data heterogeneity in the process of behavior classification by multimodal fusion, this paper proposes a low-rank multimodal data fusion method, which utilizes the complementarity between data modalities of different dimensions in order to classify and identify dangerous driving behaviors. This method uses tensor difference matrix data to force low-rank fusion representation, improves the verification efficiency of dangerous driving behaviors through multi-level abstract tensor representation, and solves the problem of output data complexity. A recurrent network based on the attention mechanism, AR-GRU, updates the network input parameter state and learns the weight parameters through its gated structure. This model improves the dynamic connection between modalities on heterogeneous threads and reduces computational complexity. Under low-rank conditions, it can quickly and accurately classify and identify dangerous driving behaviors and give early warnings. Through a large number of experiments, the accuracy of this method is improved by an average of 1.76% compared with the BiLSTM method and the BiGRU-IAAN method in the training and verification of the self-built dataset.

Keywords: FMCW radar; dangerous behavior driving detection; low-rank tensor learning; multimodal fusion; AR-GRU; attentional mechanisms

1. Introduction

Hazardous driving is a major contributor to vehicle crashes and fatalities, including violent and emotional driving behaviors and a variety of other poor driving behaviors; according to the National Highway Traffic Safety Administration (NHTSA), traffic fatalities in 2021 increased by an estimated 10.5% over the previous year, with urban roadway fatalities rising by 16%. Therefore, hazardous driving behavior detection is essential for effective control strategies for drivers, and the improvement of mass transit safety is potentially valuable.

Typical sensor approaches for wireless perception include Wi-Fi, the Wi-Fi link selection continuous state decision process [1], and the Scalable Uplink Multiple-Access (SUMA) protocol for Wi-Fi backscatter systems [2]. The method improves the performance of the measurement action, in which the CSI (channel state information) [3,4] is an important part of the modeling state. Sensing information from sensors, such as body-worn sensors [5] and the Industrial Internet of Things [6], has led to the development of multi-information fusion algorithms and action-recognition technologies. Among these is video information, which contains multimodal information, such as reflected brightness information, distance, and heat distribution. However, in the complex environment of vehicle driving, deviations in behavior recognition may be caused by environmental brightness distortion and the misjudgment of distance heat information [7]. The traditional array radar can only monitor a single target object point-to-point, while the method based on the FM continuous
wave encapsulates the millimeter-wave radar antenna array, which solves the problem of multi-path and complex environments caused by the narrow space inside the vehicle. This offers an excellent solution. Therefore, FMCW has a good prospect in human body pose estimation and microwave neighborhoods [8,9].

The purpose of multimodal fusion is to maintain the original method of the model in combination with multiple effective modalities and unified high-dimensional representation; therefore, the key problem in multimodal data is the existence of data heterogeneity. The normalized high-representative data can reflect the essential features of the original recognition, such as speech recognition fusion facial emotion recognition [10–12] and self-supervised learning (SSL), to extract multiple features for emotion recognition. In the process of FMCW radar frame signal preprocessing, features such as Doppler, velocity, scattering point difference, and amplitude are usually extracted separately, and two kinds of excellent features are randomly combined [13–15], which may not solve the problem of extracting features given the coupling relationship between each feature and the balance of feature space. In the literature [16–18], other methods explore the high-dimensional representation for multi-feature fusion, use the multi-feature fusion algorithm of HSV, LBP, and LSFP to improve high-dimensional fusion representation, and binarize the DM image constructed by multi-feature fusion. Human complex targets monitor small target features, perform algorithm fusion for local entropy, average gradient strength, and other features, and perform peak normalization processing in vector space (MFVS). Micro-action multimodal classification generally uses the Hidden Markov Model [19], more complex actions or gestures using a convolutional neural network (CNN) to verify the public NTU-Microsoft-Kinect through the DisturbIoU algorithm in [20,21], and the Gesture (NTU) dataset and the Vision for Intelligent Vehicles and Applications (VIVA) dataset.

In this paper, we use the FMCW radar-parsing raw data for DAM to fuse micro-Doppler, velocity Doppler, and amplitude features in low-rank tensor representation (LMF) and use the attention mechanism to assign weights in the gating system of the ReLU activation function. This method solves the problems of heterogeneous fusion confusion and feature space balance of multimodal fusion, and makes the following contributions: (1) It fuses the micro-Doppler spectrum with radial velocity and amplitude complex features in a dynamic sparse cascade, and couples and fuses low-dimensional action information features into a high-dimensional linear vector representation that can be represented dynamically, thus improving model stability as well as classification efficiency. (2) By context modeling and using a low-rank tensor network (LMF), the tensor representation is created by computing the outer product of three different single-peaked modes, and modeling the inter-interactions achieves end-to-end learning. (3) The proposed method in GRU-gated cells uses a self-attentive-rectified-linear-unit (AReLU)-based activation in GRU cells, and the performance of FM continuous-wave tensor representations can be improved by the activation function.

The rest of this paper is organized as follows. In the Section 2, an overview of the related methods is provided. In the Section 3, the related principles and technologies of radar information feature extraction and feature fusion are introduced. In addition, the Section 4 describes the implementation principle and parameters of the feature fusion method and classification model LMF-AR-GRU in detail. The Section 5 details the experimental setup and the analysis of the experimental results, and the overall efficiency and advantages of the system are compared and analyzed.

2. Related Work

2.1. Semantic Recognition Methods for Behavior Extraction

Dangerous driving behavior recognition is one of the most intuitive human–computer interaction methods. The behavior recognition methods include visual images, biological signals, inertial sensors, and radar signals. The image-based method captures RGB video information from multiple angles through the camera to identify and make the image sequence of the target heterogeneous. Khan, F.S. et al. used the bag-of-words image
representation method in the study of static images [22]. Xin, M. et al. proposed a composite triple-loss function for high-similarity image information to solve the problem of intra-and inter-class confusion in action recognition [23]. Due to the poor light reflection effect in the car due to safety factors, the recognition accuracy is weakened from the input layer. Behavior recognition based on biological signals includes EEG signals, EMG signals, etc., and mainly uses the current information generated during muscle movement in the EMG for identification. Song, J.Y., and others still apply the wearable EMG acquisition system through the exoskeleton. LSTM was used to study the effect of different features of EMG signals on behavioral accuracy [24]. Based on sensors such as inertial sensors, gyroscopes, etc., Sun, P.S. et al. used a quadratic discriminant analysis (QDA) classifier and support vector machine (SVM) classifier through cascade classification using inertial sensors [25]. The degree of dependence is still a problem in in-vehicle perception.

Wu Zhu et al. used the channel state information of a non-wearable and low-cost Wi-Fi R.F. signals are more difficult to use in the feature extraction of Doppler information. Therefore, Zhang, Z.L. et al. used a millimeter-wave radar to extract human motion Doppler features and proposed an adaptive extreme learning machine (ELM) method to improve the recognition rate in high-noise environments and obtained good results [27]. In this paper, a 77 GHz millimeter-wave radar was used to solve the limitations of other sensing input layers. The advantages of FMCW radar are simple, including an accurate range and speed measurement, independent of the environment, and a UHF high-bandwidth FM continuous wave, which lead to good speed resolution and distance resolution in in-vehicle target sensing. Figure 1 represents the millimeter-wave identification method equipment and environment.

2.2. Feature Fusion and Classification Methods

Feature fusion is not simply the process of stacking modal features, but also involves decomposing multidimensional patterns to generate a high-quality fusion feature for heterogeneous low-frequency fusion [28]. Dynamic cohesion within modalities is a key issue for classification. Zhang, Z.F. proposed a memory fusion network, a convolutional long short-term memory network with different attention (Conv-LSTM), and a fully connected LSTM (FC-LSTM) fusion proposed Spatial-Temporal Dual Attention Network (STDAN) to address the problem of human action recognition [29]. Du, P.F. proposed the fusion method of Dual Attention Modal Separation Fusion Network (BAMS), which fuses and separates the representations of each modality so that the independent modalities can be directly fused and represented [30]. Abavisani, M. proposed an adversarial encoder–decoder-
classifier framework to perform multidimensional clustering of multimodal branches in a convolutional neural network, and introduced adversarial training to make the expression layers of each modality consistent [31]. Since differences can lead to data ambiguity and data loss during the normalization process, Li, M. proposed a tensor fusion network (DTFN), which introduced the proposed features into the tensor fusion layer to increase the spatial correlation. The real-time performance of dangerous driving warnings is extremely strong, and the high-dimensional tensor representation caused by the outer product leads to an exponential increase in the computational complexity of the system. Nie, W.Z. proposed a modality fusion network for individual local dependencies, which contains a bidirectional corpus multi-connection LSTM-FFN fusion network, thus improving the local dependency transfer between feature blocks [32] and avoiding the problem of exploding gradients. Z. Bai proposed a low-rank tensor fusion method whose tensor representation is based on a GRU-gated structure to fuse context-aware multimodal features [33]. Therefore, the context modeling of low-rank tensors can solve the situation of the high complexity of the model and has end-to-end fusion features, leading to a good performance in emotion recognition, action recognition, and gesture recognition [34–37]. Contextual multimodal integration of advantageous resources allows data flow to be effectively modeled dynamically based on feature modality in LSTM and GRU units. By adjusting the parameters of the fine-tuning layer gating unit, video features can be efficiently fused and identified [38].

Multimodal fusion, differentiated from dynamic connection, and the high complexity of multimodality cause ultra-high latency, which is a key problem. Therefore, this paper proposes a method for recognizing dangerous driving behavior based on millimeter-level frequency-modulated continuous-wave radar. In this paper, the RDM method is used to map the radar data in the velocity dimension to generate micro-Doppler eigenvectors, radial velocity eigenvectors, and amplitude eigenvectors through successive frames. The three vectors are represented by a low-rank tensor fusion representation through local dependency correlation, and a method of combining main features and effective features is adopted, and the high-dimensional data completely show the completeness of dangerous driving actions within the time window while reducing the complexity. The overall model uses the improved attention mechanism LMF-GRU model to improve the reliability of the system and uses the AReLU activation function to extract the fusion representation for target detection in the time series.

3. Driving Behavior Signal Processing

3.1. In-Vehicle Millimeter-Wave Radar Echo Detection

The method of using FMCW radar for human pose estimation involves firstly signal preprocessing and the optimal setting of radar parameters. Secondly, for the micro-motion feature extraction, the echo signal and the IF signal use fast Fourier transform for any one FM period, and the slow time distance Doppler feature of the time slice is obtained by determining the distance FFT of the IF signal in the frame period, and the radial velocity feature is obtained in the extraction of the specific distance and real-time window. The combined amplitude features are obtained by mapping in the horizontal direction, at the same distance and velocity index according to the stitching of the frame time [9].

Linear FM continuous-wave radar makes a fixed difference frequency signal between the echo signal and the transmit signal by modulating the frequency of the transmit signal, and uses the difference frequency signal to obtain the distance information and speed information of the target. The linear FM continuous-wave radar receiver is highly sensitive, and the transmitter power is low in the same noise environment with the same target detection capability. The target echo time delay is much smaller than the time width of the transmit signal, and unlike in the pulse radar, the radar transmitter and receiver are working at the same time, so the FMCW radar is not operating in a blind zone. For close-range detection of driving target behavior, millimeter-level error can be achieved. The basic principle of perception is to determine the distance in that direction from the transmitter
to the receiver using the delay of the transmitting and returning waves, so the following expression represents the initial phase of the returning waves.

\[ P_x(t) = 2\pi \left[ f_0(t - \tau) + \frac{1}{2T} B(t - \tau)^2 \right] + \phi_0 \] (1)

The instantaneous phase of the differential frequency signal is the phase difference between the transmit phase and the return, expressed as the following expression.

\[ P_m(t) = 2\pi f_0 \tau + \frac{2\pi}{T} B t \tau + \frac{\pi}{T} B t^2 \] (2)

\( B \) is the maximum bandwidth, \( T \) is the period of the high-speed FM continuous wave, \( T_c \) represents the pulse width of the linear FM wave, \( A_T \) represents the amplitude of the transmit signal, and \( \tau \) is the difference frequency signal after mixing the time delay of the transmit signal and the return signal. The phase noise of the mixed differential frequency signal in detecting human motion can be neglected, so the differential frequency signal is expressed as the following expression using millimeter-wave matrix sampling.

\[ Y[n, m] = A_R e^{i(2\pi f m T_f + \frac{4\pi}{T^2} R(n T_f + m T_s))} \] (3)

3.2. Feature Vector Extraction

Due to the scattering characteristics of the FM continuous wave, the human behavior facing the radar is reflected by the waveform to generate echoes, the reflected echoes are put into the frequency modulation of the radar and time-frequency analysis, and the kurtosis of the deviated signal distribution in the time domain can be expressed as

\[ F_{kur} = \frac{1}{TT_f} \int T \left| r(t) \right|^4 \text{dt}, \text{The skewness information for the signal symmetry measure in the time domain can be expressed as} \]

\[ F_{skew} = \frac{1}{TT_f} \int T \left[ \left| r(t) \right| - F_n \right]^3 \text{dt}, \text{where} \ t(t) \text{denotes the original data of the received echo, and} \ T \text{denotes the length of the original data time slice.} \]

Performing multidimensional Fourier transform in the frequency domain data can prevent the time domain feature quantities from overlapping in the time slice, and feature extraction is performed in the frequency domain. The Fourier transform of the continuous frequency modulation signal of the radar echo is expressed as

\[ F(\omega) = \mathcal{F}[f(t)] = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} \text{dt} \]

In the unstable and irregular signal, the short-time Fourier transform is performed on the signal via the linear transformation method in the extraction process of the driver’s micro-Doppler feature, and the collected original echo signal is defined as \( x(t) \in L^2(\mathbb{R}) \). The short-time Fourier is expressed as the following expression.

\[ STFT_{x}(t, \Omega) = \int_{-\infty}^{+\infty} x(\tau) g^*(\tau - t) e^{-j\Omega \tau} d\tau = \left< x(\tau), g(\tau - t) e^{-j\Omega \tau} \right> \] (4)

The resolution in the time domain during processing is determined by the window function \( w(t) \), and the frequency resolution in the frequency domain is determined by the window function \( W(\omega) \). The enantiomorph relationship is that the longer the effective length of the window function \( w(t) \) is, the more concentrated the integration efficiency over the time series will be, leading to an increase in Doppler resolution. In this paper, micro-Doppler extraction is performed using pre-processed FM continuous waves by adding time function windows (STFT) for dangerous driving behavior, and micro-Doppler features are obtained by Fourier transform. The micro-Doppler features are used as the main features to identify the dangerous actions, and the micro-Doppler sparse matrix is obtained by mapping the velocity in a time slice sequence. The sensitivity accuracy of micro-Doppler for actions is higher than that of video frames and skeleton key points. The Doppler spectrum dataset at different distances responds to the tiny movements of different echo body points of the target, and the radial velocity features and amplitude information features of the driver at different distances continue to be extracted by different
raw data mappings. Figure 2 represents the mapped micro-Doppler spectra of key features, combined into feature datasets at different distances.

\[
\mathcal{S}(t, \omega) = \mathcal{F} = \int x(\tau) g^*(\tau-t) e^{-j2\pi \omega \tau} d\tau = \mathcal{F}(x(\tau), g(\tau-t) e^{-j2\pi \omega \tau})
\]

(4)

The resolution in the time domain during processing is determined by the window function \( w(t) \), and the frequency resolution in the frequency domain is determined by the window function \( W(\omega) \). The enantiomorphic relationship is that the longer the effective length of the window function \( w(t) \) is, the more concentrated the integration efficiency over the time series will be, leading to an increase in Doppler resolution. In this paper, micro-Doppler extraction is performed using pre-processed FM continuous waves by adding time function windows (STFT) for dangerous driving behavior, and micro-Doppler features are obtained by Fourier transform. The micro-Doppler features are used as the main features to identify the dangerous actions, and the micro-Doppler sparse matrix is obtained by mapping the velocity in a time slice sequence. The sensitivity accuracy of micro-Doppler for actions is higher than that of video frames and skeleton key points. The Doppler spectrum dataset at different distances responds to the tiny movements of different echo body points of the target, and the radial velocity features and amplitude information features of the driver at different distances continue to be extracted by different raw data mappings. Figure 2 represents the mapped micro-Doppler spectra of key features, combined into feature datasets at different distances.

Figure 2. Distance target of 0.30 m to 1.10 m for the micro-Doppler characteristic spectrum.

In the effective range of radar detection, the motion of the target is described by the direction and velocity at the same distance from the radar. Two-dimensional FFT is carried out on the pre-processed data of the radar to extract the radial velocity of the radar, and the distance of the target point from the radar is determined by the CFAR method according to the fluctuation degree of the signal to decide the set threshold, and the radial velocity of the target point is measured according to the Doppler series information points to find the peak point. A single frame of the time series contains a 128 × 128 sparse matrix, and the original data size will be extracted first without compression. In the original data, the key frames are captured, and no obvious echo state frames or time redundant frames are lost, and the amplitude features are extracted from the original data, which are divided into real and imaginary parts, and the amplitude features will reflect the dense area of the action and have a predictive role for the start and end time series of the action.

3.3. Doppler Vector Fusion Representation

The key problem of multimodal fusion is the dynamic articulation between the features and the normalization problem, which is used to fuse the modalities for the classification normalization problem of the lower layer network, where the method is used to represent the tensor representation, but because of the large number of data collected in the paper to calculate the original multi-peak data of RDM mapping Doppler features, the individual parameters are redundant, and the relevant parameters are non-correlated with multiple data form inputs, so the fusion method of low-rank multimodal can be solved, and the tensor representation is expressed using the outer product between each modality and the interconnection between the modalities, and the data peak \( X \) of a single time slice is expressed by the following expression.

\[
X = \otimes_{m} m_{u}, m_{u} \in \mathbb{R}^{d_u}
\]

(5)
where the outer product on the set $\otimes_{u=1}^{U} \mathbb{R}$ is represented by the index vector $u$, and $m_u$ denotes the number plus one. The input tensor is represented as a linear layer vector $g(\cdot)$ through $X \in \mathbb{R}^{d_1 \times d_2 \cdots \times d_r}$. The tensor is represented as the following expression.

$$h = g(X; W, b) = W \cdot X + b, \ b \in \mathbb{R}^{d_y} \quad (6)$$

In the expression, $W$ represents the weight of the layer and $b$ represents the bias. When the number of modes is determined, $X$ is represented as a tensor, and the weight of the $W$ order is a $U + 1$-order tensor of $\mathbb{R}^{d_1 \times d_2 \cdots \times d_r}$. In this method, high-efficiency low-rank multimodal fusion of parallel decomposition is used in vector fusion, and the original collected Doppler information is converted from tensor $X$ to $\{M_u\}_{u=1}^{U}$. This process is decomposed in parallel with the decomposition process of low-rank tensor factors. Tensor representations are expressed in parallel networks through simplification and factorization.

$$h = \left( \sum_{i=1}^{r} \left( \sum_{u=1}^{U} w_u^{(i)} \right) \right) \cdot X$$

$$= \sum_{i=1}^{r} \left( \sum_{u=1}^{U} w_u^{(i)} \right) \cdot X$$

$$= \wedge_{u=1}^{U} \left( \sum_{i=1}^{r} \left( \sum_{u=1}^{U} w_u^{(i)} \cdot m_u \right) \right) \quad (7)$$

Equation (7) analyzes the complexity of the output tensor and the linear symmetric projection. Using this model prevents the disadvantage of inefficient system, and the complexity of the traditional method is $O(d_y \prod_{m=1}^{M} d_m)$. This model reduces the complexity of the system to $O(d_y \times r \times \sum_{m=1}^{M} d_m)$, where $\wedge_{u=1}^{U}$ represents the product of the elements on the tensor sequence $\wedge_{u=1}^{U} = x_1 \cdot x_2 \cdot x_3$. This method represents the tensor fusion of three identical time slice single-peak modes of micro-Doppler features, radial velocity features, and amplitude features. The tensor fusion method for the two modes is represented by the following expressions.

$$h = \left( \sum_{i=1}^{r} \left( \sum_{u=1}^{U} w_u^{(i)} \otimes v_u^{(i)} \right) \right) \cdot X$$

$$= \left( \sum_{i=1}^{r} \left( \sum_{u=1}^{U} w_u^{(i)} \otimes m_u \right) \right) \cdot \left( \sum_{i=1}^{r} \left( \sum_{u=1}^{U} v_u^{(i)} \otimes m_v \right) \right) \quad (8)$$

3.4. Recurrent Neural Network for Classification

GRU is a modified version of the recurrent neural network. GRU is a full-gated recurrent unit, and the recurrent neural network is represented using $a^{(t)}$, where $a$ is the memory unit, $g(\cdot)$ is the activation function, $x$ is the input, $b$ is the bias, and $t$ is the time point. In GRU, the memory unit is $c$, but $\tilde{c}$ is used instead of $c$.

$$a^{(t)} = g \left( W_a \left[ a^{(t-1)}, x^{(t)} \right] + b_a \right) \quad (9)$$

$$\tilde{c}^{(t)} = \tanh \left( W_c \left[ c^{(t-1)}, x^{(t)} \right] + b_c \right) \quad (10)$$

A gate that plays an important role in iterative feature memory in the gating unit of GRU is the update gate $\Gamma_u$. The odd number is 1 and the even number is 0, so the value of $\Gamma_u$ should be placed between 0 and 1. The applicable activation function is ReLU, which is in the ideal state when $\Gamma_u$ is equal to 0, making the t moment and $t - 1$ moment of the memory unit very similar. The update gate is expressed using the following expression.

$$\Gamma_u = \sigma \left( W_u \left[ c^{(t-1)}, x^{(t)} \right] + b_u \right) \quad (11)$$
The input micro-Doppler features and the other two features are replaced by other parameters in order to approach 1 without affecting the cat, and to solve this problem, the next gate is generated by the following equation.

$$C^{(t)} = \Gamma_u \ast \tilde{C}^{(t)} + (1 - \Gamma_u) \ast C^{(t-1)}$$  \hspace{1cm} (12)

4. Methods and Models

4.1. Overall Methodology Process

As shown in Figure 3, dangerous driving is determined via the radial distance change of the action and signal generation peak, the input original echo with vector $C$, and the input original echo using the $H$, $W$, and $T$ three-dimensional matrix representation $X \in \mathbb{R}^{C \times H \times W \times T}$. In the process of behavior signal acquisition, three kinds of action recognition features are used in information fusion to achieve the classification effect, three kinds of action features in the RDM radial mapping and Fourier transform are used to generate a new feature representation, and the micro-Doppler spectrum key features $K \in \mathbb{R}^{C \times H \times W \times T}$, distance time Doppler velocity spectrum hidden features, amplitude features of identification hidden features, and hidden features are uniformly expressed as $Q \in \mathbb{R}^{C \times H \times W \times T}$. This method firstly uses FMCW millimeter-wave radar to collect the original information on the dangerous people in the car, firstly by using an MTI filter for static clutter filtering, and then Fourier transform for data to accumulate velocity information for different distances. After that, Fourier transform is applied to the data so the velocity information of different distances can be accumulated in the distance–time spectrum, and the window functions of different time slices are STFT-transformed and modulated to obtain the key micro-Doppler features. The three different dimensional features are normalized to find the dynamic articulation between the modes expressed as a high-dimensional unified tensor, and the fused feature tensor is expressed as $\mathbb{R}^{d_1 \times d_2 \times \cdots \times d_u}$. The low-rank multimodal fusion method for different dimensional features is reduced and normalized so as to reduce the model time complexity, which constitutes the temporal distance feature also known as the velocity feature through temporal processing, and each position of the target driving process body is relative to the radar. The distance of each position of the target’s body relative to the transmitting end of the radar will change, due to the distance Doppler generates for the Doppler spectrum peaks. According to the location of the peaks and the temporal sequence, the relative actions can be differentially classified, and the fused multi-level feature data $\mathbb{R}^{d_1 \times d_2 \times \cdots \times d_u}$ are input to the AR-GRU through the update gate and reset gate to update the state of the network input parameters, prompting the GRU to learn the weight parameters.

Figure 3. Behavior recognition model with low-rank fusion weight analysis of multicascade Doppler features.
Since this method focuses on studying and explaining the methodical recognition of dangerous driving behavior, the method of testing is not described in too much detail. In the recognition of dangerous driving behavior under vehicle radar monitoring, the most important thing is the lightness of the system, as the system model must quickly detect dangerous driving actions that are happening or will happen soon, and so the system must have a low time complexity. Regarding the running time of the LMF low-rank multimodal model, it is important to talk about the experimentally collected self-built dataset, which is used in this method to evaluate the task. The “Verify” data subset and the “test-clean” dataset are used for the test data. The collected Doppler features are vector-decomposed and divided into multiple frames with a window size of 20 ms and a step size of 10 ms, and the corresponding features are extracted using a feature extractor, and a multidimensional filter is used to normalize the multimodal features to a mean of zero and perform mean variance analysis on the results. For the identification of all dangerous driving behaviors via the recognition of specific actions in experimental verification, the average time needed for the recognition of specific dangerous driving behaviors is 489 ms, which can fully meet the need for a timely warning.

4.2. AM-LMF and AR-GRU

4.2.1. Tensor Representations

The tensor is usually created by taking the outer product over the input modalities. Multimodal fusion is represented as the multilinear function \( f : V_1 \times V_2 \times \ldots \times V_M \rightarrow H \), where \( V_1, V_2, \ldots, V_M \) is the vector space of input modalities, and the output vector space determines a set of vector representations \( \{z_m\}_{m=1}^M \) that encode the single-peak information of different modalities, and the goal is to integrate the single-modal representations into a compact multimodal representation for downstream tasks. Tensor representation is a method of multimodal fusion, where the input representation is first transformed into a high-dimensional tensor and then mapped back to a low-dimensional output vector space. It is more effective at capturing multimodal interactions than simple concatenation or pooling. In addition, to be able to model interactions between any subset of modalities using a single tensor, this method proposes a single-peaked expanded representation that appends interaction information to single-peaked data. The input tensor \( Z \) formed by the single-peaked representation is calculated by the following equation.

\[
Z = \bigotimes_{m=1}^M z_m, z_m \in \mathbb{R}^{d_m}
\] (13)

The three unimodal two-by-two combinations are combined to form three multimodal features, and all three modalities are also stitched together to form a high-dimensional multimodal feature, and the modal combinations are used as inputs to the relational network to determine the relationship between the modalities. For the unimodal features, the relational network first performs nonlinear transformation operations on different dimensions, and the bimodal features that are combined two by two will represent the fusion results as expressions.

\[
T_b = F_4 (f_{key} \oplus f_{vel}) + F_5 (f_{key} \oplus f_{amp}) + F_6 (f_{vel} \oplus f_{amp})
\] (14)

For the three modes of dangerous driving action recognition, all features are fused into high-dimensional modes and normalized using the connection relations of the relational network, and the process and results can be expressed as the following equation.

\[
T_c = F_7 (f_{key} \oplus f_{vel} \oplus f_{amp})
\] (15)

where \( F_1, F_2, \ldots, F_7 \) denotes a series of different linear and nonlinear transformations; linear transformation is a fully connected operation, and nonlinear transformation uses the ReLu activation function; \( f_{key}, f_{vel}, \) and \( f_{amp} \) denote the micro-Doppler feature, the radial velocity
feature, and the amplitude feature, respectively; and ⊕ represents the splicing operation. Figure 4 represents the whole process of linear transformation.

\[ W_{\text{overall}} = \bigoplus W_{\text{linear} \cdot \text{unit}} \]

\[ \overline{W}_m = \sum_{i=1}^R \bigoplus_{n=1}^N w_{n,m}^{(i)}, w_{n,m}^{(i)} \in \mathbb{R}^d \]

\[ w_n^{(i)} = \left[ w_{n,1}^{(i)}, w_{n,2}^{(i)}, \ldots, w_{n,h}^{(i)} \right] \]

The unimodal represents \( Z_a, Z_b, Z_c \), which is used as the input of the multimodal fusion to generate a new unimodal, which is multimodal-normalized by passing the unimodal input to the three sub-networks \( f_{\text{context}}^{\text{context}}, f_{\text{doppler}}^{\text{doppler}}, \) and \( f_{\text{imp}}^{\text{imp}} \), respectively. Figure 5 represents the process of unimodal to multimodal normalization.

Figure 4. Multimodal attention mechanism for low-rank factor linear weight representation.

4.2.2. Low-Rank Tensor Fusion Method

In this paper, a novel low-rank multi-peak fusion model based on a self-attentive mechanism is proposed, which uses a low-rank weight tensor with an attention mechanism to make multi-peak fusion more effective and more globally relevant. The low-rank tensor representation solves the problem arising from feature fusion representation by introducing a weight factor \( W \) to decompose the low-rank factor \( W \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N} \) to represent the tensor of order \( d \). The principle is to perform weight decomposition into \( N \)-specific low-rank factors by weighting the weight factor \( W \). The exact decomposition for the tensor is expressed as the following expression.

\[ \overline{W}_m = \sum_{i=1}^R \bigoplus_{n=1}^N w_{n,m}^{(i)}, w_{n,m}^{(i)} \in \mathbb{R}^d \]
The tensor is highly expressive but does not scale well to a large number of modalities, so dependencies between modalities to improve the correlation and local dependence between the various modalities. The attention mechanism equipped with attentional neural networks with parallelizable computations, lightweight structures, and the ability to capture remote and local dependencies. The key to the attention mechanism approach is the measurement of the correlation between $Z_w$ and $Z_q$. The compatible function $g(z_n, q)$ generates a score, $k$, which reflects the dependencies between $Z_w$ and $q$. The score is transformed into a probability by the function softmax, and finally, this probability is used as a weight and applied to the input modalities to improve the correlation and local dependence between the various modalities. The attention mechanism uses the dot product attention compatibility function, expressed as the following expression, where $w^{d1}$ and $w^{d2}$ are the learnable parameters.

$$g(z_n, q) = \langle w^{d1} z_n, w^{d2} q \rangle$$

Figure 5. Multimodal low-rank factor normalized representation process.

4.2.3. Compatibility Study of Attentional Mechanisms

The method proposed in this paper is an improvement of the low-rank multimodal fusion approach and is an effective integration of the input modality approach based on the low-rank multimodal fusion approach. The method proposes a novel self-attentive mechanism equipped with attentional neural networks with parallelizable computations, lightweight structures, and the ability to capture remote and local dependencies. The key to the attention mechanism approach is the measurement of the correlation between $Z_w$ and $q$. The compatible function $g(z_n, q)$ generates a score, $k$, which reflects the dependencies between $Z_w$ and $q$. The score is transformed into a probability by the function softmax, and finally, this probability is used as a weight and applied to the input modalities to improve the correlation and local dependence between the various modalities. The attention mechanism uses the dot product attention compatibility function, expressed as the following expression, where $w^{d1}$ and $w^{d2}$ are the learnable parameters.

$$g(z_n, q) = \langle w^{d1} z_n, w^{d2} q \rangle$$

Compared with the model of the traditional attention mechanism, the present model does not introduce redundant parameters and the model has lower complexity and a faster running speed due to its ability to improve accuracy while maintaining low complexity. The tensor is highly expressive but does not scale well to a large number of modalities, so the weights are decomposed into low-rank factors, which reduces the number of parameter bases in the modalities. The low-rank decomposition computes tensor-based fusion by using a parallel decomposition of the low-rank weight tensor and the input tensor. The method in this paper can scale linearly with the number of modes.

4.2.4. AReLU-GRU

The common activation function of the GRU gating unit based on the attention mechanism is the hyperbolic tangent ($\text{tanh}$), although the effective activation function has a high degree of integration of the computational effect, and so the algorithm complexity is not suitable for use in the in-vehicle computing model. This model uses rectified linear units based on attention AReLU is the activation function and AReLU as the use of time-indexed element-by-element detection. Negative factors will be adaptive parameters for the target data to amplify by changing the weight factor method. Using this activation function can prevent the disappearance of Doppler data gradients in tensor processing, and there are
also small learning samples with very few parameters and two parameters for fast network training, AReLU in multimodal fusion also helps to capture the interaction information and dependencies between the information, where the attention mechanism can be used to help GRU to focus on the most significant information for the tensor representation, and the use of the ReLU activation function in the GRU gating unit significantly improves the system performance because of the excellent performance that can be demonstrated for this data volume. Therefore, the model optimizes the alpha and beta values of AReLU by starting with alpha = 0.03 and beta = -3. The default values are suitable for tensor fusion and model efficiency, and the parameters are set so that AReLU is similar to the ReLU-like activation function. Since alpha is the weighting factor controlling negative values, a value of alpha = 0.03 with simultaneously negative values of the threshold function has minimal impact on the model. When the validation model of the activation function AReLU in GRU is feasible, the model turns the parameter beta = -3 into beta = 2, and by iteratively conducting experimental validation and using the formula derivation alpha = 0.57, the generalizability of the model in application to behavior recognition is not limited.

5. Experimental Design and Analysis

5.1. Experimental Setup and Environment Variables

The experimental equipment used in this paper included a millimeter-wave radar clock with an amplitude of 77–81 GHz, produced by Texas Instruments, corresponding to a wavelength of about 4 mm and a bandwidth of 4 GHz. The hardware had an ultra-high-frequency bandwidth that is superior to that of other wireless sensing devices, ensuring high-precision recognition of movements. The hardware experimental environment equipment used mainly included an IWR1642 radar sensor device and DCA1000EVM data capture card, sensor bracket, Lenovo R7000 i5-7300HQ GTX1050 802.11AC wireless network card laptop, and two cars and SUVs. In the experimental scenario, on sunny or rainy days, after experimental comparison of the same parameters of the state of the weather and environmental factors, if the millimeter-wave echoes are minimal, they do not merit discussion. An experiment was conducted for vehicles of different sizes, including a three-compartment sedan and off-road cars. The radar sensor was fixed to the B pillar of the vehicle, and at a height of 85 cm from the vehicle chassis, a radar transmitter and receiver were placed perpendicular to each other facing the target driver, and the driver location was just within the radar scanning range so that the target dangerous behavior characteristics could be obtained. The experimental setup is shown in Figure 6.

![Figure 6. Multimodal low-rank factor normalized representation process.](image)

The experimental equipment and environment used for collecting information included a Windows 10 Intel Core i7-6700 CPU processor, with a main frequency of 3.4 GHz, a memory size of 32 GB, a GTX1080 graphics card, and PyCharm2018 software; the programming environment was Python 3.7. In this model, in the compiled environment Python 3.7, MemTotal indicated the actual amount of RAM occupied by the system, us-
ing physical memory minus some reserved bits and the size of the kernel’s binary code. `MemTotal` produced 816,136 kb in the environment command in the proc file using `cat /proc/partitions`. The partitions of the system itself were identified in the proc file as `cat /proc/partitions`. In the 10 partitions, from 0 to 9, the partition capacity was summed to obtain the ROM capacity of 1,563,645 kb. The data acquisition equipment used included a Logitech-wired Wi-Fi surveillance camera with 720P/1080P high-definition wide-angle images and was easy to install. After comparing and analyzing the data collected from multiple placement angles and positions, the camera installed in the front of the co-pilot position was found to have the best effect, and the sitting posture image data for the front-seat driver and occupant in the car were obtained.

Before setting the parameters, the speed and distance of the moving target in the car were first estimated, because an estimated maximum speed or distance that is too large will lead to a significant decrease in speed resolution. A suitable sampling time setting can ensure the completeness of an action sample and will not produce redundancy. The parameters of the millimeter-wave radar were set, as shown in Table 1. In the sampling process, to ensure the integrity of the behavior, the frame period was set to 50 ms, and the number of frames was set to 100 to achieve a sampling time of 5 s for each action, which ensured the integrity of the action and reduced the blank redundant time slices, so each unit was set to 256 samples and 128 Chirp.

<table>
<thead>
<tr>
<th>Radar Parameter</th>
<th>Units</th>
<th>Numerical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective bandwidth</td>
<td>MHz</td>
<td>3411.25</td>
</tr>
<tr>
<td>Maximum distance error</td>
<td>cm</td>
<td>4</td>
</tr>
<tr>
<td>Maximum measurement</td>
<td>m/s</td>
<td>6</td>
</tr>
<tr>
<td>Speed resolution</td>
<td>m/s</td>
<td>0.095</td>
</tr>
<tr>
<td>Frequency modulation slope</td>
<td>MHz/µs</td>
<td>66.626</td>
</tr>
<tr>
<td>Frame period</td>
<td>ms</td>
<td>50</td>
</tr>
<tr>
<td>Number of samples</td>
<td>1</td>
<td>256</td>
</tr>
<tr>
<td>Chirp</td>
<td>1</td>
<td>128</td>
</tr>
</tbody>
</table>

The data in the table are the optimal parameters for the experiment.

In terms of the basic conditions, the experiment was conducted with 6 known experimental subjects, 3 males and 3 females with different heights and weights, with an age range of 23 to 55 years, a height range of 152 cm to 190 cm, and a weight range of 43 kg to 97 kg. Six random subjects with unknown information were used as comparisons. The sample set for this experiment consisted of 9 risk behavior datasets, with a sample acquisition time of 5 s and one specific action that was repeated 60 times. The dataset contained 6480 samples, and each data sample acquisition set was $255 \times 255 \times 3$, the Doppler distance was set to 0.3–1.1 m, the unit was 0.1 m, and target sensing was performed at 9 distances.

Due to the special nature of in-vehicle recognition, radar recognition of dangerous movements can only be realized for the upper body due to obscurity, and leg movements could not be identified due to obscurity.

5.2. Experimental Analysis
5.2.1. Model Performance Analysis for Visual Verification

While capturing the driver’s wireless FM signal, the dual-channel processing of RGB image information followed by determining human skeletal key points is essential for describing human posture. Therefore, human skeletal key point detection is the basis for many computer vision tasks, such as motion classification, abnormal behavior detection, and autonomous driving, among others. This experiment validated the use of surveillance equipment to capture images of front-row drivers and passengers, and used the AlphaPose model to obtain the coordinates of 13 skeletal key points in the upper body: left ear, left eye, nose, right ear, right eye, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, and right hip. These skeletal key points and the line segments
In terms of the basic conditions, the experiment was conducted with 6 known experimenters. In the off-road vehicle experiment, due to the presence of people’s shadows and poor lighting environment, the average overall recognition rate for the nine behaviors was 73.14%; however, the millimeter-wave radar had the highest recognition accuracy compared to other techniques used to recognize dangerous behavior within vehicles. Due to the larger internal space of the off-road vehicle, the noise rate of the recognition effect was slightly reduced, and the recognition accuracy inside the off-road vehicle according to experimental comparison was 0.641% higher than the accuracy obtained for the same individuals and the same parameters in the small car. Figure 8 presents a comparison of the experimental environment and equipment used in the car vs. the off-road vehicle. The off-road vehicle was obviously larger than the car, and there was a difference in the recognition effect.

To extract the information of the key points of the skeleton, the change in distance between the key points is used to determine whether any of the nine dangerous actions can be observed, but the recognition rate is low, easily affected by other people’s shadows and poor lighting environment, and the average overall recognition rate for the nine behaviors was 73.14%; however, the millimeter-wave radar had the highest recognition rate at 95.74%. Therefore, the model has better recognition accuracy than other techniques used to recognize dangerous behavior within vehicles. Due to the larger internal space of the off-road vehicle, the noise rate of the recognition effect was slightly reduced, and the recognition accuracy inside the off-road vehicle according to experimental comparison was 0.641% higher than the accuracy obtained for the same individuals and the same parameters in the small car. Figure 8 presents a comparison of the experimental environment and equipment used in the car vs. the off-road vehicle. The off-road vehicle was obviously larger than the car, and there was a difference in the recognition effect.

Figure 7. (a) Right turn (H) key point marker movement prediction; (b) answering the phone (E) key point marker action prediction.
5.2.2. Analysis of Cross-Person Results

The experimental setup included six people with different physical characteristics differing widely in height from 152 cm to 190 cm, including three males and three females. In FMCW-radar-based cross-person movement recognition, drivers with different physiques and heights may produce large differences in recognition accuracy because the radar is sensitive to distance and has different responses to driving targets regarding the three features. The six participants were given different letter labels, a, b, c, d, e, and f, to indicate different people. For each experiment, the dispersion of each set of data was demonstrated by visualizing the data nodes of the six targets to improve the generality of the experiment and determine a set of data for the six unknown participants for comparison, and a box plot graph was created to show each target median and magnitude. From the box plot diagram, it can be seen that good behavior recognition was achieved for every individual. Figure 9 presents the box plot diagram.

![Box plot diagrams for six known experimental subjects; (b) box plots for the six unknown experimental subjects.](image)

5.2.3. Analysis of Results at Different Distances

In behavior detection, the physiological characteristics of the tested participants will directly affect the recognition accuracy, including the weight and height of the participants, which will affect the distance of the target to the radar transmitter. The experiment aimed to reflect the millimeter-wave radar FM continuous-wave signal for different distances and determine the speed and the distance to the target. This method used nine time detection nodes, the unit of distance was 10 cm, and the resolved data for the distances of 0.3 m, 0.4 m, 0.5 m, 0.6 m, 0.7 m, 0.8 m, 0.9 m, 1.0 m, and 1.1 m were experimentally. The results obtained after collecting the Doppler spectrum of different distances show that the Doppler spectrum of the distance of 0.3–0.5 m had no obvious echo and radial velocity, indicating that there was no reflected signal of the target at this distance, and thus the actual action state of the driver cannot be reflected, so the data stream and tensor of this distance were eliminated through data screening. However, at the distance of 0.5–1.1 m from the radar transmitter, obvious behavioral Doppler feature data could be detected, and the recognition rate for the distance of 0.3–0.5 m did not exceed 20% as reflected by the phenomenon shown in Figure 10a, and the Doppler spectrum of this distance cannot be used as evaluation data; however, the recognition rate for driver behavior at the distance of 0.5–1.1 m reached an average of 92.1%. The CDF plot in Figure 10b reflects that under the same experimental conditions, three groups of targets with different physiological characteristics were set: target a, target b, and target c. Target a was taller and heavier than
the other two targets, so it can be concluded that target a achieved a higher recognition accuracy at a smaller distance. However, target c was smaller than the other two targets and also lighter, so the best recognition accuracy was achieved at the sensing distance of 0.8–1.1 m. Therefore, cross-person hazardous driving perception must consider the uncertainty of target body characteristics.

5.2.4. Analysis of Classification Models under Different RNNs

Both GRU and LSTM are sub-branches of RNN. However, the fatal disadvantage of RNN is that the gradient disappears, and data explosion occurs in the process of processing data flow, so the introduction of a gating mechanism can effectively solve this problem. GRU controls the gradient of the input information of the upper layer and calculates the information flow of the hidden layer by updating the gate. In terms of parameters, the activation function that causes the gradient to disappear is replaced by a rectified linear (ReLU) function, which is initialized by an unbounded activation function and identities weights. LSTM is composed of two gate parameters (excluding gate and input gate), and the parameter control is completely different from that of GRU. In information processing, neither the information from the previous moment nor the information from the current hidden layer is selected, while GRU can choose. Additionally, the most important thing in the perception of human body posture in the car is a timely early warning to reduce the time complexity. Although GRU has one less gate than LSTM, GRU has faster convergence and fewer parameters. The accuracy of the same action of GRU reflected from the confusion matrix is slightly higher than that of LSTM parameters. The highest accuracy rate was 96%. The lowest accuracy rate was 88%, and the comparison of the confusion matrix shows that there were similar correlations between different actions, including similar misjudgment outputs, with the highest being 19%, and the output classification of the model may overlap. Additionally, multi-parameter LSTM resulted in higher-similarity output than GRU for the classification of adjacent actions. From the following matrix, it can be seen that the advantages of GRU are more obvious, and it is more suitable for the processing of radar information collected in the car. The confusion matrix is shown in Figure 11. The confusion matrix indicates the effect of the classifier on the test dataset, and the accuracy of the recognition effect for different behaviors is indicated using a gradient from white to blue. High accuracy for the same behavior is indicated using blue.
was 84.6% and the highest recognition rate was 89.8% for double dominance modal fusion, and the lowest recognition rate was 84.9% for unimodality, the lowest recognition rate was 84.6% and the highest recognition rate was 89.8% for double dominance modal fusion, and the lowest recognition rate was 76.2% and the highest recognition rate was 95.7% for multimodal multi-peak representation. The three-fold line data statistics indicate that the single modality is compared and analyzed overall with the two-by-two combined modalities and the fused multimodal, and it can be derived from the three line graphs in the figure that the single modality recognition was the worst, followed by the double dominant modal fusion, and finally the multimodal multi-peak representation. For the unimodal features, the relational network first transforms them nonlinearly before stitching them together. The model puts the single modality into the GRU model alone for the attention weight decomposition method for recognition, and the single modality is compared and analyzed overall with the two-by-two combined modalities and the fused multimodal, and it can be derived from the three line graphs in the figure that the single modality recognition was the worst, followed by the double dominant modal fusion, and finally the multimodal multi-peak representation. The lowest recognition rate was 76.2% and the highest recognition rate was 84.9% for unimodality, the lowest recognition rate was 84.6% and the highest recognition rate was 89.8% for double dominance modal fusion, and the lowest recognition rate was 91.7% and the highest recognition rate was 95.7% for multimodal multi-peak representation. The three-fold line data statistics indicate that there were no overlapping points, so the accuracy for the same action was in line with the overall modal accuracy, so the line graph indicates the quantitative relationship between the modal fusion method and the number, and the influence of the dependency between the modalities on the accuracy rate is shown in Figure 12.

Figure 11. (a) Accuracy under GRU parameters; (b) accuracy under LSTM parameters.

5.2.5. Analysis of Combined Modalities

A single mode cannot respond to the integrity movement of physiological action in the human body recognition process, so the feature input of the model in this paper was a three-channel input, with micro-Doppler spectral features as the main recognition input features and radial velocity and amplitude features as hidden features, and three single modes were combined two by two to form three multimodal peaks, and all three modes were also spliced together to form a high-dimensional multimodal feature, with the modal combination as the input to the relational network to determine the relationship between the modes. For the unimodal features, the relational network first transforms them nonlinearly before stitching them together. The model puts the single modality into the GRU model alone for the attention weight decomposition method for recognition, and the single modality is compared and analyzed overall with the two-by-two combined modalities and the fused multimodal, and it can be derived from the three line graphs in the figure that the single modality recognition was the worst, followed by the double dominant modal fusion, and finally the multimodal multi-peak representation. The lowest recognition rate was 76.2% and the highest recognition rate was 84.9% for unimodality, the lowest recognition rate was 84.6% and the highest recognition rate was 89.8% for double dominance modal fusion, and the lowest recognition rate was 91.7% and the highest recognition rate was 95.7% for multimodal multi-peak representation. The three-fold line data statistics indicate that there were no overlapping points, so the accuracy for the same action was in line with the overall modal accuracy, so the line graph indicates the quantitative relationship between the modal fusion method and the number, and the influence of the dependency between the modalities on the accuracy rate is shown in Figure 12.
AReLU method to its optimal state. The model uses this set parameter value, and the accuracy of the LMF-AR-GRU model is higher than that of the BiGRU-IAAN and BiLSTM models both in terms of UA and WA. The method of integrating GRU units with a multimodal fusion of Attention ReLU activation proposed in this paper can improve the interaction between features, capture driver actions in remote interaction recognition, and solve the problem of BiLSTM gradient disappearance and information explosion in a series of experimental validation work to optimize the AReLU value so that the AReLU is similar to the ReLU. The control unit components of AR-GRU are AReLU ($\alpha = 0.01, \beta = -4$), the specific parameter value that returns the AReLU method to its optimal state. The model uses this set parameter value, and the data representation can reach up to 84.6% unweighted accuracy (UA) and 89.7% weighted accuracy (WA). This variant of the model parameters improves UA and WA by only 0.6% and 0.7%, respectively, relative to the IAAN baseline. The results of this experiment suggest that the AReLU unit does help to improve the overall accuracy of the model, but requires empirical determination of the ideal parameter values. This ensures that negative values close to 0 do not adversely affect performance. The effectiveness of AR-GRU in a dangerous driving behavior data classification setting is demonstrated by the validation of the experimental accuracy rates. Figure 13 represents the visual representation of the accuracy of different technical methods, both underweighted and unweighted, and it can be seen that the proposed method in this paper demonstrates good recognition results.
In this paper, we explore the method of multi-featured low-rank multimodal fusion based on the use of FMCW radar, which decomposes Doppler information into different dimensions of information data by DAM, and the method generates multidimensional Doppler spectral information via Fourier change and then mapping to different vectors. Due to a large amount of information and the poor fit of the Doppler matrix, the information is normalized using specific low-rank multimodal (LMF) processing. The low-rank multimodality is linearly proportional to the number of modes and achieves competitive results in different multimodal tasks. The data of the tensor are fed into the proposed AR-GRU approach, and then the nonlinear fusion of the weight decomposition is performed by the attention unit. Based on multiple low-rank tensors to represent the dense weight matrix within the RNN layer, the present model is evaluated based on the millimeter-wave action self-built dataset, which significantly outperformed the conventional model in terms of dynamic articulation of low-rank fusion. The three Doppler features exhibited excellent weight cascades, the fusion tensor completely interpreted the complete limb echoes for the 5 s acquisition time from the beginning to the end of the action, and the redundant parameters could be reduced in the driving environment to reduce temporal complexity and achieve a timely warning and correction of dangerous driving behaviors. After this study, we will consider various variables to enrich the model input, and try out some new network architectures by manually adjusting the parameters and selecting the network architecture to improve the finer granularity of the recognition effect and reduce the overfitting effect of the recognition network.

**Author Contributions:** Conceptualization, Z.H. and Z.L.; methodology, Z.L.; software, Z.L.; validation, Z.L., X.D., Z.M. and G.L.; formal analysis, Z.L.; investigation, Z.L.; resources, Z.L.; data curation, Z.L.; writing—original draft preparation, Z.L.; writing—review and editing, Z.L.; visualization, Z.L.; supervision, Z.H.; project administration, Z.H.; funding acquisition, Z.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (Grant 62262061, Grant 62162056), Key Science and Technology Support Program of Gansu Province (Grant 20YF8GA048), 2019 Chinese Academy of Sciences “Light of the West” Talent Program, Science and Technology Innovation Project of Gansu Province (Grant CX2J037, 17CX2J039), 2019 Lanzhou City Science and Technology Plan Project (2019-4-44), 2020 Lanzhou City Talent Innovation and Entrepreneurship Project (2020-RC-116), and Gansu Provincial Department of Education: Industry Support Program Project (2022CYZC-12).

![Figure 13. (a) Model weighted accuracy (WA) histogram; (b) model unweighted accuracy (UA) histogram.](image-url)