IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 9, 2020, accepted July 6, 2020, date of publication July 17, 2020, date of current version August 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010063

Enhanced Multi-Channel Feature Synthesis for Hand Gesture Recognition Based on CNN With a Channel and Spatial Attention Mechanism

CHUAN DU[1], LEI ZHANG[1], XIPING SUN[1], JUNXU WANG[1], AND JIALIAN SHENG[2]
[1]School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou 510275, China
[2]Shanghai Radio Equipment Research Institute, Shanghai 200090, China

Corresponding author: Lei Zhang (zhanglei57@mail.sysu.edu.cn)

This work was supported in part by the National Natural Sciences Foundation of China under Grant 61771372 and Grant 61771367, and in part by the Natural Science Foundation of Shanghai under Grant 17ZR1428700.

ABSTRACT Millimeter-wave (MMW) radar hand gesture recognition technology is becoming important in many electronic device control applications. Currently, most existing approaches utilize the radical and micro-Doppler features from single-channel MMW radar, which ignores the different importance of the information contained in the micro-Doppler feature background or target areas. In this paper, we propose an algorithm for hand gesture recognition jointly using multi-channel signatures. The algorithm blends the information of both micro-Doppler features and instantaneous angles (azimuth and elevation) to accomplish hand gesture recognition performed with the convolutional neural network (CNN). To have a better features fusion and make CNN focus on the most important target signal regions and suppress the unnecessary noise areas, we apply the channel and spatial attention-based feature refinement modules. We also employ gesture movement mechanism-based data augmentation for more effective training to alleviate potential overfitting. Extensive experiments demonstrate the effectiveness and superiorities of the proposed algorithm. This method achieves a correct classification rate of 96.61%, approximately 5% higher than that of the single-channel-based recognition strategy as measured based on MMW radar datasets.

INDEX TERMS Hand gesture recognition, multi-channel signatures, channel and spatial attention mechanism, convolutional neural network, data augmentation.

I. INTRODUCTION

It is essential for human hand gesture recognition to be extensively applied for several important tasks in areas such as electronic device control, biomechanics research, and virtual reality gaming [1], [2]. It is convenient for users to control equipment using hand gestures. For instance, for safe driving, the hand gesture is designed as a control method in vehicles that can avoid undesirable physical touching of buttons [17].

In the last decade, various techniques have been applied in the area of hand gesture recognition. Multitudinous computer vision-based optical methods utilizing RGB cameras [10] and depth cameras [3], [4] perform well with respect to gesture tracking [5], [6]. The fusion information of the RGBD gesture data and upper-body skeletal motion data is fully utilized through a CNN for Italian sign language gesture classification in [7]. Nevertheless, because of the very large variance of light intensity, dependability in extreme environments remains an issue, for instance, under conditions of strong light and darkness [17]. Moreover, when the hand moves rapidly, it is necessary to use more numerous pixels and frame rate optical sensors for the recognition task, and stronger computing power is also required [14]. In addition to optical sensing methods, some passive sensing methods, including pyroelectric infrared sensing and WiFi signal sensing, etc., also exhibit good performance [8], [9].

Compared with passive sensing methods, since the transmitting waveforms can be artificially designed according to different tasks, active sensing methods, such as sound [11], magnetic field [12] and radio frequency (RF) [13]

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

technologies, are more robust under complex circumstances, which leads to widespread interest [21]. Magnetic sensing performs well for gesture tracking; however, the demand for fingers to be equipped with sensors is usually undesirable [12]. Ultrasonic active sensing can acquire Doppler and range profile features from the echoes, which are appropriate for the hand gesture recognition task [13]–[15], [21]. Nevertheless, the detection distance limitation of ultrasonic sensing technology would restrict its applications in different realistic scenarios.

Unlike the methods above, the performance of MMW radar applied in hand gesture recognition is not as restricted by the lighting conditions and sensing distance. Moreover, owing to its capability to penetrate objects, the miniaturized radar can be embedded within the equipment to achieve robust and convenient operation. Micro-Doppler signatures on spectrograms have been used to train a CNN [17], [25]–[28], which reflects the powerful feature extraction ability of the CNN. The temporal information within the gesture process is also utilized to achieve dynamic continuous hand gesture recognition based on sequential models in [29]–[33]. Radar I-Q channel signals or signals received by multiple sensors are jointly studied in [16], [34], where successful classification performance is obtained under widely varying lighting conditions. These existing schemes of gesture recognition only use the micro-Doppler signatures measured by radar, which would be unfavorable to strong directional gesture discrimination. Moreover, the presence of noise and clutter in target echoes will affect both the feature extraction and correct recognition.

On the other hand, there are many different situations for gesture recognition, such as different people's gestures differing slightly from each other, different positions of hands, different speeds of hand movements, etc., so it is difficult to obtain measured data covering various situations. Several data augmentation strategies have been put forward to hinder CNNs from overfitting when trained by a variety of restricted training data [18]–[20]. Krizhevsky *et al.* [18] employ three kinds of image enhancement methods on the training images to train a CNN for generalization improvement and better classification. Similar spatial data augmentation is also applied on the video data to reduce the overfitting for human activity recognition [14], [20]. Because there are neither strong physical mechanisms nor interpretability for these data augmentation methods, we propose a data augmentation method based on the gesture movement mechanism for effective training and interpretability which is also suitable for the measured MMW radar dataset.

To solve the problems mentioned above, in this paper, we propose a gesture recognition algorithm using attention mechanism and multi-channel MMW radar features to adapt to the complex hand gesture recognition scenes. Multi-channel wide-band MMW radar is used to acquire echo signals of gestures. Clutter cancellation is applied to eliminate the influences of stationary objects and background scenes. To focus on the important features and suppress the unnecessary ones along the channel and spatial dimensions of CNN, we apply the attention mechanism [37] to blend cross-channel and spatial information together and increase the model's representation power. Furthermore, considering the strong directionality of most hand gestures, we blend the azimuth and elevation angle information, as well as the micro-Doppler signatures from multi-channel MMW radar acquired via the CNN, to fully extract the discriminative features, which can effectively distinguish the directional characteristics of gestures and improve the recognition performance. Moreover, according to certain physical mechanisms, we augment the data by corresponding scaling transformation to improve the model's generalization and prevent CNN overfitting. Compared with conventional single-channel methods, the proposed multi-channel MMW hand gesture recognition offers superiorities of both competitive recognition performance and adaptation to complex scenes.

The main contributions of this work can be concluded as follows. (1) To emphasize the meaningful features and weaken the impacts of unimportant areas along the spatial dimensions and improve the adaptive feature fusion along the channel of the CNN, we introduce the channel and spatial attention modules. (2) To fully utilize the gesture direction information, we synthesize a multi-channel feature by integration of the micro-Doppler, elevation and azimuth angle information, effectively enhancing the recognition performance and robustness. (3) To account for different micro-Doppler features in various complex scenes and reduce the extra work of training data acquisition, we propose a two-dimensional scaling-based data augmentation method according to the gesture movement mechanism.

This paper is organized as follows. We first introduce multi-channel MMW radar signal processing in Section 2. Our enhanced hand gesture recognition method is then presented in Section 3. In Section 4, our proposed models and reported experimental results on measured hand gesture data are evaluated. Section 5 concludes our work.

II. RELATED WORK

Recently, there have been several attempts [40]–[44] to utilize attention mechanism to improve the performance of models in the optical image or video gesture recognition task. Peng, *et al.* [40] utilize the Residual Attention Network which has an encoder-decoder style attention module to perform gesture recognition. By refining the feature maps, the model can perform well even with noisy inputs. Instead of directly computing the $3D$ attention map, we decompose the process to learn the channel and spatial attention separately. This separate attention generation process can have much less computational and parameter overhead.

Several works [41]–[44] try to introduce different compact modules to exploit the spatial-temporal relationship in the video gesture recognition task. But their attention modules mainly contribute to the temporal fusion along with the recurrent steps to learn long-term spatiotemporal features when taking spatial or spatiotemporal features as input. In work [41], G. Zhu, *et al.* use global average-pooled features to compute the channel-wise attention. However, it is shown that those are suboptimal features, so we use max-pooled features as well to infer a better channel attention.

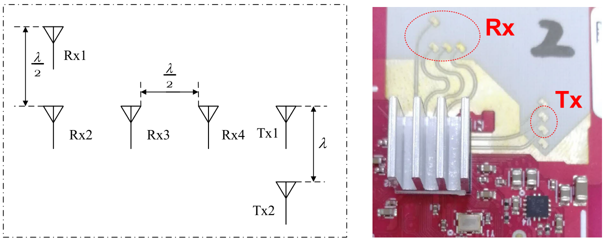
In our model, we exploit a channel-wise attention to achieve a more efficient multi-channel radar echo features fusion. Different from the optical image gesture recognition task, the clutters and noise always have an obvious influence on the performance, so we utilize a spatial attention to make the CNN focus on the most important target signal regions and suppress the unnecessary clutter areas.

III. MULTI-CHANNEL MILLIMETER WAVE RADAR SIGNAL PROCESSING

This section mainly introduces the multi-channel MMW radar system settings and data preprocessing for radar gesture recognition.

A. SPECIAL ARRANGEMENT OF THE RADAR SYSTEM

The radar system for hand gesture recognition data acquisition is a two-transmitter and four-receiver system. Through the special arrangement of transmitting and receiving antennas, the receiving antenna has freedom with respect to azimuth and elevation, which can measure the azimuth and elevation angle of a hand gesture at a certain time. The specific arrangement of the antennas is shown in Figure 1. In this radar system, receiving antennas Rx_1 and Rx_2 are placed vertically with wavelength distance between them, so the freedom of elevation angle measurement is obtained. In addition, receiving antennas Rx_2, Rx_3 and Rx_4 are arranged in parallel with the interval of one-half of a wavelength, and therefore, the freedom of azimuth angle measurement is achieved.



(a) (b)

FIGURE 1. Diagram of the specific antenna arrangement: (a) antenna layout array with two transmitting antennas and four receiving antennas. (b) Physical map of the radar system.

It is well known that the time of the gesture echo reaching different antennas is related to the spatial positioning of antenna and hand. When a hand is not directly in front of two radars, there will be a wave path difference in the distance of the echo signal to each antenna, which is reflected in the different initial phase modulations of the signals received by each antenna. Considering the previous analysis, the phase differences between different channels can be obtained according to the micro-Doppler results of the echo signal. Then, the azimuth and elevation angles corresponding to the gesture can be acquired by using the spatial analysis method according to the phase difference of azimuth and elevation.

In this work, frequency modulated continuous wave (FMCW) radar is used to obtain gesture echoes. We set the initial frequency of the radar, the frequency modulation slope and the sampling rate of the intermediate frequency (IF) signal after mixing as f_c, K and f_s, respectively. The number of signal samples in each FMCW transmission cycle is N_{sp}, while the effective bandwidth of the radar transmission signal is $B = \frac{N_{sp}}{f_s}K$ and the range resolution of the radar is $R_{res} = \frac{c}{2B}$, where c is the speed of light. We design the transmission signal of the radar as follows:

$$S_T\left(t, \hat{t}\right) = exp\left(j2\pi\left(f_c t + \frac{1}{2}\gamma\hat{t}^2\right)\right), \tag{1}$$

where γ is the frequency modulation slope and $\hat{t}$ is the fast time, i.e., the time in a transmitting waveform cycle. The bandwidth of the transmitted signal is $B = \gamma * \hat{t}$.

Supposing that there are N_s stationary targets and N_m moving gesture targets within the radar detection range, the echo received by the radar is $S_R = S_s + S_m$, where S_s is the echo summation of the stationary targets and S_m is the echo summation of the moving targets. S_s can be expressed as

$$S_s = \sum_{s=1}^{N_s} a_s exp\left(j2\pi\left(f_c\left(t - \frac{2R_s}{c}\right) + \frac{1}{2}\gamma\left(\hat{t} - \frac{2R_s}{c}\right)\right)\right), \tag{2}$$

where R_s is the distance of the sth stationary target to the radar and a_s is the echo amplitude of the sth stationary target. S_m can be calculated as

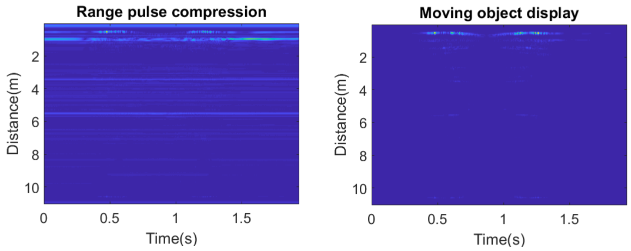
$$S_m = \sum_{m=1}^{N_m} a_m exp\left(j2\pi\left(f_c\left(t - \frac{2R_m}{c}\right) + \frac{1}{2}\gamma\left(\hat{t} - \frac{2R_m}{c}\right)\right)\right), \tag{3}$$

where R_m is the distance of the mth moving target to the radar and a_m is the echo amplitude of the mth moving target.

The mixer in FMCW radar mixes the received signal S_R with the transmitted signal S_T to obtain the IF signal, i.e. $S_{IF} = S_R \times conj\left(S_T\right)$, where $conj\left(\cdot\right)$ denotes conjugation. In this way, the IF signal after mixing can be expressed as

$$\begin{aligned} S_{IF} = & \sum_{s=1}^{N_s} a_s exp\left(j\left(-\frac{2\pi\gamma\left(2R_s\right)}{c}\hat{t}\right)\right) \\ & + \sum_{s=1}^{N_s} a_s exp\left(j\left(-\frac{2\pi\left(2R_s\right)}{\lambda}\right)\right) \\ & + \sum_{m=1}^{N_m} a_m exp\left(j\left(-\frac{2\pi\gamma\left(2R_m\right)}{c}\hat{t}\right)\right) \\ & + \sum_{m=1}^{N_m} a_m exp\left(j\left(-\frac{2\pi\left(2R_m\right)}{\lambda}\right)\right). \end{aligned} \tag{4}$$

Pulse compression is then carried out. Figure 2(a) shows the range compression results of double hand shaking measured by MMW radar in complex scenes. It contains the echo information of static scenes and moving hands.



(a) (b)

FIGURE 2. (a) The range compression results of double hand shaking in complex scenes. (b) The static scene cancellation result of double hand shaking IF echo.

B. PREPROCESSING OF CLUTTER SUPPRESSION AND MOVING TARGET INDICATION

In the process of data acquisition, the echoes of radar include the echoes of moving gesture targets and static scenes, among which static scenes will degrade the gesture recognition performance and belong to clutter. Therefore, in order to enhance the signal to clutter ratio, it is necessary to eliminate the radar echoes of static scenes. Since the distance between the static scene and the radar remains unchanged, the Doppler frequency is zero. Moreover, because the position of the static target is constant with respect to the radar, the echoes of the static targets are the same for the two adjacent radars, except for noise, while the echoes of the moving target exhibit the wave path difference of the range change. Utilizing these characteristics, the static scene echoes can be eliminated by subtracting the echoes received by the two adjacent radars, while the echoes of the moving target can be retained. Figure 2(b) shows the static scene cancellation result of the double hand shaking IF echoes. Compared with Figure 2(a), we can find that most of the static scene echoes have been cancelled, and only the echoes of the double hand shaking are retained.

IV. MULTI-CHANNEL MMW RADAR HAND GESTURE RECOGNITION

Through experimentation, we find that it is difficult to recognize gestures with strong directionality when using only micro-Doppler features, since the general structures of different directional gesture micro-Doppler images are very similar. Furthermore, we propose a recognition method based on multi-channel fusion features of micro-Doppler, elevation and azimuth angle information. To make the model adaptively focus on the important target signal region and make better use of the features, we introduce the attention mechanism to the CNN.

A. MICRO-DOPPLER FEATURES

In this paper, inspired by the phenomenon of signal Doppler frequency shift, we transform the signal Doppler frequency shifts of different gestures into gesture Doppler information images utilizing pulse compression and moving target indication (MTI) in FMCW radar technology. When the transmitting signals encounter the gesturing human hands, there will be Doppler shifts in the echo signals. If the direction or speed of the hand movement changes, the Doppler shift will inevitably change. Namely, the Doppler shift can successfully express the information of hand movements.

B. 3-D FEATURES SYNTHESIS

For strong directional gestures, because of the similar general structures of their micro-Doppler images, it is difficult to recognize them using only their micro-Doppler features. Thus, considering their strong directionality, we introduce elevation and azimuth angle information, which is conducive to gesture recognition. First, we acquire the micro-Doppler information of each channel gesture echo through the multi-channel MMW radar with special placement. Then, through static scene cancellation, only the echo information in the perception area of the moving target is analyzed by the micro-Doppler channel. According to the phase difference information of the micro-Doppler channel, the corresponding azimuth and elevation angle of the gesture can be obtained and analyzed. Finally, inspired by RGB channels of natural images, these three different feature images can be concatenated as a three-dimensional matrix to express the spatial direction information of the gesture motion and to achieve more precise gesture recognition.

C. CNN FOR MULTI-CHANNEL HAND GESTURE RECOGNITION

For the task of data-driven hand gesture recognition, given a set of multi-channel features $\mathcal{X} = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(D)}\right\}$, suitable features $\mathcal{Z} = \left\{z^{(1)}, z^{(2)}, \ldots, z^{(D)}\right\}$ are automatically extracted via a CNN with the network parameters W which will then be fed into a classifier to predict the corresponding labels $\mathcal{C} = \left\{c^{(1)}, c^{(2)}, \ldots, c^{(D)}\right\}$, $c^{(i)} \in \{0, 1\}^{S}$. For convenience of calculation, we down-sample the original 3 × 128 × 128 pixel features to 3 × 32 × 32 pixels. To enable our gesture classifier to converge more rapidly, we normalize each channel of three-dimensional features to fall within the range of 0 to 1.

Our CNN classifier is composed of two convolution layers, each of which is followed by the max pooling operator. The outputs of the second convolutional layer are fed into the following two fully-connected layers (FCLs) with 120 and 84 neurons, as introduced in [14], [22]. Finally, the corresponding class probabilities $p\left(C\,|x, W\right)$ for the ten kinds of gestures are achieved by the softmax layer. The architecture of the CNN is shown in Figure 3. The class label is predicted by

$$c^{*} = \arg\max p\left(C\,|x, W\right). \tag{5}$$

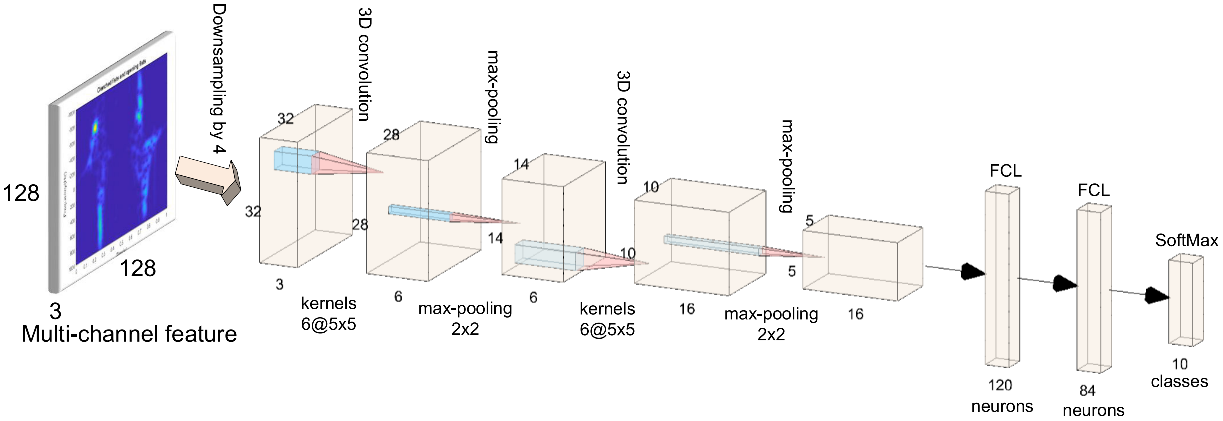


FIGURE 3. The employed CNN architecture.

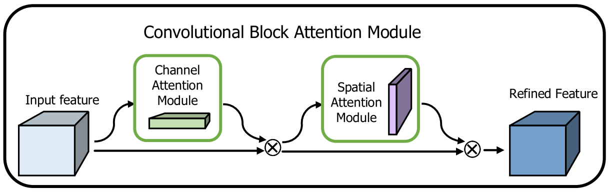


FIGURE 4. The overview of CBAM. The module has two sequential sub-modules: channel and spatial.

Rectified linear unit (ReLU) activation functions are applied in all layers of the CNN, except for the softmax layer:

$$f(z) = \max(0, z). \tag{6}$$

We compute the corresponding class probabilities as:

$$p(C|x, W) = \frac{\exp(z_C)}{\sum_k \exp(z_k)}, \tag{7}$$

where z_q is the qth neuron of the output layer.

Afterward, the network is trained by minimizing a cost function with respect to the parameters W over the training dataset D_{tr}. The cross entropy between the true labels and the output of the softmax classification is selected as the cost function:

$$L(W, D_{tr}) = \sum_{i=0}^{|D_{tr}|} y^{(i)} \log\left(p\left(c^{(i)} \middle| x^{(i)}, W\right)\right), \tag{8}$$

where $y^{(i)}$ is the corresponding true label.

D. CONVOLUTIONAL BLOCK ATTENTION MODULE

To further reduce the impacts of noise and clutter, the attention mechanism is integrated with the original CNN to enable it to adaptively focus on important features and suppress unnecessary ones. Specifically, each layer of the CNN described in the previous section is replaced by the convolutional block attention module, as follows.

Given an intermediate feature map $z \in \mathbb{R}^{B\times H\times V}$ as input, we infer a channel attention map $M_c \in \mathbb{R}^{B\times 1\times 1}$ and a spatial attention map $M_s \in \mathbb{R}^{1\times H\times V}$ as shown in Figure 4, which can be described as

$$\begin{aligned} z' &= M_c(z) \odot z \\ z'' &= M_s(z') \odot z, \end{aligned} \tag{9}$$

where $\odot$ denotes element-wise multiplication, and B denotes the number of channels of the feature map. For the element-wise multiplication, the attention values are broadcasted (copied) in a way that the spatial attention values are broadcasted along the channel dimension, and vice versa. The operation process is shown in Figure 5. The details of the process are described in the following.

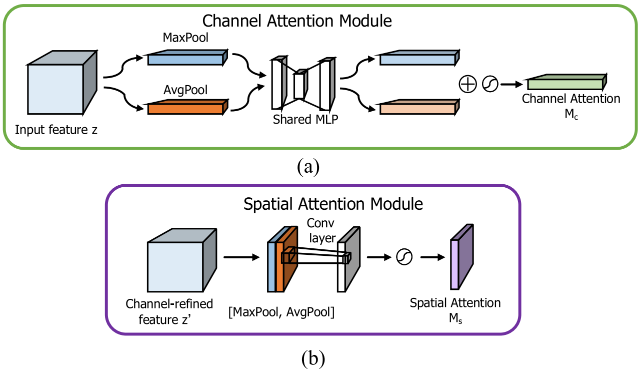


FIGURE 5. Diagram of channel and spatial attention modules. (a) The process of channel attention. (b) The process of spatial attention.

1) CHANNEL ATTENTION MODULE

Since each channel of a feature map can be regarded as a feature detector [39], we apply channel attention to focus on 'what' is meaningful in an input 3-D gesture feature. The average pooling and max pooling operation are used simultaneously to infer finer channel-wise attention, as in a previous work [37].

Taking one layer of CNN as an example, given the output feature map z of the previous layer, we use average pooling and max pooling operations to obtain two different spatial context features z^c_{avg} and z^c_{max}, respectively. They are then sent together into the same multilayer perceptron (MLP), which has one hidden layer, to generate the channel attention map $M_c \in \mathbb{R}^B$. Supposing that the hidden layer size is h, the channel attention can be described by the following equation:

$$\begin{aligned} M_c(z) &= \sigma\left(MLP\left(AvgPool(z)\right) + MLP\left(MaxPool(z)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(z^c_{avg}\right)\right) + W_1\left(W_0\left(z^c_{\max}\right)\right)\right), \end{aligned} \tag{10}$$

where σ denotes the sigmoid activation function, and $W_0 \in \mathbb{R}^{h\times B}$ and $W_1 \in \mathbb{R}^{B\times h}$ denote the MLP weights.

2) SPATIAL ATTENTION MODULE

To focus on 'where' an informative region is in the input 3-D gesture feature, we produce a spatial attention map by extracting the inter-spatial relationships of features. First, we obtain the average-pooled and max-pooled features z^s_{avg} and z^s_{max} along the channel dimension, the effectiveness of which has been proven in highlighting informative regions [37]. We then concatenate them and apply a one layer convolution to obtain a spatial attention map $M_s \in \mathbb{R}^{H\times V}$, which can represent the region to be emphasized or suppressed. The spatial attention can be obtained as

$$\begin{aligned} M_s(z) &= \sigma\left(f^{7\times 7}\left(\left[AvgPool(z); MaxPool(z)\right]\right)\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[z^s_{avg}; z^s_{\max}\right]\right)\right), \end{aligned} \tag{11}$$

where σ denotes the sigmoid activation function and $f^{7\times 7}(\cdot)$ denotes a convolution layer with the convolution kernel size of 7×7.

E. DATA AUGMENTATION

Since the number of the CNN's parameters is huge, a large quantity of training data covering diverse situations is needed to avoid over-fitting. However, there are only 20 samples for each kind of gesture. At the same time, if the elevation angles, users or speeds of the same gestures are different, the gesture echoes will also exhibit some differences. It is difficult to take all possible situations into account when we collect the measured data. Therefore, the training dataset of the same gesture needs to be augmented according to the mechanism of gesture movement. In particular, we perform certain scale transformations of the micro-Doppler characteristics of the hand gesture to simulate the same gesture under different circumstances.

1) DIFFERENT ELEVATION ANGLES

When the elevation angles of the same gestures are different relative to the radar, the measured radial velocities of the gestures are different. That is, the smaller the elevation angle of a hand gesture is at the same height relative to the radar, the larger the radial velocity component, so their speeds reflected in the micro-Doppler features are distinct. The scale transformation of the micro-Doppler feature in the velocity dimension can be used to simulate the changes on account of the same gestures with different elevation angles, as shown in Figure 6.

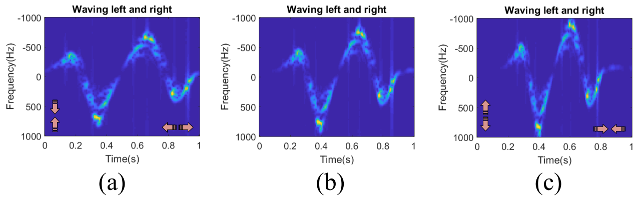


FIGURE 6. The data augmentation method for the case of different elevation angles.

2) DIFFERENT USERS

When different people make the same kind of gestures, due to the different lengths of the fingers, although the angular speeds of the finger movements are the same, the speeds of the finger tips are diverse and the motion cycles are the same. Therefore, the effect can be reflected by the scale transformation of the micro-Doppler feature in the Doppler dimension and constancy in the time dimension, as shown in Figure 7.

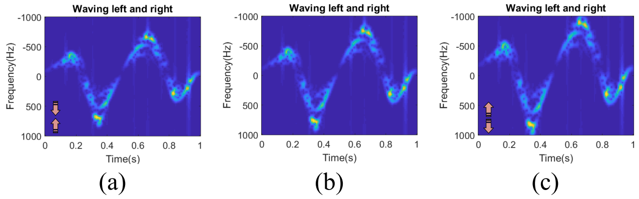


FIGURE 7. The data augmentation method for the case of different users.

3) DIFFERENT SPEEDS

When the same person performs the same gestures at different speeds, as the spatial range of the gestures is fixed, the gesture cycles become different. Thus, the compression of the micro-Doppler feature in the speed dimension and elongation in the time dimension, or the opposite, are carried out to cover these situations, as shown in Figure 8.

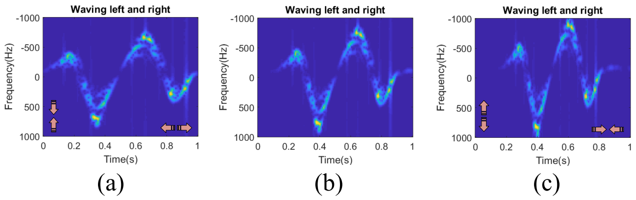
As discussed above, we augment the training dataset in the following way. First, we concatenate a 128×128 micro-Doppler signature of a gesture cycle with the corresponding azimuth and elevation angle signature as a 3-D feature image. Then, based on the image center, we remove a $128 \times n \times 3$ and a $128 \times m \times 3$ part along the time dimension and velocity dimension, respectively, where $m, n \in [0.6, 1]$ and the change interval is 0.1. Then, for each original 3-D feature image, we can obtain 24 augmented images.

V. EXPERIMENTS

A. MEASURED MULTI-CHANNEL MMW RADAR GESTURE DATASET

1) MEASURED DATA

We employ MMW radar to obtain micro-Doppler features of ten hand gestures from a single participant or multiple



(a) (b) (c)

FIGURE 8. The data augmentation method for the case of different speeds.

participants. In our experiments, IWR1642 radar operating at 77 Ghz is employed, the configuration of which is shown in Table 1 [23]. The average output power of this radar is 12 dBm. The radial velocity range this radar can sense covers 2.6 cm/s to 2.6 m/s, and the antenna beam width is 120 degrees, which is suitable for hand gesture measurement [17]. The hand motions are measured in the main lobe of the radar antenna, the average distance from which to the radar is approximately 30 cm.

TABLE 1. IWR1642 radar chirp parameters.

Parameter	Value
Chirp BW	∼4 GHz
Chirp repetition interval	400 μs
Number of chirps per frame	128
Range Resolution	∼4 cm
Velocity resolution	0.032 m/s

2) EXPERIMENTAL SETUP

The ten hand gestures employed in this experiment are (*a*) swiping from left to right, (*b*) swiping from left to right, (*c*) swiping from lower left to upper right, (*d*) swiping from lower right to upper left, (*e*) swiping forward, (*f*) swiping backward, (*g*) waving left and right, (*h*) double finger clicking, (*i*) clenching and opening fists and (*j*) snapping fingers. The employed gestures are depicted in Figure 9.

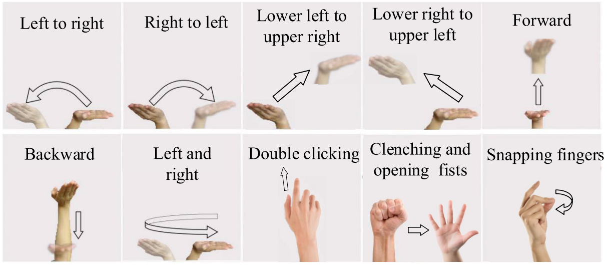
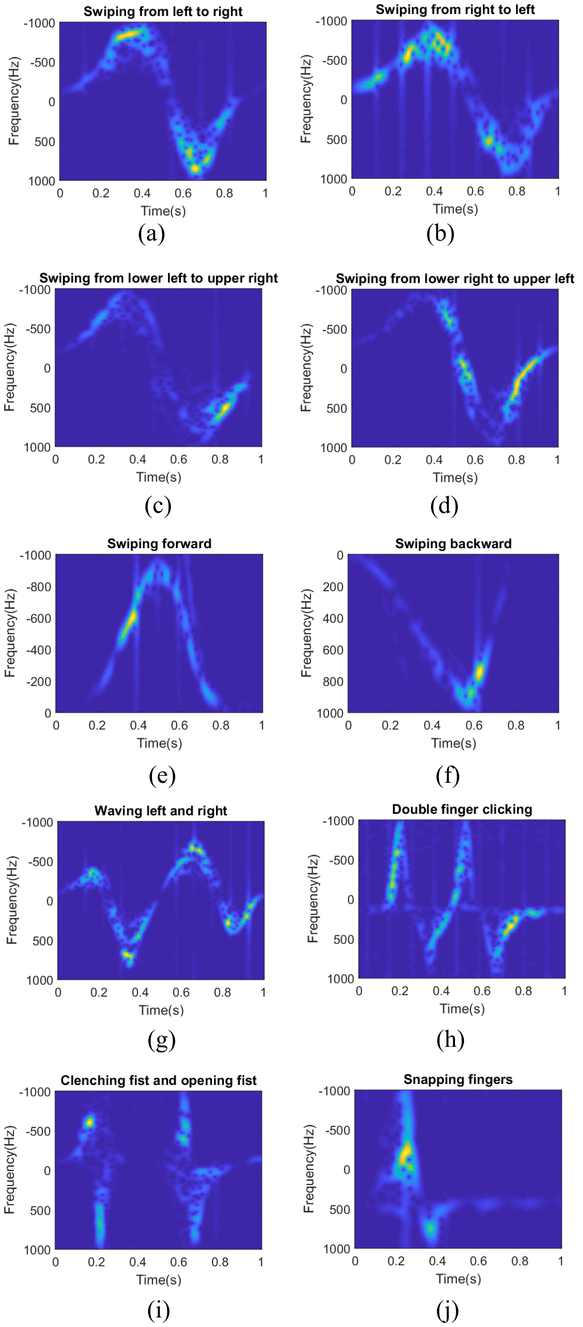


FIGURE 9. The ten hand gestures employed.

To analyze micro-Doppler features, we employ short-time fast Fourier transform (FFT) to acquire the spectrograms of gestures. The window size of the FFT and the time step of the non-overlapping samples are set as 256ms and 1ms, respectively. The acquired spectrograms of the ten gestures are shown in Figure 10. We can find that the micro-Doppler features of different gestures represent some diversities in the joint time-frequency domain. In particular, the general micro-Doppler image structures of the gestures (*a*) ∼ (*d*) and (*g*) are very similar: only the intensity variations of the signal amplitudes are different. This is because the radial velocities are analogous, even though the directions of the motions are distinct. In the experiment, each kind of gesture is measured 50 times for each of the 10 participants; in this way, we can obtain an augmented dataset with 120000 3-D feature images in total by means of the data augmentation in section 3.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

FIGURE 10. Examples of the spectrograms of the ten gestures.

B. SEPARABILITY OF THE EXTRACTED FEATURES

To further demonstrate the prominent separability of the features extracted by our algorithm, Figure 11 compares the

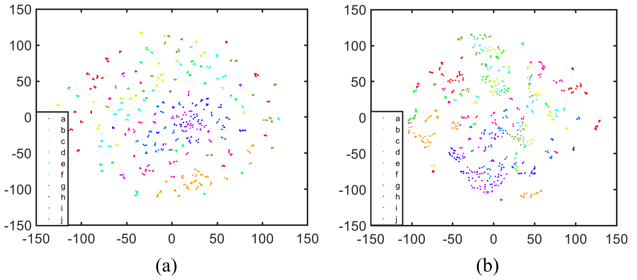


FIGURE 11. The separation of (a) the original data samples and (b) feature samples extracted by our method.

separation of the original data and the features extracted by our CNN utilizing t-SNE [24] to reduce them to two dimensions. Compared with the original data, our CNN can fully extract the discriminative features to reduce the gesture sample distances within a class and increase the sample distances between classes.

C. RECOGNITION PERFORMANCE

Among the 120000 3-D feature images from ten participants, 60% of them are used as training data and 40% as test data. Valid accuracy is assessed via the 5-fold validation method, which divides the measured data into five different training datasets and test datasets. This leave-one-subject-out cross validation method is utilized to evaluate our hand gesture recognition algorithm's performance on the measured gesture dataset.

1) WITH OR WITHOUT MULTIPLE CHANNELS

It is proven that CNNs are effective in combining data from different sources [16]. In this experiment, we compare the classification performance of our model with that of another model, [17], which does not utilize the azimuth and elevation information of gestures. In Table 2, we present the average classification accuracy of our method trained with different input modalities. We can see that, individually, our 3-D feature gesture recognition algorithm (accuracy $=$ 96.61%) performs better than the only micro-Doppler feature (accuracy $=$ 91.34%) method, which proves that the introduction of azimuth and elevation information is helpful for the gesture recognition.

TABLE 2. The average classification accuracy of our algorithm trained with different input modalities.

Features	Micro-Doppler feature	3-D feature
Average classification accuracy	91.34%	**96.61%**

2) WITH OR WITHOUT DIRECTIONALITY

To further analyze how the introduction of azimuth and elevation angle information affects gesture recognition, we divide the gesture dataset into two categories. The gestures in Figure 10 $(a) \sim (g)$ represent strong directional gestures, and the others in Figure 10 belong to the class of weak directional gestures. The training and testing processes of these two datasets are respectively carried out, and the results are shown in Table 3. We find that the recognition performance of our algorithm applied on the strong directional gesture dataset is more obviously improved by introducing multi-channel information than in the case of the weak directional gestures. Since the azimuth and elevation information mainly reflect the directional characteristics of gestures, the gestures with strong directionality are made more separable.

TABLE 3. The effect of introducing multi-channel features to strong and weak directional gestures.

	Strong directional gestures	Weak directional gestures
Micro-Doppler features	92.14%	90.61%
Multi-channel features	98.58%	94.13%

3) WITH OR WITHOUT TRAINING DATASET AUGMENTATION

In Table 4, we present the correct classification rates of our method with and without data augmentation. The results show that the training error increases upon enabling data augmentation, while the test error decreases. This demonstrates that the proposed data augmentation method can reduce overfitting and improve the generalization of the gesture classifier by simulating the mechanisms of different gesture movements and covering a wider range of possible situations.

TABLE 4. The training and test classification accuracy of our algorithm with and without data augmentation.

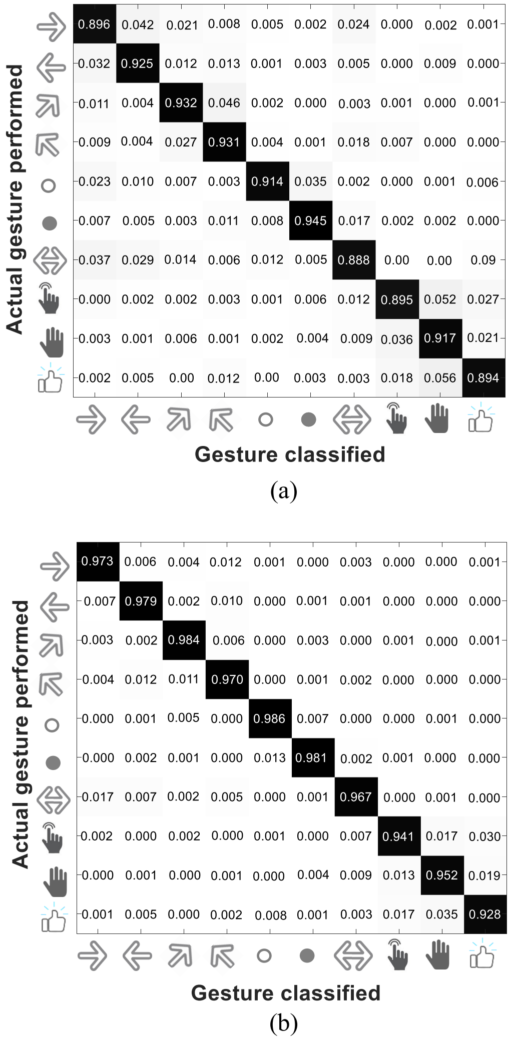
	With data augmentation	Without data augmentation
Average training classification accuracy	98.89%	**99.56%**
Average test classification accuracy	**96.61%**	94.97%

4) AVERAGE CLASSIFICATION ACCURACY

For the ten investigated kinds of gestures, the average classification accuracy of the CNN trained by the 3-D gesture feature dataset is 96.61%. The 5-fold validation accuracies are shown in Table 5. To analyze the misclassification, the classification accuracy confusion matrices (%) of the single-channel and multi-channel methods are presented in Figure 12. In Figure 12, we can see that when utilizing the single-channel classification method, gestures $(a) \sim (d)$ are easily confused with each other because they have similar micro-Doppler structures, as shown in Figure 10. However, when using the multi-channel method, the classification accuracies are greatly improved. Due to the introduction of azimuth and elevation angle information, their directional characteristics can be better distinguished.

TABLE 5. The 5-fold validation accuracies of our algorithm.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Classification Accuracy	97.72%	96.33%	96.47%	97.38%	95.16%	96.61%



(a)

(b)

FIGURE 12. The confusion matrices of the classification accuracy. (a) The single-channel method; (b) the multi-channel method.

TABLE 6. The results of our model compared with some competitive methods.

Method	Average classification accuracy
KNN	88.93
SVM	90.21
CNN	91.34
CNN+3D	**96.61**
CNN+3D +Attention	**97.17**

Table 6 compares the results of our model with some competitive methods with respect to our challenging measured gesture dataset with the same conditions. As shown in Table 6, because of CNN's deep nonlinear mapping, strong feature extraction and data representation ability [17], the CNN-based method performs better than the methods based on K-nearest neighbor (KNN) [35] and support vector machine (SVM) [36]. Our method exhibits the best performance, since it offers the CNN's strong feature extraction ability and the excellent azimuth and elevation information separability. Moreover, with the attention mechanism, our model can learn what and where to emphasize or suppress and refine intermediate features effectively.

D. THE ROLE OF THE ATTENTION MECHANISM

For qualitative analysis, we apply Grad-CAM [38] to our trained models, whose results can clearly represent the attended regions. Through analyzing the regions in the 3-D gesture features that the models have considered as important for correct prediction, we attempt to find out how the models make full use of the features.

In Figure 13, we can see that the Grad-CAM masks of the CNN with attention mechanism cover the target object regions more reasonably than the original CNN. Specifically, the original CNN devotes more attention to the noise and

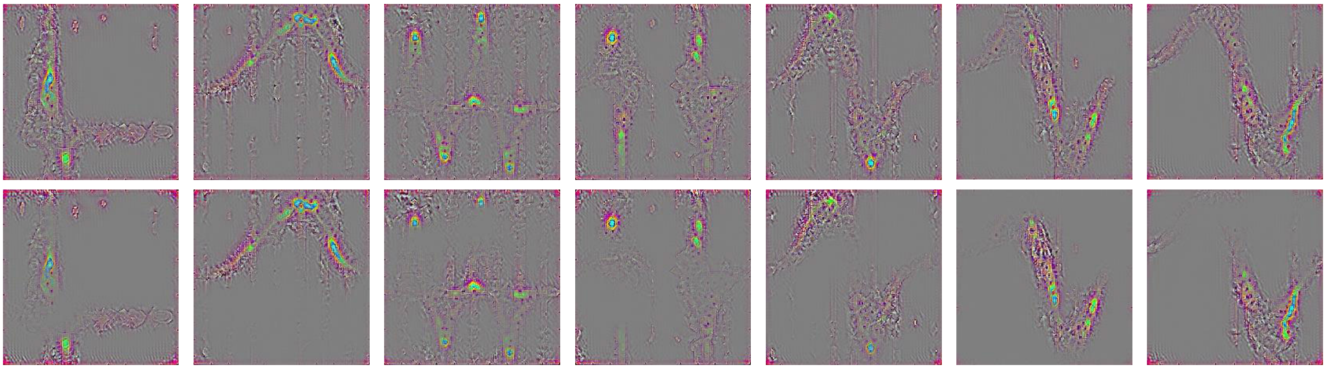


FIGURE 13. The micro-Doppler feature regions of different gestures considered important by different models. The figures above show the results of the original CNN. The figures below show the results of the CNN with attention mechanism.

weak echo areas of micro-Doppler features than the CNN with attention mechanism. That is, the CNN with attention mechanism effectively learns to utilize the important information in the target signal regions and suppress the nonessential information from the noise areas.

VI. CONCLUSION

We propose an enhanced 3-D micro-Doppler feature synthesis method for hand gesture recognition with attention mechanism based on CNN. The proposed classifier uses fused multi-channel micro-Doppler features, along with elevation and azimuth angle information. To focus on the important features and suppress the unnecessary ones, we apply the attention mechanism to blend the cross-channel and spatial information together. Furthermore, movement mechanism-based data augmentation is developed to cover different complex measurement situations and alleviate overfitting.

By means of extensive evaluation, we demonstrate that the combination of multi-channel features improves classification accuracy considerably. We further demonstrate that the proposed data augmentation technique plays an important role in achieving superior performance. For the challenging measured dataset, our algorithm achieves a classification accuracy of 96.61%, approximately 5% higher than that of the single-channel method. Our future work will investigate and utilize the temporal information during the hand gesture process to further improve the performance.

ACKNOWLEDGMENT

The authors would like to thank the editor and reviewers for their constructive comments and suggestions, which improve the quality of this paper.

REFERENCES

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 3, pp. 311–324, May 2007.

[2] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.

[3] Z. Ren, J. Meng, and J. Yuan, "Depth camera based hand gesture recognition and its applications in Human-Computer-Interaction," in *Proc. 8th Int. Conf. Inf., Commun. Signal Process.*, Dec. 2011, pp. 1–5.

[4] P. Trindade, J. Lobo, and J. P. Barreto, "Hand gesture recognition using color and depth images enhanced with hand angular pose data," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2012, pp. 71–96.

[5] B.-W. Min, H.-S. Yoon, J. Soh, Y.-M. Yang, and T. Ejima, "Hand gesture recognition using hidden Markov models," in *Proc. IEEE Int. Conf. Syst., Man, Cybernetics. Comput. Cybern. Simulation*, Oct. 1997, pp. 4232–4235.

[6] F.-S. Chen, C.-M. Fu, and C.-L. Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models," *Image Vis. Comput.*, vol. 21, no. 8, pp. 745–758, Aug. 2003.

[7] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multiscale deep learning for gesture detection and localization," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2014, pp. 474–490.

[8] J. Gong, Y. Zhang, X. Zhou, and X.-D. Yang, "Pyro: Thumb-tip gesture recognition using pyroelectric infrared sensing," in *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2017, pp. 553–563.

[9] O. Zhang and K. Srinivasan, "Mudra: User-friendly fine-grained gesture recognition using WiFi signals," in *Proc. 12th Int. Conf. Emerg. Netw. EXp. Technol.*, Dec. 2016, pp. 83–96.

[10] Y. Yao and C.-T. Li, "Hand gesture recognition and spotting in uncontrolled environments based on classifier weighting," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3082–3086.

[11] S. Gupta, D. Morris, S. Patel, and D. Tan, "SoundWave: Using the Doppler effect to sense gestures," in *Proc. ACM Annu. Conf. Hum. Factors Comput. Syst. (CHI)*, 2012, pp. 1911–1914.

[12] K.-Y. Chen, K. Lyons, S. White, and S. Patel, "UTrack: 3D input using two magnetic sensors," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2013, pp. 237–244.

[13] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, Oct. 2016, pp. 851–860.

[14] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 1–7.

[15] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-based dynamic hand gesture recognition using micro-Doppler signatures," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2017, pp. 0928–0931.

[16] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Ljubljana, Slovenia, May 2015, pp. 1–8.

[17] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[19] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 572–578.

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

[21] Y. Sang, L. Shi, and Y. Liu, "Micro hand gesture recognition system using ultrasonic active sensing," *IEEE Access*, vol. 6, pp. 49339–49347, 2018.

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[23] P. Goswami, S. Rao, S. Bharadwaj, and A. Nguyen, "Real-time multi-gesture recognition using 77 GHz FMCW MIMO single chip radar," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2019, pp. 1–4.

[24] M. L. V. Der and G. E. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[25] Y. Wang, S. Wang, M. Zhou, Q. Jiang, and Z. Tian, "TS-I3D based hand gesture recognition method with radar sensor," *IEEE Access*, vol. 7, pp. 22902–22913, 2019.

[26] Y. Kim and B. Toomajian, "Application of Doppler radar for the recognition of hand gestures using optimized deep convolutional neural networks," in *Proc. 11th Eur. Conf. Antennas Propag. (EUCAP)*, Paris, France, Mar. 2017, pp. 1258–1260.

[27] B. Dekker, S. Jacobs, A. S. Kossen, M. C. Kruithof, A. G. Huizing, and M. Geurts, "Gesture recognition with a low power FMCW radar and a deep convolutional neural network," in *Proc. Eur. Radar Conf. (EURAD)*, Nuremberg, Germany, Oct. 2017, pp. 163–166.

[28] J. Zhang, J. Tao, and Z. Shi, "Doppler-radar based hand gesture recognition system using convolutional neural networks," in *Communications, Signal Processing, and Systems*. 2019, pp. 1096–1113. [Online]. Available: https://arxiv.org/pdf/1711.02254.pdf, doi: 10.1007/978-981-10-6571-2_132.

[29] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, Apr. 2018.

[30] J. S. Suh, S. Ryu, B. Han, J. Choi, J.-H. Kim, and S. Hong, "24 GHz FMCW radar system for real-time hand gesture recognition using LSTM," in *Proc. Asia–Pacific Microw. Conf. (APMC)*, Kyoto, Japan, Nov. 2018, pp. 860–862.

[31] G. Malysa, D. Wang, L. Netsch, and M. Ali, "Hidden Markov model-based gesture recognition with FMCW radar," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Washington, DC, USA, Dec. 2016, pp. 1017–1021.

[32] S.-J. Ryu, J.-S. Suh, S.-H. Baek, S. Hong, and J.-H. Kim, "Feature-based hand gesture recognition using an FMCW radar and its temporal feature analysis," *IEEE Sensors J.*, vol. 18, no. 18, pp. 7593–7602, Sep. 2018.

[33] J.-W. Choi, S.-J. Ryu, and J.-H. Kim, "Short-range radar based real-time hand gesture recognition using LSTM encoder," *IEEE Access*, vol. 7, pp. 33610–33618, 2019.

[34] T. Sakamoto, X. Gao, E. Yavari, A. Rahman, O. Boric-Lubecke, and V. M. Lubecke, "Radar-based hand gesture recognition using I-Q echo plot and convolutional neural network," in *Proc. IEEE Conf. Antenna Meas. Appl. (CAMA)*, Tsukuba, Japan, Dec. 2017, pp. 393–395.

[35] R. Lionnie, I. K. Timotius, and I. Setyawan, "An analysis of edge detection as a feature extractor in a hand gesture recognition system based on nearest neighbor," in *Proc. Int. Conf. Electr. Eng. Informat.*, Bandung, Indonesia, Jul. 2011, pp. 17–19.

[36] C.-C. Hsieh and D.-H. Liou, "Novel Haar features for real-time hand gesture recognition using SVM," *J. Real-Time Image Process.*, vol. 10, no. 2, pp. 357–370, Jun. 2015.

[37] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19.

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.

[40] M. Peng, C. Wang, and T. Chen, "Attention based residual network for micro-gesture recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 790–794.

[41] G. Zhu, L. Zhang, L. Yang, L. Mei, S. A. A. Shah, M. Bennamoun, and P. Shen, "Redundancy and attention in convolutional LSTM for gesture recognition," *IEEE Trans. Neural Netw.*, vol. 31, no. 4, pp. 1323–1335, Jun. 2019.

[42] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 273–286.

[43] Y. Li, Q. Miao, X. Qi, Z. Ma, and W. Ouyang, "A spatiotemporal attention-based ResC3D model for large-scale gesture recognition," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 875–888, Jul. 2019.

[44] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, and W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0206049.

CHUAN DU was born in Henan, China, in 1988. He received the B.S. degree in communication engineering from the China University of Geosciences, Wuhan, in 2012, and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, in 2019. He is currently a Postdoctoral Research Fellow with the School of Electronics and Communication Engineering, Sun Yat-sen University. His major research interests include radar target recognition, machine learning, and deep learning.

LEI ZHANG was born in Zhejiang, China, in 1984. He received the Ph.D. degree from Xidian University, in 2012. He is currently working as an Associate Professor with the National Laboratory of Radar Signal Processing, Xidian University. He is also working with the School of Electronics and Communication Engineering, Sun Yat-sen University. His research interests include radar imaging (SAR/ISAR) and motion compensation.

XIPING SUN was born in Henan, China, in 1993. He received the B.S. degree from the Ocean University of China, in 2016, and the M.S. degree from Xidian University, in 2019. He is currently pursuing the Ph.D. degree in signal processing with Sun Yat-sen University. His major research interests include radar imaging and MIMO radar.

JUNXU WANG was born in Guangdong, China, in 2000. He is currently pursuing the B.S. degree with the School of Electronics and Communication Engineering, Sun Yat-sen University. His research interests include computer vision and machine learning.

JIALIAN SHENG received the Ph.D. degree from Xidian University, Xi'an, China, in 2016. Since 2016, she has been working with the Shanghai Radio Equipment Research Institute, China. Her current research interests include (inverse) synthetic aperture radar (SAR/ISAR) imaging, MIMO radar signal processing, terahertz radar imaging, and so on.