IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 October 2023, accepted 9 November 2023, date of publication 15 November 2023,
date of current version 1 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333036

RESEARCH ARTICLE

RF-CSign: A Chinese Sign Language Recognition System Based on Large Kernel Convolution and Normalization-Based Attention

HUANYUAN XU[1], YAJUN ZHANG[1], ZHIXIONG YANG[2], HAOQIANG YAN[1], AND XINGQIANG WANG[1]

[1]School of Software, Xinjiang University, Ürümqi, Xinjiang 830046, China
[2]School of Future Technology, Xinjiang University, Ürümqi, Xinjiang 830017, China

Corresponding author: Yajun Zhang (zyj@xju.edu.cn)

This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2022D01C54, and in part by the Doctoral Research Start-Up Fund of Xinjiang University under Grant 202212120001.

ABSTRACT Hearing impaired people use sign language for communication, which relies on the movement gestures of body parts and plays a vital role in human-computer interaction. Most wireless sensing-based gesture recognition studies have recognized simple gestures but overlooked the recognition of complex activities, such as sign language. In addition, cross-domain recognition often requires a large amount of data to train classifiers for each environment. Therefore, we propose RF-CSign, which aims to achieve high accuracy in sign language recognition and cross-domain recognition. First, we use Radio Frequency Identification (RFID) to collect signals and obtain denoised signals through data pre-processing, so that they can be processed in a neural network. Second, the RF-CSign network is proposed with the inclusion of large kernel convolution to reduce the complexity of the model and to make the model with long-range correlations, thereby enhancing recognition accuracy. Third, RF-CSign employs a pixel Normalization-based Attention Module (NAM) to enhance the stability of the model, thereby addressing the problem of model overfitting. Finally, RF-CSign achieves high accuracy in cross-domain environments through a migration learning approach. The experimental results showed that the average recognition accuracy of RF-CSign reached 99.17%, and the average recognition accuracy for new users and new environments recorded 96.67% and 97.50%, respectively.

INDEX TERMS Sign language recognition, RFID, large kernel convolution, transfer learning.

I. INTRODUCTION

Communication is vital in human life; through communication, people acquire and exchange knowledge, interact, form contacts, and express emotions and needs. While spoken language is the standard method for societal communication, for individuals with hearing impairment, sign language (SL) emerges as their primary method of interaction [1]. According to the World Health Organization (WHO), 1.5 billion people are currently affected by hearing impairment worldwide, and this number is expected to rise to 2.5 billion by 2050 [2]. There are 27.8 million hearing impaired people in China, which is a significant population [3].

With the development of wireless sensing technology, human-machine interaction (HMI) [4], activity recognition [5], location tracking [6], and various other intelligent applications [7] have provided different solutions. This technology also provides new solutions to human-to-human communication barriers using wireless sensing technology. Sign language recognition based on wireless sensing technology [8] allows hearing impaired people to overcome various

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

communication barriers and enables them to benefit from technological advancements.

Approaches to sign language recognition can be classified as vision-based [9], [10], sensor-based [11], [12], [13], [14], [15], [16], [17], and a blend of both [18]. Vision-based gesture recognition techniques use a camera to capture the visual data of a gesture, and then process the visual data to complete the recognition. However, vision-based recognition techniques may require further pre-processing of the raw video stream for feature extraction and may deal with the high visibility and error sources that typically arise in computer vision systems. Meanwhile, sensor-based recognition techniques have the advantage of reducing the impact of gesture detection and segmentation to recognize gestures with lower processing power, and these sensors can track the user's movement and obtain information about the user's movement in space and time. However, most of these methods require users to wear sensor devices [12], [13], [14], contributing to the inconvenience of using them on a daily basis.

In contrast, gesture recognition techniques based on wireless-sensing, present a more extensive range of possible applications than those reliant on vision and wearable sensors. Wireless sensing-based methods realize gesture recognition without the need for users to wear a device, and do not consider environmental factors. Prior studies have employed various wireless sensing technologies for human activity recognition, such as Wi-Fi [15], [19], RFID [20], radar [21], [22], [23], and ultrasound [16], [17]. The Wi-Fi-based recognition technology has the advantages of low cost and easy expansion, but in the recognition of fine-grained actions, such as interactive gestures [24], the recognition effect is not good. Radar-based recognition systems [21], [23] have produced good results in sign language recognition; however, these methods often require specialized and expensive equipment, which may result in high equipment costs [25].

Unlike other sensing technologies, RFID has the advantages of no battery sensitivity, simple structure, flexible coverage, low hardware cost, easy deployment, and no-need-to-wear feature [26], which makes wireless sensor technology suitable for fine-grained activity identification in a wide range of application scenarios. However, the current RFID-based gesture recognition research has some limitations. For example, RF-Finger [27] described the influence of fingers on tag arrays at centimeter-level resolution by extracting fine-grained reflection features from raw RF signals. RF-Pen [28] constructed a four-antenna system, collected phase, and RSSI data, and then mixed the differences between the two data in spatial coordinates to determine spatial coordinates, thereby restoring the trajectory of user actions. Yu et al. [20] combined a convolutional neural network (CNN) and long- and short-term memory (LSTM) networks to achieve the accurate recognition of eight traffic command gestures. However, the gestures recognized by these systems are coarse-grained sign languages such as traffic command gestures or finger movement trajectories of letters written in a wide range of tag arrays of more than 10 cm. In the Chinese sign language, there are several sign languages whose gestures simultaneously involve using the hand, wrist, and arm. Among these, the features of hand and wrist movements are more finely tuned than those of the arm movements. Additionally, comparable arm gestures in various sign languages (such as ''give'' and ''leave'') can produce very similar signal shifts. On the other hand, certain sign languages (such as ''have'' and ''come'') rely on single finger or wrist movements to indicate meaning, leading to slight signal differences. Furthermore, in the past, there were frequently unsatisfactory results in the cross-domain of new users and new environments for gesture recognition. As a result, we must perform fine-grained sign language identification on sign languages with very comparable signal changes and small amplitudes, while maintaining high accuracy for new users and environments.

Overall, this paper discusses several key aspects. First, we explore ways to capture complex sign language movements in sign language. For a complex gestural movement like Chinese Sign Language, even a single-handed sign language involves finger, palm, wrist, and arm movements. Second, for the data collected by RFID to be processed by the CNN and at the same time to ensure high recognition accuracy, we have to design a data preprocessing algorithm to process the data and convert the data into a 2D image. Third, to achieve device diversity, we must design a lightweight model to accomplish sign language recognition and maintain the accuracy of sign language recognition. Fourth, we use less data to classify multiple sign language gestures on already trained networks for new users and new environments.

To address these highlighted problems, we put forth a cross-domain model for the recognition of Chinese sign language, utilizing deep learning and RFID in this field. First, based on the principle of signal sensing, we found that the 2D multi-tag array can improve the data collection capacity through extensive experiments. In the experiments, a 2×2 tag array was used to collect the movement data of fingers, palms, wrists, and arms from space. Second, owing to the presence of ambient and thermal noise during data collection, we performed phase unwrapping, filtering, interpolation, and normalization on the collected data to restore the changing characteristics of the RF signals, and then plotted the signals into a two-dimensional image. Third, we propose an RF-CSign network, which uses depth-wise convolutions (DW-Conv) and depth-wise dilated convolutions (DW-D-Conv) to reduce the complexity of the model, provide the model with a larger receptive field and enable the network to have long-range correlation, and a Normalization-based Attention Module (NAM) was added to each module to ensure the performance of the model, to solve the problem of performance degradation of lightweight networks. Fourth, to realize sign language recognition for new users and new environments on the already trained network, we use transfer learning to make the pre-trained network applicable to new users and new environments.

RF-CSign is based on the COTS RFID implementation. The acquired data exhibits that the average classification accuracy of RF-CSign for sign language identification was 99.17%. The average recognition accuracy for new users and new environments was 96.67% and 97.50%, respectively.

The remainder of this paper is organized as follows. In Section II, we provide an overview of the conclusions of previous studies on sign language recognition. In Section III, we present demonstrations of the RFID principle, the selection of sign language features, and the reasons behind the tag array's construction. In Section IV, the RF-CSign structure is described. In Section V, we assess the performance of RF-CSign and compare it with other approaches. Finally, in Section VI, we conclude the paper.

II. RELATED WORK

A. SIGN LANGUAGE RECOGNITION

Current techniques for sign language recognition primarily fall into two groups [1]: those relying on computer vision and those based on sensors. With the significant development of deep neural networks, computer vision-based sign language recognition methods have been widely used. Studies have used RGB cameras to capture sign language videos and then recognize sign language through bidirectional VTNs [29]. In addition, certain prior studies used Kinect [10], [30] and Leap Motion [31] to collect data on sign language, which produced better recognition results. For instance, Sun et al. [30] used a Kinect device to collect a large number of consecutive Chinese sign languages and achieved accurate sign language image matching using an Extenics Immune Neural Net. However, these systems are susceptible to environmental conditions as well as the risk of long video sequence data, excessive training resources, and disclosure of user privacy.

Meanwhile, wearable sensor-based sign language recognition utilizes the sensors worn on the user's hand to capture hand changes, which are represented as features. For example, Literature has also revealed the development of a wearable smart band with integrated nanocomposite pressure sensors to detect contraction/relaxation of arm muscles before data are provided to the machine learning algorithms to classify American Sign Language digital gestures through the selection of features, weights, and biases [12]. In addition, IMU sensors in smartwatches have been used to achieve hand and arm motion tracking within the context of SoM [32]. Although the wearable sensor approach can be implemented effectively for sign language recognition, it has significant limitations. First, wearable devices inevitably rely on fixed devices. Second, wearable devices are in contact with the human body, and collecting data for an extended period can cause discomfort. Sign language recognition based on wireless sensing has gradually gained popularity owing to its ubiquitous sensing signals and the low cost of wireless devices.

In addition, wireless technologies like Wi-Fi [15], [19], RFID [20], and radar [8], [11], [33], [34] have been used for sign language recognition. For instance, the channel state information of Wi-Fi was used to collect data for the recognition of American Sign Language within the contexts of SignFi [15] and WiSign [19]. Related literature has also revealed the use of UWB Radar to collect data on British Sign Language and the use of VGG16 model to recognize six sign languages that express emotions [11]. There was also the use of CW Radar to collect spectrograms of sign language signals, in which five sign languages were classified based on the KNN algorithm [33]. Despite the fact that these works have achieved excellent results in sign language recognition, they still suffer from poor cross-domain recognition effect and expensive cost, which prevents them from being used in practical situations. RFID-based wireless sensing systems are contactless, convenient, and low-cost compared with wireless sensing technologies. RF-CSign builds commercially available RFID devices, and the proposed system was developed to enable fine-grained sign language recognition without the need to wear any sensing device.

B. RFID-BASED SENSING

RFID, which is used in various fields, transmits signals from a signal transmitter to irradiate the target action and uses the phase and RSSI characteristics of the reflected signals to identify it because of this characteristic. Most previous studies have demonstrated the use of RFID in the fields of location tracking [6], [35], activity recognition [26], [27], [28], [36], health monitoring [37], privacy protection [5], and so on. For example, in the case of POLO, a mobile robot was operated to localize tagged items using NRP algorithms, with an array of tags carried by itself [6]. Meanwhile, in the case of SILoc [35], multiple antennas on top of the robot were utilized to gather phase data from fixed tags to calculate the robot's position in 2D and 3D spaces and to solve the localization problem of inconsistent robot speeds. In another study that involved Au-Id [26], an array of RFID tags was deployed to take advantage of spatial diversity, especially in recording diverse physical and behavioral features from human movements, with the goal of allowing the RFID system to conduct user identification via human movements. The purpose was to facilitate the RFID system to carry out user identification via human movements. Another study used a combination of RFID tags and a public cloud service to process the reflected data from RFID tags on the public cloud for real-time fall detection, specifically for older people [37]. Meanwhile, in the case of RFace [7], the face's 3D geometric and biomaterial features were extracted from the reflection characteristics of the RFID tag matrix to resist various spoofing attacks and realize authentication. GR-fid [38] compared the stability of the phase and RSSI, selected phase as a gesture feature, and accurately identified six gestures by Weighted Dynamic Time Wrapping. To accurately recognize

traffic command gestures, SGRS [39] first extracts fine-grained phase characteristics from RF signals and then matches the gestures using the k-means algorithm. Another study [40] employed eight antennas to accurately track the movement in a smart house. RFnet [41] recognized 26 letters of American Sign Language and dynamic gestures from 0-9 through a multi-branch 1D-CNN network. In the literature [42], edge machine learning (EML) achieved high-precision hand motion recognition. ReActor [43] used the random-forest approach to classify 18 different gestures successfully.

Existing device-less RFID solutions mainly focus on recognizing large-sized movements and simple gesture movements or localization tracking over a wide area, rather than recognizing sign language movements with several parts linked and fine-grained. Unlike the aforementioned studies, the current study used fewer tags, instead of complex tag arrays or antenna arrays, to obtain signal features of sign language in specific combinations to achieve high-precision sign language recognition. In addition, better recognition results were obtained under cross-domain conditions.

III. PRELIMINARIES

This section introduces the technical principles of RFID identification systems. Preliminary experiments were conducted to demonstrate the effectiveness of the signal acquisition model proposed in this study.

A. RFID PRINCIPLE

In radio frequency identification technology, the reader transmits radio frequency signals through the antenna. Tags activated by the signal sent by the reader serve as information transmitted back to the reader through the backscatter link. General commercial readers, such as Impinj Speedway r420, in addition to the ID information, will be obtained to the label, but also to the label that will be reflected in the signal indicators, namely the Received Signal Strength Indication (RSSI), phase, and Doppler shift. As the reader provides very noisy Doppler shift [52], both RSSI and phase metrics are typically used for perceptual identification.

The signal must travel between the reader and the tag through two transmissions in free space: a forward link from the reader to the tag and a backscatter connection from the tag to the reader. Phase offsets from the signal propagating in the air, the reader, and the tag itself make up the majority of the phase of the signal returned by the tag in free space:

$$\theta = \left(\frac{4\pi d}{\lambda} + \theta_R + \theta_T\right) mod 2\pi \tag{1}$$

where d represents the length of the propagation path; θ_R and θ_T, which stand for the phase offsets brought about by the reader and the tag, respectively, are constants for the same reader and tag.

The reflection route of the tag varies as the user moves between the antenna and the tag when the tag and antenna are deployed into the environment, causing variations in the RSSI and phase of the received signal. The signals received in this instance may be separated into static and dynamic pathways. The term ''static path signal'' refers to a signal reflected by a static item, whereas the term ''dynamic path signal'' refers to a signal reflected by a moving body. The received signal can be expressed as follows:

$$\begin{aligned} S = S_s + S_d &= A_s e^{-j\theta_s} + \sum_{k=1}^{M} A_k e^{-j\frac{2\pi d_k}{\lambda}} \\ &= A_s e^{-j\theta_s} + A_d e^{-j\theta_d} \end{aligned} \tag{2}$$

where d_k is the length of the kth dynamic path; M denotes the number of dynamic paths; S_s refers to static signal (can be regarded as a constant) or specifically the superposition of all static path signals; S_d refers to dynamic path signal that changes with time.

As dynamic path signal varies with time, the total signal S changes accordingly. In other words, when the user does an action, the RSSI and phase of the total signal S change accordingly, and these changes correspond to the corresponding action.

B. RFID PHASE

Recognition methods based on phase features have been reported to be significantly more resistant to multipath and noise than RSSI, so much of the wireless sensing recognition work revolves around the phase [38]. When the user performs an action, a phase offset is introduced in space owing to the reflective properties of the tag, in addition to the balances introduced by the reader and the tag itself. The total phase offset can be expressed as:

$$\theta = \frac{4\pi d}{\lambda} + \theta_R + \theta_{TAG} + \theta_T \tag{3}$$

where θ_{TAG} signifies the phase variation induced by the reflectivity properties of the tag.

This equation shows that the changes in θ_{TAG} changes the phase value θ of the tag. When a user performs an action in the space between the tag and the antenna, θ_{TAG} undergoes a transformation, consequently leading to a shift in the phase measurement θ. Therefore, the recognition of sign language actions can be accomplished by constructing a relationship model of sign language actions θ_{TAG}, and θ.

C. SIGN LANGUAGE FEATURE SELECTION

We used dynamic time warping (DTW) to compare the similarity of different time series of the same gesture from three users to choose the most efficient feature for human sign language recognition. The results are shown in Figures 1(a) and 1(b). We discovered that, for a given user, the average Euclidean distance between each gesture phase's time series is significantly smaller than the average Euclidean distance between the time series of the RSSI information. This finding suggests that the similarity of the phase information to the time series is much higher than that of the RSSI information in the sign language sensing data. Therefore, the phase information is more stable and robust than RSSI information.

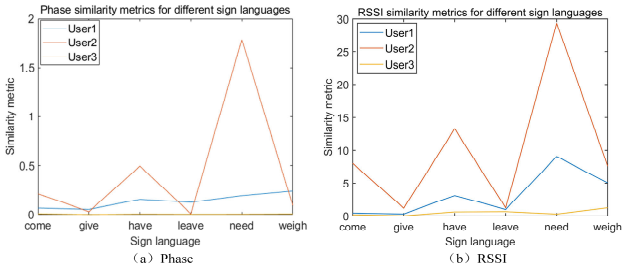


FIGURE 1. Similarity metric of phase and RSSI under three users.

D. PRELIMINARY EXPERIMENTS AND ANALYSIS

Prior to the sign language recognition experiments, deployed experiments were designed to address the first challenge of detecting complex sign language movements. Based on previous research, H47UHF tags were selected and formed into 2 × 2 tag arrays. The tag arrays can simultaneously capture the movements of fingers, palms, wrists, and arms of sign language movements and realize the complete collection of signal changes generated by sign language movements. However, 2 × 2 tag arrays can eliminate the coupling effect better, and fewer tags can reduce the potential deployment difficulty. The schematics for the H47UHF tag that is being utilized and the antenna-tag for our system are shown in Figure 2(a) and 2(b), respectively.

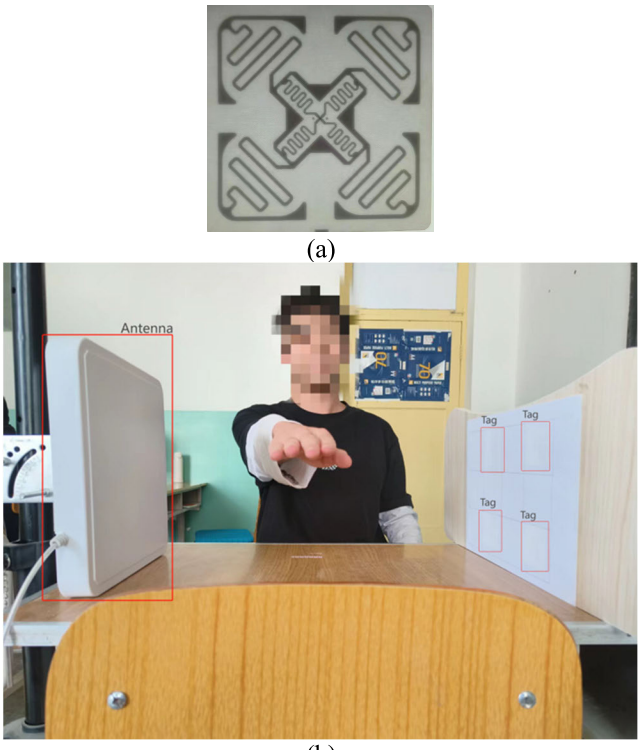


FIGURE 2. (a) H47UHF tag. (b) Experimental setup.

In this study's proposed system, the participating volunteers were required to only make movements of Chinese sign language according to the requirements of sign language movements. In this regard, antennas and tags were set up on either side of a table, and a 48cm gap was maintained between the antennas and tags during the RF-CSign experiments. Two preliminary experiments were conducted. The phases of the action under a label and the phase under a 2 × 2 label array were specifically organized. In particular, tags are deployed into the environment. Subsequently, a comparative analysis was conducted by collecting the phases of the sign language movements.

Figure 3 shows the phase collected by two different sign language actions under a label. The phase features managed by the two sign language actions of "give" and "leave" under a label appeared to be highly similar. As a result, the recognition of the two actions is likely to be confused. Both movements included the sub-action of holding the hand. Nonetheless, as hand movements were spatially different, a tag array was necessary to collect the sign language data in this study.

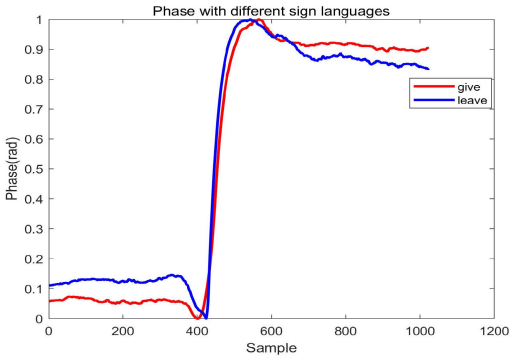


FIGURE 3. Phase of different sign language actions under a single label.

As the four tags were arranged into a tag array in a 2 × 2 array, the data of sign language movements on an empty surface, as well as the movements of each part (i.e., finger, palm, wrist, and arm) in terms of the reflective properties of the tags were collected. Figures 4(a) and 4(b) correspond to the phases of different tags for the sign language actions of "give" and "leave" under the 2 × 2 tag array, respectively. Based on the figures, the collected phases of each tag appear to be different. The most noticeable change occurred when the hand passed through tag 3. In addition, other sign languages also vary in phase. These results suggest that the phases of the 2 × 2 tag array exhibit physical properties that characterize the behavior of sign language actions, which holds potential for application in sign language identification.

IV. SYSTEM DESIGN

Figure 5 presents the proposed system, which consists of four main modules: a signal collection module, a data pre-processing module, a pre-training module, and a sign language recognition module. First, the signal acquisition module captures the original phase data using a 2 × 2 array of RFID tags. Subsequently, within the signal pre-processing module, the initial phase underwent a preliminary

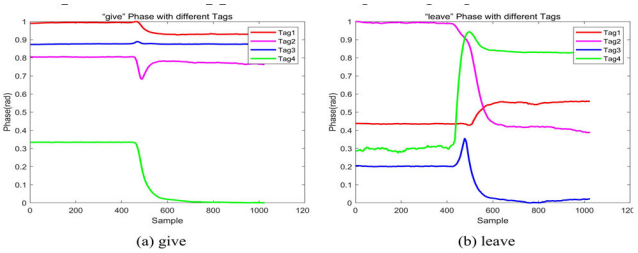


FIGURE 4. Phases of different sign language actions under different labels.

unwrapping operation concerning its phase. The manipulated data were subsequently subjected to a denoising process to mitigate noise-induced disruptions. In addition, the denoised data were interpolated to keep the data smoother and exhibit the same length. The smoothed data were converted into an image format for output after a normalization operation.

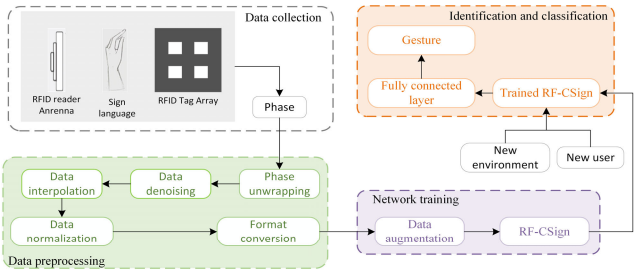


FIGURE 5. System design.

Next, the pre-training module imported the pre-processed data into a deep learning model after data enhancement, and the model was pre-trained. The model mainly consists of lightweight blocks based on large kernel convolution, residual blocks, and attention modules (NAM). The pre-processed data were sent to the model, where the signal features of the sign language were first extracted, and the pre-training weights of the model were continuously adjusted to obtain a satisfactory pre-trained model (RF-CSign). Finally, the RF-CSign is migrated to perform recognition and classification through the output layer. The generalization performance of the model was verified by recognizing and classifying the data from new users and new environments.

A. DATA PRE-PROCESSING MODULE

Raw signals should not be directly input into the pre-training, recognition, and classification modules for processing because of various noises and interferences in the data acquisition process. The pre-processing operation was conducted on the raw signals before the processed data were recognized and classified.

1) PHASE UNWRAPPING

The initial phase, which is directly captured by the RFID reader, is termed a recurring function or the enveloped phase. It jumps at each transmission cycle node, and this jump is known as phase mutation. As the phase value decreased to 0, it leaped to 2π. This phase mutation can critically affect the system's judgment of the change in the tag reflection characteristics. Therefore, it was deemed necessary for the current study to obtain the real phase value using the phase unwrapping algorithm. In particular, a phase de-entanglement algorithm, which comes with MATLAB, was used. The results are shown in Figure 6.

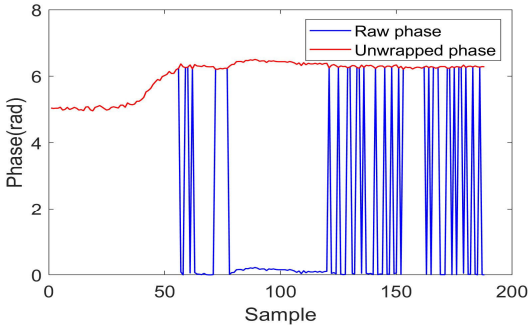


FIGURE 6. Comparison of raw phase and phase-after-phase unwrapping.

2) DATA DENOISING

The received signal contains many useless signals owing to the effect of the white noise. The original phase of the acquisition fluctuated randomly above and below the valuable signal. It is necessary to adopt targeted filtering methods based on this randomness to suppress useless signals and enhance valuable signals. Therefore, this study chose to denoise the original phase using a standard Kalman filter [49] (for white noise).

3) DATA INTERPOLATION AND NORMALIZATION

In the process of data collection, even if the same individual performs the same sign language, it is not guaranteed that the speed of the sign language movement is the same, resulting in different lengths of the data collected by the reader, long and short received signals, and data loss in the received data. Hence, to preserve the completeness of the dataset, it is essential to initially standardize the dimensions of the dataset and retrieve the missing data, enabling the data to achieve a form of consistent time-domain sampling. In this study, a third-order Hermite interpolation polynomial was used to interpolate the data to establish uniformly sampled data in the time domain. The third-order Hermite interpolating polynomial is expressed as:

$$H_3(x) = \left[\left(1+2\frac{x-x_0}{x_1-x_0}\right)y_0 + (x-x_0)\,y_0'\right]\left(\frac{x-x_1}{x_0-x_1}\right)^2 + \left[\left(1+2\frac{x-x_1}{x_0-x_1}\right)y_1 + (x-x_1)\,y_1'\right]\left(\frac{x-x_0}{x_1-x_0}\right)^2 \tag{4}$$

where x_0 and x_1 are the positions of the two neighboring points that need to be interpolated; y_0 and y_1 are the dependent

variables corresponding to the independent variables of x_0 and x_1; y'_0 and y'_1 are the derivatives at the corresponding positions; x_0, x_1, y_0, and y_1 are already known information; y'_0 and y'_1 are estimated values based on the known information.

Accordingly, different lengths of collected data for each tag were unified to the same 1,024 data lengths, according to the third-order Hermite interpolation method. The integrity of the data was preserved to the maximum extent possible.

To simultaneously enhance the comparability between the data and improve the accuracy of sign language recognition, it is necessary to carry out a normalization operation on the interpolated data. Therefore, we used min-max normalization to transform the interpolated data linearly. After data normalization, the phase data were converted into 1000×1000 pixel image formats for the output. Subsequently, the images were randomly classified into the training, validation, and test sets.

B. PRE-TRAINING MODULE

After pre-processing, all data were passed to the pre-training module. The neural network realized the feature extraction function, and the classifier in the output layer recognized the images in the verification set. The overall comparison accuracy is returned, whereas the comparison recognition accuracy updates the shared parameters. Finally, a pre-trained model (RF-CSign) was obtained. Figure 7 shows the network structure diagram.

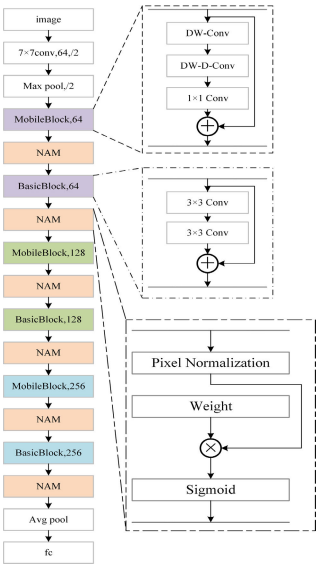


FIGURE 7. RF-CSign network structure.

1) MOBILEBLOCK MODULE

For common residual networks, using the same convolution kernel to operate on the input features can effectively utilize parameter sharing and reduce the number of parameters in the model. However, the same convolution kernel also brings a fixed receptive field, which leads to the inability to capture sufficient contextual information to build the relationship between different features. To build remote dependencies between different features, in CBAM [50], large kernel convolution is utilized to capture the spatial relationships of different channel features to generate the attention graph. Meanwhile, in the literature [51], researchers utilized large kernel convolution to construct large kernel attention, which realizes self-attention and captures the long-range relationship. However, if we directly add the big kernel attention to RF-CSign, it inevitably incurs extra computational overhead and parameters to our model.

To avoid increasing the complexity of the model and to take advantage of the large kernel attention self-attention and long-range relationship, we redesigned some of the residual modules of the residual network as MobileBlock, and reconstructed the residual modules using depth-wise convolution (DW-Conv) and depth-wise dilation convolution (DW-D-Conv). As shown in Figure 5, MobileBlock can be represented as:

$$F(x) = f^{1\times Á1}(DWDConv(DWConv(x))) \tag{5}$$

$$H(x) = F(x) + x \tag{6}$$

where x denotes the input, $DWDConv()$ denotes the 7×7 depth-wise convolution with dilation three operations, $DWConv()$ denotes the 5×5 depth-wise convolution operation, $f^{1\times Á1}$ denotes the 1×1 convolution operation, $F(x)$ denotes the feature map after the large kernel convolution process, $H(x)$ denotes the output feature map.

2) RESIDUAL BLOCK

In 2015, He et al. [44] proposed a deep convolutional neural network, specifically ResNet-18. ResNet-18 solves the problem of gradient vanishing in deep neural networks, and its distinguishing feature is its pioneering Residual Block structure. The basic idea of the Residual Block is to introduce cross-layer connections between the inputs and outputs, which allows the residuals to be updated directly during training without the need to consider updates in all layers. This approach eliminates the gradient weakening phenomenon owing to deep networks, which allows deeper networks to be trained, resulting in better performance. Figure 5 shows the residual basic-block. The representation of the residual module is given by Equation (4).

3) NORMALIZATION-BASED ATTENTION MODULE

We define a pixel normalization method for normalizing each pixel point of the input data such that the sum of squares of each pixel point in the channel direction is 1. Specifically, it computes the average of the squares of the input data in the channel dimension and then divides the original input data by the computed average of the squares to obtain the pixel normalization result. This process causes the sum of squares of each pixel point in the channel direction to be

1. This guarantees that the model's interaction with each pixel is equitable, preventing excessive reliance on certain distinct pixel points, thereby enhancing the model's ability to generalize and maintain stability. Our method can be written as:

$$v_{norm} = \frac{v_c}{\sqrt{mean\left(v_c^2\right) + \epsilon}} \tag{7}$$

where v_c denotes an element of the channel dimension, $mean()$ averages the result of squaring v_c over the channel dimension, and ϵ is a very small constant.

In the network we designed, we used large kernel convolution to improve the performance of the network, but this also made our model unstable. Therefore, we introduced a Normalization-based Attention Module (NAM) [45] to our model, which measures the importance of pixels through pixel normalization, and the scale factor quantifies the fluctuations in spatial pixels and signifies their significance, which improves the stability of the model and maintains its performance. As shown in Figure 5, NAM can be represented as:

$$W_\lambda = \frac{\lambda_i}{\sum_{j=0} \lambda_j} \tag{8}$$

$$M_S = \sigma\left(W_\lambda\left(PN\left(F\right)\right)\right) \tag{9}$$

where λ is the scaling factor, W_λ is the weights, and σ is the sigmoid function.

C. RECOGNITION AND CLASSIFICATION MODULE

After pre-training the model, a test set of training data was subjected to validation, and the average accuracy and overall accuracy for each class were output. Transfer learning is used for new users and for data in new environments. Transfer learning is a prevalent machine learning strategy that conveys the categorization capability of a model from an established environment to a dynamic one. In migration learning [46], a source domain typically encompasses a vast quantity of labeled data and knowledge, and the objective is to employ information from the source domain to annotate instances within the target domain. The source domain was linked to the initial arrangement in relation to sign language identification. In addition, the target domain represents the new user and the new environment. Therefore, cross-domain recognition was implemented in the pre-trained model with regard to the differences between users and environments. The migration process is illustrated in Figure 8.

V. RESULTS

This section presents the hardware and software used in the experiments. This section also discusses the verification results of the model's performance based on data collected from RFID devices and multi-tag arrays.

A. EXPERIMENTAL SETUP

This subsection describes the details of the experimental environment, hardware environment, software facilities,

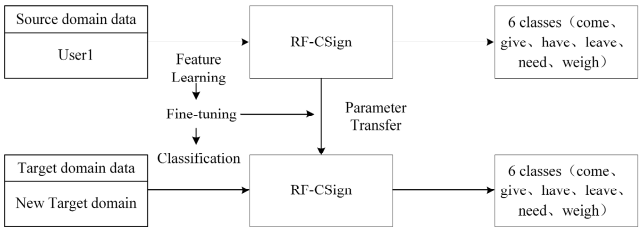


FIGURE 8. Transfer process.

datasets, and evaluation indicators. Overall, there were three experimental scenarios. As shown in Figure 9(a), Experimental Scenario 1 involved a neat arrangement of chairs, podiums, and other furniture within a classroom measuring 10 m in length and 7 m in width. Referring to Figure 9(b), Experimental Scenario 2 was a meeting room measuring 8.5 m in length and 7.2 m in width. The conference room has an elliptical conference table, sofas, chairs, and iron cabinets. Finally, as shown in Figure 9(c), Experimental Situation 3 was arranged within a student's living quarters with dimensions of 8.5 m in length and 4.5 m in width. The dormitory had four iron bunk beds, two iron closets, and several electronic devices. The tag array designed in this study was a 2 × 2 2D multi-tag array. The target was positioned within the array of the antenna and tag, with a gap of 48 cm between the antenna and tag. In the context of sign language recognition, the user was asked to be positioned at the front of the table and place their hands amidst the array of the antenna and tag to execute the sign language. The sensing data, which were obtained when the user performed the sign language action, were then transmitted to a PC via Ethernet for data processing and recognition.

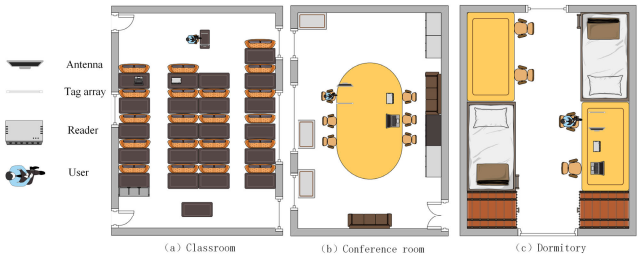
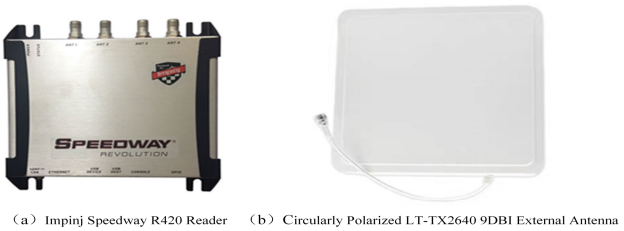


FIGURE 9. Experimental scene.

For the hardware environment, this study used a Lenovo laptop, Impinj Speedway R420 reader, and circularly polarized LT-TX2640 9DBI external antenna (operating frequency 923.875 MHZ). Figure 10 presents the Impinj Speedway R420 reader and the circularly polarized LT-TX2640 9DBI external antenna used in this study. Four 50 × 50 mm UHF passive electronic tags were affixed in the middle of a curved wooden board (52 × 24 cm). With the spacing of the tags at 6 cm and the antenna fixed on top of a tripod, both the antenna and the center of the tag array were ensured to be on the same line.



（a）Impinj Speedway R420 Reader （b）Circularly Polarized LT-TX2640 9DBI External Antenna

FIGURE 10. Experimental equipment.

Regarding the software setup, the research framework was utilized on a Lenovo machine equipped with a 3.20 GHz AMDR7 and 16 G RAM for data gathering and preliminary processing. The RFID reader was linked to a portable computer using an Ethernet cord. Concurrently, the Low-Level Reader Protocol (LLRP) served as the medium for interaction. The procedure was executed in C. MATLAB was used for the implementation of the data pre-processing algorithm, whereas Python was used for the implementation of the neural model.

Referring to Figure 11, this study acquired six experimental Chinese sign languages from the National General Sign Language Dictionary (edited by the National Sign Language and Braille Research Center of the Chinese Association of the hearing impaired). In particular, 5400 data samples were collected from five volunteers who were required to perform standard sign language movements to assess the validity of sign language recognition. All four male volunteers and one female volunteer were adults. Two male volunteers had normal body weights, whereas one male volunteer was thin. The remaining male volunteers had heavier body weights.

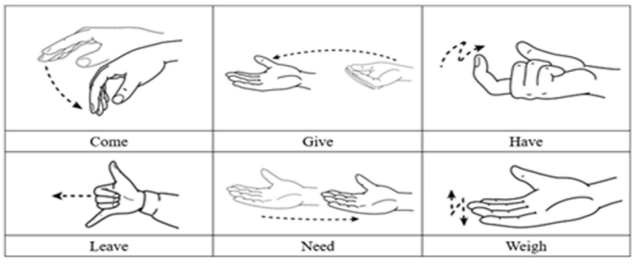


FIGURE 11. Recognized sign language.

This study utilized two assessment criteria, specifically accuracy and F1-score, to assess the effectiveness of the model. The outcome of the sign language classification considered four scenarios: (1) the count of instances that were genuinely positive and forecasted as positive (N_{TP}); (2) the count of instances that were genuinely negative and forecasted as negative (N_{TN}); (3) the count of instances that were genuinely negative yet forecasted as positive (N_{FP}); and (4) the count of instances that were genuinely positive yet forecasted as negative (N_{FN}).

In this study, accuracy was defined as a metric of the probability of accurately identifying a user's sign language movement:

$$A_{cc} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \tag{10}$$

Concurrently, the F1-score is presented as the balanced average of precision (P) and recall (R) within the model:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

B. ACCURACY OF SIGN LANGUAGE RECOGNITION

This section describes the accuracy of sign language recognition in this study. In Experimental Scenario 1, User 1 performed 200 experiments for each sign language. After the data refinement, 80% of the information was utilized as a learning set for model training, whereas the remaining 20% was employed as a testing set to ascertain the effectiveness of the model. Figure 12 depicts the confusion matrix for sign language identification, which illustrates the recognition accuracy for each sign language gesture.

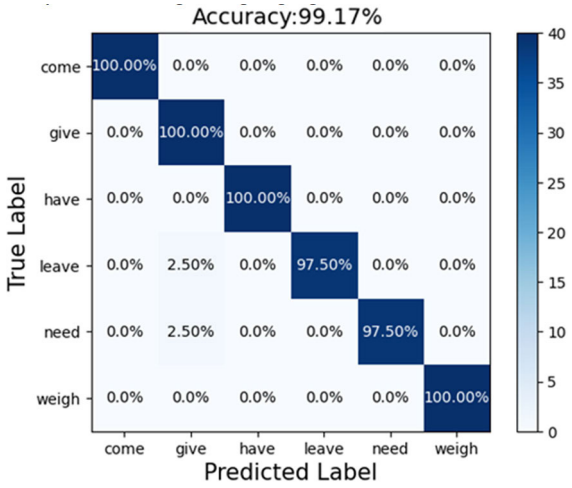


FIGURE 12. Confusion matrix for different sign language classifications.

RF-CSign recorded an average accuracy of 99.17% with a test set of 240 data samples. Among them, the sign language of "come," "give," "have," and "weigh" were all recognized accurately in the test set due to their distinctive features. The recognition accuracy of "leave" and "need" sign language was 97.50%. From Figure 9, we can see that the hand movements of the sign language of "give" and the sign language of "need" are almost the same, except that the hand movements of these two movements are in opposite directions, and one of them rises, and the other falls in the phase, but sometimes the hand positions are too high or too low in the tag array, resulting in recognition of the sign language of "need" as the sign language of "need." However, sometimes, because the position of the hand is too high or too low in the label array, the rise and fall in the phase are not obvious, leading to the recognition of the sign language of "need" as the sign language of "give." For both the sign language of "give" and the sign language of "need," it can be seen in Figure 11 that in some cases, due to the

TABLE 1. Ablation study of different modules in RF-CSIGN.

RF-CSign	Params. (M)	FLOPS(G)	Acc(%)
Resnet-14	2.78	1.41	97.92
w/o DW-Conv	1.70	5.45	98.33
w/o DW-D-Conv	1.79	1.04	96.67
w/o NAM	1.81	1.08	98.75
RF-CSign	1.81	1.08	99.17

fingers being too close together, this may be recognized as "give" rather than "leave."

C. ABLATION EXPERIMENT

Ablation experiment is designed to verify the performance of the proposed model. The backbone networks of the compared models were all RF-Csign networks.

We verified that each component of our design is necessary by deleting one component at a time in the RF-CSign network, and the experimental results are shown in Table 1. DW-Conv convolves each channel, which reduces the parameter sharing between channels and reduces the complexity of the model, thereby increasing the computational overhead. DW-D-Conv takes advantage of a larger receptive field to obtain a long-range relationship, which improves classification performance by 3.5%. The NAM improved the stability of the model, which improved the classification performance by 0.41%. Overall, our designed RF-CSign network improves the classification performance by 1.25% over the similarly structured ResNet-14 network, while reducing the computational overhead and number of parameters of the model.

D. COMPARISON OF DIFFERENT SOLUTIONS

This section compares the accuracy of the sign language recognition model of RF-CSign with five sign language recognition models: ResNet-18 [44], ResNet-34 [44], Tgt [47], TL [46], ResNet-6 [48], and VGG-16 [11]. These models were trained on the RFID sign language dataset and the accuracy of each validation set was recorded. The results are shown in Figure 13. The number of training sets was 768 and the number of validation sets was 192.

The obtained curve of the training result in Figure 13 reveals that the RF-CSign model started to fit gradually after 10 epochs of training. The highest accuracy of the validation set was 99.80%. Although VGG16 incorporates a substantial number of parameters, it was observed to be prone to overfitting because it predominantly focused on the fully connected layer, leading to continuous fluctuations in the accuracy of its validation set identification. ResNet-18 and ResNet-34 demonstrated their capacity to effectively solve the problem of disappearing or exploding gradients following the introduction of the residual structure. However, the lack of the attention module resulted in poorer performance than the RF-CSign model in terms of the convergence speed of the model training. Three models, namely Tgt, TL, and ResNet-6,

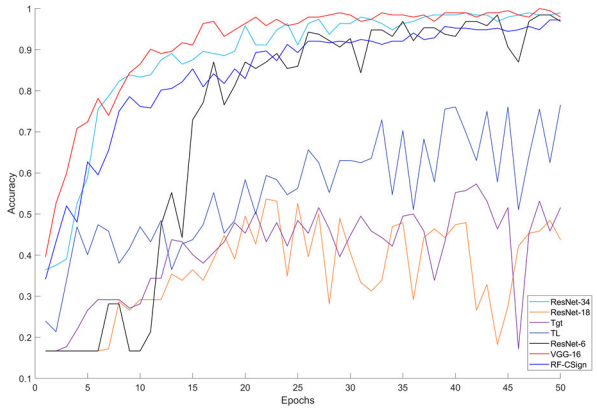


FIGURE 13. Training results for different models.

suffered from severe overfitting problems because the actual number of network layers was not very high. Furthermore, the maximum accuracy does not exceed 80%.

E. PERFORMANCE OF DIFFERENT USERS

This section focuses on the effect of new users using the same sign language on model recognition. Although the same sign language actions are performed, different users may exhibit different reflective signals in drawn sign language because of their different body sizes. The collected signals are not the same, resulting in varying effects on the model. To verify the accuracy of the RF-CSign model for new users' sign language recognition, four volunteers (three men and one woman) performed six sign language actions, each sign language 50 times, in Experimental Scenario 1. Figure 14 presents the F1-score and recognition accuracy of the RF-CSign model for the sign language recognition of new users.

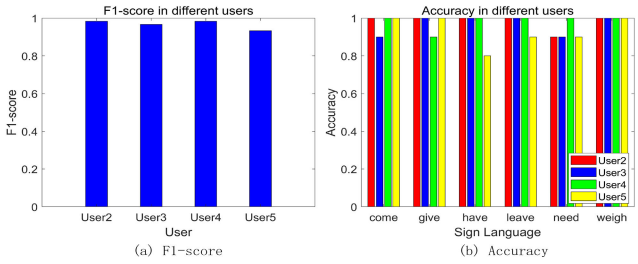


FIGURE 14. Sign language recognition for different users.

Regarding the F1-score of the RF-CSign system for new users' sign language recognition, User 2, User 3, User 4, and User 5 recorded 98.33%, 96.67%, 98.33%, and 93.33%, respectively. The recorded F1-score for User 5 was lower than 95% because the pre-training model was trained using the data of the male volunteers. There was an insufficient amount of data in the training set for female users, resulting in a lower recognition effect than for male users. In terms of recognizing each sign language for each user, the sign language of "weigh" had the best recognition accuracy, with an average recognition accuracy of 100%. The sign

language of "come," "give," "leave," and "need" recorded recognition rate of more than 90% for different users. The sign language of "have" recorded a lower recognition rate due to the close similarity in its movements with the sign language of "weigh." Nonetheless, the system still provides a recognition rate of more than 80% for this sign language.

F. PERFORMANCE OF DIFFERENT ENVIRONMENTS

Experimental Scenarios 2 and 3 were executed to confirm the effectiveness of the model in unfamiliar settings. For each new environment, the user performed the sign language 300 times (i.e., 50 times for each action). Figure 17 presents the F1-score and recognition accuracy of the RF-CSign model for sign language recognition in new environments.

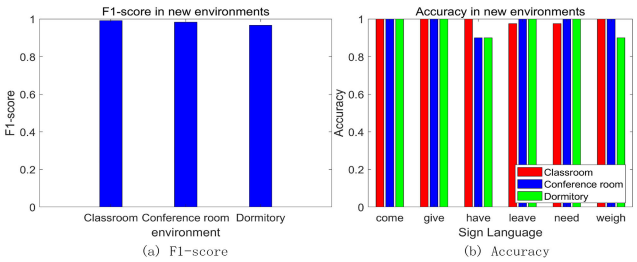


FIGURE 15. Sign language recognition for different environments.

The F1-score of the model for sign language recognition in the conference room and dormitory were 98.33% and 96.67%, respectively, which may be attributed to the presence of many cabinets and other clutters in the conference room and dormitory. These may cause environmental noise in the signal, resulting in a slightly lower recognition effect compared to the outcomes of Experimental Scenario 1. The recognition accuracy for the sign language of "have" was poorer due to the close similarity between the action of sign language of "have" and the action of sign language of "weigh." Nonetheless, the recognition accuracy for each sign language action exceeded 90%.

G. COMPARISON WITH OTHER GESTURE RECOGNITION ALGORITHMS

The complicated Chinese sign language motions are sensed by RF-CSign using a single antenna and tag array. The phase data is used to construct the phase image, which transforms the signal recognition problem into an image recognition problem. The RF-CSign network achieves high-precision sign language recognition by categorizing various Chinese sign language phase images. As illustrated in Figure 16, we compare RF-CSign to other gesture recognition systems using wireless sensing (SignFi [15], WIHF [24], and RFnet [41]). SignFi recognizes ASL using Wi-Fi signals, with average recognition rates of 98.01% and 90.74% in in-domain and cross-domain scenes, respectively. WIHF deduces gesture movements using Wi-Fi signals, and the average recognition rates in the in-domain and cross-domain scenes were 97.65% and 96.74%, respectively. RFnet uses a 7×7 tag array for gesture detection, and the average recognition rates of dynamic gestures in the in-domain and cross-domain scenes are 94.80% and 94.75%, respectively. In contrast, RF-CSign performs better, with average accuracies of 99.17% and 97.09% in the in-domain and cross-domain scenes, respectively.

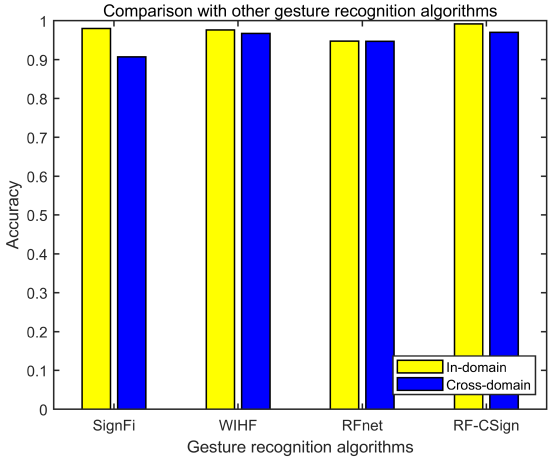


FIGURE 16. Comparison with other gesture recognition algorithms.

H. IMPACT OF TAG-ANTENNA DEPLOYMENTS

We evaluated the RF-CSign from the perspective of different antenna deployments. We specifically changed the 42 cm to 60 cm distance between the tag and the antenna. The results are shown in Figure 17. The recognition accuracy of RF-CSign was more than 95% in all cases, and it was the highest when the distance between the tag and antenna was 48 cm. This is because the body width of most Chinese people is within 48 cm, and the closer the distance between the tag and the antenna, the stronger the power of the collected sign language signal, and the better the performance of the method. The reason why a distance of 42 cm between the tag and the antenna is not as optimal as a distance of 48 cm lies in the fact that, at a distance of 42 cm, users may experience difficulties in performing sign language effectively, resulting in an inadequate display of various actions. The results show that the recognition effect is best when the distance between the tag and the antenna is 48 cm, considering the human body condition and perception principle.

I. PERFORMANCE AT DIFFERENT OPERATIONAL DISTANCES

The performance of RF-CSign was evaluated at three operating distances: 12 cm, 24 cm, and 36 cm. The distance is the distance between the user's hand and the tag. The result is shown in Figure 18. Different sign languages can be accurately recognized by RF-CSign, with a recognition rate higher than 93%. The results suggest that the recognition effect is optimal when the distance is 24 cm. The closer or farther the hand is from the tag, the lower the recognition accuracy. This is due to the fact that using sign language too

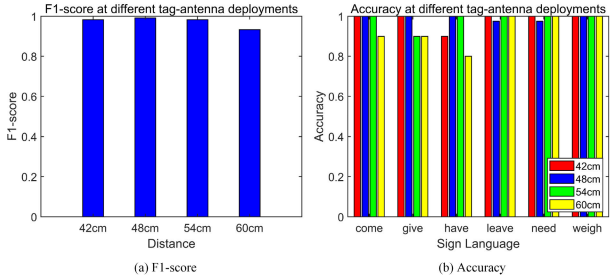


FIGURE 17. Sign language recognition for different tag-antenna deployments.

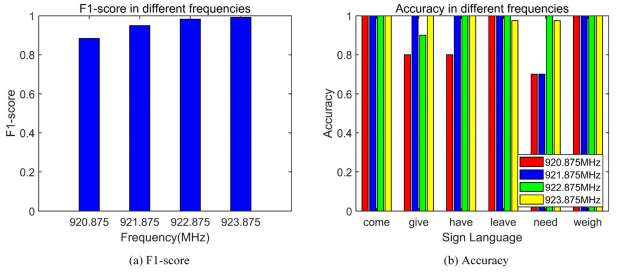


FIGURE 19. Sign language recognition for different frequencies.

closely to the antenna would block a portion of the signal and weaken the phase change brought on by the tag reflection. If you are too far away, the tag's phase change is not readily apparent. Therefore, a moderate distance can lead to better system performance.

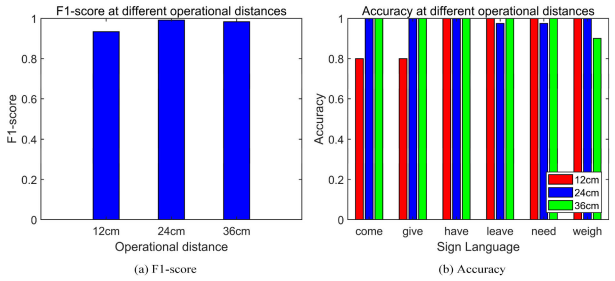


FIGURE 18. Sign language recognition for different operational distances.

J. PERFORMANCE OF DIFFERENT FREQUENCIES

We evaluated performance at different reader acquisition frequencies. We studied the recognition effect of the phase collected at four frequencies (including 920.875 MHz, 921.875 MHz, 922.875 MHz, and 923.875 MHz) in the frequency range of "China 920-925 MHz." The results are shown in Figure 19. In general, the higher the frequency of phase acquisition, the more information that can be read per unit time. We collected 50 out of the six sign languages at each acquisition frequency. It can be seen that the F1-score of RF-CSign under four acquisition frequencies are 96.67%, 98.31%, 98.33%, and 99.17%, respectively, and the F1-score under the acquisition frequency of 923.875 MHz is the highest. Because our sign language gestures are fine-grained gestures, the higher frequency of phase acquisition results in more information being collected, and thus, a better recognition effect.

VI. DISCUSSION

Although all experiments in this study produced relatively good results, the proposed system encountered several limitations. First, a relatively large amount of data was required to train the model in order to meet its generalization performance. Therefore, the reliability of the recognition accuracy reported in this study for the system may not be adequate since this study only gathered small samples. With that, a more reliable neural network will be designed, and the feature distance between the sample data and the sample data will be calculated by learning from small samples to achieve accurate user sign language recognition in small samples.

Second, recognizing sentence-level sign language is a new research direction, which is a new challenge due to the existence of more complex and fine-grained individual sign language words to recognize sentence-level sign language. Considering that, the combination of data segmentation and target detection techniques will be considered to split the whole sentence-level sign language into different parts. Following that, the split content will be recognized before splicing in order to achieve the recognition of sentence-level sign language. Additionally, recurrent neural networks will be used to analyze other sentences to complete the translation of the sentence-level sign language of the sensed data for sentence-level sign language recognition.

VII. CONCLUSION

This study primarily aimed to propose an RFID-based Chinese sign language recognition model and construct an RFID-based Chinese sign language dataset based on sign language collection sensing data from the National Sign Language Dictionary. The RF-CSign model was proposed to address the shortcomings of the sensing data and sign recognition model. In the RFID-based Chinese sign language recognition task, the RF-CSign model recorded a higher accuracy than the conventional model. Meanwhile, in the cross-domain sign language recognition task, the average F1-score for new users and new environments were 96.67% and 97.50%, respectively. The accurate classification of sign language sensing data using the RF-CSign model is of great significance for the popularization and application of wireless sensor-based interactions for hearing impaired people.

This study contributed three significant implications. First, this study served as the first to use Chinese Sign Language (CSL) as a gesture action for RFID-based gesture classification. Second, we propose an RF-CSign network that reduces the computational overhead and number of parameters of the model using depth-wise convolution (DW-Conv), which uses depth-wise dilation convolution to obtain remote dependencies between different features. Third, we propose a pixel normalization method applied in the

Normalization-based Attention Module (NAM) to normalize the pixels in the channel and ensure the performance of the model. Fourth, this study explored the effect of migration learning on new users and environments and verified the robustness of the model.

REFERENCES

[1] E.-S.-M. El-Alfy and H. Luqman, "A comprehensive survey and taxonomy of sign language research," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105198.

[2] C. Shelly and C. Alarcos, "Introduction," in *World Report on Hearing*. Geneva, Switzerland: WHO, 2021, pp. 5–6.

[3] M. Y. Hu, "In China, 27.8 million people have their lives muted," 2020. [Online]. Available: https://mp.weixin.qq.com/s/cnEjHFYUEW6jYDdWO3s6kw

[4] H. Kong, L. Lu, J. Yu, Y. Chen, and F. Tang, "Continuous authentication through finger gesture interaction for smart homes using WiFi," *IEEE Trans. Mobile Comput.*, vol. 20, no. 11, pp. 3148–3162, Nov. 2021.

[5] H. Fei, F. Xiao, J. Han, H. Huang, and L. Sun, "Multi-variations activity based gaits recognition using commodity WiFi," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2263–2273, Feb. 2020.

[6] D. Xie, X. Wang, A. Tang, and H. Zhu, "POLO: Localizing RFID-tagged objects for mobile robots," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Vancouver, BC, Canada, May 2021, pp. 1–10.

[7] W. Xu, J. Liu, S. Zhang, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "RFace: Anti-spoofing facial authentication using COTS RFID," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Vancouver, BC, Canada, May 2021, pp. 1–10.

[8] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using RF sensing," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3763–3775, Feb. 2021.

[9] J. Hu, Y. Liu, K.-M. Lam, and P. Lou, "STFE-Net: A spatial–temporal feature extraction network for continuous sign language translation," *IEEE Access*, vol. 11, pp. 46204–46217, 2023.

[10] C. O. Sosa-Jiménez, H. V. Ríos-Figueroa, and A. L. Solís-González-Cosío, "A prototype for Mexican sign language recognition and synthesis in support of a primary care physician," *IEEE Access*, vol. 10, pp. 127620–127635, 2022.

[11] H. Hameed et al., "Privacy-preserving British sign language recognition using deep learning," in *Proc. EMBC*, Glasgow, U.K., 2022, pp. 4316–4319.

[12] R. Ramalingame, R. Barioul, X. Li, G. Sanseverino, D. Krumm, S. Odenwald, and O. Kanoun, "Wearable smart band for American sign language recognition with polymer carbon nanocomposite-based pressure sensors," *IEEE Sensors Lett.*, vol. 5, no. 6, pp. 1–4, Jun. 2021.

[13] J. Wu, L. Sun, and R. Jafari, "A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 5, pp. 1281–1290, Sep. 2016.

[14] B. G. Lee and S. M. Lee, "Smart wearable hand device for sign language interpretation system with sensors fusion," *IEEE Sensors J.*, vol. 18, no. 3, pp. 1224–1232, Feb. 2018.

[15] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign language recognition using Wi-Fi?" in *Proc. IMWUT*, New York, NY, USA, 2018, pp. 1–21.

[16] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. MoBiCom*, New York, NY, USA, 2016, pp. 82–94.

[17] Y. Iravantchi, M. Goel, and C. Harrison, "BeamBand: Hand gesture sensing with ultrasonic beamforming," in *Proc. CHI*, New York, NY, USA, 2019, pp. 1–10.

[18] K. Guo, H. Zhou, Y. Tian, W. Zhou, Y. Ji, and X.-Y. Li, "Mudra: A multi-modal smartwatch interactive system with hand gesture recognition and user identification," in *Proc. INFOCOM*, London, U.K., 2022, pp. 100–109.

[19] L. Zhang, Y. Zhang, and X. Zheng, "WiSign: Ubiquitous American sign language recognition using commercial Wi-Fi devices," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–24, Jun. 2020.

[20] Y. Yu, D. Wang, R. Zhao, and Q. Zhang, "RFID based real-time recognition of ongoing gesture with adversarial learning," in *Proc. 17th Conf. Embedded Netw. Sensor Syst.*, New York, NY, USA, Nov. 2019, pp. 298–310.

[21] H. Wu and J. Ma, "A scalable gesture interaction system based on mm-wave radar," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Atlanta, GA, USA, Jun. 2021, pp. 466–469.

[22] W. Jiang, Y. Ren, Y. Liu, Z. Wang, and X. Wang, "Recognition of dynamic hand gesture based on mm-wave FMCW radar micro-Doppler signatures," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 4905–4909.

[23] G. Park, V. K. Chandrasegar, and J. Koh, "Accuracy enhancement of hand gesture recognition using CNN," *IEEE Access*, vol. 11, pp. 26496–26501, 2023.

[24] C. Li, M. Liu, and Z. Cao, "WiHF: Gesture and user recognition with WiFi," *IEEE Trans. Mobile Comput.*, vol. 21, no. 2, pp. 757–768, Feb. 2022.

[25] Y. Chen, J. Yu, L. Kong, Y. Zhu, and F. Tang, "RFPass: Towards environment-independent gait-based user authentication leveraging RFID," in *Proc. 19th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Stockholm, Sweden, Sep. 2022, pp. 289–297.

[26] A. Huang, D. Wang, R. Zhao, and Q. Zhang, "Au-Id: Automatic user identification and authentication through the motions captured from sequential human activities using RFID," in *Proc. IMWUT*, New York, NY, USA, 2019, pp. 1–26.

[27] C. Wang, J. Liu, Y. Chen, H. Liu, L. Xie, W. Wang, B. He, and S. Lu, "Multi–touch in the air: Device-free finger tracking and gesture recognition via COTS RFID," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Honolulu, HI, USA, Apr. 2018, pp. 1691–1699.

[28] H. Wang and W. Gong, "RF-pen: Practical real-time RFID tracking in the air," *IEEE Trans. Mobile Comput.*, vol. 20, no. 11, pp. 3227–3238, Nov. 2021.

[29] W. Qin, X. Mei, Y. Chen, Q. Zhang, Y. Yao, and S. Hu, "Sign language recognition and translation method based on VTN," in *Proc. Int. Conf. Digit. Soc. Intell. Syst. (DSInS)*, Chengdu, China, Dec. 2021, pp. 111–115.

[30] Y. Sun, T. Yuan, J. Chen, and R. Feng, "Chinese sign language key action recognition based on extenics immune neural network," in *Proc. IEEE Int. Conf. Adv. Electr. Eng. Comput. Appl. (AEECA)*, Dalian, China, Aug. 2020, pp. 187–191.

[31] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A modified LSTM model for continuous sign language recognition using leap motion," *IEEE Sensors J.*, vol. 19, no. 16, pp. 7056–7063, Aug. 2019.

[32] T. Zheng, C. Cai, Z. Chen, and J. Luo, "Sound of motion: Real-time wrist tracking with a smart watch-phone pair," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, London, U.K., May 2022, pp. 110–119.

[33] Y. Lu and Y. Lang, "Sign language recognition with CW radar and machine learning," in *Proc. 21st Int. Radar Symp. (IRS)*, Warsaw, Poland, Oct. 2020, pp. 31–34.

[34] H. Kulhandjian, P. Sharma, M. Kulhandjian, and C. D'Amours, "Sign language gesture recognition using Doppler radar and deep learning," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[35] J. Zhang, X. Liu, T. Gu, X. Tong, S. Chen, and K. Li, "SILoc: A speed inconsistency-immune approach to mobile RFID robot localization," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Vancouver, BC, Canada, May 2021, pp. 1–10.

[36] Y. Bu, L. Xie, Y. Gong, C. Wang, L. Yang, J. Liu, and S. Lu, "RF-dial: Rigid motion tracking and touch gesture detection for interaction via RFID tags," *IEEE Trans. Mobile Comput.*, vol. 21, no. 3, pp. 1061–1080, Mar. 2022.

[37] K. Takatou and N. Shinomiya, "IoT-based real-time monitoring system for fall detection of the elderly with passive RFID sensor tags," in *Proc. ITC-CSCC*, Nagoya, Japan, 2020, pp. 193–196.

[38] Y. Zou, J. Xiao, J. Han, K. Wu, Y. Li, and L. M. Ni, "GRfid: A device-free RFID-based gesture recognition system," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 381–393, Feb. 2017.

[39] B. Chen, Q. Zhang, R. Zhao, D. Li, and D. Wang, "SGRS: A sequential gesture recognition system using COTS RFID," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Barcelona, Spain, Apr. 2018, pp. 1–6.

[40] K. Bouchard, S. Giroux, B. Bouchard, and A. Bouzouane, "Regression analysis for gesture recognition using passive RFID technology in smart home environments," *Int. J. Smart Home*, vol. 8, no. 5, pp. 245–260, Sep. 2014.

[41] H. Ding, L. Guo, C. Zhao, F. Wang, G. Wang, Z. Jiang, W. Xi, and J. Zhao, "RFnet: Automatic gesture recognition and human identification using time series RFID signals," *Mobile Netw. Appl.*, vol. 25, no. 6, pp. 2240–2253, Dec. 2020.
[42] M. Merenda, G. Cimino, R. Carotenuto, F. G. D. Corte, and D. Iero, "Edge machine learning techniques applied to RFID for device-free hand gesture recognition," *IEEE J. Radio Freq. Identificat.*, vol. 6, pp. 564–572, 2022.
[43] S. Zhang, Z. Ma, C. Yang, X. Kui, X. Liu, W. Wang, J. Wang, and S. Guo, "Real-time and accurate gesture recognition with commercial RFID devices," *IEEE Trans. Mobile Comput.*, vol. 22, no. 12, pp. 7327–7342, Dec. 2023.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[45] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, "NAM: Normalization-based attention module," in *Proc. NIPS*, Sydney, NSW, Australia, 2021, pp. 1–5.
[46] S. A. Rokni, M. Nourollahi, and H. Ghasemzadeh, "Personalized human activity recognition using convolutional neural networks," in *Proc. AAAI*, New Orleans, LA, USA, 2018, pp. 8143–8144.
[47] T. Gong, Y. Kim, J. Shin, and S. Lee, "MetaSense: Few-shot adaptation to untrained conditions in deep mobile sensing," in *Proc. SenSys*, New York, NY, USA, 2019, pp. 110–123.
[48] Z. Ma, S. Zhang, J. Liu, X. Liu, W. Wang, J. Wang, and S. Guo, "RF-Siamese: Approaching accurate RFID gesture recognition with one sample," *IEEE Trans. Mobile Comput.*, early access, Oct. 27, 2022, doi: 10.1109/TMC.2022.3217487.
[49] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
[50] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Munich, German, 2018, pp. 3–19.
[51] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Jul. 2023.
[52] H. Ding, L. Shangguan, Z. Yang, J. Han, Z. Zhou, P. Yang, W. Xi, and J. Zhao, "FEMO: A platform for free-weight exercise monitoring with RFIDs," in *Proc. Sensys*, New York, NY, USA, Nov. 2015, pp. 141–154.

HUANYUAN XU was born in Zhanjiang, Guangdong, China, in 1997. He received the B.S. degree in optoelectronic information science and engineering from Qingdao University, China, in 2020. He is currently pursuing the master's degree in software engineering with Xinjiang University, China. His research interests include computer vision, pattern recognition, and signal processing.

YAJUN ZHANG was born in Nanyang, Henan, China, in 1983. He received the B.S. and master's degrees from Xinjiang University, China, in 2007 and 2010, respectively. He is currently an Associate Professor with Xinjiang University. His research interests include indoor localization technology, pattern recognition, and signal processing.

ZHIXIONG YANG was born in Ningde, Fujian, China, in 1998. He received the bachelor's degree in software engineering from the Changchun University of Science and Technology, China, in 2019. He is currently pursuing the Ph.D. degree in computer science and technology with Xinjiang University, China. His research interests include indoor positioning techniques, pattern recognition, and signal processing.

HAOQIANG YAN was born in Karamay, Xinjiang, China, in 1997. He received the B.S. degree in software engineering from Southwest Petroleum University, China, in 2019. He is currently pursuing the master's degree in software engineering with Xinjiang University, Xinjiang. His research interests include deep learning, pattern recognition, and signal processing.

XINGQIANG WANG was born in Tai'an, Shandong, China, in 1999. He received the B.S. degree in computer science and technology from Anhui University, China, in 2021. He is currently pursuing the master's degree in software engineering with Xinjiang University, China. His research interests include pattern recognition and signal processing.