Hindawi
Mobile Information Systems
Volume 2023, Article ID 7959916, 11 pages
https://doi.org/10.1155/2023/7959916

*Research Article*

# A Multitask Sign Language Recognition System Using Commodity Wi-Fi

**Zhongjian Gao,[1,2,3] Chien-Cheng Lee,[2] Lianhui Zheng,[1] Ruige Zhang,[1,3] and Xiaofu Xu[1]**

[1]*School of Mechanical and Electrical Engineering, Sanming University, Sanming, Fujian 365004, China*
[2]*Department of Electrical Engineering, Yuan Ze University, Taoyuan 320, Taiwan*
[3]*Key Laboratory of Equipment Intelligent Control of Fujian Higher Education Institute, Sanming, Fujian 365004, China*

Correspondence should be addressed to Chien-Cheng Lee; cclee@saturn.yzu.edu.tw

Received 6 April 2022; Revised 20 July 2022; Accepted 10 August 2022; Published 10 February 2023

Academic Editor: Liping Zhang



Wi-Fi sensing for gesture recognition systems is a fascinating and challenging research topic. We propose a multitask sign language recognition framework called Wi-SignFi, which accounts for gestures in the real world associated with various objects, actions, or scenes. The proposed framework comprises a convolutional neural network (CNN) and *K*-nearest neighbor (KNN) module. It is evaluated on the public SignFi dataset and achieves 98.91%, 86.67%, and 99.99% average gesture recognition accuracies on 276/150 activities, five users, and two environments, respectively. The experimental results show that the proposed gesture recognition method outperforms previous methods. Instead of converting the channel state information (CSI) data of multiple antennas into three-dimensional matrices (i.e., color images) as in the existing literature, we found that the CSI data can be converted into matrices (i.e., grayscale images) by concatenating different channels, allowing the Wi-SignFi model to balance between speed and accuracy. This finding facilitates deploying Wi-SignFi on Nvidia's Jetson Nano edge embedded devices. We expect this work to promote the integration of Wi-Fi sensing and the Internet of Things (IoT) and improve the quality of life of the deaf community.

## 1. Introduction

Sign language is an indispensable special language in the deaf community's daily life [1, 2]. Communication barriers often occur between deaf and normal people who are not familiar with the sign language [3, 4]. Wearable-based sign language interpreters aim to solve the abovementioned communication difficulties, but they are often expensive, low in versatility, and inconvenient to carry [5–8]. Vision-based sign language translators can overcome the shortcomings of portable sign language translators since it only needs a camera for natural interactions [1, 2, 9, 10]. However, users must place themselves within the field of view (FOV) of the camera, which may cause personal privacy information to be disclosed. At some point, their sign language gesture recognition system was susceptible to lighting conditions and obstacles.

Compared with wearable-based and video-based sensing solutions, wireless sensing can cover a wider detection range with fewer privacy concerns [11]. Due to the low cost and ease of deployment of reusable wireless communication infrastructure, Wi-Fi based wireless sensing solutions are rapidly developing [11, 12]. Currently, Wi-Fi sensing solutions mainly adopt two indicators: received signal strength indicator (RSSI) and channel state information (CSI) [13, 14]. Compared with RSSI, CSI can measure more fine-grained information and is suitable for capturing smaller movements such as heartbeats and gestures [3, 11, 12]. In 2018, Ma et al. [3] released a CSI-based sign language gesture dataset called SignFi, which collected 276 sign language words in the daily life of deaf people through Wi-Fi signals. They used a nine-layer convolutional neural network (CNN) model to recognize these gestures. However, gestures in the real world usually may correspond to different objects,

actions, and scenes. In our work, the main contributions of the proposed work are summarized as follows:

(1) We propose a multitask framework called Wi-SignFi that can not only recognize gestures but also identify users and environments. The Wi-SignFi model is a lightweight and end-to-end architecture consisting of an eight-layer CNN and a KNN module. Unlike existing references, the CSI data fed to Wi-SignFi does not require preprocessing such as denoising and unwrapping. Experimental results show that our proposed method achieves an average gesture recognition accuracy of 98.91%, which significantly outperforms previous works. Therefore, our proposed method is simple and effective.

(2) The experiments demonstrate that the accuracy of the model for recognizing gestures is affected by the resolution of the input data. In previous reports, the CSI data collected from three antennas on the Wi-Fi transmitter is normally converted into RGB CSI color images, which are then fed into CNNs. This approach does not increase the resolution of the input data, resulting in poor gesture recognition performance of the model. Conversely, the training time of the model grows proportionally to the resolution of the input data. In this regard, the model's time-consuming and gesture recognition accuracy can be balanced by extending the CSI data from different antennas into single-channel grayscale images as the input data. This finding facilitates the deployment of Wi-SignFi on edge embedded devices.

(3) Wi-SignFi can be deployed on a Nvidia's Jetson Nano device with 4 G memory. To the best of our knowledge, this is the first WiFi-based gesture recognition system to be applied to embedded devices. Wi-SignFi on the Jetson Nano device achieves an inference speed of 27 CSI instances per second.

The rest of this article is organized as follows: We review the existing literature in Section 2. Section 3 introduces the SignFi dataset and proposes the Wi-SignFi framework. Experiments and results are explained in Section 4. Wi-SignFi running on Jetson Nano devices is illustrated in Section 5. Section 6 concludes this work and gives future directions.

## 2. Related Works

At present, device-free sign language recognition systems are mainly divided into two categories: computer vision-based methods and wireless technology-based methods. Nath and Arun utilized the convex hull algorithm and template matching algorithm in the OpenCV software package for sign language recognition [1]. They implemented a real-time sign language translation on the ARM processor board. All sign language recognition systems in [2, 9, 10] were implemented by Microsoft's Kinect device. Aly et al. [2] combined a principal component analysis network (PCA-Net) and support vector machine (SVM) to recognize sign language gestures of different users. Huang et al. [9] found that it is difficult to obtain reliable features for hand-crafted features to adapt to various sign language gestures, so they proposed a 3D convolutional neural network (CNN) to automatically extract significant spatiotemporal features. A Brazilian sign language dataset named LIBRAS-UFOP was recognized by a two-stream convolutional network with a recognition accuracy of 74.25% [10]. In addition, Pu et al. [15] proposed a weakly supervised continuous sign language recognition system consisting of two modules: a 3D convolutional residual network (3D-ResNet) and an encoder-decoder sequence network. The system was verified on two large datasets RWTH-PHOENIX-Weather and CSL [15]. Cui et al. [16] utilized CNN and bi-directional recurrent neural network (RNN) to extract spatiotemporal information from raw sign language video datasets.

With the rise of the Internet of Things (IoT) and autonomous driving technology, there is growing interest in wireless sensing technology [17]. Wireless sensing solutions based on Wi-Fi have been extensively investigated due to their low cost and ease of deployment [3, 11–13, 18].

Wi-Fi sensing solutions have two indicators: RSSI and CSI [13, 14]. Since RSSI is easily accessible, many researchers have extracted human motion features from RSSI in the early days of Wi-Fi wireless sensing. Sigg et al. [19] analyzed the static and dynamic properties of the collected RSSI to recognize human gestures. Abdelnasser et al. [20] proposed an RSSI-based gesture recognition system WiGest. The system focuses on changes in Wi-Fi signal strength to recognize user's air gestures. In [21], a one-dimensional convolutional neural network (1D-CNN) general framework for RSSI dynamic gesture detection and recognition was built. Experimental results showed that the recognition accuracy of the seven complex dynamic gestures was 93.03%.

RSSI is the result of the superposition of multipath signals, which cannot effectively distinguish the multipath signals in the process of Wi-Fi signal propagation [22]. Hence, RSSI-based applications need to deploy multiple wireless links to reduce the impact of multipath effects [21]. For complex environments, the stability and reliability of RSSI fluctuate greatly, and it is impossible to capture real signal changes caused by human movements [13]. Nevertheless, CSI can distinguish multipath signals through the orthogonal frequency division multiplexing (OFDM) technology [23]. Compared with RSSI, CSI is more stable under static conditions and more sensitive under dynamic signals [13].

In 2011, Halperin et al. released the CSI tool, which greatly facilitated the acquisition of CSI information on commercial Wi-Fi devices [24]. The CSI tool attracted a large number of researchers to utilize CSI for Wi-Fi activity sensing research [12, 25–27]. WiFinger is designed to recognize 9-digit finger gestures from the American Sign Language (ASL) [12]. WiSign is an indoor sign language recognition system that can recognize five gestures with an accuracy of 93.8% [25]. DF-WiSLR [26] can recognize 19 dynamic and 30 static sign gestures. Experimental results showed that gesture direction and environment had a great influence on recognition performance. In [27], a dual-stream

convolutional network was used to extract spatiotemporal information from six CSI action datasets.

Reference [28] describes that the mapping relationship between gestures and CSI data is not unique, which differs from traditional gesture image data. The CSI data generated by the same gesture can vary greatly by person, location, orientation, and scenarios. Gao et al. [28] used dynamic phase index (EDP-index) error to remove the influence of different positions and orientations on gestures to improve the quality of CSI-based wireless sensing. In [29], spatiotemporal information from CSI gesture data was extracted via a parallel long short-term memory fully convolutional network (LSTM-FCN) to accommodate user differentiation and gesture diversity. The gesture recognition system identified 50 common gestures from 5 users with 98.9% accuracy. WiGRUNT [30] realized domain-independent features based on CSI gestures through a spatiotemporal dual attention mechanism and validated it on the Widar3 dataset.

SignFi [3] achieved the average recognition accuracy of 98.01%, 98.91%, 94.81%, and 86.66% in the lab276, home276, lab + home276, and lab150, respectively. In 2020, reference [31] compared the three types of deep learning: long short-term memory (LSTM), CNN, and attentive bi-directional LSTM (ABLSTM). The experimental results showed that the CNN model had the best recognition performance on the SignFi dataset. The average recognition accuracy of the proposed CNN model for Lab276, Home276, Lab + Home276, and Lab150 was 99.855%, 99.674%, 99.734%, and 93.84%, respectively. In the same year, Lee and Gao [4] applied dual-output two-stream to the SignFi dataset and obtained good recognition results. In 2021, Ahmed et al. [32] used an LSTM framework with 150 hidden units to identify sign language in the SigniFi dataset. In addition to taking advantage of deep learning methods, Farhana Thariq Ahmed et al. [33] also adopted machine learning methods to manually extract high-order statistical (HOS) features from the SignFi dataset and implemented gesture classification via support vector machines (SVMs).

## 3. Materials and Methods

*3.1. Channel State Information.* In wireless communications, CSI describes how a signal propagates information from the sender to the receiver and represents the combined results of reflection, scattering, fading, and power attenuation over distance [34]. Let $X(f, t)$ and $Y(f, t)$ be the frequency domain representations of the transmitted and received signals with the carrier frequency $f$ at the time $t$. Then, the relationship between the transmitted signal and the received signal can be expressed as [14]:

$$Y(f,t) = H(f,t) \times X(f,t), \tag{1}$$

where $H(f, t)$ is the channel frequency response (CFR) of the carrier with the frequency $f$ at the time $t$. The CSI is composed of the CFRs corresponding to different frequency subcarriers for each antenna. Each CSI includes the amplitude and phase relationship of each subcarrier in the orthogonal frequency division multiplexing (OFDM) link. Each CSI can be represented as follows [14]:

$$H(k) = \|H(k)\| e^{j\angle H(k)}, \tag{2}$$

where $H(k)$ is the CSI of the $k^{\text{th}}$ subcarrier, $\|H(k)\|$ and $\angle H(k)$ are the amplitude and the phase of the $k^{\text{th}}$ subcarrier, respectively. They represent the important characteristics of CSI.

Using the CSI tool released by Halperin et al. [24], the raw CSI data can be obtained from each received data packet of a commercial Wi-Fi network interface card (NIC). The amplitude and phase of each CSI on the subcarrier $k$ sampled at the time $i$ can be obtained from the following equations:

$$\|H_k(i)\| = \sqrt{(Re)^2 + (Im)^2}, \tag{3}$$

$$\angle H_k(i) = \operatorname{actan}\left(\frac{Im}{Re}\right), \tag{4}$$

where $Re$ and $Im$ are the real and imaginary parts of the CSI on the subcarrier $k$ sampled at the time $i$. Thus, each subcarrier of the CSI provides amplitude and phase information that can be calculated at any time.

*3.2. SignFi Dataset.* The SignFi dataset contains CSI traces of 276 sign gestures that are commonly used in daily life. The dataset has been gathered through a receiver with one internal antenna and a transmitter with three external antennas. Figure 1 shows a schematic diagram of the laboratory and home environment.

As shown in Figure 1, the user is not standing in the line-of-sight (LOS) between the Wi-Fi transmitter (AP) and the receiver (STA). In comparison with LOS, the non-line-of-sight (NLOS) signals reflected by human behavior are much smaller, which makes it more difficult for sign language gestures to be recognized. For home and lab environments, the distance between STA and AP is 1.30 m and 2.30 m, respectively. The transmitting antenna array is orthogonal to the main transmission and receiving directions in the home environment. However, the angle between the transmit antenna array and the direct path differs by about 40 degrees in the laboratory environment. It can be seen from Figure 1 that the layout of the laboratory environment is more complex than the layout of the home environment. There is a substantial difference between these two environments, which results in completely different CSI signals received for the same gesture.

The SignFi dataset contains two parts. The first part of the dataset contains 8,280 instances divided into 276 gesture categories. Among them, 5,520 instances and 2,760 instances are from the laboratory environment and the home environment, respectively. There are 20 and 10 instances of each gesture in the lab and home environment, respectively. The second part of the dataset includes 7500 instances with 150 gesture categories, which means there are 50 instances for each gesture and only 10 instances for each user. The first part of the dataset was collected by one user, and the second

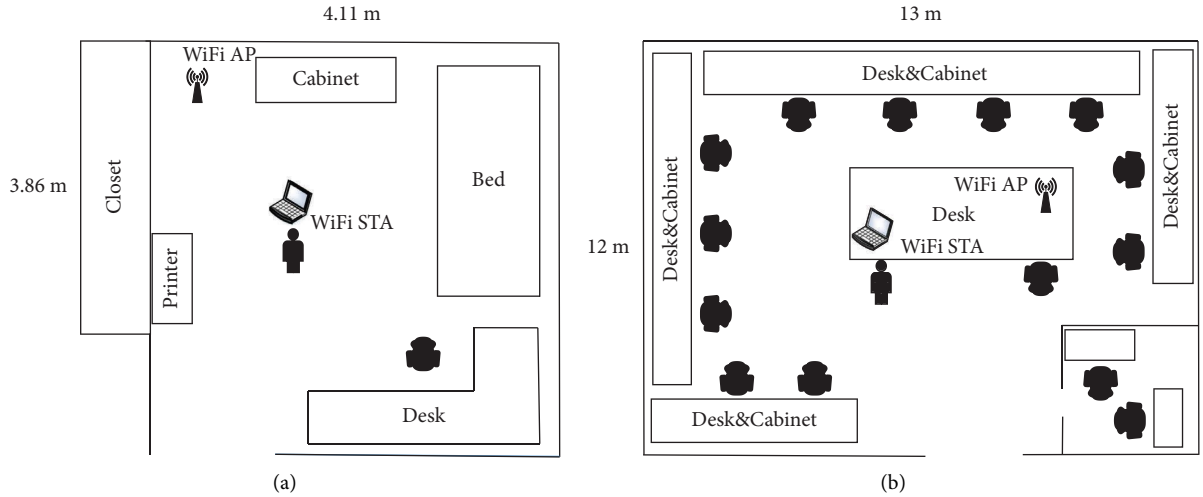


Figure 1: Schematic diagram of the laboratory and home environment. (a) Home environment and (b) lab environment.

part of the dataset was collected by five users. The statistics of the SignFi dataset are shown in Table 1.

*3.3. System Overview.* The flow chart of the proposed multitask sign language recognition method is shown in Figure 2. We obtained the raw CSI data from the SignFi dataset in Figure 2. The magnitude and phase of each raw CSI sample can be extracted, normalized, and transformed into a CSI image of size $200 \times 60 \times 30$ as described in [3, 31]. Unlike the abovementioned literature, we do not denoise and unwrap the amplitude and phase information.

Figure 2(a) shows that we resize each $200 \times 60 \times 3$ CSI image to $224 \times 224 \times 3$ images and use RandomVerticalFlip data augmentation technology. The RandomVerticalFlip data augmentation technology only increases the diversity of samples during the training phase without increasing the number of data samples. Figure 2(b) shows flattening the third dimension of a color CSI image of $200 \times 60 \times 3$ to obtain a grayscale CSI image of $200 \times 180$. The number of input channels in the first layer of the Wi-SignFi framework depends on the input depth of the CSI image.

After completing the abovementioned steps, the CSI image can be fed into the Wi-SignFi framework for tasks such as sign word, user, and environment recognition. Our experiments used nonrepetitive 5-fold cross-validation, which is consistent with the SignFi [3]. As can be seen from Table 1, the Lab150 and Lab + Home276 datasets are collected by five users and two environments. In the Wi-SignFi framework, the CNN module is used to recognize sign language gestures from the whole SignFi dataset, while the *K*-nearest neighbor (KNN) module performs user recognition on the Lab150 dataset and environment recognition on the Lab + Home276 dataset, respectively.

*3.4. Wi-SignFi Framework.* The proposed Wi-SignFi framework consists of an eight-layer CNN and a KNN module, as shown in Figure 3.The input size of a CSI colour image data is 224 x 224 x 3, so the first layer of Wi-signfi's convolution kernel requires 3 channels. Nevertheless, when the input data size is $200 \times 180$ CSI images, the first convolutional layer of Wi-SignFi only needs one channel. To prevent the loss of features caused by the deepening of the network layer, the Wi-SignFi network adopts the shortcut structure. The shortcut branch includes $1 \times 1$ convolution kernels and batch normalization (BN). A concatenation fusion is applied to the input of the last convolutional layer, which can fuse multilevel image features and reduce the loss of important information during the convolution process. The branch of the concatenation fusion adopts $3 \times 3$ max-pooling. A $3 \times 3$ max-pooling can reduce the data dimension, enhance the local receptive field, and improve the translation invariance of features.

The CNN module in the Wi-SignFi framework includes seven convolutional layers and one fully connected layer. The CNN module involves recognizing sign language gestures and the KNN module involves identifying different users or environments. The CNN module covers all datasets, while the KNN module is limited to Lab150 and Lab + Home276 datasets. The KNN module shares the feature maps extracted by the CNN module instead of manually extracting feature maps. Therefore, Wi-SignFi is a lightweight and end-to-end model.

## 4. Experimental Results on the SignFi Dataset

*4.1. Network Training and Test Settings.* We performed all experiments on sign language recognition tasks on a PC equipped with Intel (R) Xeon (R) CPU E5-2650 v3 @ 2.30 GHz CPU and GeForce GTX 2080 GPU with 8 GB of memory. We used the adaptive moment estimation (Adam) optimization algorithm with an initial learning rate of 0.0001 to train the network and update the weights and biases. The batch size is set to 16, and the training epochs are 250. We choose the rectified linear unit (ReLU) as the network activation. The experiments adopt nonrepetitive 5-fold cross-validation and follow the SignFi training and evaluation scheme. In other words, the ratio of training samples to test

Table 1: Statistics of the SignFi dataset.

| Data groups | Number of gesture categories | Number of gesture instances | Number of instances of each gesture per user | Number of users |
|---|---|---|---|---|
| Home276 | 276 | 2760 | 10 | 1 (user 5) |
| Lab276 | 276 | 5520 | 20 | 1 (user 5) |
| Lab + Home276 | 276 | 8280 | 20 + 10 | 1 (user 5) |
| Lab150 | 150 | 7500 | 10 | 5 (user 1, 2, 3, 4, and 5) |

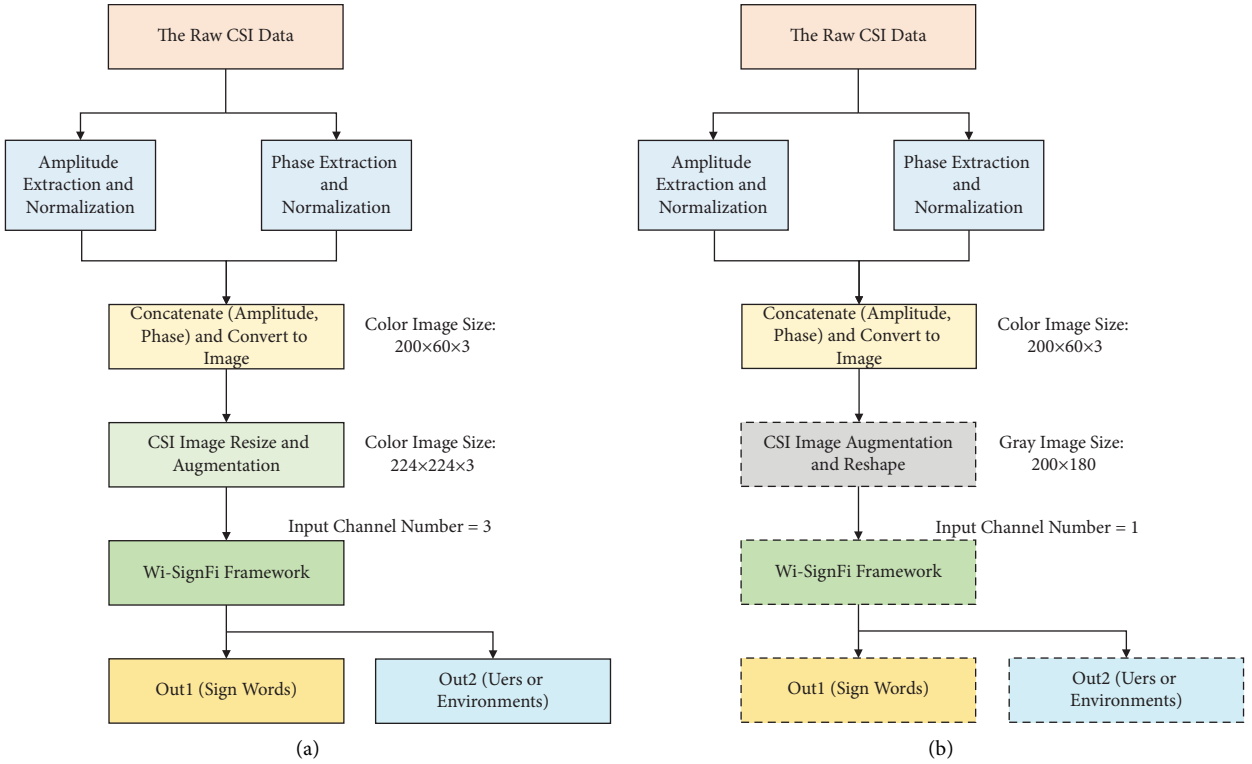


Figure 2: Flowchart of the proposed multitask sign language recognition method. (a) Color image size, $224 \times 224 \times 3$. (b) Gray image size, $200 \times 180$.

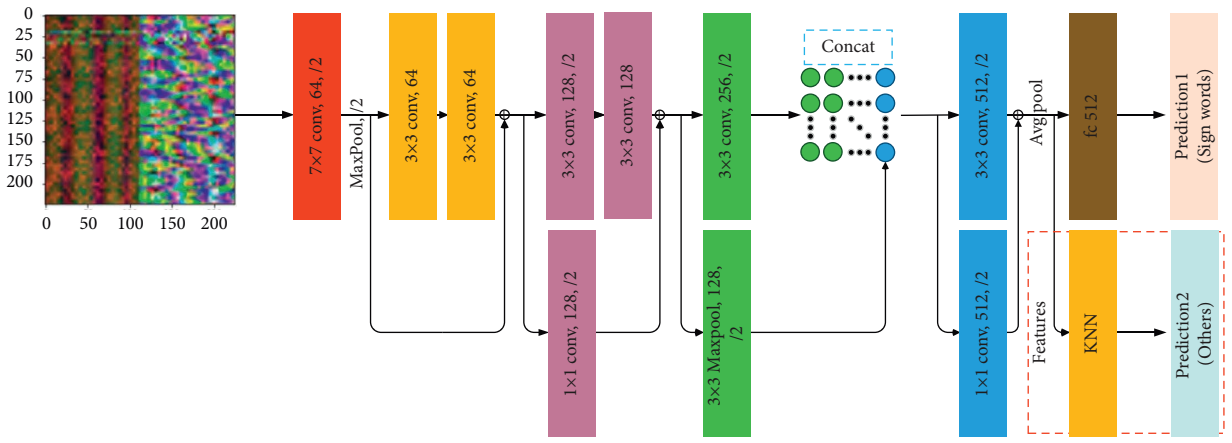


Figure 3: Our proposed Wi-SignFi framework.

samples is 8 : 2. The CNN module of the Wi-SignFi framework performs sign language gesture recognition for all datasets.

*4.2. Evaluation of the Different Input Sizes.* The file format of the raw CSI data is xxx.dat. Reference [24] mentions that CSI samples are extracted from the raw CSI data using Linux CSI Tool. Each CSI sample is a set of complex numbers. Equations (3) and (4) are applied to each CSI sample for obtaining magnitude and phase information. However, the data format of the SignFi dataset is xxx.mat, which contains the extracted data. Generally, there are three input data sizes for amplitude and phase information: (1) $200 \times 60 \times 3$; (2) $224 \times 224 \times 3$; and (3) $200 \times 180$. Figure 4 shows the combination matrices of three different input sizes for sign language "CONTINUE" in the laboratory environment.

The SignFi dataset provides a CSI sample for only 30 subcarriers. A receiver with one internal antenna needs to simultaneously receive three sets of the CSI data from a transmitter with three external antennas. There are 200 CSI instances for each sign gesture. Therefore, the size of each CSI matrix of SignFi is $200 \times 30 \times 3$. Amplitude and phase information is obtained from the raw CSI measurements of the SignFi dataset. They can be combined and reshaped into combined matrices of size $200 \times 60 \times 3$ as shown in Figure 4(a). The $Y$-axis of Figure 4(a) represents the 200 CSI instances for each gesture. The first half (0–29) of the $X$-axis of Figure 4(a) is amplitude information, and the second half (30–59) is the phase information.

The color channels of Figure 4(a) correspond to the three antenna signals. We resized the height and width of the combined matrix of $200 \times 60 \times 3$ to 224 to get the combined matrix of $224 \times 224 \times 3$, as shown in Figure 4(b). We flattened the third dimension of the combined matrix of $200 \times 60 \times 3$ to get the combined matrix of $200 \times 180$, as shown in Figure 4(c). Every 60 pixels on the $X$-axis of Figure 4(c) is the amplitude and phase combination matrix of one antenna of the wireless transmitter. The input data of $200 \times 60 \times 3$ and $224 \times 224 \times 3$ can be normalized and multiplied by 255 to convert them into the color image required by the CNN. The input data of $200 \times 180$ is the grayscale image after the same operation as above. Figure 4(c) is a two-dimensional matrix different from the three-dimensional matrix of Figures 4(a) and 4(b). When Figure 4(c) is used as the input data of the Wi-SignFi model, the channel size of the input layer of the Wi-SignFi model should be set to 1.

Next, let us explore and evaluate the impact of different input data sizes on model recognition performance. The evaluations for different input sizes are shown in Table 2.

It can be seen from Table 2 that when the input resolution is $200 \times 60 \times 3$, the recognition accuracy of the Wi-SignFi model is the lowest. The recognition result of the model increases with the increase of the input resolution. Compared with the model recognition results of the input data with a resolution of $224 \times 224 \times 3$, the model recognition results of the input data with a resolution of $200 \times 180$ are only slightly lower, excluding the Lab150 dataset. Input data of $200 \times 180$ resolution is suitable for multiuser datasets such as the Lab150 dataset. We consider the recognition accuracy of the model is easily affected by the resolution of the input data. When the resolution of the input data increases, the training time of the model increases accordingly, as shown in Figure 5.

To visually explore the impact of the increased resolution on the training time cost, we express the time at different resolutions as a percentage and use Wi-SignFi ($224 \times 224 \times 3$) as the benchmark. Figure 5 shows that Wi-SignFi ($200 \times 60 \times 3$) takes about half as long as Wi-SignFi ($224 \times 224 \times 3$) in all SignFi datasets. Wi-SignFi ($200 \times 180$) and Wi-SignFi ($200 \times 60 \times 3$) take similar time, excluding the Home276 dataset. Because the number of samples in the Home276 dataset is relatively small, it has a greater impact on the slightly increased time overhead. According to Figure 5 and Table 2, we conclude that Wi-SignFi ($200 \times 180$) can balance the time-consuming and recognition accuracy of the model.

*4.3. Impact of Data Augmentation.* The data augmentation introduced in this article performs RandomVerticalFlip processing on the CSI image input data without increasing data samples. Figure 6 shows the impact of the RandomVerticalFlip operation on the Wi-SignFi gesture recognition performance.

It can be seen from Figure 6 that the data augmentation of RandomVerticalFlip has little effect on the recognition of Wi-SignFi gesture recognition in the Home276, Lab276, and Lab + Home276 datasets. However, this data augmentation operation improves the recognition accuracy of Wi-SignFi gesture recognition by more than 3% on the Lab150 dataset. The Lab150 dataset is collected from five different users. Therefore, we believe that the data augmentation of RandomVerticalFlip helps to improve the Wi-SignFi model's sign language recognition accuracy for multiple users, but has a little effect on improving the single-user sign language recognition accuracy.

*4.4. Comparison of Existing Sign Language Recognition Models on the SignFi Dataset.* There are five sign language recognition models on the SignFi dataset. We summarize the comparison results of different sign language recognition technologies on the SignFi dataset in Table 3.

Table 3 shows the recognition results of different kinds of literature on the SignFi dataset from 2018 to 2021. The models in this literature can be divided into three categories: CNN, LSTM, and SVM. LSTM [32] only needs amplitude values to achieve good recognition performance, except for the Lab150 dataset. In contrast, HOS-Re [33] achieves good performance on the Lab150 dataset by manually extracting sign language gesture features from CSI traces and then using SVM as a classifier.

The other methods are all CNN methods. The input data modality of the Wi-SignFi model is the same as that of CNN [31] and SignFi [3], both are concatenations of amplitude and phase information. The input data of dual-output two-stream with ResNet50 [4] contains not only amplitude and

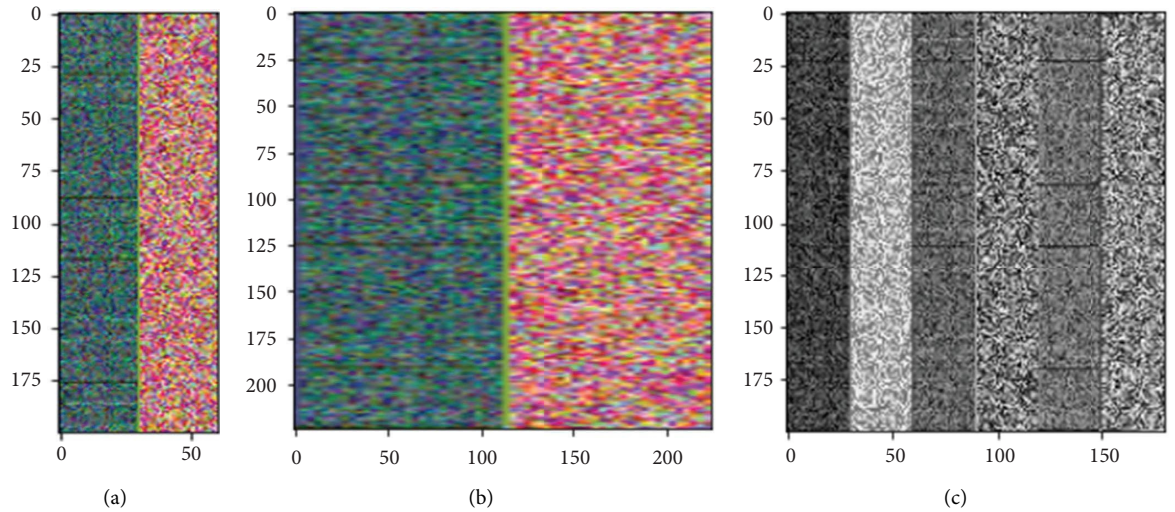


Figure 4: Combination matrices of three different input sizes for the sign word "CONTINUE" in a laboratory environment. (a) The combined matrix of amplitude and phase information from the raw CSI data. Its size is $200 \times 60 \times 3$. (b) A combined matrix that resizes the height and width of (a) to 224. Its size is $224 \times 224 \times 3$. (c) The combined matrix formed by flattening the third dimension of (a). Its size is $200 \times 180$.

Table 2: Evaluation of the different input sizes.

| Data groups | Wi-SignFi ($200 \times 60 \times 3$) (%) | Wi-SignFi ($200 \times 180$) (%) | Wi-SignFi ($224 \times 224 \times 3$) (%) |
|---|---|---|---|
| Home276 | 98.91 | 99.71 | **99.75** |
| Lab276 | 99.61 | 99.67 | **99.80** |
| Lab + Home276 | 98.90 | 99.40 | **99.69** |
| Lab150 (five users) | 94.95 | **96.79** | 96.41 |
| Mean | 98.09 | 98.89 | **98.91** |

The bold values given in Table 2 represent the best action recognition results for each of the four datasets.

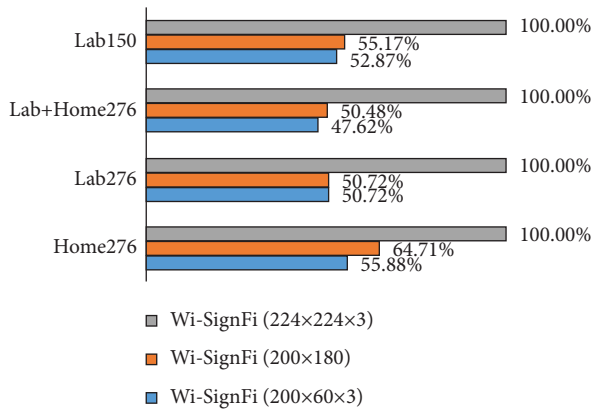


Figure 5: Time consumption of the different input sizes.

phase information, but also difference information including the gesture motion. The input data for the abovementioned CNN models are all preprocessed, except for the Wi-SignFi model. The input data resolution of CNN [31] and SignFi [3] are both $200 \times 60 \times 3$, while the dual-output two-stream with ResNet50 [4] is $224 \times 224 \times 3$. CNN [31] achieved the best recognition results in Lab276 and Lab + Home276, 99.855% and 99.73%, respectively. However, Wi-SignFi outperforms other methods in recognition accuracy on the Home276 and Lab150 datasets, 99.75% and 96.41%, respectively. Meanwhile, Wi-SignFi ranks first in the four datasets with an average accuracy of 98.91%. The Lab + Home276 dataset with mixed multienvironment data and the Lab150 dataset with multiuser data resulted in a significant drop in the recognition accuracy of the SignFi model. In contrast, our model can maintain good performance, which indicates that our proposed model has a certain generalization ability in complex environments.

*4.5. Comparison with Existing Neural Networks.* Wi-SignFi is a CNN with only eight convolutional layers, so we also chose some lightweight neural networks for comparison. The input data resolution for these lightweight models is fixed at $224 \times 224 \times 3$. Since the CSI data is very different from the ImageNet data, the existing neural networks are trained from scratch. The evaluation results of existing neural networks in sign language recognition are shown in Table 4.

According to Tables 3 and 4, we conclude that Wi-SignFi is suitable for the SigniFi dataset. For small sample data like SignFi, the network layer of the sign language gesture recognition model does not need to be very deep. lightweight networks such as shuffleNet [35], MnasNet [36], and MobileNet [37] applied to mobile terminals are not very good at recognizing the SignFi dataset.

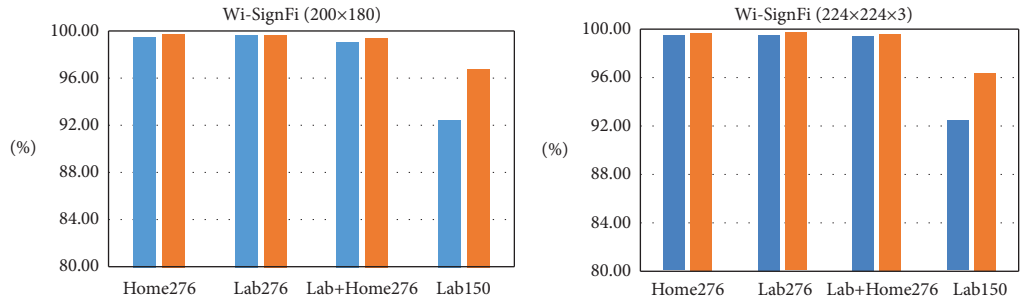


Figure 6: Wi-SignFi gesture recognition performance with/without data augmentation. (a) Wi-SignFi ($200 \times 180$) and (b) Wi-SignFi ($224 \times 224 \times 3$).

Table 3: Recognition accuracy of Wi-SignFi and other methods.

| Methods | Modality | Home276 (%) | Lab276 (%) | Lab + Home276 (%) | Lab150 (%) | Mean (%) | Rank |
|---|---|---|---|---|---|---|---|
| SignFi [3] (2018) | A + P | 98.91 | 98.01 | 94.81 | 86.66 | 94.60 | 6 |
| HOS-Re [33] (2019) | — | 98.26 | 97.84 | 96.34 | 96.23 | 97.17 | 5 |
| CNN [31] (2020) | A + P | 99.64 | **99.855** | **99.73** | 93.84 | 98.27 | 3 |
| Dual-output two-stream with ResNet50 [4] (2020) | A + P + Difference | 99.13 | 96.79 | 97.08 | 95.88 | 97.22 | 4 |
| LSTM [32] (2021) | A | 99.50 | 99.80 | 99.40 | 78 | 94.18 | 7 |
| **Wi-SignFi ($224 \times 224 \times 3$)** | A + P | **99.75** | 99.80 | 99.69 | 96.41 | **98.91** | **1** |
| **Wi-SignFi ($200 \times 180$)** | A + P | 99.71 | 99.67 | 99.40 | **96.79** | 98.89 | **2** |

The bold values given in Table 3 represent the best action recognition results for each of the four datasets.

Table 4: Evaluation of the existing neural networks.

| Data groups | Resnet18 (224* 224* 3) (%) | MnasNet (224* 224* 3) (%) | ShuffleNetV2 (224* 224 * 3) (%) | MobileNetV2 (224* 224* 3) (%) | Wi-SignFi (224* 224* 3) (%) |
|---|---|---|---|---|---|
| Home276 | 99.13 | 92.25 | 92.43 | 97.91 | **99.75** |
| Lab276 | 99.59 | 98.70 | 97.95 | 99.31 | **99.80** |
| Lab + Home276 | 99.19 | 98.67 | 96.92 | 98.50 | **99.69** |
| Lab150 | 95.28 | 89.61 | 74.03 | 88.00 | **96.41** |
| Mean | 98.30 | 94.81 | 90.33 | 95.93 | **98.91** |

*4.6. Classification Results for Users and Environments.* In the multitask Wi-SignFi framework, we use SVM, random forest (RF), and KNN methods for user identification on the Lab150 dataset and environment identification on the Lab + Home276 dataset, respectively. It can be seen from Table 5 that there is little difference in the classification accuracy of users and environments between Wi-SignFi ($200 \times 180$) and Wi-SignFi ($200 \times 60 \times 3$). The environment recognition accuracy of SVM, RF, and KNN on the Lab + Home276 dataset is very well. It may be that the CSI data collected in the home environment and the laboratory environment are significantly different and easy to distinguish. This assumption is consistent with that described in [4]. However, KNN achieves the best result with 86.68% user recognition accuracy on the Lab150 dataset. We believe that the same gestures made by different people have a certain similarity.

*4.7. Cross-Domain Sensing Based on Different Scenarios.* Reference [28] describes that the mapping relationship between gestures and the CSI data is not unique, which differs from the traditional gesture image data. The CSI data generated by the same gesture can vary greatly by person, location, orientation, and scenarios. The target domain contains only one of the home or laboratory environment data to evaluate the general performance of the Wi-SignFi model in different scenarios. As shown in Figure 7(a), Lab276 and Home276 are the training and testing datasets for the Wi-SignFi ($200 \times 180$) model, whereas Figure 7(b) shows the opposite. The test dataset contains 1656 examples, according to Table 5.

Figure 7(a) shows that when only Lab276 samples are used as the training dataset, the recognition accuracy of Home276 is only 0.36 percent. According to references [4, 26, 28], the poor results may be attributed to the significant difference between the CSI data of the two environments. Inspired by the few-shot learning method [30], few samples from the target domain are added to the source domain. By adding a Home276 sample for each action ($5520 + 276$) in the training dataset, the test accuracy can be reached to 85.39%. With more Home276 data samples in the training set, the recognition accuracy of the test dataset

Table 5: Classification results for users and environments.

| Data groups | Methods | Wi-SignFi (200 × 180) (%) | Wi-SignFi (224 × 224 × 3) (%) |
|---|---|---|---|
| Lab + Home276 (two environments) | SVM | 99.99 | 99.99 |
| | RF | 99.79 | 99.91 |
| | KNN | 99.99 | 99.99 |
| Lab150 (five users) | SVM | 85.95 | 85.95 |
| | RF | 80.84 | 81.00 |
| | KNN | 86.68 | 86.67 |

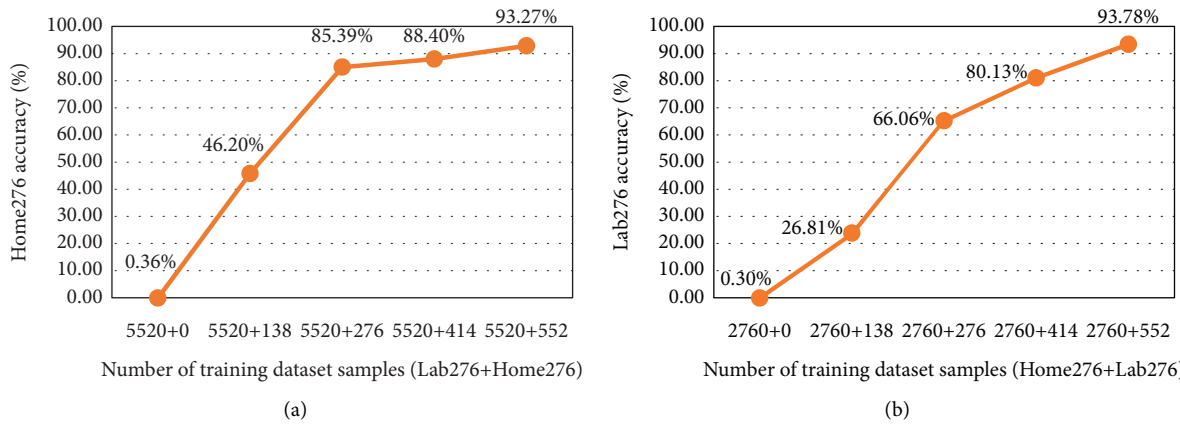


Figure 7: Cross-domain gesture recognition performance based on different scenarios. (a) Cross-domain (lab⟶home). (b) Cross-domain (home⟶lab).

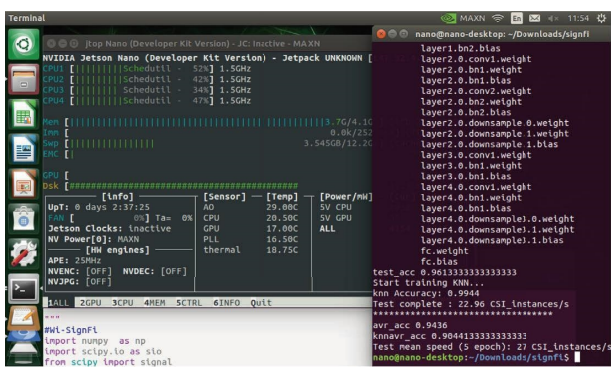

Figure 8: Wi-SignFi runtime interface on a Jetson Nano device.

improves. When the training dataset contains two Home-276 samples per action (5520 + 552), the recognition accuracy of Home276 is close to Lab + Home-276 in Table 5. Figures 7(b) and 7(a) show similar results.

## 5. Wi-SignFi Framework Deployed on a Jetson Nano Device

Jetson Nano is a single-board computing platform from Nvidia [38]. Figure 8 shows Wi-SignFi (200 × 180) is applied to a Jetson Nano device for the gesture and user recognition on the Lab150 dataset. After saving the Wi-SignFi (200 × 180) model parameters, we randomly select 1500 CSI instances from the Lab150 dataset for testing. The results obtained by averaging five tests of our model on the Jetson Nano are shown in Figure 8.

As shown in Figure 8, although our model runs on a Jetson Nano device on almost the full 4G RAM, it achieves an average inference speed of 27 CSI instances per second. Gesture recognition accuracy for the Wi-SignFi framework running on Jetson Nano devices dropped slightly, while user recognition improved. Overall, the deployment of the Wi-SignFi framework on embedded devices performed well.

## 6. Conclusions

In this article, we implement a multitask sign language recognition system based on the Wi-SignFi framework. The system not only recognizes gestures, but also the user and the environment to which the gesture corresponds. Experimental results on SignFi datasets are as follows: (1) The CSI data fed to Wi-SignFi does not require tedious preprocessing such as denoising and unwrapping, but our model still outperforms previous works on gesture recognition. (2) For the Lab + Home276 dataset with mixed multienvironment data and the Lab150 dataset with multiuser data, our model can keep acceptable accuracy, which indicates that our proposed model has a certain generalization ability in complex environments. Meanwhile, we used KNN to perform user and environment recognition tasks on these two datasets, respectively. (3) According to the experimental results, the gesture recognition accuracy of the model is greatly affected by the resolution of the input data. Although Wi-SignFi (224 × 224 × 3) achieves the best results, its training time is greatly increased. The experimental results show Wi-SignFi (200 × 180) can balance the time-consuming and recognition accuracy of the model. (4) Wi-

SignFi is a lightweight and end-to-end model. Wi-SignFi on the Jetson Nano device achieves an inference speed of 27 CSI instances per second. Our proposed sign language recognition system is expected to integrate with the IoT to improve the lives of the deaf community.

Our proposed Wi-SignFi has some limitations, which will be the direction of our further research in the future. Firstly, Wi-SignFi is a lightweight CNN model that is suitable for small-scale datasets and cannot learn the complex temporal dynamics information of gesture actions. Combining Wi-SignFi with LSTM or a transformer would be an effective way for predicting the gesture activity.

Secondly, cross-domain wireless sensing is a difficult topic, so it has always been a research hotspot. In our experiments, we evaluate the cross-domain recognition performance of Wi-SignFi via the few-shot learning method. We expect to eliminate all samples from the target dataset in the future to facilitate the development of domain adaptation.

Finally, existing references tend to map the CSI data from the three antennas on the Wi-Fi transmitter to three RGB channels and then convert them to CSI color images. However, we extend the CSI data from different antennas to a single-channel grayscale image as the input data of the model and obtain satisfactory gesture recognition accuracy. We will investigate the correlation between transmitting or receiving antennas in future work. Additionally, we will actively explore more efficient ways to acquire the CSI data and real-time applications of deep learning on embedded devices.

## Data Availability

Previously reported SignFi data were used to support this study and are available at https://yongsen.github.io/SignFi/. These prior studies (and datasets) are cited at relevant places within the text as references [3].

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Authors' Contributions

C.-C.L and Z.G. were responsible for conceptualization, methodology, validation, formal analysis, investigation, visualization, and resource collection. C.-C.L supervised and reviewed and edited the manuscript. Z.G. was involved in software and preparation of the original draft. R.Z. collected the data. L.Z and X.X were responsible for funding acquisition. All authors have read and agreed to the published version of the manuscript.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (11803015); Provincial Natural Science Foundation of Fujian (2018J05009, 2020J01387, and JZ160476); Training Plan for Outstanding Young Scientists in Colleges and Universities in Fujian (Minjiaoke (2018) 47); Doctoral Research Fund of Sanming University (16YG09); Education and Research Project for Young and Middle-Aged Teachers in Fujian (JAT200615 & B202006); and National Fund Cultivation Program Project of Sanming University (PYT2104&PYT2108); Fujian Provincial Education Science "14th Five-Year Plan" 2022 (FJJKBK22-172).

## References

[1] G. G. Nath and C. Arun, "Real time sign language interpreter," in *Proceedings of the 2017 IEEE International Conference on Electrical*, pp. 1–5, Instrumentation and Communication Engineering (ICEICE), Karur, India, April 2017.

[2] W. Aly, S. Aly, and S. Almotairi, "User-independent American sign language alphabet recognition based on depth image and PCANet features," *IEEE Access*, vol. 7, pp. 123138–123150, 2019.

[3] Y. Ma, G. Zhou, S. Wang, H. Zhao, W. Jung, and W. Mobile, "SignFi: sign language recognition using WiFi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–21, 2018.

[4] C. C. Lee and Z. Gao, "Sign Language recognition using two-stream convolutional neural networks with wi-fi signals," *Applied Sciences*, vol. 10, no. 24, 2020.

[5] N. Praveen, N. Karanth, and M. Megha, "Sign language interpreter using a smart glove," in *Proceedings of the 2014 International Conference on Advances in Electronics Computers and Communications*, pp. 1–5, IEEE, Bangalore, India, October 2014.

[6] A. Z. Shukor, M. F. Miskon, M. H. Jamaluddin, F. b. Ali@ Ibrahim, M. F. Asyraf, and M. B. Bahar, "A new data glove approach for Malaysian sign language detection," *Procedia Computer Science*, vol. 76, pp. 60–67, 2015.

[7] T. Kanokoda, Y. Kushitani, M. Shimada, and J. I. Shirakashi, "Gesture prediction using wearable sensing systems with neural networks for temporal data analysis," *Sensors*, vol. 19, no. 3, p. 710, 2019.

[8] M. A. Al-qaness, A. Dahou, M. Abd Elaziz, and A. Helmi, "Multi-ResAtt: multilevel residual network with attention for human activity recognition using wearable sensors," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, 2022.

[9] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3d convolutional neural networks," in *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, Turin Italy, June 2015.

[10] L. R. Cerna, E. E. Cardenas, D. G. Miranda, D. Menotti, and G. Camara-Chavez, "A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a microsoft Kinect sensor," *Expert Systems with Applications*, vol. 167, Article ID 114179, 2021.

[11] J. Liu, G. Teng, and F. Hong, "Human activity sensing with wireless signals: a survey," *Sensors*, vol. 20, no. 4, p. 1210, 2020.

[12] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "WiFinger: talk to your smart devices with finger-grained gesture," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 250–261, Heidelberg, Germany, September 2016.

[13] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi channel state information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, 2017.

[14] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: indoor localization via channel Response," *ACM Computing Surveys*, vol. 46, no. 2, pp. 1–32, 2013.

[15] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4165–4174, Long Beach, CA, USA, June 2019.

[16] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880–1891, 2019.

[17] S. Kumar, P. Tiwari, and M. Zymbler, "Internet of Things is a revolutionary approach for future technology enhancement: a review," *Journal of Big data*, vol. 6, no. 1, pp. 111–121, 2019.

[18] R. Alazrai, M. Hababeh, B. A. Alsaify, M. Z. Ali, and M. I. Daoud, "An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals," *IEEE Access*, vol. 8, pp. 197695–197710, 2020.

[19] S. Sigg, U. Blanke, and G. Tröster, "The telepathic phone: frictionless activity recognition from wifi-rssi," in *Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 148–155, Budapest, Hungary, March 2014.

[20] H. Abdelnasser, M. Youssef, and K. A. Harras, "Wigest: a ubiquitous wifi-based gesture recognition system," in *Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 1472–1480, IEEE, Hong Kong, China, April 2015.

[21] X. Pan, T. Jiang, X. Li, X. Ding, Y. Wang, and Y. Li, "Dynamic hand gesture detection and recognition with WiFi signal based on 1d-CNN," in *Proceedings of the 2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, Shanghai, China, May 2019.

[22] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, pp. 617–628, New York, NY, USA, September 2014.

[23] E. J. I. C. M. Perahia, "IEEE 802.11 n development: history, process, and technology," *IEEE Communications Magazine*, vol. 46, no. 7, pp. 48–55, 2008.

[24] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: gathering 802.11 n traces with channel state information," *ACM SIGCOMM - Computer Communication Review*, vol. 41, no. 1, p. 53, 2011.

[25] J. Shang and J. Wu, "A robust sign language recognition system with multiple Wi-Fi devices," in *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*, pp. 19–24, New York, NY, USA, August 2017.

[26] H. F. T. Ahmed, H. Ahmad, K. Narasingamurthi, H. Harkat, S. K. Phang, and D. F. Wi, "DF-WiSLR: device-free wi-fi-based Sign Language recognition," *Pervasive and Mobile Computing*, vol. 69, Article ID 101289, 2020.

[27] B. Sheng, Y. Fang, F. Xiao, and L. Sun, "An accurate device-free action recognition system using two-stream network," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7930–7939, 2020.

[28] R. Gao, W. Li, Y. Xie et al., "Towards robust gesture recognition by characterizing the sensing quality of WiFi signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–26, 2022.

[29] Z. Tang, Q. Liu, M. Wu, W. Chen, and J. J. C. C. Huang, "WiFi CSI gesture recognition based on parallel LSTM-FCN deep space-time neural network," *China Communications*, vol. 18, no. 3, pp. 205–215, 2021.

[30] Y. Gu, X. Zhang, Y. Wang et al., "WiGRUNT: WiFi-enabled gesture recognition using dual-attention network," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 736–746, 2022.

[31] M. R. M. Bastwesy, N. M. ElShennawy, and M. T. F. Saidahmed, "Deep learning Sign Language recognition system based on wi-fi CSI," *International Journal of Intelligent Systems and Applications*, vol. 12, no. 6, pp. 33–45, 2020.

[32] H. F. T. Ahmed, H. Ahmad, S. K. Phang, H. Harkat, and K. Narasingamurthi, "Wi-fi CSI based human Sign Language recognition using LSTM network," in *Proceedings of the 2021 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pp. 51–57, Bandung, Indonesia, July 2021.

[33] H. Farhana Thariq Ahmed, H. Ahmad, S. K. Phang, C. A. Vaithilingam, H. Harkat, and K. J. S. Narasingamurthi, "Higher order feature extraction and selection for robust human gesture recognition using CSI of COTS wi-fi devices," *Sensors*, vol. 19, no. 13, p. 2959, 2019.

[34] A. E. Kosba, A. Saeed, and M. Youssef, "Robust WLAN device-free passive motion detection," in *Proceedings of the 2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 3284–3289, IEEE, Paris, France, April 2012.

[35] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, Salt Lake City, Utah, June 2018.

[36] M. Tan, B. Chen, and R. Pang, "Mnasnet: platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2820–2828, Long Beach, California, June 2019.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, Utah, June 2018.

[38] N. Developer, "Getting started with jetson nano developer kit," 2020, https://developer.nvidia.com/embedded/learn/get-started-jetson-nano-devkit#:%7E:text=Setup%20Steps%26text=Power%20on%20your%20computer%20display,power%20on%20and%20boot%20automatically.