RF-Siamese: Approaching Accurate RFID Gesture Recognition With One Sample

Zijing Ma, Shigeng Zhang, *Member, IEEE,* Jia Liu, Xuan Liu, Weiping Wang, Jianxin Wang, *Senior Member, IEEE,* and Song Guo *Fellow, IEEE*

Abstract—Performing accurate sensing in diverse environments is a challenging issue in wireless sensing technologies. Existing solutions usually require collecting a large number of samples to train a classifier for every environment, or further assume similar sample distribution between different environments such that a model trained in one environment can be transferred to another. In this paper, we propose RF-Siamese, an RFID-based gesture sensing approach that achieves comparable accuracy to existing solutions but requires only a few samples in each eivironment. RF-Siamese leverages Siamese networks to distinguish different gestures with only a small number of samples and is enhanced by several novel designs to achieve high accuracy in diverse environments. First, the network structure and parameters (e.g., loss function and distance metric) are carefully designed to be suitable for RFID gesture recognition. Second, a permutation-based dataset generation strategy is proposed to make full use of the collected samples to enhance the recognition accuracy. Third, a template matching method is proposed to extend the Siamese network to classify multiple gestures. Extensive experiments on commercial RFID devices demonstrate that RF-Siamese achieves a high accuracy of 0.93 with only one sample of each gesture when recognizing 18 different gestures, while state-of-the-art approaches based on transfer learning and meta learning achieve an accuracy of only 0.59 and 0.70, respectively.

Index Terms—Gesture Recognition, RFID, Few-shot Learning, Siamese Network

1 Introduction

Wireless sensing has emerged as an amazing technique for various smart applications, e.g., human-machine interaction (HMI) [1]–[4], activity recognition [5]–[8], and ubiquitous computing [9]–[11]. Compared with sensing technology based on Wi-Fi [12] and millimeter wave [13], wireless sensing systems based on radio frequency identification (RFID) can track multiple users simultaneously [14] with low-cost hardware, making them easy to deploy and suitable for users to contact with computers or smartphones without wearing redundant devices. These wireless sensing applications become prosperous because they enable users to enjoy a better quality of service without using tedious hardware.

One challenging issue in current wireless sensing technologies is how to perform robust and accurate sensing in different environments. Many wireless gesture recognition systems have been proposed, however their performance heavily depends on the environment. For instance, the performance of a model trained in an environment A might significantly degrade when the model is deployed in a different environment B [15]. This is usually ascribed to the complex propagation of wireless signals, especially in indoor environments, which makes the received signals diverse in different environments even when the user performs the same gesture. Wireless signal propagation is affected by many factors, including the distance between the user and the antenna, the orientation of the user, and the obstacles in the environment. The collected signals used to train the gesture recognition model are usually the superposition of signals traversing different paths. Consequently, when a model trained with data collected from a meeting room is utilized to classify the samples from a laboratory, the model performance might degrade. This problem limits most wireless sensing systems to work well only in the strict experimental environment, which seriously hinders the commercial development of wireless sensing systems and becomes an urgent problem to be solved [16].

Existing solutions to achieve accurate gesture recognition in diverse environments fall into two types. The first type of solutions tries to retrain a model for each new environment [17]–[19]. However, training a new sensing model from scratch requires collecting a large number of training data, which is time-consuming and laborious. Recently, some works use deep learning to leverage its ability in automatically extracting features to achieve high recognition accuracy. However, sensing models based on deep learning usually require a large number of training samples, which further increases the burden in collecting samples. The second type of solutions, which has attracted much research attention in recent years, employs transfer learning to fine-tune a model pre-trained in the source environment to adapt to the target environment with samples from the latter [15], [20], [21]. Such works are usually based on deep neural networks, in which the shallow layers extract common features that can be transferred across different

- *Zijing Ma, Shigeng Zhang, Weiping Wang, and Jianxin Wang are with the School of Computer Science and Engineering, Central South University, China. E-mail: {mazijingcsu, sgzhang, wpwang, jxwang}@csu.edu.cn.*
- *Jia Liu is with the Department of Computer Science and Technology, Nanjing University, China. E-mail: jialiu.cs@gmail.com.*
- *Xuan Liu is with the College of Computer Science and Electronic Engineering, Hunan University, China, 410082. E-mail: xuan_liu@hnu.edu.cn.*
- *Song Guo is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. Email: song.guo@polyu.edu.hk.*

Manuscript received January 1st, 2021; revised XX, 2021.

environments. Then the weights of deep layers are fine-tuned by using a few number of samples from the target environment [22]. However, this type of approaches requires samples from different environments following similar distributions. Otherwise, the transfer learning might fail or even cause negative transfer [23]. Moreover, approaches based on transfer learning do not necessarily reduce the number of samples. It has been shown that several dozens or even hundreds of samples from both the source and the target environments are needed to achieve high accuracy in such approaches [15].

In this paper, we ask the following question: ***Can we perform accurate gesture recognition in different environments with only a few samples of each gesture and no assumptions on the sample distribution?*** In one word, we want to devise a gesture sensing system that can achieve high accuracy in a new environment, given only a few samples of each gesture collected from the new environment and no samples from source environment. As our answer, we propose RF-Siamese, an RFID-based gesture recognition system that leverages the Siamese network to recognize different gestures with only a few samples of each gesture. The Siamese network is a promising model in few-shot learning [24]–[27]. It trains a pair of weight-sharing networks that can maximize the distance between samples from different categories and minimize the distance between samples from the same category, which makes it be able to distinguish samples from different categories with only a small number of training samples.

It is not trivial to put RF-Siamese into practice and we need to address the following challenges. First, the Siamese network is originally designed for image recognition. However, the environment dependency problem in RFID sensing is quite different from that in image recognition. It is necessary to adequately and uniquely modify the Siamese network in image recognition before it can be applied in RFID sensing. Second, as the samples of each gesture are rare, how to make full use of the samples to achieve high recognition accuracy should be considered. Third, how to extend the binary classification Siamese network to classify multiple gestures (18 gestures in this paper) in an effective manner is also a challenging problem. We make the following designs to address these challenges. First, we carefully devise the network structure and parameters such as loss function and distance metric to make it suitable for RFID sensing. Second, we propose a new permutation-based dataset generation strategy to make full use of collected samples, avoiding the drawback of the traditional strategy where some samples are randomly neglected. Third, we modify the original binary classification Siamese network to a multiple classification model based on template matching, enabling it to classify multiple gestures.

We implement RF-Siamese based on commercial RFID devices and conduct extensive experiments to evaluate its performance. The results show that RF-Siamese demonstrates superior performance: it achieves a recognition accuracy of 0.93 with only one sample of each gesture when recognizing 18 different gestures. As a comparison, the state-of-the-art approaches based on transfer learning and meta learning only achieve accuracy of 0.59 and 0.70 in the same case, respectively. On average, RF-Siamese can reduce the number of samples by around 70% to achieve the same accuracy as state-of-the-art solutions.

The rest of this paper is organized as follows. In Section 2, we review related work. Some preliminaries are presented in Section 3. The detailed design of RF-Siamese is described in Section 4, including the design of the network structure, the selection of loss functions and the template-based multiple gesture recognition. The performance of RF-Siamese is evaluated and compared with state-of-the-art solutions in Section 5. Finally, we conclude this paper in Section 7.

2 Related Work

2.1 Gesture Recognition

Gesture recognition plays a significant role in HMI and has drawn massive attention in research. Early works rely on sensors integrated in wearable devices to capture gestures' features. For instance, uWave [28] leverages an accelerometer to collect the acceleration of user's gestures and classifies the gestures using template matching with DTW distance. The work in [29] utilizes the magnetic sensors of smartwatches to track the user's postures. Moreover, fine-grained hand activity recognition has been realized in [30]. These approaches require the user to wear dedicated devices, which are inconvenient to use in daily life.

Contactless gesture recognition has been proposed to improve the convenience. Computer-vision-based approaches [31], [32] use cameras to record the images of user's gestures, and classify different gestures with deep neural networks. Such approaches achieve high accuracy and real-time in gesture recognition, whereas their performance depends on illumination in the environment, which limits their commercial deployment. Wireless sensing systems do not suffer from insufficient light, and can also recognize gestures without wearing redundant devices. Most existing works leverage WiFi [33], [34], RFID [3], [35], [36], and millimeter wave [37], [38] to realize gesture recognition. For instance, the work in [34] recognizes 8 finger gestures with channel state information (CSI) and improves its robustness by removing environmental noise, reaching an accuracy higher than 0.9. Although these works try to remove the environmental factors when recognizing different gestures, they all require a large number of training samples, which is laborious and time-consuming and hindering the deployment of these systems in practical environments.

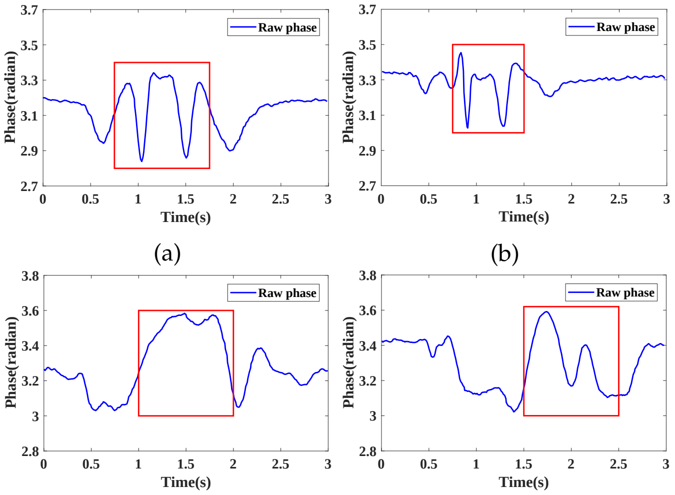
2.2 RFID-based Gesture Recognition

Recent years have witnessed a vigorous interest in leveraging RFID to recognize users' gestures because RFID devices are contactless, convenient, and low-cost. Compared with computer-vision-based approaches [39], RFID-based approaches do not suffer from light insufficiency. Moreover, RFID-based approaches can recognize multiple users at the same time, which is still a difficult problem for Wi-Fi-based approaches [14].

GRfid [35] calculates the DTW distance between two samples and recognizes them using a template matching method with DTW distance. However, calculating DTW distance is extremely time-consuming and thus GRfid is not real-time. To reduce latency and achieve real-time recognition, ReActor [36] combines 13 statistical features (e.g., *min*, *max*, *mode*, etc) and wavelet decomposition coefficients to achieve accurate gesture recognition. The accuracy of ReActor is nearly 0.95, and its time cost in the recognition process is less than 100 ms. An extended version of ReActor [40] further considers how to remove reflections from static obstacles to enhance recognition accuracy. With the rapid development of deep learning, recent works use deep neural networks to classify gestures without extracting features manually. In the work [41], the authors calculate each tag's probability of the gesture performing above the tag, transform the probability matrix to an image, and use the CNN to classify gestures. In the work [42], the authors propose an ongoing gesture recognition approach using adversarial learning. They leverage a LSTM network to classify gestures, and thus can output the results before the gesture finishes. In [18], the authors develop a multimodal CNN to aggregate the RSS and phase data to discriminate different gestures, and propose an adversarial model to remove domain-specific information. However, these approaches commonly require a large number of samples to train the model.

2.3 Accurate Sensing with A Few Samples

Although the aforementioned RFID-based gesture recognition approaches achieve good performance, most of them are environment-dependent, which means that a trained model cannot be deployed to a new environment directly otherwise its performance degrades sharply. The reason is that the propagation distances of wireless signals are various in different environments due to distinct reflection objects. Recently, some works have been proposed to address this problem by using transfer learning [15], [20], [43]–[45]. These works mainly try to train the new model in a new environment using the knowledge of models trained in similar environments. CrossSense [15] proposes a data roaming model based on ANN to generate synthetic Wi-Fi data, and leverages transfer learning to transfer the shallow layers of ANN when deployed to a new environment such that it can use a few samples to fine-tune the model. In [20], the authors propose RF-EATS, which utilizes Variational Autoencoder (VAE) to generate synthetic samples to train the model. It also leverages transfer learning to transfer the trained VAE to a new environment. OneFi [45] leverages velocity distribution to generate synthetic gestures' spectrogram with distinct angles to train the model, and uses transfer learning to transfer the classification model, reducing the number of training samples. MetaSense [43] and RF-net [44] leverage meta-learning, also known as *learn to learn*, to decrease the number of training samples. They separate the source dataset into several tasks, each representing a distinct environment, and train the model with a few epochs using tasks to teach the model how to adapt to a new environment. Hence, they only use a few samples to fine-tune the model when the model is deployed to a new environment. The works in [7], [46] leverage domain adaptation, which originates from transfer learning, to realize cross-environment sensing. Domain adaptation exploits adversarial learning to learn domain-invariant features with a number of source samples and a few target samples. The work in [47] leverages Siamese network based on CNN-LSTM and distribution discrepancy to solve environment dependency problem. However, it still needs a number of samples in source environment to pre-train the model and cannot fix the flaw of traditional generation strategy. These works still suffer from quantity and quality of source domain data. They can partly solve the environment dependency problem, however, they all need to collect a great number of data from similar environments, which still causes high time cost.



(a) (b)

(c) (d)

Fig. 1: The phase profiles when a user performs the 'Knock' and 'Down' gestures in different environments: (a) 'Knock' in a meeting room; (b) 'Knock' in a laboratory; (c) 'Down' in a meeting room; (d) 'Down' in a laboratory. The profiles for the same gesture are different in different environments.

3 Preliminaries

In this section, we investigate the environment dependency problem of RFID recognition systems. Moreover, we conduct some preliminary experiments to show the performance degradation of existing solutions when the environment changes and illustrate that approaches based on transfer learning cannot fully address this problem.

3.1 RFID Communication Model

In RFID communication, the received RF signal at the reader side can be represented as

$$S(t) = \sum_{i=1}^{N} A_i(t) e^{J[\frac{2\pi}{\lambda} \times 2d_i(t) + \gamma] \mod 2\pi} + n(t), \quad (1)$$

where $A_i(t)$ and $d_i(t)$ are the attenuation factor and the propagation distance of the i-th path at time t respectively, N is the number of reflection paths, $n(t)$ is the Gaussian noise, λ is the wavelength of RF signals, and γ is the phase offset caused by inherent characteristics of RFID tags and antenna.

Equation (1) implies that in different environments with different reflection objects, even though the user performs the same gesture, the received signals at RFID reader might vary. For example, the distance between the user and the antenna, the antennas' view angle, and the ambient reflection objects all result in distinct $d_i(t)$, and thus the signals might change. To illustrate this point, we ask a user to perform two gestures in a meeting room and a laboratory with the same setting and plot the phase profiles in different cases in Fig. 1. It can be observed that even though the user performs the same gesture, the profiles of the received phase values are significantly different in different environments, which means phase values collected from two different environments are out-of-distribution (OOD).

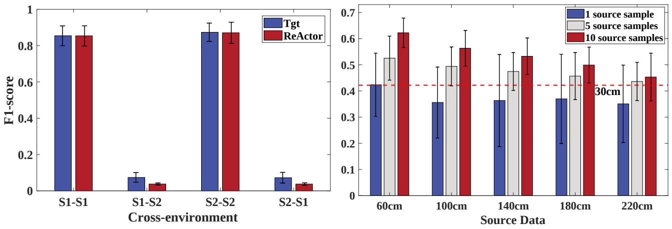
3.2 Environment Dependency

As discussed in Section 3.1, due to the multipath effects, the received signals might vary greatly even when the user performs the same gesture in different environments. Therefore, the performance of a system trained in an environment will degrade if it is used to classify samples from another environment. We conduct two experiments to validate the performance degradation of existing approaches when used in cross-environment cases.

To this end, we implemented a state-of-the-art RFID gesture recognition system *ReActor* [36] and a model called *Tgt* [43] which is based on CNN. We evaluated the F1-score of these approaches in two environments: $S1$ is an open meeting room, where the operating distance is 60 cm and the antenna view angle is 30°, and $S2$ is a laboratory, where the operating distance is 30 cm and the antenna view angle is 0°. We collect 10 samples of each gesture in the two environments (We considered 18 different gestures in this paper, as listed in shown in Fig. 11b). Note that each sample contains m phase series corresponding to m different tags. For example, a sample is a time-series data in the form of $\{p^1_{t(1,1)}, \dots, p^1_{t(1,k1)}; \dots; p^m_{t(m,1)}, \dots, p^m_{t(m,km)}; G_l\}$, where $p^i_{t(i,j)}$ is the j-th phase reading of the i-th tag and G_l is the label of the l-th gesture. Note that for different tags the time at which the phases are collected are different due to the randomness in RFID communication.

The cross-environment performance of the two approaches is plotted in Fig. 2a. In the figure, we denote *training set-test set* as a cross-environment scenario. Specifically, *S1-S2* denotes 70% samples from *S1* are used as training set and 30% samples from *S2* are used as test set, and *S1-S1* denotes 70% samples from S1 are as training set and 30% samples from S1 are as test set. As shown in Fig. 2a, the two models' F1-score are higher than 0.87 when training and testing with data from the same environment. However, when the test environment is different from the training environment, the F1-score drops sharply to below 0.2. This clearly demonstrates the dependency on the environment of existing solutions.

Some recent works tried to leverage transfer learning to retain the high accuracy when the systems are deployed to a new environment [15], [20]. Based on the theory of transfer learning [22], for similar tasks such as classifying the same set of gestures that have distinct data distribution, the shallow layers of neural network can extract task-general features that are the same across different tasks. On the contrary, the deep layers extract task-specific features which cannot generalize to other tasks. However, the environment dependency problem cannot be fully addressed by transfer learning. First, transfer learning needs a large number of samples in the source domain to pre-train the models, while collecting data is laborious and time-consuming. Second, the source data greatly influence the performance of transfer learning. If the data distributions of source domain and target domain are dissimilar, the transferred knowledge might be harmful to the models [23] and even result in negative transfer.

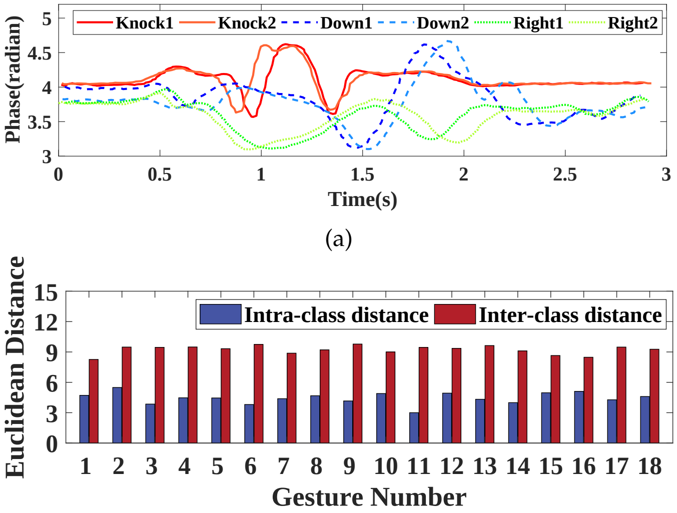


(a) F1-score at different distances.

(b) F1-score of transfer learning with different source data.

Fig. 2: The environment dependency problem: (a) the model performance deteriorates when used in cross-environment cases, and (b) approaches based on transfer learning cannot fully address the environment dependency problem.

We implement the transfer learning method proposed in [20] to validate this point. We collect 10 samples of each gesture from 6 different operating distances: 30 cm, 60 cm, 100 cm, 140 cm, 180 cm and 220 cm. Among these datasets, we define the dataset collected from 30 cm as the target dataset, and other datasets as source datasets. Next, we use each source dataset to train an independent model *Tgt* with 1, 5, and 10 samples of each gesture respectively. For the target dataset, we randomly select 1 sample of each gesture as training set, and 9 samples of each gesture as test set. After training with source data, we freeze the parameters of convolutional layers, and then use the training set of the target dataset to fine-tune the fully connected layers. Finally, we utilize the model to classify the test set of the target dataset. The results are plotted in Fig. 2b, in which the red dotted line denotes the F1-score of the model trained from scratch with one sample from 30 cm, which is 0.42. The figure shows that as the number of source samples increases, the F1-score also increases. However, users have to collect at least 5 samples such that transfer learning achieves an F1-score higher than 0.42, which is laborious. Moreover, the F1-score with data from 220 cm is 0.43, which is nearly equal to 0.42 even when we use 5 samples to pre-train the model because the data from the two distances are dissimilar and the model learns negative knowledge from the source data. These results indicate that if the source data are not similar to the target data or the source data are not enough, transfer learning's performance degrades. Moreover, the transferred knowledge might be deleterious in the target environment. How to select source data with better quality remains a problem in transfer-learning-based systems. These results motivate us to devise an accurate sensing system that can achieve high accuracy with only few samples in the new environment while avoiding negative transferring effects from the source data.



(a)

(b)

Fig. 3: The distance between the same gestures is shorter than that between different gestures: (a) The signals of 'Knock', 'Down', and 'Right'. Each gesture has two samples. (b) Intra-class distance and inter-class distance of each kind of gesture.

4 Detailed Design of RF-Siamese

4.1 Overview of RF-Siamese

We leverage the Siamese network to realize RF-Siamese. The Siamese network calculates the Euclidean distance between two inputs' features and discriminates them by the distance. The intuition is that samples of the same class from the same environment follow a similar data distribution. Hence, the distance between two samples from the same class is shorter than that from different classes. As shown in Fig. 3a, we plot two samples of three different gestures, namely *Knock*, *Down* and *Right*. The signals of the same gesture are more similar than that of different gestures. Furthermore, in Fig. 3b, we plot the intra-class distance and the inter-class distance of each gesture. The intra-class distance means the average distance among all the samples belonging to the same gesture, while the inter-class distance means the average distance among samples in one gesture and all other samples[1]. It can be clearly observed that for each kind of gesture, the inter-class distances of all gestures are larger than the intra-class distance. In spite of this, when training samples are limited, it is hard for the traditional classifier to learn a clear decision boundary to classify samples.

Fig. 4 shows the framework of RF-Siamese. First, RF-Siamese preprocesses the raw phase values to smooth the data, mitigate environmental noises, segment the active signals, and then transforms the processed phase values to spectrograms via short-time fourier transform. After preprocessing, a dataset generation strategy is designed to generate input to the Siamese network which can make full use of the samples. RF-Siamese then trains the Siamese network with the generated dataset, in which we carefully select the loss functions and distance metrics that are most suitable for RFID gesture recognition. Finally, RF-Siamese constructs a test dataset with all training samples and test samples, inputs the test data to the trained Siamese network, and classifies the samples by template matching.

4.2 Signal Preprocessing

4.2.1 *Filtering and Smoothing*

The raw phase values contain noises caused by ambient reflection objects. We leverage the Savitzky-Golay filter to filter out these noises and smooth the phase values as in the existing work. Savitzky-Golay filter is based on the method of polynomial least square, and has been extensively utilized in signals denoising and smoothing since it can preserve the signals' original shape and width after filtering [35], [36].

4.2.2 *Signal Normalization*

To mitigate the influence of absolute values, we leverage the Min-Max normalization method to map the phase values to the range of $[0, 1]$. Min-Max normalization can magnify the signals' fluctuation caused by gestures and rid of the effect of absolute values. For a given tag T, we denote its phase values ϕ_T by $\{\phi_1, \ldots, \phi_n\}$, where n is the length of ϕ_T. Take the i-th phase ϕ_i $(1 \leq i \leq n)$ as an example, the normalized value is calculated as

$$\widetilde{\phi_i} = \frac{\phi_i - \phi_{min}}{\phi_{max} - \phi_{min}}, \tag{2}$$

where ϕ_{max}, and ϕ_{min} are maximum and minimum of ϕ_T, respectively.

4.2.3 *Interpolation and Segmentation*

Due to collision, the lengths of phase series collected from different tags might vary. In our experimental setting, the sample rate of the reader in the *MaxThroughput* mode is around 86Hz. Hence, we leverage linear interpolation to interpolate the phase series with a 86Hz sample rate. Furthermore, to segment the activate signal, i.e., the signal when user is performing the gestures, we leverage the Modified Varri method used in [35]. The method calculates two values $A_m = \sum_{n=1}^{N} |x_k|$ and $F_m = \sum_{n=1}^{N} |x_k - x_{k-1}|$ within a sliding window, where N is the length of the window and x_k is the k-th phase value. By sliding the window, we segment the active signal with the largest $\mathcal{G}$, which is denoted as

$$\mathcal{G}(m) = \mathcal{C}_{\mathcal{A}}|\mathcal{A}_{m+1} - \mathcal{A}_m| + \mathcal{C}_{\mathcal{F}}|\mathcal{F}_{m+1} - \mathcal{F}_m|, \tag{3}$$

where m is the m-th window. Empirically, we set $\mathcal{C}_{\mathcal{A}}$ and $\mathcal{C}_{\mathcal{F}}$ to 1. The length of window is set to 90 since it takes about 1 second to finish a gesture and the reader collects about 90 phase values per second.

4.2.4 *Short-time Fourier Transform*

To extract features of time domain and frequency domain, we utilize short-time fourier transform to transform the time series data to time-frequency spectrograms based on

$$STFT(t, f) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j2\pi f\tau} d\tau, \tag{4}$$

[1] The equations to calculate intra-calss distance and inter-class distance are given in Eq. (8) and Eq. (9), respectively.

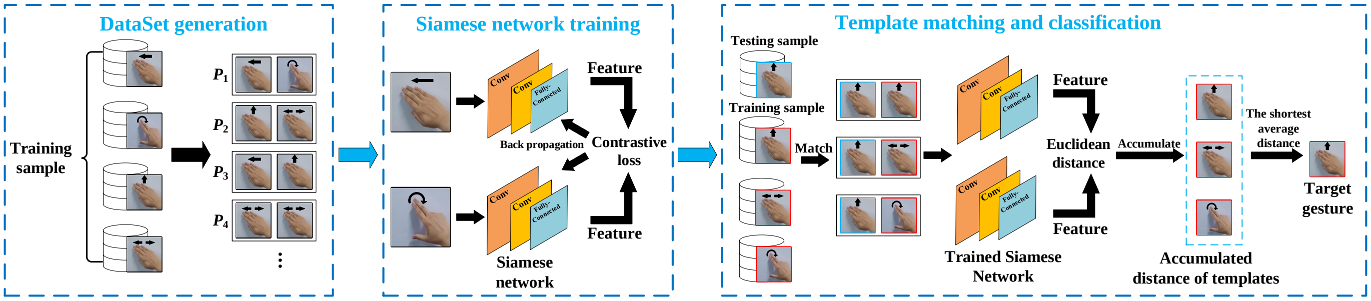


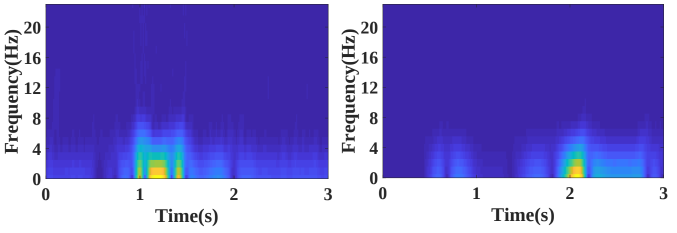
Fig. 4: Framework of RF-Siamese.

where $x(\tau)$ denotes the phase values, $h(\tau-t)$ is the window function. In RF-Siamese, we utilize a Hamming window whose length L_{window} is 48, and the length to overlap between segments is $L_{window}-1$. The length of FFT is 48. Fig. 5 shows two spectrograms of *Knock* and *Push* respectively. To fully use characteristics of tags' spectrograms, RF-Siamese concatenates the spectrograms along the horizontal direction as input to the Siamese network.

4.3 Structure of the Siamese Network

In RF-Siamese, we utilize the Siamese network to calculate the Euclidean distance between two samples' features. Siamese network has been extensively used to detect whether two samples are in the same class or not in various fields [24]–[27]. As shown in Fig. 6, the Siamese network takes two samples as input simultaneously, extracts features from the samples, computes the Euclidean distance between the features, and updates the shared parameters with the contrastive loss function.

The Siamese network consists of two neural networks that share the parameters. For each network, it processes one input and maps the input to the feature space. If the networks do not share the parameters, they might map the samples from the same gesture to a different feature space due to the difference of models' initial parameters. By sharing the network's parameters, the two networks can extract similar features. The order of two samples does not matter (i.e., Siamese network's output of input (X_1, X_2) is the same as input (X_2, X_1). The backbone network extracts local features in both time domain and frequency domain from the spectrograms and uses fully connected layers to flatten the features. We discuss the concrete backbones in Section 5.3. Then the Siamese network computes the Euclidean distance of the features, which is regarded as the similarity between two inputs. Theoretically, the distance should be smaller if the inputs are from the same gesture and becomes larger if the inputs are from different gestures. In this way, the Siamese network can discriminate samples with high accuracy even though when there are only a few samples of each gesture available.



(a) (b)

Fig. 5: (a) Spectrogram of *Knock*. (b) Spectrogram of *Push*. The spectrograms vary in both time domain and frequency domain.

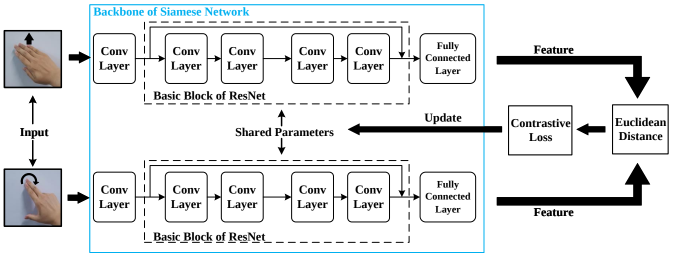


Fig. 6: Structure of the Siamese network used in RF-Siamese, whose backbone is a 6 layers ResNet.

4.4 Dataset Generation Strategy

Because only a few training samples are available for each gesture, how to efficiently use the samples to generate a training dataset is one challenging problem. Traditional Siamese network randomly chooses two samples from the training samples and combines them into a pair denoted as (Y, X_1, X_2), where X_1 and X_2 are the samples and Y denotes whether X_1 and X_2 are from the same class. We give an example of traditional strategy in the left side of Fig. 7. Suppose that there are 4 samples. For each sample, the strategy first decides whether to choose the same gesture or different gestures with equal probability, and then randomly selects a gesture from the same or different gestures to pair with it. In this intuitive strategy, however, some training samples might not be chosen because samples are selected randomly and each sample is selected only once, resulting in performance degradation when classifying samples of classes that are not selected.

To address this problem, we propose a new dataset generation strategy. As shown in the right side of Fig. 7, we utilize the full permutation to derive the dataset. For a sample in the training sample set, we pair it with all the samples in the training set once, including itself. Note that since the order of the samples in the pair does not matter, we skip the same pair. For instance, we only select either the pair of (A, B) or the pair of (B, A). The full permutation-based strategy avoids the drawback in the traditional strategy (i.e., some samples are ignored) and guarantees all the

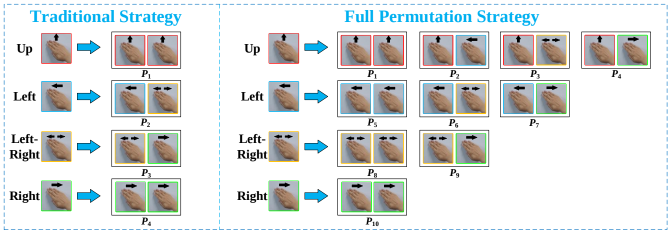


Fig. 7: Traditional strategy vs. Full permutation strategy: The traditional strategy neglects some sample pairs, while dataset generated by the full permutation strategy covers all possible sample pairs.

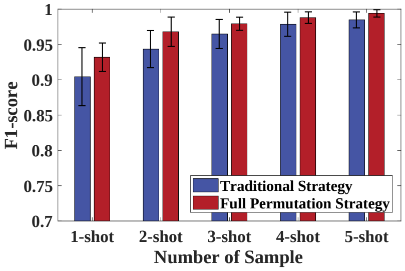


Fig. 8: F1-score with traditional strategy and permutation-based strategy.

samples are used efficiently. The full permutation might slightly increase cost in preparing the training set. However, because only a few samples are available, e.g., one or two samples of each gesture, the size of the generated dataset is acceptable. We present the time delay in TABLE 1, in which the time delay of generating the dataset denotes the time spent in generating all pairs and the time delay of each epoch spans from the time when the data loader loads the first batch to the time when the data loader loads the last batch of data. With one sample for each gesture, the time delay of the dataset generation process is 0.04 second and each epoch lasts 0.11 second on average. Even with five samples of each gesture, the time delay is 0.93 second and 1.10 second respectively. Since we train the network for 50 epochs, the total time is about 5 second with one sample of each gesture and less than 1 minute with 5 samples of each gesture, which is acceptable. Fig. 8 plots the F1-score with traditional strategy and the full permutation-based strategy. In all cases, the performance improves when the proposed full permutation-based strategy is used. The F1-score is improved by 3 percent when only one sample is used.

TABLE 1: The average time cost of each epoch during the training process with different training samples.

Shot Delay(s)	1	2	3	4	5
DataSet Generation	0.04	0.15	0.34	0.59	0.93
Epoch	0.11	0.23	0.44	0.73	1.10

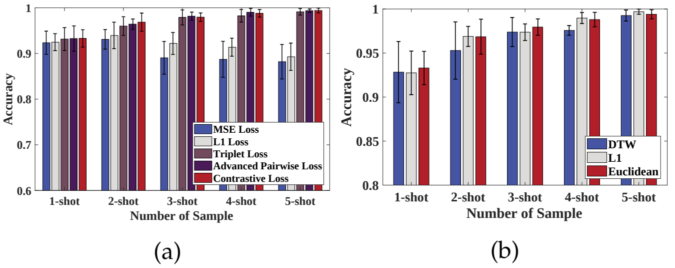


Fig. 9: (a) Accuracy of RF-Siamese with different loss functions. (b) Accuracy of RF-Siamese with different distance metrics in template matching.

4.5 Contrastive Loss Function

As we mentioned in Section 4.1, the key of RF-Siamese is to drive the distance of samples from the same gesture to be smaller than that from different gestures. However, most prevalent loss functions in regression problem (e.g., L_1 loss and MSE loss) do not satisfy this requirement. Therefore, we leverage the contrastive loss function in RF-Siamese, which is defined as

$$L(W) = \sum_{n=1}^{N} L(W, (Y_n, S_n))$$
$$= \frac{1}{2}\sum_{n=1}^{N}(1-Y_n)(D(W,S_n))^2 + Y_n[\max(0, m - D(W,S_n))]^2, \tag{5}$$

where W is the parameters of the Siamese network, N is the number of input pairs, $S_n = (X_n^1, X_n^2)$ is the n-th input pair consisting of two inputs X_n^1 and X_n^2, Y_n denotes whether X_n^1 and X_n^2 belong to the same gesture (i.e., $Y_n = 0$ if X_n^1 and X_n^2 belong to the same gesture and $Y_n = 1$ if not), $D(W, S_n)$ represents the Euclidean distance of features of two inputs, and m is the *margin* of contrastive loss, restricting the loss to be smaller than m.

RF-Siamese attempts to find out the network's parameters W to minimize the contrastive loss with the Adam optimizer [48]

$$W = \underset{W}{argmin} \sum_{n=1}^{N} L(W, (Y_n, S_n)). \tag{6}$$

The contrastive loss function is divided into two parts by Y_n. If $Y_n = 0$ (i.e., two samples belong to the same gesture), the loss function is $(D(W,S_n))^2$. If $(D(W,S_n))^2$ is large, it means the network is not suitable, so W is updated to make $(D(W,S_n))^2$ smaller. If $Y_n = 1$, the function is $[\max(0, m - D(W,S_n))]^2$. Hence, if $D(W,S_n)$ is small, it deviates from the principle that the distance of samples from different gestures is large, so the loss increases. After training with contrastive loss, the features' distance of the same gesture will be much smaller than that of different gestures. Therefore, RF-Siamese can differentiate samples by the distance.

We compare the accuracy of contrastive loss with L_1 loss, MSE loss, advanced pairwise loss [47], and triplet loss [49] and plot the result in Fig. 9a. It can be observed that the accuracy of these loss functions is close with one sample for each gesture. When the number of samples

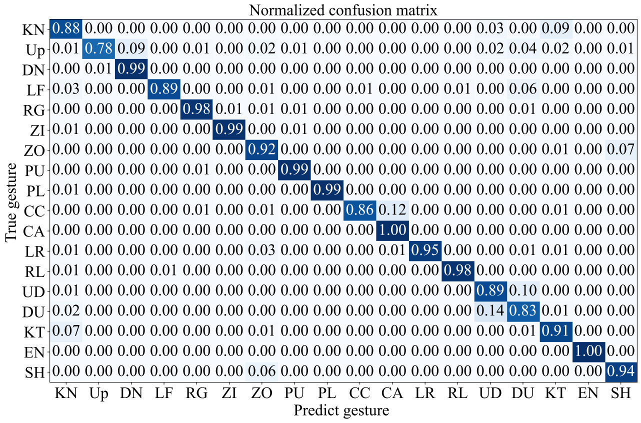


Fig. 10: Confusion matrix of RF-Siamese with one sample.

increases, the accuracy of advanced pairwise loss, triplet loss and contrastive loss keeps increasing while others decrease. The reason is that these 3 loss functions can drive the gap between the same gesture and different gestures larger, and thus informative features can be extracted to recognize the gestures. However, other loss functions do not satisfy this bias. Therefore, even though the number of samples increases, the distances between the same gesture and distinct gestures are still too close to classify the sample, causing a drop in accuracy .

Compared with contrastive loss, advanced pairwise loss need to balance the importance of 2 hyperparameters, named m and b, causing excessive training cost. Moreover, the accuracy of triplet loss and contrastive loss is close, but we select contrastive loss as the loss function in RF-Siamese due to two reasons. First, thanks to the full-permutation strategy, the contrastive loss can push all different samples away, which fixes the flaw of contrastive loss with traditional generation strategy. Therefore, with the full-permutation strategy, the performance of contrastive loss can reach as high as the performance of triplet loss. Second, the number of data pairs generated by the full-permutation strategy in contrastive loss is less than that in triplet loss since a triplet contains an anchor, a positive sample, and a negative sample. Thus the training overhead of triplet loss is higher than that of contrastive loss.

4.6 Multiple Gesture Classification

The traditional Siamese network is a binary classification model. By computing the distance between features of two inputs, it discriminates whether these two inputs belong to the same class or not. Hence, it cannot be directly applied to classify multiple gestures. In RF-Siamese, we modify the Siamese network to a multi-classification model based on template matching. In specific, after model training, all training samples are selected as templates. For a test sample, it pairs with all templates. These pairs are fed into the trained Siamese network, and the Siamese network will output the corresponding distance of each pair. Then we select the class with the minimum average distance as the target class of the test sample. Fig. 4 shows an example in *template matching and classification* part. Assume that there is a test sample of *Up* and 3 templates of *Up*, *Left-right*, and *Circle clockwise*. The test sample first matches with the templates, and thus 3 input pairs are generated. The pairs are input to the trained Siamese network and the network outputs distance of each pair. Then we accumulate the distance for each gesture of templates. Finally, the average accumulated distance of *Up* is the shortest, and thus *Up* is the target gesture. The time complexity of template matching is $O(N)$, where N is the number of templates. Compared with traditional template matching [35], this method can bind the network parameters to the distance. Concretely, the distances between templates and test samples are not computed on original data directly, but on the features extracted by the network, where the Siamese network has reshaped the distance via loss.

To determine which type of distance metric is the most suitable, we investigate three commonly used distance metrics and plot corresponding F1-score in Fig. 9b, namely L1 distance, Euclidean distance, and dynamic time wrapping (DTW) distance [35], [50], respectively. The figure shows that the F1-score of these 3 matching distances are close. However, the time complexity in computing DTW distance is much higher than the other two distances, and thus it costs several seconds to classify a test sample in our experiment, which is not feasible for real-time application [36]. Hence, we decide to accumulate the Euclidean distance of the pairs to corresponding gestures of training samples, and finally classify the test sample as the class with the minimum average Euclidean distance. Note that using Manhattan distance is also adequate in RF-Siamese. The confusion matrix of RF-Siamese with one sample of each gesture is shown in Fig. 10. All the gestures' accuracy is higher than 0.83 except the gesture "Up". The accuracy of some gestures such as *circle anticlockwise* and *enlarge*, can reach an accuracy as high as 1. The average accuracy of all gestures is 0.93.

5 Performance Evaluation

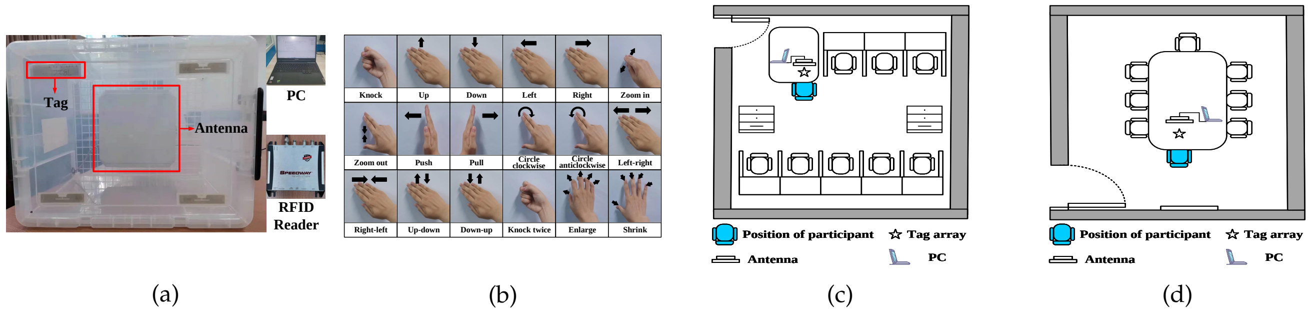
5.1 Experiments Setup and Dataset Collection

Implementation: The hardware of RF-Siamese is shown in Fig. 11a. It consists of a Lenovo Laptop, an Impinj Speedway R420 reader, and a circularly polarized Laird S9028PCL antenna. We attach four Monza AZ-9654 RFID passive tags at the corners of a transparent cover of a plastic box, whose size is 53cm*39cm*32cm, and fix the antenna at the bottom of the box such that the antenna and the tags are in line-of-sight. Empirically, it takes about 1 second to complete a gesture, so we collect phase values for 3 seconds with 20dBm antenna power in the *MaxThroughput* mode. We use Octance Sdk 3.4.0.0 to collect phase values and implement RF-Siamese using the PyTorch 1.2.0 framework. The model is trained in a server equipped with 4 NVIDIA TITAN V GPUs and 256 GB memory with Intel(R) Xeon(R) Silver 4114 2.20GHz processors.

Metrics: We use accuracy and macro F1-score to evaluate the performance of RF-Siamese, which can be formulated as

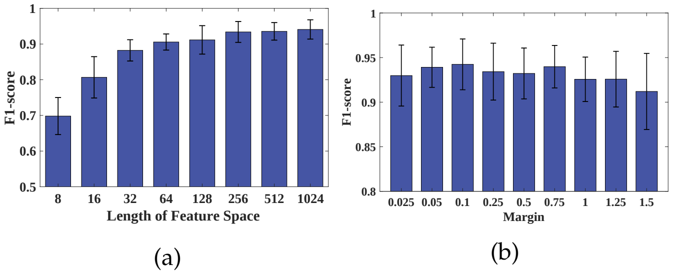
$$F1 = \frac{1}{K}\sum_{k=1}^{K} 2 \times \frac{P_k \times R_k}{P_k + R_k}, \tag{7}$$

where K is the number of gesture classes, P_k and R_k is the *precision* and *recall* of the k-th class respectively.



(a) (b) (c) (d)

Fig. 11: (a) RF-Siamese's hardware. (b) 18 gestures. (c) Layout of the laboratory. (d) Layout of the meeting room.



(a) (b)

Fig. 12: F1-score with different parameters of Siamese network: (a) F1-score vs. length of feature space. (b) F1-score vs. different margin value (m).

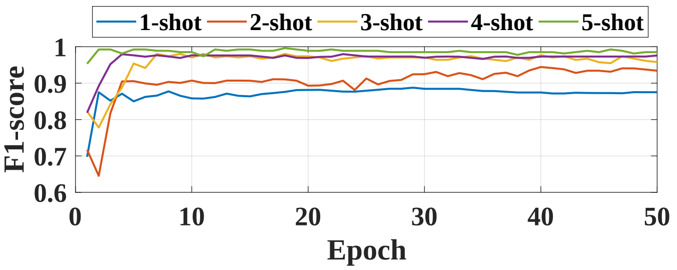


Fig. 13: F1-score of RF-Siamese vs number of epochs.

For the p-th gesture G_p, we define *inter-class distance* and *intra-class distance* of G_p as

$$d_{inter}(p) = \frac{1}{K-1}\sum_{q\neq p}\Big[\frac{1}{N_p N_q}\sum_{X_i\in S_p, X_j\in S_q}||f(X_i)-f(X_j)||_2^2\Big], \tag{8}$$

and

$$d_{intra}(p) = \frac{1}{N_p^2}\sum_{X_i, X_j\in S_p}||f(X_i)-f(X_j)||_2^2, \tag{9}$$

where S_p is the set of samples corresponding to G_p, N_p is the number of samples in S_p, and $||\cdot||_2^2$ denotes the Euclidean distance between the features of two samples X_i and X_j after processing by the Siamese network. We also leverage the ratio $d_{intra}(p)$ of to $d_{inter}(p)$ to illustrate the performance of RF-Siamese, which is denoted as

$$Ratio(p) = \frac{d_{intra}(p)}{d_{inter}(p)}. \tag{10}$$

The smaller $Ratio(p)$, the higher capability of the approach in classifying different gestures with small number of samples. All the results reported here are averaged over 100 independent runs.

Parameters setup: We set the kernel size of CNN to 3 with stride 1. As for the length of features, some information might be lost if the length is too short, while it costs more computational cost if the length becomes longer. As shown in Fig. 12a, as the length of feature space increases, RF-Siamese's F1-score also increases. We set the length to 512 since the longer length of feature causes more training overhead while the increase in F1-score is marginal. The margin m in Eq. 5 is set to 0.1 as the F1-score of RF-Siamese diminishes when $m \geq 1$ as shown in Fig. 12b. In RF-Siamese, we set the training epoch to 50 as it converges to the best performance as shown in Fig. 13. The batch size and learning rate are set to 128 and 0.001, respectively.

Dataset: The 18 gestures to be classified are shown in Fig. 11b. Ten users are invited to participate in our experiments, and their ages range from 22 to 28. Participants are required to perform 18 gestures 10 times repeatedly in 4 operational distances, 5 antenna view angles, and 2 distinct environments. We perform the experiments in two different rooms, one laboratory and one meeting room, the layout of which are shown in Fig. 11. The participants perform the gestures based on their own interpretation without guidelines. For each environment, we randomly select s samples of each gesture as training set, where s ranges from 1 to 5. When s samples are used, the corresponding results are denoted as s-shot learning. After selecting the training set, the remaining samples of each gesture are used as test set.

5.2 Baselines

We compare RF-Siamese with 7 state-of-the-art solutions, which are *Tgt* [43], *ReActor* [36], *Prototypical Network* (PN) [51], *Matching Networks* (MN) [52], *Transfer Learning (TL)* [21], *MetaSense* [43], and *RF-Net* [44].

Tgt: *Tgt* is a 4 layers ResNet-like network with 1 basic block, and the kernel size is 3.

ReActor: *ReActor* extracts 13 statistical features (e.g., *min*, *max*, *mode*, etc) and wavelet decomposition coefficients manually as the features, and leverages random forest to classify the gestures.

TL: TL is a transfer-learning-based method [21], and the backbone is *Tgt*. After training with source samples, the CNN is frozen and the fully connected layers are finetuned with target samples.

PN: Prototypical network generates prototypes of each gesture in an embedding space and calculates the Euclidean

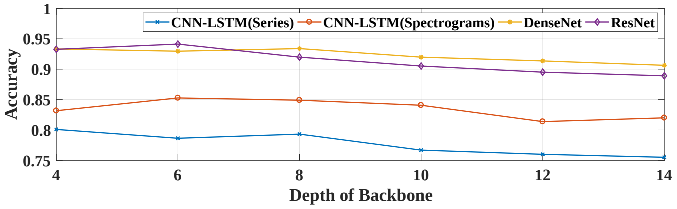


Fig. 14: Accuracy of RF-Siamese with different backbones and network depths.

distance between the test samples and prototypes, and then classifies the samples to the closest gesture of prototypes.

MN: Matching networks embed the input, including training samples and test samples, to a high-dimensional feature space via attention mechanism modules, and calculating the Cosine distances between features of training and test samples.

MetaSense: MetaSense is a meta-learning-based method. It leverages MAML [53] as its meta-learning scheme. By splitting source environments' dataset into sub-tasks and combing the sub-tasks as a new environment, MetaSense enables model to learn how to adapt to a new environments with several samples of each activity.

RF-Net: RF-Net is a meta-learning-based method. It adopts a metric-based meta-learning framework, including a dual-path network to extract features from time domain and frequency domain.

5.3 Performance of Different Backbones

To evaluate the performance of different RF-Siamese's backbones, we leverage 3 typical network models as the backbone, namely CNN-LSTM, ResNet [54], and DenseNet [55]. We test the performance of CNN-LSTM with two different types of input, namely phase series and spectrograms. For phase series, we use one-dimensional CNN to extract the features. The experiment is 1-shot learning, which means for each environment we choose one sample of each gesture randomly to train the model. The results are shown in Fig. 14. It can be observed that as the depth increases, the performances of backbones increase and then drop. When the backbone is a 6 layers ResNet, the accuracy is 0.94, which is the highest. The reason is that in few-shot learning, the training samples are rare. Even though we use dataset generation strategy to create more inputs, the network still easily gets overfitted. Consequently, it is not a good choice to construct a deep backbone in few-shot learning. Furthermore, we observe the accuracies of ResNet and DenseNet are close, which are both higher than the accuracy of CNN-LSTM. We consider ResNet and DenseNet both can fully use the previous information of inputs, e.g., ResNet directly inputs the input to deeper layers, and DenseNet reuses features extracted by previous layers, and thus can extract more informative features to reach a higher accuracy. It can be observed that the accuracy of CNN-LSTM is the lowest. The reason is two-fold. First, phase series only reveals features of time domain and lacks features of frequency domain, resulting in low accuracy. Second, using a LSTM layer to extract features from spectrograms directly might obtain irrelevant features and cause a relatively low accuracy since LSTM is more effective in time series prediction.

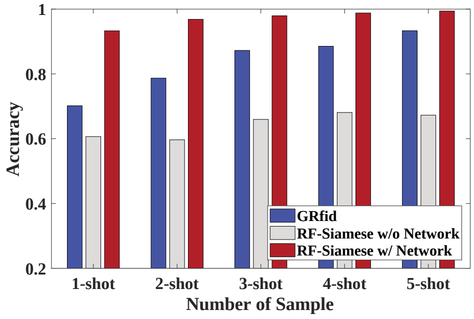


Fig. 15: Accuracy of GRfid [35] and RF-Siamese.

5.4 Performance of Short-time Fourier Transform

To validate the effectiveness of short-time fourier transform, we compare the performance of RF-Siamese when inputting phase series and spectrograms collected at 60cm respectively. The backbone remains unchanged and we use one-dimensional convolutional neural networks, whose kernel size is 5, to extract features from phase series. As shown in TABLE 2, the performance of using spectrograms is higher than that of using phase values. Specifically, the F1-score increases at least by 0.025 with 1 training sample of each gesture while at most by 0.067 with 5 training samples of each gesture. Since spectrograms can provide more fine-grained features in frequency domain without losing too much information in time domain, the Siamese network can extract more features to classify the samples. The results indicate the effectiveness and necessity of short-time fourier transform.

5.5 Performance of Contrastive Loss

To further figure out the effectiveness of contrastive loss, we calculate the ratio of raw data and features extracted by Siamese network based on Eq.10. We plot the ratio of raw data and ratio of features of 18 classified gestures in TABLE 3. It is noted that features' ratios of all gestures are smaller than that of raw data, which means the intra-class distance decreases and the inter-class distance increases via contrastive loss. Specifically, the average ratio of features is 0.329 while the ratio of raw data is 0.484, which means RF-Siamese decreases the ratio by 0.155. We also compare RF-Siamese with a straightforward template matching method GRfid [35], and explore the performance of RF-Siamese if we remove the Siamese network. As shown in Fig. 15, GRfid's accuracy is 0.7 with one sample of each gesture, which is 0.2 lower than that of RF-Siamese. This is because the template

TABLE 2: Accuracy with phase values and spectrograms respectively at 60cm.

Input	Number of Sample				
	1	2	3	4	5
Phase Values	0.776	0.818	0.849	0.872	0.884
Spectrograms	0.801	0.880	0.907	0.931	0.951

TABLE 3: Ratio of raw data and ratio of extracted features.

	Knock	Up	Down	Left	Right	Zoom in	Zoom out	Push	Pull
Ratio of raw data	0.570	0.579	0.407	0.471	0.478	0.391	0.493	0.507	0.425
Ratio of features	0.313	0.312	0.246	0.408	0.373	0.256	0.371	0.404	0.332
	CC	CA	Left-right	Right-left	Up-down	Down-up	Knock twice	Enlarge	Shrink
Ratio of raw data	0.543	0.317	0.527	0.449	0.438	0.574	0.602	0.451	0.493
Ratio of features	0.318	0.233	0.423	0.311	0.283	0.345	0.298	0.358	0.335

matching of RF-Siamese is correlated with the network and contrastive loss. The distance between test samples and templates is not computed on original data directly, but on features extracted by the network, which means if test sample and template belong to the same gesture, their feature's distance is closer than that of original data, and vice versa. Furthermore, if we remove the Siamese network, RF-Siamese deteriorates into traditional template matching method since the features' distance between samples will not be optimized by the Siamese network. Due to the random selection of templates, the performance of RF-Siamese without network degrades to 0.6 with one sample of each gesture, and increases marginally as the number of samples increases.

These results indicate that the Siamese network with contrastive loss can effectively gather samples of the same class while pushing samples of different classes away, and this is the biggest difference between RF-Siamese and traditional template matching method GRfid, where distances between templates and test samples are not optimized by the network. This enables RF-Siamese can extract more distinct and informative features, and thus RF-Siamese can recognize gestures with even one sample of each gesture. Therefore, RF-Siamese cannot work well without the Siamese network.

5.6 Impact of Tag-Antenna Deployments

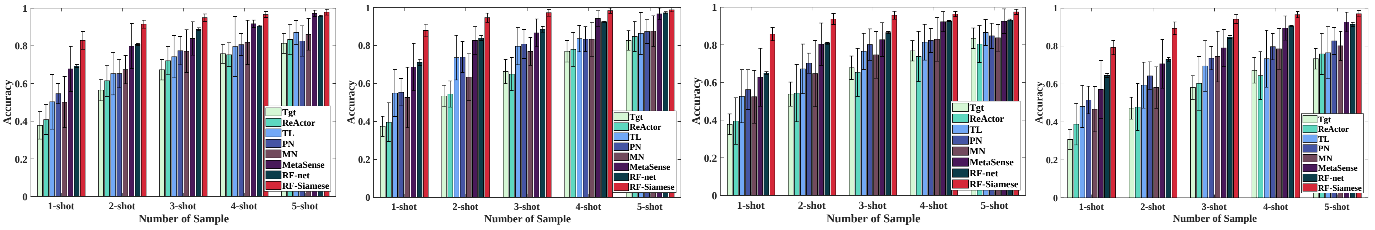
We evaluate RF-Siamese from the perspective of distinct tag-antenna deployments. Specifically, we vary the distances between tags and antenna from 60 cm to 150 cm. The results are reported in Fig. 16. RF-Siamese's average accuracy is 0.84, which is the highest among baselines with one sample of each gesture. Specifically, with one training sample of each gesture, the average accuracy of traditional sensing methods, i.e., *Tgt* and *ReActor*, is around 0.38, the accuracies of *TL*, *PN* and *MN* are 0.52, 0.54, and 0.51, respectively. Compared with *MetaSense* and *RF-Net*, RF-Siamese lifts the accuracy by 0.21 and 0.18. Note that performance of approaches degrades because of power attenuation. The results show that RF-Siamese can achieve a high accuracy under different tag-antenna deployments circumstances.

5.7 Performance at Different Operational Distances

We test the performance of the baselines and RF-Siamese at 4 different operational distances: 30 cm, 60 cm, 100 cm, and 140 cm. The distance is the distance between the user's hand and the cover of the plastic box. The results are shown in Fig. 17. It can be observed that the F1-score decreases as the distance increases due to power attenuation. RF-Siamese outperforms all baselines in all cases. With one sample of each gesture at 30cm, RF-Siamese's F1-score is 0.93, while the F1-score of *Tgt* and *ReActor* are 0.44 and 0.36 respectively. This is because *Tgt* and *ReActor* have to learn sufficient knowledge from a large number of training data, and thus it is difficult for them to achieve a high F1-score from one training sample of each gesture. Compared with *TL*, *PN* and *MN*, whose F1-score are 0.54, 0.58 and 0.62 with one sample of each gesture, RF-Siamese improves the F1-score by around 0.3. The F1-score of *MetaSense* and *RF-Net* are about 0.7, which are lower than RF-Siamese by 0.2 since the performance of meta-learning is highly correlated with source data. Moreover, we figure out that the error bars of *Tgt* and *PN* are relatively high, which means these methods suffer from vibration of recognition performance. We consider when training samples are limited, these two methods are easily overfitted with training set, causing performance degradation in test set. The Siamese network makes full use of limited training samples by dataset generation strategy and separates samples from different classes as much as possible with contrastive loss, so it suffers less from overfitting even though training samples are limited and is more stable in the application. Furthermore, as the distance increases, the F1-score of all methods decreases. Since we attach the RFID tags on the cover of the plastic box but not the user's hands, the RF signals are reflected by gestures indirectly. Hence, as the distance increases, the signals reflected by gestures become weaker, resulting in gestures' feature loss in RF signals. It is a bottleneck of contactless wireless sensing systems. Generally speaking, RF-Siamese outperforms the state-of-the-art solutions obviously with different number of samples of each gesture, implying RF-Siamese can achieve a high performance with only a few samples.

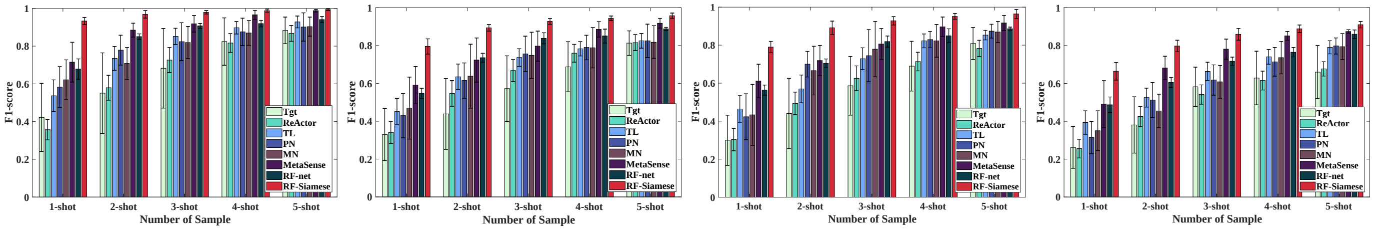
5.8 Impact of Antenna Angles

We evaluate the performance with different antenna angles. The antenna angle refers to the included angle of the antenna and the horizontal axis. For example, when the antenna is opposite to the user, the antenna angle is 0. We investigate 5 antenna angles at 30cm, including 30°, 60°, 0°, $-30°$, $-60°$. The results are shown in Fig. 18. Generally, RF-Siamese outperforms other baselines from 1 to 5 training samples of each gesture. With one sample of each gesture, RF-Siamese's average F1-score at 5 angles is 0.89, while the F1-score of *Tgt* and *ReActor* are 0.37 and 0.36, the F1-score of *TL*, *PN* and *MN* are 0.53, 0.54 and 0.56, and the scores of *MetaSense* and *RF-Net* are 0.73 and 0.69. With five samples of each gesture, RF-Siamese's average F1-score is 0.99, which is the highest among the baselines. The antenna angles do not influence the F1-score greatly since RF-Siamese uses a circularly polarized antenna to transmit RF signals, whose cover area is an orb. Therefore, even though the antenna



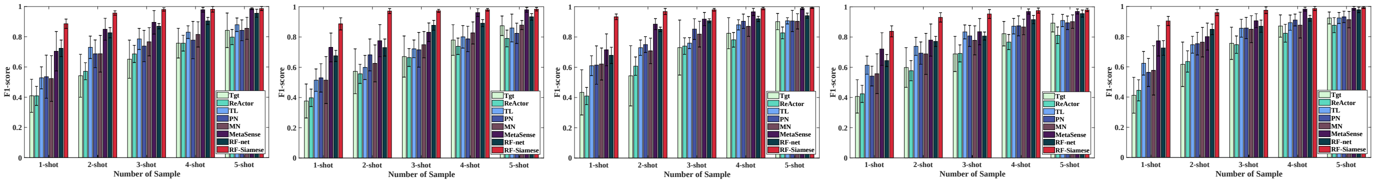
(a) 60cm. (b) 90cm. (c) 120cm. (d) 150cm.

Fig. 16: Accuracy of the RF-Siamese and other approaches at different tag-antenna deployments.



(a) 30cm. (b) 60cm. (c) 100cm. (d) 140cm.

Fig. 17: F1-score of the RF-Siamese and other approaches at different operational distances.



(a) 30°. (b) 60°. (c) 0°. (d) −30°. (e) −60°.

Fig. 18: F1-score of the RF-Siamese and other approaches at different antenna angles.

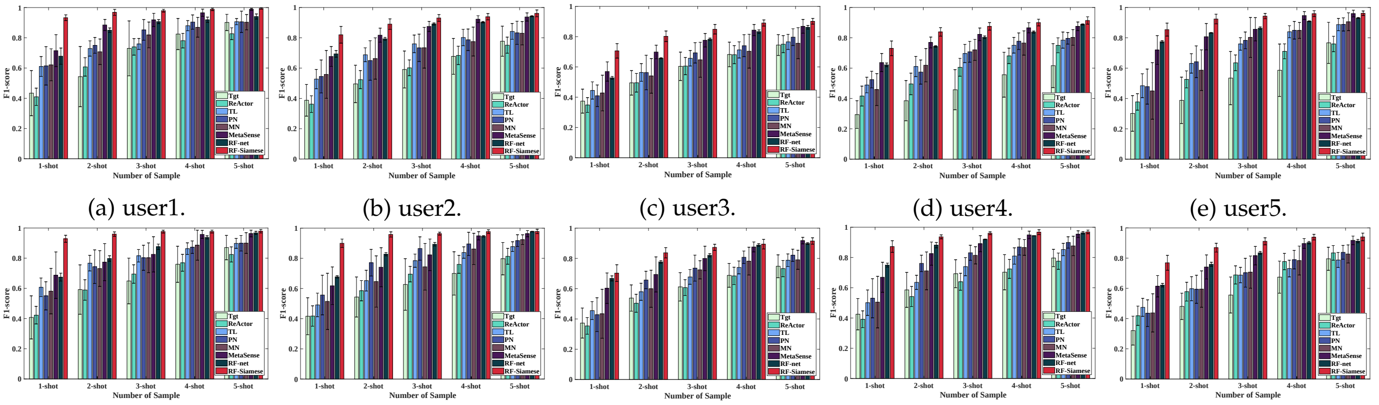
angle changes, the sensing area still covers users and the RF signals can be reflected by gestures. The results indicate RF-Siamese can adapt to environments with various antenna angles in an efficient way.

5.9 Performance of Different Users

Users' habits are unique when performing gestures. For instance, some users are left-handed while others are right-handed. Different habits cause different received signals. Fig. 19 plots the RF-Siamese's F1-score of 10 users at 30cm and 0°antenna angle, and the users perform the gestures based on their interpretation of gestures. The figure shows that RF-Siamese outperforms the baselines even though the F1-score of some users is lower than average. With one training sample of each gesture, RF-Siamese's F1-score is 0.70 at least, and 0.93 at most, which outperforms *Tgt* and *ReActor* by nearly 0.5, *TL*, *PN* and *MN* by 0.3, and *MetaSense* and *RF-Net* by 0.2. With five training samples of each gesture, RF-Siamese can achieve as high as 0.99 with user 1, and the lowest F1-score is 0.90 with user 3. Note that there is a fluctuation of F1-score in different users since users are required to perform the gestures based on their interpretation, and thus some gestures might be performed ambiguously, resulting in performance fluctuation. Nevertheless, RF-Siamese reaches a higher F1-score compared with other baselines. The results indicate RF-Siamese can be applied to different users efficiently.

5.10 Performance of Different Environments

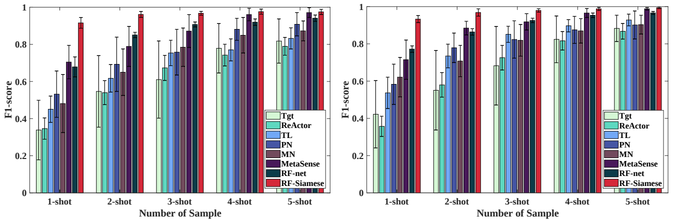
There are unique reflection objects in distinct environments, resulting in the difference of received wireless signals. Therefore, we investigate RF-Siamese in two distinct environments, which are an open meeting room and a crowded laboratory respectively. The meeting room represents simple multipath environment and the laboratory represents complex multipath environment since there are more reflectors around the antennas and tag array. Note that we use data from these two environments to train each other in transfer learning, and we set the distance to 30cm and the antenna angle to 0°. The F1-score in the laboratory is shown in Fig. 20a. With one sample of each gesture, RF-Siamese's F1-score is 0.91, and 0.97 with five samples of each gesture. *ReActor*'s F1-score is 0.35 with one sample of each gesture and 0.79 with five samples of each gesture. Interestingly, transfer learning's F1-score is only 0.45 with one sample of each gesture and 0.83 with five samples of each gesture. It performs worse than *PN* and *MN*, whose F1-score is 0.53 and 0.50 respectively. This is because the source data distribution is not similar to the target data distribution, and thus the model learns harmful prior knowledge from the data collected from the meeting room. The scores of *MetaSense* and *RF-Net* are 0.70 and 0.68 respectively with one sample of each gesture. As shown in Fig. 20b, RF-Siamese achieves as high as 0.93 and 0.99 with one and



(a) user1. (b) user2. (c) user3. (d) user4. (e) user5.

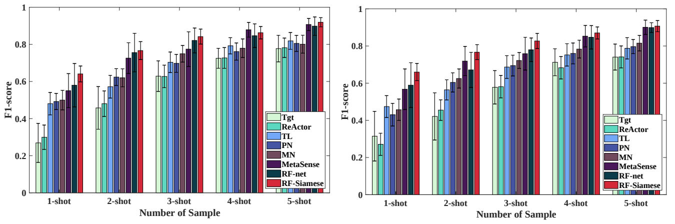
(f) user6. (g) user7. (h) user8. (i) user9. (j) user10.

Fig. 19: F1-score of the RF-Siamese and other approaches with different users.



(a) (b)

Fig. 20: F1-score in different environments: (a) F1-score in the laboratory. (a) F1-score in the meeting room.



(a) (b)

Fig. 21: F1-score with external interference: (a) One participant walks around the user. (a) Two Participants walk around the user.

five samples of each gesture in the meeting room, which outperforms *ReActor* by nearly 0.50 and 0.12, respectively. Transfer learning's F1-score is 0.53 with one sample of each gesture, and 0.92 with five samples of each gesture. Moreover, the F1-score of *MetaSense* and *RF-Net* are 0.72 and 0.77 respectively. The two experiments show that although transfer learning and meta-learning have been extensively used to handle the environment dependency problem, they might cause degradation of models when data distribution of source domain is dissimilar to data distribution of target domain, and thus it is not always the best solution.

5.11 Performance with External Interference

We investigate the performance of RF-Siamese with external interference. Specifically, we ask participants walk casually behind the user, and the distance between participants and the user is about 1.5 m. The number of walking people varies from 1 to 2. Since the walking people also reflect the RF signals, they can be regarded as external interference. The results are shown in Fig. 21. We note that due to the external interference, the performance of all baselines, including RF-Siamese drops greatly. For example, the F1-score of RF-Siamese is only 0.65 in 1-shot learning and 0.92 in 5-shot learning. The reason is that all these approaches do not take external interference into consideration. They mainly focus on how to reduce training cost. Hence, when external interference happens, the performance of these approaches decreases. Nevertheless, RF-Siamese still reaches a relatively high performance compared with other baselines since its strong ability of extracting informative features still works even with external features.

6 Limitations and Future Works

RF-Siamese realizes an accurate gesture recognition method with only one a few samples, yet there are still some important aspects which can be further addressed. We discuss the limitations of RF-Siamese as follows.

Cross-environment performance. The goal of RF-Siamese is to recognize gestures accurately with a few samples. However, while achieving high accuracy with only a few samples, RF-Siamese needs to retrain the model from scratch in each environment. It does not exploit the potential knowledge embedded in samples collected from other environments to further enhance the robustness and accuracy in gesture recognition. Some works based on transfer learning [15], [20] or meta-learning [43], [44] try to reuse such samples, but their accuracy are limited because their performances are highly correlated with the quantity and quality of source data. To overcome this problem, one possible solution is to find domain-independent features and reuse the features to reduce the training cost in the new environment. In the future, we plan to combine few-shot learning and domain-independent features to enhance the robustness and cross-environment performance of RF-Siamese in different environments.

Speed Variance. RF-Siamese cannot well handle very fast/slow gestures because it uses a fixed-length sliding window in the current implementation. If the user performs the gestures too fast, the length of active signals will be shorten and thus irrelevant signals are introduced to the sliding window. Likewise, a sliding window with fixed size might lose active signals if user performs the gestures too slow. An intuitive method is resizing the sliding window's size dynamically, however, resizing the window's size is not a good method in our model because we use STFT to process the active signals, in which the size of spectrograms depends on the length of active signals. If the size of spectrograms changes, it will result in a failed matching of the Siamese network. A potential method is to use NLP networks such as Transformer since these models can deal with inputs with various sizes.

7 Conclusion

We propose RF-Siamese, an RFID gesture recognition system based on the Siamese network. RF-Siamese achieves a high accuracy with only a few samples, which can reduce the time cost in data collection. To adapt to RFID sensing, we carefully devise the structure and parameters of the Siamese network, propose a new dataset generation strategy, and modify the Siamese network to a multi-classification model based on template matching such that it can classify 18 gestures. We implement RF-Siamese in the real environment and conduct extensive experiments to investigate its performance. RF-Siamese achieves an accuracy of 0.93 with only one sample of each gesture, while the state-of-the-art traditional method only achieves an accuracy of 0.44, the approach based on transfer learning only achieves an accuracy of 0.59, and two meta-learning-based approaches achieve an accuracy of 0.7.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Grant Nos. 61772559, 62172154, 61872310, 62072231), the Hunan Provincial Natural Science Foundation of China under grant No. 2020JJ3016, Shenzhen Science and Technology Innovation Commission (JCYJ20200109142008673).

References

[1] Ju Wang, Jianyan Li, Mohammad Hossein Mazaheri, Keiko Katsuragawa, Daniel Vogel, and Omid Abari. Sensing finger input using an RFID transmission line. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 531–543, 2020.

[2] Han Ding, Lei Guo, Cui Zhao, Fei Wang, Ge Wang, Zhiping Jiang, Wei Xi, and Jizhong Zhao. RFnet: Automatic gesture recognition and human identification using time series RFID signals. *Mobile Networks and Applications*, 25(6):2240–2253, 2020.

[3] Haoyu Wang and Wei Gong. RF-pen: Practical real-time rfid tracking in the air. *IEEE Transactions on Mobile Computing*, 2020.

[4] Shigeng Zhang, Zijing Ma, Chengwei Yang, Xiaoyan Kui, Xuan Liu, Weiping Wang, Jianxin Wang, and Song Guo. Real-time and accurate gesture recognition with commercial RFID devices. *IEEE Transactions on Mobile Computing*, pages 1–16, 2022.

[5] Lina Yao, Quan Z Sheng, Xue Li, Tao Gu, Mingkui Tan, Xianzhi Wang, Sen Wang, and Wenjie Ruan. Compressive representation for device-free activity recognition with passive RFID signal strength. *IEEE Transactions on Mobile Computing*, 17(2):293–306, 2017.

[6] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall S Burd. Deep learning for RFID-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, pages 164–175, 2016.

[7] Fangxin Wang, Jiangchuan Liu, and Wei Gong. Multi-adversarial in-car activity recognition using RFIDs. *IEEE Transactions on Mobile Computing*, 20(6):2224–2237, 2020.

[8] Shigeng Zhang, Zijing Ma, Kaixuan Lu, Xuan Liu, Jia Liu, Song Guo, Albert Y. Zomaya, Jian Zhang, and Jianxin Wang. HearMe: Accurate and Real-time Lip Reading based on Commercial RFID Devices. *Accepted to appear in IEEE Transactions on Mobile Computing*, PP(99):1–13, 2022.

[9] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. RFID tattoo: A wireless platform for speech recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–24, 2019.

[10] Chao Feng, Jie Xiong, Liqiong Chang, Fuwei Wang, Ju Wang, and Dingyi Fang. RF-identity: Non-intrusive person identification based on commodity RFID devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–23, 2021.

[11] Yuancan Lin, Lei Xie, Chuyu Wang, Yanling Bu, and Sanglu Lu. DropMonitor: Millimeter-level sensing for RFID-based infusion drip rate monitoring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–22, 2021.

[12] Biyun Sheng, Fu Xiao, Letian Sha, and Lijuan Sun. Deep spatial–temporal model based cross-scene action recognition using commodity WiFi. *IEEE Internet of Things Journal*, 7(4):3592–3601, 2020.

[13] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. msense: Towards mobile material sensing with a single millimeter-wave radio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–20, 2020.

[14] Lei Xie, Chuyu Wang, Yanling Bu, Jianqiang Sun, Qingliang Cai, Jie Wu, and Sanglu Lu. Taggedar: An RFID-based approach for recognition of multiple tagged objects in augmented reality systems. *IEEE Transactions on Mobile Computing*, 18(5):1188–1202, 2018.

[15] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. Crosssense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, MobiCom '18, page 305–320, New York, NY, USA, 2018. Association for Computing Machinery.

[16] Ju Wang, Liqiong Chang, Omid Abari, and Srinivasan Keshav. Are RFID sensing systems ready for the real world? In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 366–377, 2019.

[17] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 313–325, 2019.

[18] Cao Dian, Dong Wang, Qian Zhang, Run Zhao, and Yinggang Yu. Towards domain-independent complex and fine-grained gesture recognition with RFID. *Proceedings of the ACM on Human-Computer Interaction*, 4(ISS):1–22, 2020.

[19] Xun Wang, Ke Sun, Ting Zhao, Wei Wang, and Qing Gu. Dynamic speed warping: Similarity-based one-shot learning for device-free gesture signals. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 556–565. IEEE, 2020.

[20] Unsoo Ha, Junshan Leng, Alaa Khaddaj, and Fadel Adib. Food and liquid sensing in practical environments using RFIDs. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 1083–1100, Santa Clara, CA, February 2020. USENIX Association.

[21] Seyed Ali Rokni, Marjan Nourollahi, and Hassan Ghasemzadeh. Personalized human activity recognition using convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[22] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[23] Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11293–11302, 2019.

[24] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R Scott. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6728–6737, 2020.

[25] Kunal Dahiya, Ananye Agarwal, Deepak Saini, K Gururaj, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, and Manik Varma. Siamesexml: Siamese networks meet extreme classifiers with 100m labels. In *International Conference on Machine Learning*, pages 2330–2340. PMLR, 2021.

[26] Chen Yang and Shuyuan Yang. Deep ensemble siamese network for incremental signal classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3200–3204. IEEE, 2021.

[27] Wenshuo Yang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. Mscnn: A monomeric-siamese convolutional neural network for extremely imbalanced multi-label text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6716–6722, 2020.

[28] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. uwave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6):657–675, 2009.

[29] Sheng Shen, He Wang, and Romit Roy Choudhury. I am a smartwatch and i can track my user's arm. In *Proceedings of the 14th annual international conference on Mobile systems, applications, and services*, pages 85–96, 2016.

[30] Gierad Laput and Chris Harrison. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[31] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016.

[32] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In *European Conference on Computer Vision*, pages 769–786. Springer, 2020.

[33] Chenning Li, Manni Liu, and Zhichao Cao. WiHF: Enable user identified gesture recognition with WiFi. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 586–595. IEEE, 2020.

[34] Sheng Tan, Jie Yang, and Yingying Chen. Enabling fine-grained finger gesture recognition on commodity WiFi devices. *IEEE Transactions on Mobile Computing*, 2020.

[35] Yongpan Zou, Jiang Xiao, Jinsong Han, Kaishun Wu, Yun Li, and Lionel M Ni. GRfid: A device-free RFID-based gesture recognition system. *IEEE Transactions on Mobile Computing*, 16(2):381–393, 2016.

[36] Shigeng Zhang, Chengwei Yang, Xiaoyan Kui, Jianxin Wang, Xuan Liu, and Song Guo. Reactor: Real-time and accurate contactless gesture recognition with RFID. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2019.

[37] Haipeng Liu, Anfu Zhou, Zihe Dong, Yuyang Sun, Jiahe Zhang, Liang Liu, Huadong Ma, Jianhua Liu, and Ning Yang. M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar. *IEEE Internet of Things Journal*, 2021.

[38] Haipeng Liu, Yuheng Wang, Anfu Zhou, Hanyue He, Wei Wang, Kunpeng Wang, Peilin Pan, Yixuan Lu, Liang Liu, and Huadong Ma. Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–28, 2020.

[39] Benjia Zhou, Yunan Li, and Jun Wan. Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3563–3571, 2021.

[40] Shigeng Zhang, Zijing Ma, Chengwei Yang, Xuan Liu, Xiaoyan Kui, Weiping Wang, Jianxin Wang, and Song Guo. Real-time and Accurate Gesture Recognition with Commercial RFID Devices. *Accepted to appear in IEEE Transactions on Mobile Computing*, PP(99):1–15, 2022.

[41] Chuyu Wang, Jian Liu, Yingying Chen, Hongbo Liu, Lei Xie, Wei Wang, Bingbing He, and Sanglu Lu. Multi-touch in the air: Device-free finger tracking and gesture recognition via COTS RFID. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1691–1699. IEEE, 2018.

[42] Yinggang Yu, Dong Wang, Run Zhao, and Qian Zhang. RFID based real-time recognition of ongoing gesture with adversarial learning. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pages 298–310, 2019.

[43] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. Metasense: Few-shot adaptation to untrained conditions in deep mobile sensing. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, SenSys '19, page 110–123, New York, NY, USA, 2019. Association for Computing Machinery.

[44] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. RF-net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, SenSys '20, page 517–530, New York, NY, USA, 2020. Association for Computing Machinery.

[45] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. Onefi: One-shot recognition for unseen gesture via cots wifi. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 206–219, 2021.

[46] Xinyi Li, Liqiong Chang, Fangfang Song, Ju Wang, Xiaojiang Chen, Zhanyong Tang, and Zheng Wang. CrossGR: Accurate and low-cost cross-target gesture recognition using Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–23, 2021.

[47] Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. Learning gestures from wifi: A siamese recurrent convolutional architecture. *IEEE Internet of Things Journal*, 6(6):10763–10772, 2019.

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[49] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

[50] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[51] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[52] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[53] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[55] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Zijing Ma Zijing Ma received the BSc degree in computer science and technology from South China Agricultural University in 2020. He is currently working towards his MSc degree in computer science and technology from Central South University, China. His research interests include wireless sensing, RFID, and the Internet of Things. He has published papers in journal and conference including TMC and WCNC.

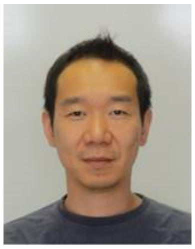
Shigeng Zhang Shigeng Zhang received the BSc, MSc, and DEng degrees, all in Computer Science, from Nanjing University, China, in 2004, 2007, and 2010, respectively. He is currently a Professor in School of Computer Science and Engineering at Central South University, China. His research interests include Internet of Things, mobile computing, RFID systems, and IoT security. He has published more than 70 technique papers in top international journals and conferences including Ubicomp, Infocom, Mobihoc, ICNP, TMC, TC, TPDS, TOSN, and JSAC. He is on the editorial board of International Journal of Distributed Sensor Networks, and was a program committee member of many international conferences including ICC, ICPADS, MASS, UIC and ISPA. He is a member of IEEE and ACM.

Jia Liu Jia Liu is an associate professor with the Department of Computer Science and Technology at Nanjing University, Nanjing, China. Before that, he received the B.E. degree in software engineering from Xidian University, Xi'an, China, in 2010. He received the Ph.D. degree in computer science and technology from Nanjing University, Nanjing, China, in 2016. His research mainly focuses on RFID systems. He is a member of the IEEE and ACM.

Xuan Liu Xuan Liu is currently a Professor in the College of Computer Science and Electronic Engineering at Hunan University. She received the BSc degree in information and computing mathematics from XiangTan University in 2005, the MSc degree in Computer Science from National University of Defense Technology in 2008, and the PhD degree in the Hong Kong Polytechnic University in 2015, respectively. Her research interests include Multi-agent reinforcement learning, RFID systems and Internet of Things. She has published more than 40 technique papers in top international journals including JSAC/ToN/TMC/TC/TPDS and top conferences including Infocom/ICNP/Mobihoc/Ubicomp.

Weiping Wang Weiping Wang received the Ph.D. degree in computer science from Central South University, China, in 2004. Currently, she is a professor in the School of Information Science and Engineering, Central South University. Her current research interests include network coding and network security. She has published more than 60 papers in refereed journals and conference proceedings.

Jianxin Wang Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in2001. He is the Dean of and a professor in School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, Bioinformatics and computer network. He is a senior member of the IEEE.



Song Guo Song Guo is a Full Professor at Department of Computing, The Hong Kong Polytechnic University. He also holds a Changjiang Chair Professorship awarded by the Ministry of Education of China. Prof. Guo is a Fellow of the Canadian Academy of Engineering, Member of Academia Europaea, and Fellow of the IEEE (Computer Society). His research interests are mainly in federated learning, edge AI, mobile computing, and distributed systems. He published many papers in top venues with wide impact in these areas and was recognized as a Highly Cited Researcher (Clarivate Web of Science). He is the recipient of over a dozen Best Paper Awards from IEEE/ACM conferences, journals, and technical committees. Prof. Guo is the Editor-in-Chief of IEEE Open Journal of the Computer Society. He was an IEEE ComSoc Distinguished Lecturer and a member of IEEE ComSoc Board of Governors. He has served for IEEE Computer Society on Fellow Evaluation Committee, Transactions Operations Committee, Steering Committee of IEEE Transactions on Cloud Computing, Editor-in-Chief Search Committee, and been named on editorial board of a number of prestigious international journals like IEEE TC, IEEE TPDS, IEEE TCC, IEEE TETC, ACM CSUR, etc. He has also served as chairs of organizing and technical committees of many international conferences.