2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS) | 979-8-3503-2433-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/MASS58611.2023.00016

MAA: Modulation-adaptive Acoustic Gesture Recognition

Yingxin Shan[†], Peng Liao[†], Xuyu Wang[‡], Lingling An[†§], and Shiwen Mao[♯]
[†]Guangzhou Institute of Technology, Xidian University, Xi'an, China
[‡]Knight Foundation School of Computing & Information Sciences, Florida International University, Miami, FL, USA
[†§]School of Computer Science and Technology, Xidian University, Xi'an, China
[♯]Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA
Email: yxinshan@stu.xidian.edu.cn, pengl3@stu.xidian.edu.cn, xuywang@fiu.edu, an.lingling@gmail.com, smao@ieee.org

***Abstract*—In this paper, we propose a modulation-adaptive acoustic gesture recognition system with smartphones (termed, MAA), which can achieve a high recognition accuracy under various modulation schemes and quickly adapt to a new modulation at low cost. Specifically, MAA creates an acoustic channel model to capture temporal and spatial features, and leverages a domain adversarial network to eliminate the difference among modulation schemes when performing the same gesture. The proposed framework includes a data collection module, a signal preprocessing module, a channel construction module, and a domain adaptation module. For data collection, we determine the appropriate signal length and bandwidth for transmitted acoustic signals. In the signal preprocessing module, channel estimation and background noise removal are incorporated. Then, we develop a tensor reconstruction network and a feature mapping network in the channel construction module to directly map features to specific gestures. For domain adaptation, we train the above two networks in the source domain, and use an adversarial network to adapt to the target domain. Experimental results show that the proposed MAA achieves a good performance on gesture recognition with different modulation schemes, with better adaptation to new modulation schemes than several state-of-the-art baselines.**

***Index Terms*—Gesture recognition, acoustic sensing, modulation schemes, domain adaptation, channel model.**

I. Introduction

Recently, Internet of Things (IoT) devices have been used everywhere in peoples daily lives to offer various sensing applications, including device-based and device-free schemes. In general, device-free sensing applications are well received by users because of their convenience and flexibility in use. Hand gesture recognition (HGR) is a typical device-free sensing application that plays an important role in human-computer interaction (HCI). Radio Frequency (RF) techniques have been developed for HGR with contactless devices [1], [2]. However, the existing schemes usually suffer high deployment cost and and poor adaptability, which hinder their wide deployment in real world scenarios. Unlike RF-based techniques, acoustic sensing provides an essential solution to the above problems. Many IoT devices are equipped with acoustic front-ends, i.e., speakers and microphones, which can transmit and receive acoustic signals for contactless sensing applications [3]–[5].

Various acoustic signal modulation schemes have been developed for acoustic sensing applications, incorporating frequency-modulated continuous wave (FMCW), orthogonal frequency division multiplexing (OFDM), single sine wave, and other types of acoustic modulation methods. For example, PDF [6] leveraged the time delay of reflected FMCW signal to track the trajectory of a moving object. AudioGest [7] transmited a 19kHz acoustic sine wave from a speaker, which was received by a microphone, allowing it to accurately estimate hand in-air time, average waving speed, and hand moving range. The authors in [8] designed a Zadoff-Chu (ZC)-based OFDM signal to obtain channel impulse response (CIR) sequence for extracting the structure-borne component, which is highly related to the sliding gestures.

The aforementioned works demonstrate that although comparable gesture recognition tasks are performed, the signals acquired are from different environments, distinct users, and diverse acoustic modulations. As low-cost acoustic sensing devices continue to proliferate and the rapid development of deep learning, leveraging deep learning to extract features and adapt to new domains for acoustic-based gesture recognition systems is a direction worth exploring. However, there are currently few public datasets for acoustic sensing, and at the same time, laborious data collection and labeling are required for new settings. Insufficient data will affect the deeper network design, bringing performance and robustness limitations to the target task. For example, under a certain modulation, a sufficient amount of data needs to be collected to ensure the performance of the model. Even though there exists a dataset collected with a modulation scheme, exploring a gesture recognition system with a different modulation scheme requires to collect data and train the model from scratch. To this end, this paper proposes MAA, a generalized cross-modulation gesture recognition framework. The key is to build the HGR system with a limited number of acoustic samples for a new modulation using the knowledge learned from the acoustic dataset collected in a known modulation. Thus the overhead of training data collection can be greatly reduced.

To design such a system, two challenges should be addressed. *Challenge I* is how to select appropriate channel characteristics to accurately model acoustic features shadowed by a performed gesture for different modulation schemes. Generally, channel model construction methods mathematically quantify the relationship between channel characteristics and certain gestures, which require domain knowledge and

2155-6814/23/$31.00 ©2023 IEEE
DOI 10.1109/MASS58611.2023.00016

also result in a high cost to design specific channel modules for each type of modulation scheme. *Challenge II* is how to eliminate the channel characteristic differences among modulation schemes under the same performed gesture. Acoustic signals are modulated for transmission, with different time and frequency resolutions. Thus it is hard to guarantee the same performance in a unified channel construction method.

In this paper, we propose a Modulation-Adaptive Acoustic gesture recognition system with smartphones, termed MAA, to achieve a high recognition accuracy under various modulation schemes and quickly adapt a new modulation at low cost. Specifically, to tackle *challenge I*, we design data collection, signal preprocessing, and channel construction methods. For different types of modulation schemes, we first determine the appropriate signal length and bandwidth in data collection, and then extract channel frequency response (CFR) as the basic acoustic channel feature in frequency domain, which is *modulation independent*. Considering both time and frequency domain, we use CIR to capture more time domain information. In the channel construction model, we create a tensor construction network to learn basic channel characteristics with different weights, and a subsequent feature mapping network for further improving the gesture recognition accuracy. The networks are adjusted to make the feature distributions of different modulation schemes closer. To address *challenge II*, our main approach is domain adaptation, which performs well in adapting to different environments and subjects in prior works [9], [10]. Combined with the tensor construction network and feature mapping network in the channel construction model, we propose *a new domain adversarial network*, which preserves the best model of gesture recognition in the source domain and can be fine-tuned for target domain through adversarial learning.

To evaluate MAA, we collect acoustic datasets from different modulation schemes and compare the proposed method with several state-of-the-art methods for domain adaptation. Experiments results show that our unified channel construction module can achieve over 92% gesture recognition accuracy on each type of dataset. More important, our adversarial network can achieve over 80% accuracy in domain adaptation.

We summarize the key contributions in this paper as follows:

- To the best of our knowledge, MAA is the first modulation-adaptive acoustic gesture recognition system, which is designed to achieve a higher HGR accuracy over different acoustic modulation schemes.
- We construct a unified channel model to process the received acoustic signals and extract features in both time domain and frequency domain. We also design a new domain adversarial learning network, to minimize the distribution of features in source domain and target domain. A well-trained model is used to quickly adapt to a new modulation scheme.
- We develop a prototype of MAA with commercial smartphones to demonstrate the robustness of the proposed system over different modulations and test environments. By comparing with several state-of-the-art baselines, we verify MAA can effectively address the domain shift among modulation schemes.

The remainder of this paper is organized as follows. Section II introduces the background and problem formulation. Section III introduces the MAA design. We present our experimental study in Section IV and conclude this paper in Section V.

II. Background and Problem Statement

This section first describes three acoustic signal modulation schemes utilized for recognition of human gestures. The the problem formulation is presented for the MAA system design.

A. Modulation Schemes

1) FMCW [11], [12]: The FMCW method is to transmit and receive acoustic signals reflected by objects. The frequency of the transmitted signal is continuously increasing or decreasing over the symbol duration. The FMCW scheme can provide a fine-grained (frequency) resolution. The FMCW modulated signal is called a *chirp*, whose bandwidth, minimum frequency, and sweeping time span are respectively denoted by B, f_{min}, and T. The transmitted signal can be written as $\cos(2\pi(f_{min} + \frac{B}{2T})t)$, where t is the time within the symbol duration T, i.e., $t \in (0, T)$ [11]. At the receiver, the received signal is demodulated by first multiplying with the signal $\cos(2\pi f_c t)$, where f_c is the central frequency, and then using a low-pass filter to complete the demodulation.

2) GSM [13], [14]: The transmitter generates a 26-bit TSC that has a good auto-correlation property and modulates it through up-sampling. Then, the carrier frequency f_c is used to up-convert the TSC before being transmitted by the speaker. At the receiver, down conversion is performed on the received signal by multiplying $\cos(2\pi f_c t)$ and $-\sin(2\pi f_c t)$, respectively. After using a low pass filter, the real part and imaginary part of the received base-band signal can be obtained.

3) ZC-based OFDM [8]: Similar to GSM, this scheme utilizes TSC in the form of a ZC sequence in OFDM, and interpolates it in the frequency domain rather than in the time domain. First, fast Fourier transform (FFT) is performed on the ZC sequence to obtain a frequency sequence $x_f[n]$. Second, $x_f[n]$ and the conjugate of $x_f[n]$ are arranged according to the positive and negative frequency parts, respectively. The rest part is padded with zero. Finally, inverse fast Fourier transform (IFFT) is performed on the zero-padded complex valued sequence. The resulting real part is considered as the transmitted sequence, since smartphones can only send the real number signal. The ZC-based OFDM methods share the same demodulation process as GSM.

As communication technology continues to advance, the modulation techniques utilized in acoustic sensing are expected to expand further. To investigate the efficacy of gesture recognition models in response to emerging modulation schemes, a significant amount of data will be required. However, the process of collecting and training with such data is both time-consuming and labor-intensive. Therefore, this

study endeavors to mitigate the training cost by transferring the knowledge acquired from datasets under pre-existing modulation schemes to new modulation schemes. Specifically, we aim to design a general gesture recognition framework, to effectively learn features of acoustic signals in different modulated schemes and eliminate the discrepancies in channel features (e.g., CFR and CIR) for the same gesture. A new modulation scheme will be fast deployed on the well-trained recognition model, thus reducing the training cost as well.

B. Problem Formulation

The aim of the proposed MAA system is to construct an appropriate acoustic channel model for different modulation schemes and eliminate the differences in channel characteristics for the same performed gesture. We denote the MAA model as $\mathbb{Z}_\zeta$, which includes four components, including (i) the tensor reconstruction network P_Θ, (ii) the feature mapping network Q_γ, (iii) the gesture predictor network G_ϑ, and (iv) the domain adaptation network A_ξ, which are respectively parameterized by Θ, γ, ϑ, and ξ. Specifically, P_Θ is used to reconstruct the input tensor, and Q_γ aims to extract features from the reconstructed input tensor. The two components are termed as the channel construction model $\mathbb{Z}_{\zeta_1}$ parameterized by ζ_1. $\mathbb{Z}_{\zeta_1}$ is then wrapped into a domain adversarial framework $\mathbb{Z}_{\zeta_2}$ parameterized by ζ_2, which can be generalized to a new modulation scheme. $\mathbb{Z}_{\zeta_2}$ includes G_ϑ and Q_γ. In summary, MAA $\mathbb{Z}_\zeta$ consists of the channel construction model $\mathbb{Z}_{\zeta_1}$ and the domain adversarial framework $\mathbb{Z}_{\zeta_2}$, where $\zeta = \zeta_1 \cup \zeta_2$.

The process of training $\mathbb{Z}_\zeta$ is as follows. The received acoustic signal collected for three different modulation schemes is denoted by $D = \{D_1, D_2, D_3\}$, where D_1 is for FMCW, D_2 is for GSM, and D_3 is for ZC-based OFDM. We train the parameters $\mathbb{Z}_\zeta$ to learn how to adapt to a new modulation scheme. In particular, we sample the source and target domain data from D, denoted as D^S and D^T, respectively, where $D^S, D^T \subsetneqq D, D^S, D^T \neq \emptyset$, and $D^S \cap D^T = \emptyset$. The source domain data and the target domain data are denoted by $D^S = \{x_i^S, y_i^S\}_{i=1}^K$ and $D^T = \{x_i^T, y_i^T\}_{i=1}^K$, respectively, where x_i^S and x_i^T are the complex frequency response matrix collected from the source domain and the target domain, respectively, y_i^S and y_i^T are the corresponding gesture labels and K is the number of gesture categories. The labels of the target domain are only used to validate the performance of domain adaptation.

Due to the different data distributions of the source domain and the target domain, the distributions of their obtained features are also different. The objective function of MAA $\mathbb{Z}_\zeta$ is to minimize the gap in distributions so that a well-performed classifier G_ϑ trained on D^S can perform well for classifying samples on D^T, which is formulated as

$$\zeta^*(D^S, D^T) = \arg\min_{\zeta} \left\| \frac{1}{n_S}\sum_{i=1}^{n_S}\phi(x_i^S) - \frac{1}{n_T}\sum_{j=1}^{n_T}\phi(x_j^T) \right\|_{\mathbb{Z}_{\zeta_1}}, \tag{1}$$

where $\|\cdot\|$ is the metric of the difference between the probability distributions of D^S and D^T, ϕ is the mapping function with $x \to \mathbb{Z}_{\zeta_1}$, n_S and n_T are the number of samples in the source and target domain, respectively. We will elaborate on the MAA design in the next section on how to achieve the above goal to have a modulation-adaptive acoustic HGR.

III. MAA System Design

MAA is a general acoustic-based HGR framework to adapt to different acoustic signal modulation schemes. Fig. 1 shows an overview of the MAA system, with data collection, signal preprocessing, channel model construction, and domain adaptation. For data collection, we sample a set of gestures using acoustic signals in different signal modulation schemes. In signal preprocessing, background noise removal and channel estimation are performed on the received raw acoustic signal to extract the basic features (e.g., CFR and CIR). Then, we use the construct channel model to augment the features. Finally, by utilizing domain adaptation, we achieve a modulation-adaptive acoustic sensing HGR, which can quickly adapt to various modulation schemes. The MAA system modules are discussed in detail in the following.

A. Data Collection

In the data collection module, acoustic signals in different modulation formats are transmitted. We use three modulation schemes as introduced in Section II-A. To generate an inaudible sound signal, the transmission frequency is set between 18kHz and 22kHz with a sampling rate F_s of 48kHz. The final transmitted signals are denoted by $S = \{S_1, S_2, S_3\}$, where the length of signal is set as $N = 480$. The sound speed is $c = 343$m/s, and the sensing range for $N = 480$ is within $\frac{cN}{2Fs} \approx 1.72$m. Such a setting is sufficient for the proposed system. Signals in different modulation formats are repeated by the transmitter. To avoid interference between frames, there is a guard interval between adjacent symbols. When data section is set with a length of 312 samples, it will occupy 6.5ms in the time domain. When the speed of finger motion is below 0.5m/s, its coherence time is 8.5ms, which is longer than 6.5ms (moving speed and coherence time are inversely proportional to each other [15]). Thus, the acoustic channel in MAA can be viewed as linear time invariant (LTI) system. At the receiver, acoustic signals reflected from gestures and other surrounding objects are received with the same sampling rate as the transmitter. As shown in Fig 2, the received acoustic signal can be segmented into M frames, and the frame length is the same as the transmitted signal length N.

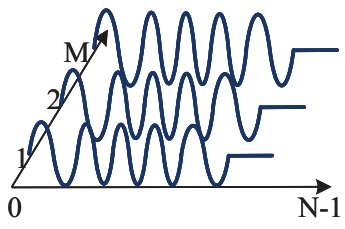

Fig. 2. Illustration of the received signal matrix.

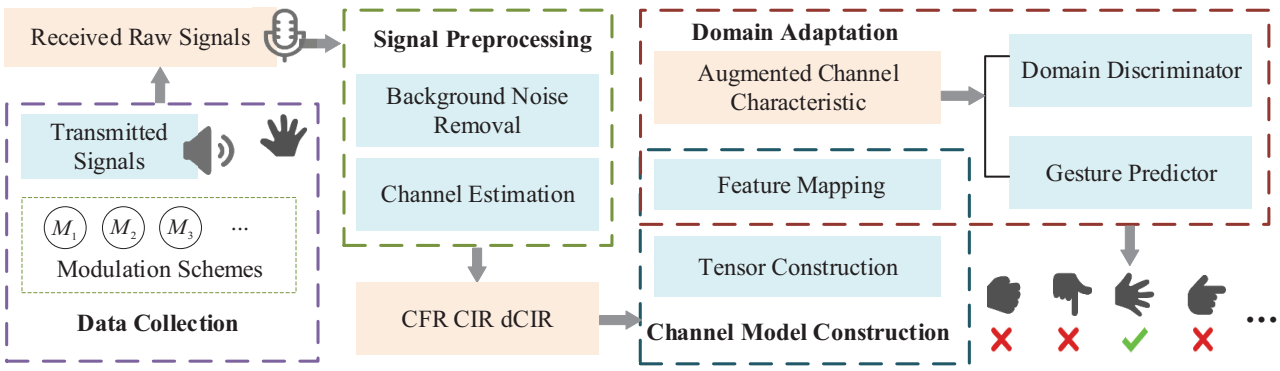

Fig. 1. Overview of the proposed modulation-adaptive acoustic gesture recognition system with smartphones.

B. Signal Preprocessing

Due to the environment noise, the received acoustic signals cannot be directly used to construct an acoustic channel model. Instead, we implement a signal preprocessing method to acquire more robust sensing information.

1) Channel Estimation: To adapt to different modulation schemes, we estimate CFR that can reflect the FSF in the frequency domain from the received signal. The received signal spectrum R^f can be represented by $R^f = S^f * H^f$, where S^f is the spectrum of the transmitted signal, H^f is the CFR in the LTI system. Note that CFR (i.e., channel information) is the ratio of the received and transmitted spectrum, which is independent to modulation [16]. Specifically, CIR can be obtained through an IFFT of CFR, which captures the corresponding environment information (e.g., the multi-path effect) in the time domain. Assume there are K paths in the sensing area and path i has a time delay τ_i, an amplitude decay a_i, and a phase offset θ_i. The CIR $H(\tau)$ is given by

$$H(\tau) = \sum_{i=1}^{K} a_i e^{-j\theta} \delta(\tau - \tau_i), \tag{2}$$

where τ is the period of transmitted signal, and $\delta(\cdot)$ is the Dirac function. The acoustic channel response in the frequency and time domain can be represented by CFR and CIR, respectively. When a subject performs a gesture, people in the surroundings may cause interference. Thus, we also use the difference between two consecutive CIR profiles at time t and $t-1$ (i.e., dCIR) to further cancel such interference and extract dynamic components in the time domain. These three basic channel features (CFR, CIR, and dCIR) serve as the input to the channel construction model.

2) Background Noise Removal: To remove the out-of-band noise in each frame, we first apply a bandpass filter to the received signal. Then, we perform FFT on the denoised signal to calculate CFR. To mitigate the background interference (i.e., from static objects), we regard the first five received frames (occupying 0.05ms) as a static state (i.e., no gestures would be performed) in our system. Since the signals reflected by the static objects occupy most of the reflected signal, and their frequencies will not change, we compute the average value of the first five frames in the frequency domain of the received signal and subtract this value from each frame.

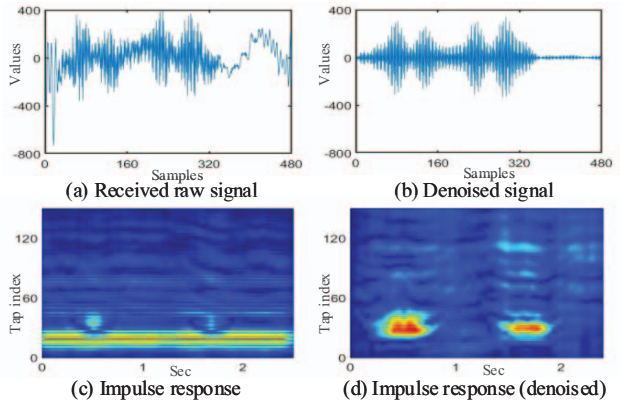

Fig. 3. Signal preprocessing workflow using GSM as an example.

To demonstrate the efficacy of our signal preprocessing model, we perform a push-pull gesture in front of the smartphone and use the GSM modulation as a example, while other modulation schemes share the same workflow. In Fig. 3, the received raw signal becomes smoother and exhibits better signal characteristics after removing out-of-band noise. Thus, the denoised signals are more suitable for the next process of feature extraction.

C. Channel Model Construction

We utilize both CFR and CIR to construct a tensor that represents information in time and frequency domains for each signal modulation scheme, and leverage dCIR to enrich time domain features. In the channel construction model, we first simply concatenate the CFR, CIR, and dCIR matrices, which are extracted from received signals in different modulation formats. We reconstruct a tensor reconstruction network, which is suitable to be generalized and will also be helpful for the subsequent domain adversarial network. The output is fed into the other network to extract temporal and spatial features. The two networks used for tensor construction are P_Θ and feature mapping Q_γ. Finally, the channel construction model $\mathbb{Z}_{\zeta_1}$ can

be formulated as $\mathbb{Z}_{\zeta_1}(x) = Q_\gamma(P_\Theta(x))$, where x denotes the matrices of CFR, CIR, and dCIR.

1) Tensor Construction P_Θ: We define the outputs of CFR, CIR, and dCIR from the signal preprocessing method as x^{cfr}, x^{cir}, and x^{dcir}, respectively, which all have the same size of $N \times M$. In order to explore a more reasonable feature presentation for each signal collected from different modulation schemes, we construct a generalized tensor with size $3 \times N \times M$. Different from simply overlapping or providing them a fixed weight, the key idea of our work is to consider channel estimation information from the three matrices as different colors, i.e., RGB values in an image, and assume they have different weights. By setting different weights to the three channels, the tensor can be reconstructed for better gesture recognition or allow a low-cost domain adversarial learning during the training process of $\mathbb{Z}_\zeta$. This process is achieved by

$$\text{ratio}_i^k(\mathbb{R}) = f_i^k(\mathbb{R}) \; / \sum_{k=1}^{K} f_i^k(\mathbb{R}), \tag{3}$$

$$\mathbb{R}' = \text{Conv2d}(\mathbb{R}, \text{ratio}(\mathbb{R})), \tag{4}$$

where $\mathbb{R}$ is the simply concatenated matrix of x^{cfr}, x^{cir}, and x^{dcir}, $f_i^k(\mathbb{R})$ is the response value of time slot i in channel k, $i \in [0, M]$, and $\mathbb{R}'$ is the final reconstructed tensor.

2) Feature Mapping Q_γ: The reconstructed tensor $\mathbb{R}'$ serves as input to the feature mapping network Q_γ, parameterized by γ. To better learn the influential features from the channel characteristics of general modulation schemes, we aim to extract deep temporal and spatial features, instead of a simple, single feature. We employ state-of-the-art convolutional neural networks (CNNs) for exploring spatial features F^{spa} and leverage gate recurrent unit (GRU) and efficient channel attention (ECA) for generating temporal features F^{tem}.

For spatial feature extraction, $\mathbb{R}'$ is regarded as an image with three input channels, with N as the height and M as the width of input planes in pixels. We directly feed $\mathbb{R}'$ into a Resnet-18 network [17]. And then rectified linear unit (ReLU) is employed in the model to incorporate non-linear factors for better learning. Finally, a high level representation of $\mathbb{R}'$ can be achieved, which is denoted by F^{spa}. Note that if we separate them into three channels and then feed them into Resnet-18, the total time will become very high. Table I provides the execution time of one epoch for different numbers of channels. The execution time will increase by over 2 seconds for just one epoch from one channel to three channels.

We then partition $\mathbb{R}'$ into x^1, x^2, and x^3, corresponding to x^{cir}, x^{dcir}, and x^{cfr}, respectively, and use low-time consuming networks to extract time-frequency features. We extract x^{time} and x^{fre} by passing the concatenation of x^2 and x^3, and x^1 into the GRU, respectively. GRU has a better performance for time series prediction, while beging lightweight than long-short term memory (LSTM). Further, we exploit ECA for learning channel attention for its simplicity and efficacy [18]. Subsequently, for producing a time-frequency joint feature, we employ ReLU activated, concatenated layers to combine x^{time} and x^{fre}, and the final output is F^{tem}. At last, by using similar concatenated layers, a fused temporal and spatial channel characteristic F is obtained for classification.

TABLE I
EXECUTION TIME OF ONE EPOCH FOR DIFFERENT NUMBERS OF CHANNELS (IN SECONDS)

Number of Channels	1	2	3
Extract	0.8569	1.7695	2.8806
Extract and update	0.9002	1.8178	2.9858

TABLE II
THE COSINE SIMILARITIES BETWEEN DOMAINS

	Mod_1^T	Mod_2^T	Mod_3^T	Mod_1^F	Mod_2^F	Mod_3^F
Mod_1^T	1.00	0.66	0.58	0.45	0.40	0.38
Mod_2^T	0.66	1.00	0.78	0.50	0.51	0.54
Mod_3^T	0.58	0.78	1.00	0.55	0.57	0.57
Mod_1^F	0.45	0.50	0.55	1.00	0.64	0.66
Mod_2^F	0.40	0.51	0.57	0.64	1.00	0.72
Mod_3^F	0.38	0.54	0.57	0.66	0.72	1.00

D. Domain Adaptation

Through the proposed signal preprocessing method and the channel construction model, high-level characteristics can be obtained for implementing high accuracy gesture recognition. *The question is whether a well trained network on one dataset will work well on another dataset in a different modulation format*. Similar to [19], we evaluate the similarities between domains (i.e., modulation schemes) by calculating the cosine similarity. Let $\{Mod_i^T\}_{i=1}^3$ and $\{Mod_i^F\}_{i=1}^3$ represent the impulse response matrices x_i^{cir} and frequency response matrices x_i^{cir} for modulation i, respectively. The cosine similarity values are presented in Table II. We find Mod_2 and Mod_3 are the two most similar domains, but their similarity is still less then 0.75. Most of the other similarity values are less than 0.6. These results show that a well trained network, without any modification, will not be effective for an unknown domain.

Our approach is to explore the domain adversarial framework $\mathbb{Z}_{\zeta_2}$ to address the above problem. The domain adaptation module is presented in Fig. 4. The data of the source domain and target domain are fed into different tensor reconstruction networks while they share the same feature mapping network. The method of applying P_Θ and Q_γ for the domain adaptation module is that the parameters Θ and γ in the network can be updated for better learning the useful features for the classifier, specifically for the source domain. In the target domain, both networks are initialized with the parameters, which are learned in the source domain while only Θ is updated for domain adversarial learning for reduced cost.

The gesture predictor G_ϑ and the domain discriminator A_ε of $\mathbb{Z}_{\zeta_2}$ are first introduced as follows. In the next subsections on domain adversarial learning, we will discuss how the proposed network addresses the domain shift problem. Essentially,

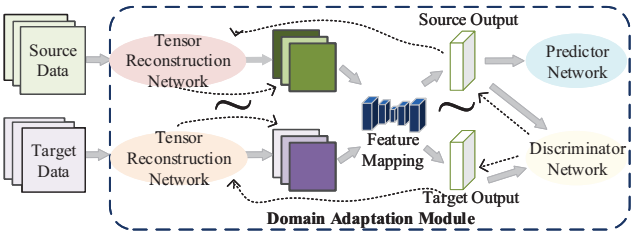

Fig. 4. Overview of the domain adaptation module.

$\mathbb{Z}_{\zeta_2}$ can be written as

$$\mathbb{Z}_{\zeta_2}(x^S, x^T) = (G_\delta(\mathbb{Z}_{\zeta_1}(x^S)), A_\xi(\mathbb{Z}_{\zeta_1}(x^S), \mathbb{Z}_{\zeta_1}(x^T))). \quad (5)$$

1) Gesture Predictor G_ϑ: As introduced in Section III-C, a high level feature F is obtained through network Q_γ. In this paper, we choose the few-shot classifier, which performs well on recognizing novel classes for predicting hand gestures [2]. We only use its training network of base classes as our gesture predictor. The predicted results are given by

$$G_\vartheta(\mathbb{Z}_{\zeta_1}(x)) = \text{softmax}(\text{W}_1 * (\mathbb{Z}_{\zeta_1}(x)) + b_1), \quad (6)$$

where W_1 and b_1 are the weight matrix and the bias of the gesture predictor network, respectively.

2) Domain Discriminator A_ε: Similar to the gesture predictor, $(F)^T$ (i.e., the transpose of F) is the input and used for classification. But the domain discriminator A_ε performs binary classification, which only needs to judge whether $(F)^T$ is from the source or the target domain. The network of domain discriminator is the same as that in the gesture predictor, but having different parameters. The domain discrimination result is given by

$$A_\xi(\mathbb{Z}_{\zeta_1}(x)) = \text{softmax}(\text{W}_2 * (\mathbb{Z}_{\zeta_1}(x))^T + b_2), \quad (7)$$

where W_2 and b_2 are the weight matrix and the bias of the domain discriminator, respectively.

3) Domain Adversarial Learning $\mathbb{Z}_{\zeta_2}$: To extract domain-independent features, domain adversarial learning is used to discard the specific features of modulation schemes, while retaining the common features useful for gesture recognition. In particular, our domain adversarial learning involves $\mathbb{Z}_{\zeta_1}$, which is trained with the gesture predictor to map specific hand gestures to features extracted from received signals. Our method adjusts parameters ζ_1, $\mathbb{Z}_{\zeta_1} = \Theta \cup \gamma$ to better work with our proposed adversarial learning network $\mathbb{Z}_{\zeta_2}$.

The process of training includes two phases. In the first phase, we aim to find an optimal model for gesture recognition for the source data D^S. We will then store the model for the source domain, which is subsequently used in the second training phase $\mathbb{Z}_{\zeta_2}$. In addition, we use the standard cross-entropy function to calculate the loss between the predicted class label and the ground truth, given by

$$\mathcal{L}_1(x, y) = \sum_{i=1}^{n} y^i \log(y^i) - \sum_{i=1}^{n} G_\vartheta(\mathbb{Z}_{\zeta_1}(x)) \log(y^i). \quad (8)$$

In the second phase, we first use the above parameters of the model to initialize Θ and γ in the target domain. Then, adversarial learning is used for training, which drives the data features from the target domain close to that from the source domain, based on (1). Similarly, the extracted features could confuse the domain discriminator, while the predictor network can effectively tell the input features are from the source or the target domain. Through adversarial learning, we can obtain a well-trained model that can adapt to the target domain through fine-tuning the saved model from the source domain. The training process is guided by two loss functions, which are given by

$$\mathcal{L}_D(D^S, D^T, \mathbb{Z}^S_{\zeta_1}, \mathbb{Z}^T_{\zeta_1}) = -E_{x^S \sim D^S}[\log(A_\xi(\mathbb{Z}_{\zeta_1}(x^S))] - E_{x^T \sim D^T}[\log(1 - A_\xi(\mathbb{Z}_{\zeta_1}(x^T)))], \quad (9)$$

$$\mathcal{L}_F(D^S, D^T, A_\xi) = -E_{x^T \sim D^T}[\log(A_\xi(\mathbb{Z}_{\zeta_1}(x^T)))]. \quad (10)$$

where $\mathcal{L}_D$ and $\mathcal{L}_F$ represent the cross-entropy losses of the domain discriminator and the gesture predictor, respectively. The total loss function $\mathcal{L}_2$ in the second phase is defined as

$$\mathcal{L}_2 = \mathcal{L}_D + \mathcal{L}_F. \quad (11)$$

Overall, we first optimize $\mathcal{L}_1$ over ζ_1 and G_ϑ by training using the source domain data. Sine we choose to learn $\mathbb{Z}_{\zeta_1}(D^T)$ by leveraging the saved $\mathbb{Z}_{\zeta_1}(D^S)$, we can then optimize $\mathcal{L}_2$ without revisiting the first phase. Specifically, $\mathbb{Z}_{\zeta_1}(D^T)$ includes P_Θ and Q_γ. Updating all the parameters Θ, γ will have a high cost of computation, because Q_γ is a complex network. Since the source domain and the target domain share the same feature mapping network, we reconstruct the input tensor of target domain through updating parameters Θ in the network P_Θ while parameters γ in Q_γ remain the same. The training goal is to minimize $\mathcal{L}_1$ and $\mathcal{L}_2$, and then the distribution of the source and target domain will be minimized.

IV. Implementation, Experiment and Evaluation

In this section, the system implementation and setup are first introduced. Then the overall evaluation of MAA is presented, including hand gesture recognition results and domain adaptation results. We will also present a micro-benchmark study and an ablation study.

A. System Implementation and Experiment Setup

1) Data Collection: Our datasets are collected from ten volunteers (numbered from 1 to 10). Each volunteer performs six kinds of hand gestures (i.e., push (G1), slide up (G2), slide left (G3), click (G4), clockwise (G5), and spread (G6)) under three different modulation schemes (domains). Fig. 5 illustrates the scenarios of our experiment. Fig. 5(a) shows a data collection scenario where a subject sits on a chair performing hand gestures in front of a smartphone. A computer is used for real-time data processing. Figs. 5(b)(c)(d) show three indoor environments where data was collected, i.e., a meeting room, a study room, and an office. We implement our system on a smartphone (i.e., OnePlus 6). The microphone of the smartphone records the received acoustic signal data, which is sent to the computer for processing and recognition.

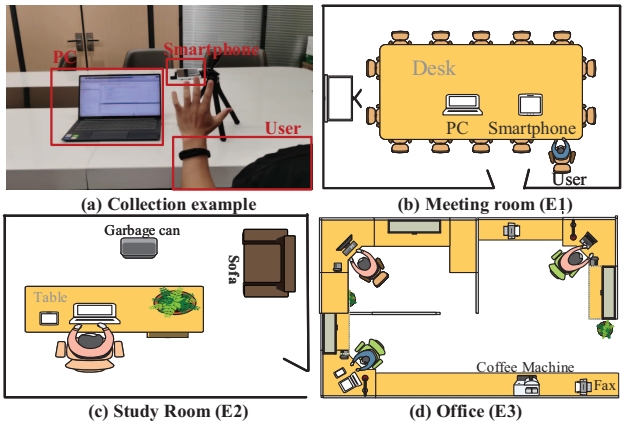

Fig. 5. Data collection scenario and the three indoor environments.

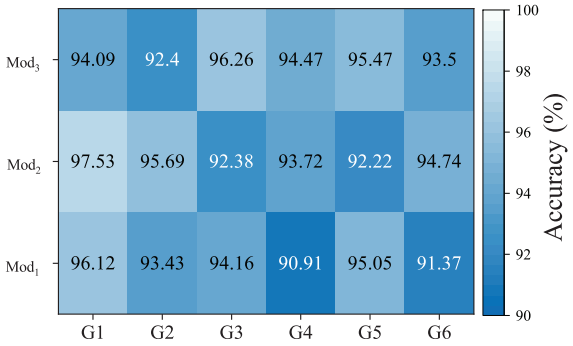

Fig. 6. HGR performance on each domain.

2) Baseline Methods: We compare our method with seven state-of-the-art learning models for HGR, which are described as follows:

- *KNN*: KNN [20] is a simple but effective method for classification, which is trained using the data of source domain and then used for testing in the target domain.
- *CNN*: CNN [21] is a model that can effectively extract 2D spatial features from input data. We use the same method as in the case of KNN.
- *TCPR*: TCPR [22] develops a domain-adaptive classifier, which does not rely on restrictive assumptions.
- *FLDA*: FLDA [23] can adapt to the differences in the marginal probability of features in the source and the target domain.
- *DANN*: DANN [24] induces the adversarial theory into domain adaptation and it aims at generating features that represent both the source domain and target domain.
- *CORAL*: CORAL [25] is a simple unsupervised domain adaptation method that uses a linear transformation to align the source and target distributions.
- *ADDA*: ADDA [26] is an domain-adversarial method, which combines the discriminative model, untied weight sharing, and a GAN loss.

We use a public platform [27], where we can send acoustic data between the PC and the smartphone in real time. Our data processing, model training, and other operations are done on the PC, which is equipped with 16GB memory, Intel i7-10700 CPU @2.90GHz, and the Nvidia GTX 1660 Graphics Card.

3) Metric: We choose the recognition accuracy to quantify the performance of MAA, which represents the percentage of the number of correctly recognized samples in the total testing samples in the target domain, which is calculated as $Acc = Num_{correct}/Num_{total}$.

B. Overall Evaluation

We first evaluate the performance of MAA from two major perspectives: (i) the results of hand gesture recognition, and (ii) the results of domain adaptation. We verify whether the channel construction model is suitable for various modulation schemes to extract hand movement features and then validate if the domain adaptation model can quickly adapt to a new modulation scheme, while still achieving a relatively acceptable recognition accuracy in the target domain.

1) Hand Gesture Recognition Results: We show the accuracy of hand gesture recognition by using MAA on each kind of dataset collected from different modulation schemes, and to verify the efficacy of our channel construction model, which is the key contributor to accurate classification.

In the first training phase of MAA, we aim to find the best classification model for the source domain. Fig. 6 presents the recognition results using the dataset collected from one of the three modulation schemes. We can see that the gesture recognition accuracy of each modulation scheme is over 90% (the lighter the color, the higher the accuracy), while Mod_1 has the poorest performance on recognizing gesture "click" and Mod_2 best recognizes gesture "push." Overall, all of these three modulation schemes can well classify "push" than the other gestures.

2) Domain Adaptation Results: Our domain adaptation design aims to make the model fast adapt to a new domain at a low cost. According to the proposed training process, we first train the channel construction model and the gesture predictor in the source domain, save this model, and then fine-tune it for the target domain. The training in the target domain is to minimize (9) and (10).

Table III shows the performance of domain adaptation of MAA and the state-of-the-art methods when choosing one dataset as source domain and another one in the remaining datasets as target domain. Our proposed MAA scheme outperforms all the baseline methods in this experiment. The average recognition accuracy of KNN is 63.51%, being the poorest performance. Among the baselines, ADDA has the closest performance to MAA, which is 73.55%. This validates the superiority of using different feature extractors for the source domain and target domain. DANN is the model leveraging a feature extractor to extract domain-independent feature, and its average accuracy is 65.41%, which is lower than ADDA and MAA. Fig. 7 shows the experimental results of leaving one dataset as target domain and others as source domain. We

TABLE III
HAND GESTURE RECOGNITION PERFORMANCE (% ACCURACY ON TARGET DOMAIN) OF DOMAIN ADAPTATION

Accuracy	1-2	1-3	2-1	2-3	3-1	3-2	Avg.
KNN	57.75	58.73	63.82	70.38	62.42	68.01	63.51
CNN	59.67	70.28	60.01	70.94	61.27	74.11	66.04
TCPR	68.07	72.05	65.29	69.14	66.61	68.46	68.27
FLDA	59.08	63.13	72.48	69.92	61.29	70.10	66.00
DANN	59.86	62.29	61.38	70.61	63.92	74.42	65.41
CORAL	62.05	70.90	62.29	79.67	60.84	78.49	69.04
ADDA	69.37	72.89	72.81	78.02	71.34	76.92	73.55
MAA	**72.13**	**75.46**	**71.47**	**82.09**	**79.63**	**85.61**	**77.73**

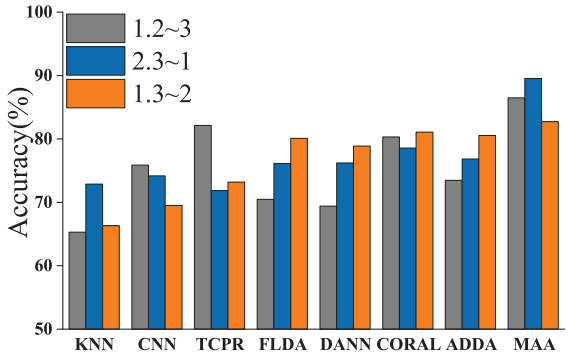

Fig. 7. Domain adaptation performance on a multi-source domain and a target domain.

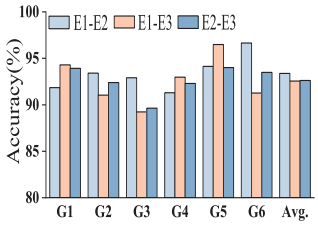

Fig. 8. Performance on changed environments.

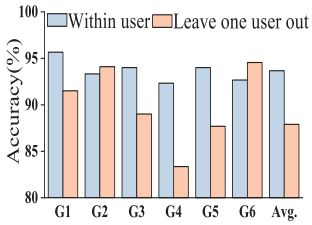

Fig. 9. Performance on within users and leave one out.

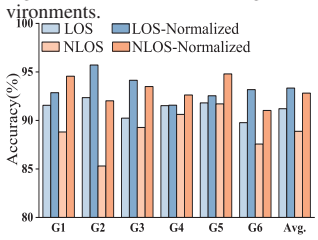

Fig. 10. Performance on different testing scenarios.

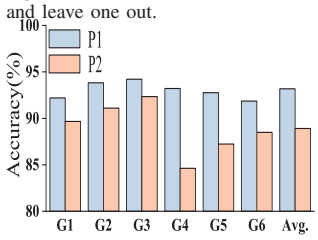

Fig. 11. Performance on different types of noise.

can see the similar trend as in Table III, which demonstrates the advantage of MAA on multiple dimensions.

C. Micro-benchmark

The following experiment is conducted to study the impact of system environments on the robustness of MAA for HGR.

1) Impact of Environments: We evaluate the MAA performance in different environments or for different subjects. Three environments are shown in Fig. 5. Some of them are surrounded by desks, computers, and small objects with rich multi-path effect. We train our model in one environment and test it in another environment. The experiment results are described in Fig. 8. We can see that even though with a new environment, the recognition accuracy is quite stable, which are 93.30%, 92.54%, and 92.62%. This is because the features we used are resilient to interference from the surrounding environment. Our signal preprocessing could better reduce the interference of static objective and background noise, which are affected by the environment.

2) Impact of Subjects: For examining the impact of different subjects, data collected from nine users are used to train our model, which is then tested on the data collected from the remaining subject. As shown in Fig. 9, we can see that the overall accuracy decreases to 87.90%, while the accuracy is 93.67% when testing on known subjects. This is because of the different gesture speeds of different subjects.

3) Impact of Distance: We explore the impact of testing scenarios, including line-of-sight (LOS) and non-line-of-sight (NLOS) on the systems. The NLOS scenario is created by adding obstacles between the subject and smartphone. Experimental results are shown in Fig. 10. We find that the gesture recognition accuracy in LOS scenarios is quite satisfactory, which is 92.21%. In NLOS scenarios, the accuracy becomes lower, probably due to the inconspicuous magnitude of input CIR matrix. We also normalize the input matrix, and find that the average recognition accuracy values in the two testing scenarios become similar after being normalized, which are 93.33% and 92.82%, respectively. Specially, we also consider the situation where subjects perform gestures sitting at different distances away from the smartphone. We find that the magnitude of signal in some modulation schemes will be seriously reduced when the distance is beyond 50cm. Hence, we consider LOS and NLOS scenarios when the distance between the subject and the smartphone is within 50cm.

4) Impact of Noise Level: We also verify the impact of noise by considering two situations, which is denoted as $P1$ and $P2$. $P1$ is to play music at a distance of $0.6 \sim 1.0$m (around 65dB) when collecting data, while $P2$ is to introduce white noise to the signal before transmission. Fig. 11 shows that the average accuracy under background music is 93.17%, which is higher than that of $P2$, which drops down to 88.90%. This experiment shows that our signal preprocessing module can filter out the music noise, whose frequency is much higher than that of the acoustic signals used in our system. Moreover, the poorer performance in $P2$ may result from the minor vibration when the smartphone is making sound.

D. Ablation Study

Our system contains several components: signal preprocessing, tensor construction, feature mapping, and domain adversarial learning. To evaluate the contribution to MAA's performance from each of them, we conduct an ablation study.

TABLE IV
ABLATION STUDY OF MAA: 'S' , 'R', 'F', AND 'D' RESPECTIVELY REPRESENT SIGNAL PREPROCESSING, TENSOR RECONSTRUCTION, FEATURE MAPPING, AND ADVERSARIAL LEARNING NETWORK

S	R	F	D	Accuracy
				42.56%
✓				69.47%
✓	✓			76.89%
✓	✓	✓		80.23%
✓	✓	✓	✓	85.61%

In particular, we directly use the received raw signals without using our designed signal preprocessing methods, use a simply concatenated matrix of CFR, CIR and dCIR to replace our constructed tensor, and the convolution or full connect network for other networks. Table IV summarizes the ablation study results. We can see that our signal preprocessing, feature mapping, and adversarial learning network modules make a big contribution to the recognition performance, while other components also help to enhance the performance of MAA.

V. CONCLUSIONS

In this paper, we proposed an MAA system to minimize the difference between different acoustic modulation schemes when performing the same gesture and fast adapt to unseen modulation schemes. First, we introduced three modulation schemes used in our system, Then, we formulated the problem and discussed in detail the four modules of MAA, including data collection, signal preprocessing, channel construction model, and domain adaptation module. Finally, we validated the proposed MAA system on our collected datasets. The results demonstrated that our proposed method could effectively achieve a superior performance on gesture recognition over several state-of-the-art methods, with great adaptability to unseen modulation schemes.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China under Grant No. 62072355. This work is also supported by NSF (CNS-2105416 and CNS-2107164).

REFERENCES

[1] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "WiDar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi," in *Proc. ACM MobiHoc'17*, Chennai, India, July 2017.

[2] R. Xiao, J. Liu, J. Han, and K. Ren, "OneFi: One-shot recognition for unseen gesture via COTS WiFi," in *Proc. ACM SenSys'21*, Coimbra, Portugal, Nov. 2021, pp. 206–219.

[3] C. Cai, R. Zheng, and J. Luo, "Ubiquitous acoustic sensing on commodity IoT devices: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 432–454, FirstQuarter 2022.

[4] Y. Bai, L. Lu, J. Cheng, J. Liu, Y. Chen, and J. Yu, "Acoustic-based sensing and applications: A survey," *Elsevier Computer Networks*, vol. 181, p. 107447, Nov. 2020.

[5] X. Wang, R. Huang, C. Yang, and S. Mao, "Smartphone sonar based contact-free respiration rate monitoring," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 2, p. Article 15, Mar. 2021.

[6] H. Cheng and W. Lou, "Push the limit of device-free acoustic sensing on commercial mobile devices," in *Proc. IEEE INFOCOM 2021*, Virtual Conference, May 2021, pp. 1–10.

[7] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "AudioGest: Enabling fine-grained hand gesture detection by decoding echo signal," in *Proc. ACM UbiComp'16*, Heidelberg, Germany, Sept. 2016, pp. 474–485.

[8] L. Wang, X. Zhang, Y. Jiang, Y. Zhang, C. Xu, R. Gao, and D. Zhang, "Watching your phone's back: Gesture recognition by sensing acoustical structure-borne propagation," *Proc. ACM Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 2, pp. 1–26, June 2021.

[9] Z. Zhou, F. Wang, J. Yu, J. Ren, Z. Wang, and W. Gong, "Target-oriented semi-supervised domain adaptation for WiFi-based HAR," in *Proc. IEEE INFOCOM 2022*, Virtual Conference, May 2022, pp. 420–429.

[10] C. Dian, D. Wang, Q. Zhang, R. Zhao, and Y. Yu, "Towards domain-independent complex and fine-grained gesture recognition with RFID," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. ISS, pp. 1–22, Nov. 2020.

[11] Y.-C. Tung and K. G. Shin, "Expansion of human-phone interface by sensing structure-borne sound propagation," in *Proc. ACM MobiSys'16*, Singapore, June 2016, pp. 277–289.

[12] W. Mao, J. He, and L. Qiu, "Cat: High-precision acoustic motion tracking," in *Proc. ACM MobiCom'16*, New York City, NY, Oct. 2016, pp. 69–81.

[13] Y. Wang, J. Shen, and Y. Zheng, "Push the limit of acoustic gesture recognition," in *Proc. IEEE INFOCOM'20*, Virtual Conference, May 2020, pp. 566–575.

[14] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proc. ACM MobiSys'17*, Niagara Falls, NY, June 2017, pp. 15–28.

[15] K. Ling, H. Dai, Y. Liu, A. X. Liu, W. Wang, and Q. Gu, "UltraGesture: Fine-grained gesture sensing and recognition," *IEEE Trans. Mobile Computing*, vol. 21, no. 7, pp. 2620–2636, July 2020.

[16] 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, Dec. 2013.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF CVPR'16*, Las Vegas, NV, June 2016, pp. 770–778.

[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Supplementary material for ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF CVPR'20*, Seattle, WA, June 2020, pp. 13–19.

[19] Z. Shi, J. A. Zhang, Y. D. R. Xu, and Q. Cheng, "Environment-robust device-free human activity recognition with channel-state-information enhancement and one-shot learning," *IEEE Trans. Mobile Computing*, vol. 21, no. 2, pp. 540–554, Feb. 2020.

[20] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE/CVF CVPR'06*, New York, NY, June 2006, pp. 2126–2136.

[21] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *Elsevier ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, Mar. 2021.

[22] W. M. Kouw and M. Loog, "Target contrastive pessimistic risk for robust domain adaptation," *arXiv preprint arXiv:1706.08082*, June 2017. [Online]. Available: https://arxiv.org/abs/1706.08082

[23] W. M. Kouw, L. J. Van Der Maaten, J. H. Krijthe, and M. Loog, "Feature-level domain adaptation," *J. Machine Learning Research*, vol. 17, no. 1, pp. 5943–5974, Sept. 2016.

[24] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, Apr. 2016.

[25] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. 2016 European Conference on Computer Vision*, Amsterdam, The Netherlands, Oct. 2016, pp. 443–450.

[26] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *IEEE/CVF CVPR'17*, Honolulu, HI, July 2017, pp. 7167–7176.

[27] Y.-C. T. Tung, D. Bui, and K. G. Shin, "Cross-platform support for rapid development of mobile acoustic sensing applications," 2018. [Online]. Available: https://github.com/yctung/LibAcousticSensing