IEEE INFOCOM 2023 - IEEE Conference on Computer Communications | 979-8-3503-3414-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/INFOCOM53939.2023.10228985

FacER: Contrastive Attention based Expression Recognition via Smartphone Earpiece Speaker

Guangjing Wang, Qiben Yan, Shane Patrarungrong, Juexing Wang and Huacheng Zeng
Department of Computer Science and Engineering, Michigan State University, USA
{wanggu22, qyan, patrarun, wangjuex, hzeng}@msu.edu

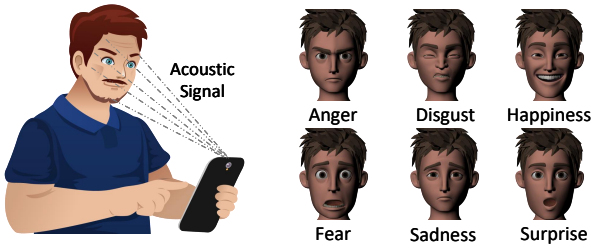
***Abstract*—Facial expression recognition has enormous potential for downstream applications by revealing users' emotional status when interacting with digital content. Previous studies consider using cameras or wearable sensors for expression recognition. However, these approaches bring considerable privacy concerns or extra device burdens. Moreover, the recognition performance of camera-based methods deteriorates when users are wearing masks. In this paper, we propose FacER, an active acoustic facial expression recognition system. As a software solution on a smartphone, FacER avoids the extra costs of external microphone arrays. Facial expression features are extracted by modeling the echoes of emitted near-ultrasound signals between the earpiece speaker and the 3D facial contour. Besides isolating a range of background noises, FacER is designed to identify different expressions from various users with a limited set of training data. To achieve this, we propose a contrastive external attention-based model to learn consistent expression features across different users. Extensive experiments with 20 volunteers with or without masks show that FacER can recognize 6 common facial expressions with more than 85% accuracy, outperforming the state-of-the-art acoustic sensing approach by 10% in various real-life scenarios. FacER provides a more robust solution for recognizing facial expressions in a convenient and usable manner.**

***Index Terms*—Acoustic sensing, expression recognition, contrastive learning, attention, domain adaptation, smartphone**

I. Introduction

With the emergence of new media in the digital era, a variety of social media services are craving for users' attention. Fine-grained emotional reaction understanding is pivotal to facilitating a user's interaction with digital content. Traditionally, crowd-sourced ratings and reviews have been used for evaluating users' feedback on services. However, they are too coarse-grained to provide real-time spontaneous feedback. To provide a more personalized service, we need an accurate and robust approach to recognizing the users' emotions and acquiring users' spontaneous feedback.

A number of techniques have been proposed to recognize emotions, expressed by various biometric features, such as facial features [1]–[3], speech features [4], or heartbeats [5]. Nevertheless, as a universal form of nonverbal communication, facial expression is recognized as the most direct way of understanding human emotions [6]. There are six widely-adopted facial expressions (FEs): anger, disgust, fear, happiness, sadness, and surprise [7]. These FEs can be modeled by the Facial Action Coding System (FACS), which includes action units (fundamental action muscles) and action descriptors (unitary movements of several muscle groups) [8]. When people make different facial expressions, different facial muscles would move, which could be captured by various sensing signals.



(a) FacER application scenario. (b) Six facial expressions [15].

Fig. 1: Facial expression recognition using a smartphone.

Existing facial expression recognition (FER) methods can be categorized as camera-based [9]–[11], radio-based [12]–[14] and acoustic-based [3] expression recognition. However, the existing camera-based methods raise privacy concerns due to their continuous video recording on a device. For instance, FaceWarehouse [9] collects RGBD data of different expressions from multiple users. Yet, serious privacy concerns impede the wide adoption of such methods in a real-world scenario. Moreover, these camera-based methods cannot accurately recognize facial expressions when the users cover their faces with masks. Other alternative methods require extra sensing devices, which affects the usability of the technology. For example, WiFace [13] uses a WiFi router with three antennas placed at a specific position for FER, but it requires additional hardware and configurations. PPGface [14] requires expensive wearable devices with photoplethysmography sensors.

In this paper, we propose *FacER*, a **Fac**ial **E**xpression **R**ecognition system based on near-ultrasound acoustic sensing on a smartphone. We utilize commodity smartphones to emit near-ultrasound signals (19-23 kHz) towards the user's face. As shown in Fig. 1a, the microphone on the smartphone will receive the reflected echoes from the face surface, which will carry the facial expression information. By examining the fine-grained echo patterns, *FacER* can differentiate six different types of facial emotional expressions as shown in Fig. 1b. We demonstrate that in a scenario where a user holds a smartphone at a relatively stable distance (*e.g.* 20-50 cm), *FacER* can recognize six universal facial expressions with more than 85% accuracy. Compared with the state-of-the-art acoustic-based SonicFace [3], *FacER* uses a commodity smartphone without requiring a customized microphone array, which presents a

more convenient and usable solution for FER.

There are two main challenges in designing *FacER*. First, acoustic interference, such as the multipath of reflected signals and environmental noises, can significantly impact recognition performance. Apart from the echoes reflected from the face, the microphones can also receive echoes from the surrounding obstacles. It is imperative to mitigate or remove any undesired signals. However, even with the application of various noise-reduction techniques, the environmental noises at a similar frequency to the emitted signal could persist. Therefore, we propose a contrastive external attention-based learning model to depict more robust facial expression feature representations. In this way, the model can distill the universal features of expressions while eliminating background noises.

Second, different users express facial expressions in different manners, and even the same user could express differently at different times. This will result in the domain adaptation issue, a common issue in machine learning (ML) models, where a model trained on a labeled dataset (source domain) cannot be successfully applied to a testing dataset (target domain). This is caused by the distribution drift between the source domain and target domain, which violates the common independent and identically distributed (i.i.d.) assumption of ML models. Therefore, we propose a domain adaptation contrastive learning algorithm to align the distribution of the source domain and target domain dataset. In this way, *FacER* can achieve consistent performance in recognizing various expressions across different users.

We evaluate *FacER* on a dataset collected from 20 volunteers of different ages, genders, and skin colors in various environments. We show that *FacER* can effectively recognize six different facial expressions from different users. Remarkably, it achieves more than 90% test accuracy when the training and testing datasets have the same distribution. It achieves more than 85% accuracy when training and testing on different sets of users. In summary, we make the following contributions:

- We design *FacER*, which uses a contrastive external attention-based acoustic facial expression recognition model to learn representative and robust facial expression features while eliminating the background noise.
- We propose the domain adaptation contrastive learning algorithm to align the training and testing data distributions, which can largely mitigate the negative effects of variations in users' facial expressions.
- We implement the smartphone-based system *FacER* and perform the evaluations in various real-life scenarios. The results show that *FacER* outperforms the state-of-the-art approaches by more than 10% in recognition accuracy with high mobility and convenience.

The rest of the paper is organized as follows. In Section II, we summarize the related work. We introduce the preliminary knowledge in Section III. In Section IV, we present *FacER* and the proposed contrastive attention model. We provide implementation details in Section V and evaluate the performance of *FacER* in Section VI. We discuss the future work in Section VII. Finally, we conclude in Section VIII.

II. Related Work

To identify the emotional status of users, researchers have proposed to use body sensors to capture physiological signals such as electromyographic (EMG) [1] and heart rate [16]. However, it normally requires a large time window such as the 30s [17] of physiological signals to analyze the profile of emotions, incurring low efficiency. ExpressEar [18] applies commercial earables augmented with inertial sensors to capture facial muscle movements associated with expressions. Similarly, to recognize facial expressions, FaceListener [19] captures facial skin deformations by transforming a headphone into an acoustic sensing device. However, it is cumbersome for users to wear these external devices to sense emotions.

Another approach, wireless and mobile sensing, has been used for behavior recognition tasks such as identifying daily activities [20]–[24], and facial expressions [13]. For instance, WiFace [13] analyzes the channel state information in WiFi signals captured by the router, which has three antennas positioned on the top of the user's head. The extracted waveform patterns can be used to recognize facial expressions. Hof *et al.* [25] propose the mm-Wave radar system with massive-antenna elements to conduct facial recognition. However, these approaches all require additional hardware and a special placement setup.

Due to the wide availability of speakers and microphones, acoustic sensing has also been widely studied. The basic idea is to use the speaker to emit acoustic signals and analyze the echo signals reflected by the sensing objects. It has a wide range of applications including breathe monitoring [26], user authentication [27], and activity recognition and tracking [28]–[30]. For example, EchoPrint [27] combines acoustic and visual signals for user authentication. It emits inaudible acoustic signals towards the user's face and extracts features from the echoes bouncing off the 3D facial contour. TeethPass [29] uses earbuds to collect occlusal sounds in binaural canals to achieve user authentication. LASense [30] achieves fine-grained activity sensing by increasing the number of overlapped samples between the emitted and received acoustic signals through signal processing, which enhances both sensing accuracy and range. SonicFace [3] detects emotional expressions by deploying a customized microphone array to capture reflected echoes. It calculates the frequency and phase shifts of pure tone signals to extract expression features. However, their method cannot compute the fine-grained facial information due to the limited frequency resolution of pure tone signals.

III. Preliminaries

In this section, we introduce the background of acoustic signals, the preliminary knowledge of attention mechanisms and contrastive learning.

A. Acoustic Signal

The acoustic signal refers to a coded chirp signal transmitted by a device. Particularly, the chirp signal can be regarded as the component of sawtooth modulation in the Frequency-Modulated Continuous Wave (FMCW), which changes its

operating frequency during the measurement. In FMCW, the frequency of the signal will periodically increase or decrease during transmission. The differences in frequency between the transmitted and received signal are proportional to the time delay Δt. Therefore, the FMCW can measure the small movement of the target, which is calculated as follows:

$$R = \frac{v_0|\Delta t|}{2} = \frac{v_0|\Delta f|T}{2B}, \tag{1}$$

where R is the distance between the sound source and the reflecting object, v_0 is the speed of sound (340 m/s) at 20 °C, Δt is the delay time, and Δf is the measured frequency difference. B is the chirp frequency bandwidth, and T is the chirp periodic time. The duration of the transmitted waveform T should be greater than the required receiving time for the distance measuring range. We use $\frac{B}{T}$ to measure the frequency shift per unit of time. Therefore, with the features of the FMCW, the chirp signal can help group reflections from various distances into multiple range bins.

B. Attention Mechanism

Similar to the human visual system, attention mechanisms [31] aim to focus limited attention on key information, which saves resources and distills the most essential information. The basic idea of attention mechanisms is to combine all of the encoded input features in a weighted fashion, with the most important features receiving the highest weights.

Given a feature map $F \in \mathcal{R}^{N\times d}$, where N is the number of elements and d is the feature dimension of each element, by multiplying three different random initialized weight matrixes, self-attention projects the F into a query matrix $Q \in \mathcal{R}^{N\times d'}$, a key matrix $K \in \mathcal{R}^{N\times d'}$, and a value matrix $V \in \mathcal{R}^{N\times d}$. The self-attention is represented as follows:

$$F_{out} = softmax(QK^T)V, \tag{2}$$

where $softmax(QK^T)$ is the attention matrix, and the F_{out} is the improved feature representation of the input F.

C. Contrastive Learning

Contrastive representation learning aims to learn an embedding space where dissimilar samples are spread out and similar samples remain close together. Normally, a positive pair refers to a pair of samples that have the same label, and a negative sample pair has different labels.

The supervised contrastive loss [32] is defined as follows when the training objective includes multiple positive and negative pairs in one batch:

$$\mathcal{L}_c = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} log \frac{exp(z_i \cdot z_p/\tau)}{\sum_{a\in A(i)} exp(z_i \cdot z_a/\tau)}, \tag{3}$$

where I is a set of samples x within a batch, $A(i) \equiv I\backslash\{i\}$, $P(i)$ is a set of indices of all positives in the multiviewed batch, $|P(i)|$ is the cardinality, $z = Proj(Enc(x))$ is the encoded feature representation by an encoder network $Enc(\cdot)$ and a projection network $Proj(\cdot)$ such as a linear layer network. The $\cdot$ denotes the inner product, and τ is a temperature parameter to adjust the final results.

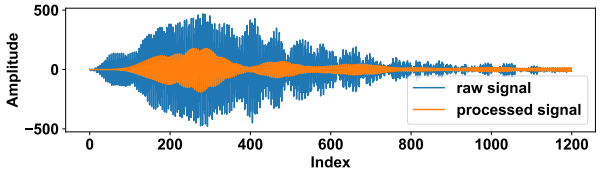


Fig. 2: The raw signal and the signal after noise removal.

IV. System Design

In this section, we design the acoustic sensing and signal pre-processing techniques to prepare the sensing data. Then, we propose the contrastive external attention-based domain adaptation model for acoustic facial expression recognition.

A. Acoustic Sensing Design

A unique facial expression contour is a distinct collection of various reflecting surfaces, which can produce a unique sum of individual echoes. As different objects absorb and attenuate sound waves to different degrees, it is possible to differentiate between the reflected echoes from objects and those from facial expressions [27].

1) Signal Generator: On a smartphone, there are usually one main speaker and microphone at the bottom or on the back, and an earpiece speaker and microphone at the top of the phone body. Considering that the earpiece speaker has a proper position to illuminate a user's face as shown in Fig. 1a, we select the earpiece speaker for emitting the acoustic signal. Similarly, considering the gesture of holding a phone, the top microphone is chosen since it is less affected by the hand.

The acoustic signal should satisfy the following properties. (i) The period of the signal should be moderate to minimize the overlap of echoes from various distances. (ii) The signal should be distinguishable from the background noise in the frequency domain, while the noise frequency is mostly under 8 kHz. (iii) The signal ought to be inaudible in real-world scenarios. Therefore, considering that the change of facial expression happens within 1 second, we choose a chirp signal of 25 milliseconds with frequency sweeping from 19-23 kHz to compose the inaudible acoustic signal with fading at the start and the end. In this way, we can better capture the expression features caused by muscle movements and filter out echoes from different obstacles. According to the Nyquist sampling theorem, the sampling rate is set as 48 kHz. *FacER* leverages the earpiece speaker to periodically emit the near-ultrasound signals and simultaneously uses the microphone to receive the signals reflected by the face. We set up the time interval as 50 milliseconds to allow all the reflected signals from the previous chirp to be received, such that we can separate two chirps before the following chirp is transmitted.

2) Noise Removal: Expression-irrelevant signals from various background noises, nearby people and obstacles should be filtered out. To achieve that, we first apply a 19-23 kHz band-pass filter to keep the desired frequency band and filter out the background noise. There remain three main types of signals in the received recording: (i) the *direct path signal* directly travels from the speaker to the microphone; (ii) the

Fig. 3: The spectrogram of six expressions in 50 milliseconds. The first row is from a woman without a mask, the second row is from a man without a mask, and the third row is from the same man with a mask.

major echo signal is the mix of echoes from the facial contour, which is the interest of *FacER*; and (iii) the *noisy echo signals* are echoes from different obstacles in the environment because of the multipath of reflected signals.

To remove the interference of the *direct path signal*, inspired by AIM [33], we use a separate speaker and microphone to record the direct transmission by positioning the devices in a clean and quiet space. Therefore, we remove the direct path signal from the received samples by minimizing $||S - cS_d||$, where S denotes the received samples, S_d denotes the pre-recorded direct signals, and c is a scaling coefficient to achieve optimal cancellation which is set as 0.9 in our experiments. In this way, we remove the *direct path signal* from the speaker to the microphone. Fig. 2 shows a segment of the raw signal (in blue) and the processed signal (in orange) after filtering out the background noise and the direct path signal noise.

Then, we proceed to remove the *noisy echo signals*. In the *FacER* application scenario, a user is facing a phone, so we assume that there is a relatively stable and static distance between the phone and the user's face. We consider filtering out the *noisy echo signals* from the nearby obstacles at different distances. The FMCW provides distance measurement, which is an important tool for differentiating among different echoes when more than one source of reflection is received. A comfortable distance between human eyes and the phone is $25-50\ cm$ [27]. Therefore, based on Eq. (1), we can calculate the desired frequency shift as:

$$|\Delta f| = \frac{2RB}{Tv_0}. \tag{4}$$

Thus, $|\Delta f|$ is between 235 Hz and 470 Hz. We further analyze the FMCW distance measurement resolution. Given the minimum measurable frequency shift $\Delta f_{min} = 1/T$, we can compute the resolution d_r that FMCW separates mixed echoes as:

$$d_r = \frac{v_0 \Delta f_{min}.T}{2B} = \frac{v_0}{2B}. \tag{5}$$

Thus, d_r is $\frac{340m/s}{2\times 4000s^{-1}} = 4.25\ cm$. The resolution of the *major echo signal* corresponding to a single sample is $\frac{v_0}{2F_s} = 3.54\ mm$, where F_s is the sampling frequency 48 kHz.

B. Contrastive Attention-based Domain Adaptation

We use the Short-Time Fourier Transform (STFT) with the Hann window to process the signal, which outputs the complex amplitude, and we compute the absolute values of the STFT values. The generated spectrogram is used as input for our proposed model in Algorithm 1. Fig. 3 shows the spectrogram of the segmented major echo signals of different expressions from two volunteers. We can observe that, for the same person, different expressions will yield different spectrograms.

However, we make two observations that would affect the modeling performance. First, it is nearly impossible to entirely remove noisy echo signals from different obstacles at various distances in some scenarios. We set up the desired frequency shift $|\Delta f|$ as 500 Hz in Fig. 3. However, there could be minimal multipath changes caused by body motions or objects between the face and the phone, which is hard to filter out since the resolution d_r for FMCW to separate mixed echoes is 4.25 cm. For example, the second row in Fig. 3 is from a man without a mask, while the third row is from the same man with a mask. We can observe subtle differences between the corresponding spectrograms (e.g., "surprise") in the two rows. Considering that the image-based expression recognition model can cope with various backgrounds around the face in an image to extract the facial expression-related features. Similarly, we aim to design the model to pay attention to the acoustic facial expression features.

Second, normal ML models face the poor generalization problem when there is a distribution shift between the training and testing datasets [34]. For example, as shown in the first and second rows of Fig. 3, for the same expression, different people will have different ways of expressing facial emotions, as manifested in the differences in the corresponding spectro-

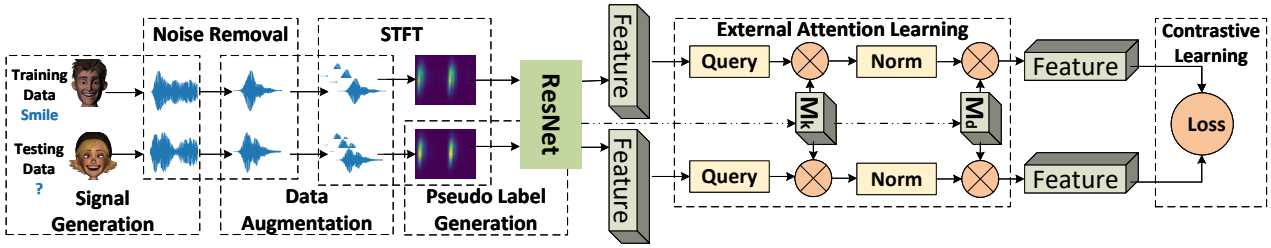


Fig. 4: The contrastive attention-based domain adaptation model for acoustic facial expression recognition.

grams. These differences will cause distribution shifts. Therefore, we should align the training and testing data distribution before training a classification model. Moreover, the model should only require a limited amount of enrollment data to improve its usability. Yet, it could be particularly challenging to extract useful features that can identify expressions from weak signals with limited acoustic samples.

To address these issues, we design the contrastive attention-based domain adaptation model to extract consistent acoustic facial expression features as shown in Fig. 4. The basic idea for domain adaptation is to minimize the expression feature representation distances across domains. However, it is challenging to gather enough data from a variety of populations for deep learning models to extract robust facial expression features. Therefore, we first propose to augment the dataset.

1) Data Augmentation: Data augmentation, which involves adding modified versions of existing data or generating new synthetic data from existing data, is a typical technique to address the data deficiency problem. First, it can mitigate the model overfitting problem when the original dataset is relatively small. Second, contrastive learning learns discriminative representations by bringing together positive pairs and separating negative pairs. The different augmented views of each sample can compose positive pairs that have the same labels, which enhances the model's discriminative ability.

We propose to use two methods for acoustic expression data augmentation. First, we shift the acoustic expression signal segment by the same distance. In accordance with the inverse square law of sound propagation, the amplitude of the signal is changed by a scale equal to the inverse square of the distance (*e.g.*, 0.3 meters). Second, for the generated spectrograms, we produce different versions of the spectrogram by multiplying the magnitudes of the spectrogram by a scalar. We generate the scalar by sampling from a Gaussian distribution with the mean being 0 and the standard deviation being 0.1. Considering our scenario when a user is holding a smartphone, a small device rotation at a fixed position creates negligible changes in the signal due to the omnidirectional nature of speakers and microphones, therefore, we only consider the changes in the device positions for acoustic signal transformation. In this way, we can grow and enrich our acoustic facial expression dataset for contrastive external attention learning.

2) Pseudo Label Generation: A domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ includes the feature space $\mathcal{X}$ and marginal probability distribution $P(X)$. If two domains are different, they have different $\mathcal{X}$ or $P(X)$, but the label space is the same. Suppose we have a source domain with a fully-labeled expression training dataset D_s, and an unlabeled testing dataset D_t is the target domain, which has the same categories as the source domain. The first problem is how to form the positive pairs from the same category of D_s and D_t when the labels in D_t are unknown.

Inspired by DeepCluster [35], we generate pseudo labels for unlabeled data in D_t according to the highest category probability. Specifically, we propose to use K-means clustering to generate pseudo labels, re-train the current model and adjust the noisy pseudo labels iteratively. For each iteration, we first calculate the centroid for each class in the target domain:

$$c_k^{(i)} = \frac{\sum_{x_t \in \mathcal{X}_t} \delta(z_t) g_t(x_t)}{\sum_{x_t \in \mathcal{X}_t} \delta(z_t)}, \tag{6}$$

where $c_k^{(i)}$ is the centroid for class k at the i_{th} iteration, $\delta(z) = \frac{exp(z)}{\sum_m exp(z_m)}$ is the softmax function that is used to output category probability. $g_t(x_t)$ outputs the representation of an acoustic expression sample x_t, and $z_t = h(g_t(x_t))$ is the linear transformed representation of x_t. The centroids can characterize the distribution of categories within the target domain. The pseudo labels are obtained via the nearest centroid:

$$y_t = \arg\min_k cos(g_t(x_t), c_k), \tag{7}$$

where $cos(\cdot)$ is the cosine distance between the representation embedding $g_t(x_t)$ and the centroid embedding c_k. As a result, we can obtain pseudo labels for the target domain dataset and compose the positive and negative pairs using both source and target domain datasets. Then, we iteratively update the model parameters by minimizing the loss function in Eq. (10) described below.

3) Attention-based Expression Learning: As mentioned before, removing all noisy echoes is almost impossible. Thus, the model should be able to distinguish the features of facial expression echoes from the background noisy echoes. The self-attention is widely used for learning robust features, but it has quadratic complexity and it does not consider the potential correlation among various samples. Multiple data samples (0.1s per sample) are generated during each expression (about 1s). Capturing correlations among these samples can facilitate the model to be focusing on the common and consistent acoustic facial expression features.

To generate the facial expression representation z, as shown in Fig. 4, we propose to use external attention [36] to focus on important features and implicitly learn correlations among

Algorithm 1: Contrastive Attention-based Cross Domain Acoustic Expression Representation Learning

Input: source dataset D_s, target dataset D_t, epoch E, iterations K per epoch, weight λ, contrastive attention-based model f

Output: source and target representations Z^s and Z^t

1 **for** $e = 1$ *to* E **do**
2 Calculate centroids in target domain using Eq. 6
3 Update pseudo labels for target data using Eq. 7
4 **for** $k = 1$ *to* K **do**
5 **for** *each batch* **do**
6 Extract features with f based on external attention in Eq. 8
7 Compute L_{ce} for each batch from D_s
8 Compute L_c from D_s and D_t using Eq. 9
9 Compute $\mathcal{L}_{ce}(\theta; D_s) + \lambda \mathcal{L}_{\mathrm{c}}^t(\theta; D_s, D_t)$
10 **end**
11 Back-propagate and update θ of model f
12 **end**
13 **end**
14 **for** *each batch* X_{batch} **do**
15 Generate source domain expression representation $Z^s_{batch} = f(X^s_{batch})$ for D_s
16 Generate target domain expression representation $Z^t_{batch} = f(X^t_{batch})$ for D_t
17 **end**
18 **return** Z^s and Z^t

all expression samples. Following symbols in Eq. (2), we first compute the attention map $A = QM_k^T$ by multiplying the query vector Q and the external learnable transposed key matrix $M_k \in \mathcal{R}^{S\times d}$. Q is projected from a feature map $F \in \mathcal{R}^{N\times d}$, where N is the size of feature elements, d and S are hyper-parameters. We normalize the attention map A and then multiply it with another external value matrix M_v. M_k and M_v are generated by additional linear layers, which can be optimized by back-propagation during training on the entire dataset. The attention-based feature map F_{out} is as follows:

$$F_{out} = Norm(QM_k^T)M_v. \tag{8}$$

In the end, we obtain the refined feature map with linear complexity $O(d \cdot S \cdot N)$, which is suitable for resource-constrained mobile devices.

4) Feature Alignment for Domain Adaptation: Reasonably, we assume that samples with the same label are closer to each other in the feature space, while samples from different categories are farther apart, no matter which domain they come from. With the augmented dataset, pseudo labels, and the attention-based learning model, we design contrastive learning to minimize the domain discrepancy by aligning facial expression features between the training and testing datasets.

Specifically, given an acoustic facial expression sample x_s from the source domain, and a sample x_t from the target domain, we minimize the distance between x_s and x_t if two samples are from the same class while maximizing the distance between two samples from different classes. The output of the model is domain-independent expression representations. Following the supervised contrastive loss in Eq. (3), we define the domain adaptation contrastive loss as follows:

$$\mathcal{L}_{\mathrm{c}}^t = \sum_{i\in I_t} \frac{-1}{|P_s(y_t^i)|} \sum_{p\in P_s(y_t^i)} log \frac{exp(z_t^i \cdot z_s^p/\tau)}{\sum_{a\in I_s} exp(z_t^i \cdot z_s^a/\tau)}, \tag{9}$$

where I_t denotes the set of target samples in a batch, I_s is the set of source samples, and $P_s(y_t^i)$ is a set of indices of all positive samples from the source domain. A positive sample means the label of the sample is the same as the pseudo label of the target anchor sample x_t. The domain adaptation contrastive loss aligns the expression representation in the target domain to the source domain. Finally, we define the acoustic expression representation learning loss function as:

$$\arg\min_{\theta} \mathcal{L}_{ce}(\theta; D_s) + \lambda \mathcal{L}_{\mathrm{c}}^t(\theta; D_s, D_t), \tag{10}$$

where $\mathcal{L}_{ce}$ is the standard cross-entropy loss applied on the D_s, λ is used to balance the two loss terms (0.5 in our experiments), and θ represents the model parameters.

Overall, the proposed contrastive external attention-based acoustic expression representation learning model is presented in Algorithm 1. After data augmentation, in each epoch, we first generate pseudo labels for target domain acoustic samples (lines 2-3). Then, we minimize the loss and back-propagate to update the model f (lines 4-12). After training, model f can align features for domain adaptation, and in turn minimize the distribution shift. We use the trained model f to generate the acoustic expression representations Z^s and Z^t (lines 14-17). Finally, we can train a classifier on source domain representations Z^s, and generate labels for the target domain representations Z^t. The proposed algorithm effectively enhances the expression recognition model performance across different users.

V. Implementation

A. Data Collection

We recruited 20 volunteers (16 males and 4 females) to participate in the acoustic facial expression data collection process. To guarantee the heterogeneity of collected facial expressions, the recruited volunteers have different skin colors and grow up in different regions of the world such as Asia, North America, and Europe. Their ages range from 20 to 38. The participants are allowed to wear glasses, a hat, and other accessories during the data collection process. To simulate different scenarios in real-life, we collect data at different locations (*e.g.*, office, dining hall, garden) with different background noise. For example, we collect data in an office with people having conversations, online meetings, and computer alarm beeping.

We show the six common facial expressions: anger, disgust, fear, happiness, sadness, and surprise as examples at the beginning of data collection. Then, a volunteer starts with a poker face and performs their preferable styles of six expressions

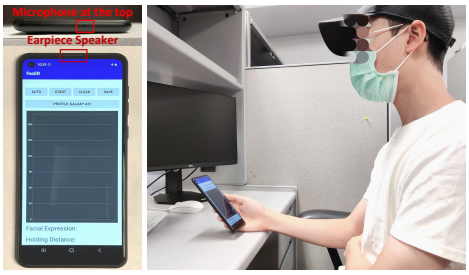


Fig. 5: A smartphone and a volunteer for data collection.

accordingly. Fig. 5 shows the example smartphone and a volunteer during the data collection process. The volunteers are allowed to hold the phones in their most comfortable gesture while watching the smartphone screen. Considering the necessity of wearing masks during the Covid-19 pandemic, we also consider expression recognition in scenarios where the participants are wearing disposable masks. For each expression, we collect about 5 seconds when the participants wear masks, and another 5 seconds when they do not.

The acoustic facial expression collection process repeats 10 times for each expression per person with rest breaks during the data collection process. An independent observer helps record the **label** of each acoustic expression sample as the ground truth. The whole data collection process takes about a week intermittently. The collected dataset is over 1 GB in plain text format. Finally, we get 20,535 samples with the sliding windows segmentation, of which the window size is 0.25s. The design consideration is that if the window size is too small, it will be hard to reflect the integral process of facial muscle movements. If the window size is too large, it will easily mix instantaneous variations of different facial expressions.

B. Experimental Setup

We use two Android phones: Samsung Galaxy A21, and OnePlus 8T to collect the acoustic signals. We develop the app to emit and collect signals based on the LibAS [37] and Chaperone [38], which provide a framework for acoustic sensing applications. LibAS can simplify the necessary signal processing procedures such as synchronization, which determines the start position of the sent signals in the received audio. The sensing signal is introduced in Section IV-A, which is a chirp signal modulated from 19-23 kHz. The app emits the signal from the earpiece speaker of a smartphone.

We employ the SciPy library for signal processing. We use ResNet-18 [39] as the backbone model for our contrastive attention-based expression recognition model, which is implemented with Pytorch. We train the model on TensorEX Ubuntu 20.04 Server with 256GB DDR4 memory, Intel(R) Xeon(R) Gold 5218R 2.10GHz CPUs, and RTX A6000 GPUs.

VI. Evaluation

In this section, we evaluate the impacts of multiple factors (*e.g.*, location, time, people, mask) on the performance of *FacER* in recognizing different acoustic facial expressions.

A. User Dependent Evaluation

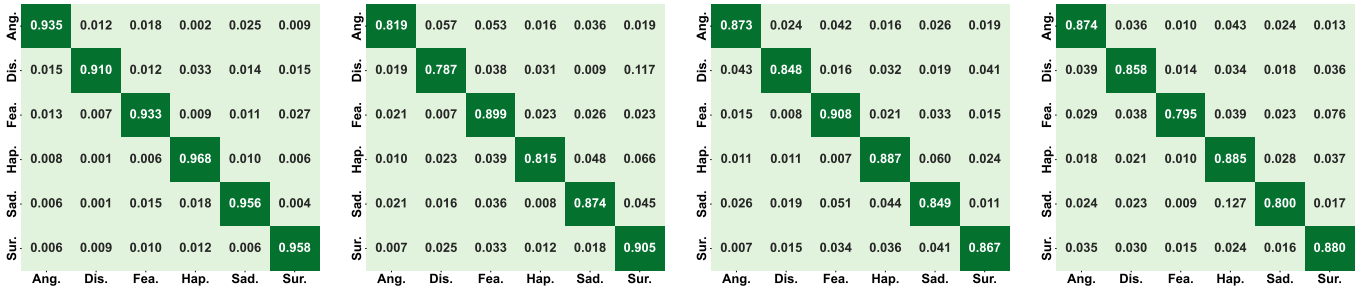
Case 1a. We first consider a simple scenario, where we can collect and label acoustic data from a group of users. Our goal is to recognize the facial expressions from this specific group of users, namely *mix testing*. We split 80% of the whole dataset as the training set and the remaining 20% as the testing set. We only use the external attention-based ResNet-18 model without cross-domain adaptation to implement classification. We use the SGD with a momentum of 0.9. The learning rate is 0.1 with a learning rate decay rate of 0.01. We show the performance of *FacER* with an accuracy heat map in Figure 6a. As we can see, the model can recognize each acoustic facial expression with more than 91% accuracy. Besides, we implement the 10-fold cross-validation, and the average testing accuracy is 94.3% with a standard deviation of 0.4%. However, the results are unsurprising because the training dataset and testing dataset belong to the same distribution. As a result, the model can easily fit in the dataset.

Case 1b. Next, we consider the impact of the environmental factors on the performance of *FacER*. We test the model performance on the data collected in three different locations, i.e., office, dining hall, and garden. We repeat the *leave-one-place-out* evaluation by training on data from two places and testing on data from another place. We implement Algorithm 1 to extract feature representations, and then we use a linear classifier with two linear layers network with hidden layer sizes 1,024 and 256 to implement expression classification.

We compare the performance of *FacER* with two baselines: (i) ResNet-18 [39], which is our backbone model trained with cross-entropy loss without adopting attention; (ii) XHAR [34], which is the adversarial training-based domain adaptation method for human activity recognition. Considering adversarial training is another major direction for cross-domain adaptation, we choose it as the baseline. We adapt their methodology to make it applicable for facial expression recognition.

We report the average accuracy, precision, recall, and F1 values across three locations in Fig. 7, which is a bar plot on the polar axis. *FacER* achieves 89.8% average accuracy, which is a bit lower than the mix testing method (94.3%). The standard deviations of the accuracy of FacER, ResNet, and XHAR from three leave-one-place-out evaluations are 0.017, 0.014, and 0.016, respectively. The results show that the location causes the distribution shift and affects the model performance because of the noise caused by different obstacles. Nevertheless, the model still achieves high performance with an F1 value of 90%, which is higher than the XHAR method (82.6%). The result proves that *FacER* can learn consistent and robust acoustic facial expression features even in a noisy environment with various types of noises.

Case 1c. We consider a more complex scenario that is related to time variation. The consideration is that the same facial expression may not be consistent over time. For example, sometimes people may show a wide smile while other times people may show a gentle smile to express happiness. Therefore, to evaluate the time factor for acoustic facial



(a) Case 1: Mix testing. (b) Case 2: Testing on Women's. (c) Case 3: Testing on Mask's. (d) Case 4: Testing on No Mask's.

Fig. 6: Evaluations of different cases. Case 1: training and testing on the mixed dataset; Case 2: training on men's dataset and testing on women's dataset; Case 3: training on users with masks and testing on users without masks; and Case 4: training on users without masks and testing on users with masks. The tick labels are angry, disgust, fear, happiness, sadness, and surprise. The x-axis contains the prediction labels, and the y-axis contains the ground truth labels.

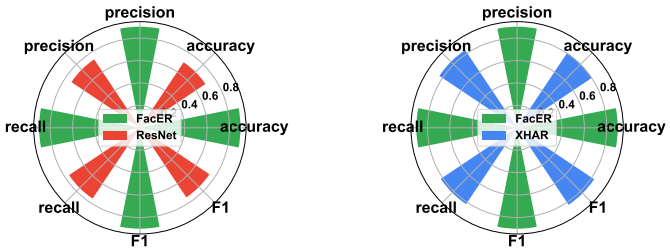


Fig. 7: The bars of location factor evaluation.

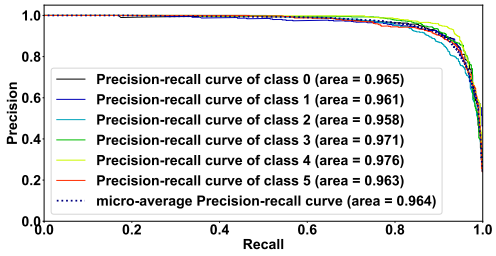


Fig. 8: The precision-recall curve of time factor evaluation.

expression recognition, we split the dataset into two subsets according to the time. The first subset is the first 8 times of data collection for each facial expression per person, which is the training dataset. The second subset is the last 2 times of data collection, which forms the testing dataset.

The results show that the time factor affects the model performance. Under the time factor condition, the accuracy of *FacER* is 90.3%. The precision and recall are 90.5% and 90.3%, respectively. The model has slightly worse performance (4% lower accuracy) than Case 1a, which shows that the changes of the same expression over time indeed degrade the recognition accuracy. Nevertheless, *FacER* can still achieve high performance, which shows the efficacy of the designed model in acoustic expression recognition. We show the precision-recall score curve in Fig. 8. Precision is $\frac{tp}{tp+fp}$ and recall is $\frac{tp}{tp+fn}$, where tp means true positive, fp means false positive, fn means false negative. The tradeoff between precision and recall for various thresholds is depicted by the precision-recall curve. The area in Fig. 8 is calculated by the average precision (AP): $AP = \sum_n (R_n - R_{n-1}) P_n$ where P_n and R_n are the precision and recall at the n_{th} threshold automatically set by Scikit-plot [40]. For different classes, *FacER* can achieve at least a 0.958 AP score. A high precision-recall (area) score under the curve shows both high recall and high precision, which indicates low false-positive and false-negative scores.

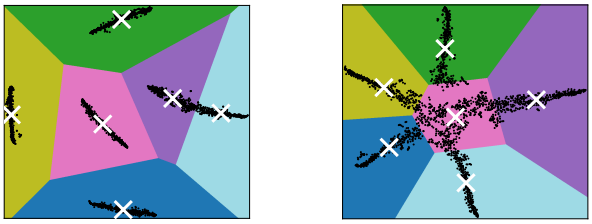
B. User Independent Evaluation

Case 2. Now we consider a more general scenario where the model is trained and tested on different sets of users. Different people have different face geometries and behaviors. As a result, they may produce distinct feature patterns. To evaluate the performance of *FacER* in the user-independent scenario, we use the dataset from 16 men as the training dataset and that from 4 women as the testing dataset.

We first show the efficacy of the K-means model in clustering the acoustic facial expression representations. We randomly select 2,048 samples from the training dataset in Fig. 9a and from the testing dataset in Fig. 9b, respectively. As we can see, the K-means model can separate the six types of expressions from both training and testing sets, which indicates the learned representations contain consistent and distinctive features. During the iterative training process, the pseudo labels of the testing data can be accurately updated. Therefore, the proposed model can effectively compose positive and negative pairs for contrastive attention learning.

The accuracy heatmap is shown in Fig. 6b, and the average accuracy is 85%, which is lower than the results from Case 1a because of the user expression differences. We notice that men and women have different ways of expressing "disgust", which causes *FacER* to only achieve 78.7% accuracy. But the result shows that men and women have high similarity in expressing "surprise", which achieves 90.5% accuracy.

To evaluate the efficacy of the domain adaptation of *FacER*, we compare *FacER* with the three baselines: XHAR [34], SonicFace [3], and ResNet [39] as shown in Fig. 10. XHAR and ResNet are set up as in Case 1b. SonicFace uses both FMCW and pure tone signals to extract different features. It focuses on tracking the movement of facial components such as the eye, eyebrow, and cheek, and it uses 1D convolution for signal processing. In contrast, we regard the spectrogram of the received echos as an image, which is the instantaneous



(a) Clustering on the training set. (b) Clustering on the testing set.

Fig. 9: K-means clustering on 2,048 sampled representations, which are processed with TSNE dimension reduction.

static facial expression. Different expressions will generate different spectrogram features as the different pixels of facial expression images. Therefore, we use the 2D convolution. Due to the key difference in the applied signals, we directly report their best performance in Fig. 10. For SonicFace, the average accuracy of the intra-session classification is 78.6%, while the best accuracy of the user-independent model with calibration is 72%. For the user-independent Case 2 scenario, *FacER* can achieve 85% accuracy and 85.2% F1 value. The adversarial training-based XHAR method only achieves 79.1% accuracy and 78.7% F1 value. The results show the superior performance of our proposed contrastive external attention-based representation learning method, as it helps extract robust and accurate acoustic facial expression features.

C. Mask Factor Evaluation

Case 3 and 4. Masks are obvious obstacles to camera-based facial expression recognition methods. It is very common for people to wear masks during the Covid-19 pandemic. We investigate the performance of *FacER* when people wear masks. As mentioned before, half of the dataset is from volunteers wearing masks (mask dataset), and the remaining half is from volunteers without wearing masks (plain dataset). We train *FacER* on the plain dataset and test on the mask dataset, and the result is shown in Figure 6c. The average accuracy is 87.2%. Then, we train *FacER* on the mask dataset and test on the plain dataset, and the result is shown in Figure 6d. The average accuracy is 84.9%. As we can see, the mask has a noticeable impact on the performance of *FacER*.

Another interesting finding is that *FacER* achieves the highest accuracy of 90.8% about the "fear" expression when training on the plain dataset, while it achieves the lowest accuracy of 79.5% about "fear" when training on the mask dataset. This can be attributed to the fact that the mask itself will generate some echoes back to the microphone, which will cloud the echoes from the facial expressions. It is harder for *FacER* to learn robust "fear" features when people wear masks. Nevertheless, for the "happiness" expression, people have a larger facial muscle movement, which could trigger the movement of masks. As a result, it can achieve 88.7% and 88.5% accuracy when testing on the mask and plain datasets, respectively. Overall, in different mask factor scenarios, *FacER* can achieve comparable accuracy with Case 2. The results

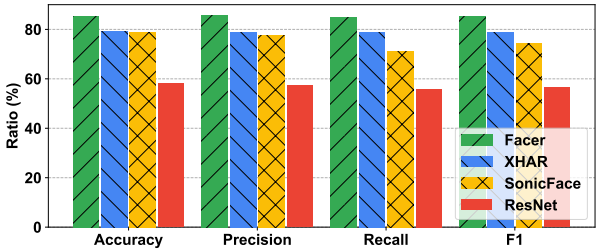


Fig. 10: The comparison across different models.

confirm the cross-domain adaptation ability of our proposed contrastive attention-based representation learning method.

VII. Discussion

Data. *FacER* is a data-driven acoustic facial expression recognition system based on the contrastive attention-based deep learning model. The quality and quantity of the acoustic sensing data are essential to the performance of *FacER*. In the future, we will collect more data from female users and underrepresented groups to reflect the diversity of facial expressions. We will also consider other forms of emotional expressions such as hand gestures. Moreover, we will collect data on facial expressions that gap a longer time such as after one week for the time factor evaluation. We will also try applying generative models to synthesize more data.

Model. In this work, we mainly consider the scenarios where the users hold the phone at a distance of 20-50 cm. In the future, we will investigate the performance of *FacER* when the distance between the user and the phone is longer. We will also study how to determine the start and end of an expression, so we can effectively segment acoustic signals to extract facial expressions for inference in real-time.

System. We understand the design difference between different phone hardware, especially its impact on the direct path from the speaker to the microphone. We will evaluate the effect of different smartphones on ultrasound signal transmission in the future. Besides, many other real-world impact factors such as slight hand shaking, and the orientation of a phone should be taken into consideration. Furthermore, we will continue optimizing resource usage on mobile devices and mitigate the security and privacy concerns of mobile sensing data.

VIII. Conclusion

In this paper, we designed *FacER*, an acoustic facial expression recognition system using a smartphone earpiece speaker and microphone. To enhance the accuracy and robustness of *FacER*, we proposed a contrastive external attention-based representation learning model to learn robust expression features across different users in various noisy scenarios. Real-world experiments show that *FacER* achieves expression recognition with more than 85% accuracy even when the users are wearing a mask, a new norm during the Covid-19 pandemic.

Acknowledgement

We appreciate the anonymous reviewers for their insightful comments on our work. This work was supported in part through National Science Foundation grant CNS-1950171.

REFERENCES

[1] C. Chen, K. Sun, and X. Zhang, "Exgsense: Toward facial gesture sensing with a sparse near-eye sensor array," in *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*, 2021, pp. 222–237.

[2] J. Nie, Y. Hu, Y. Wang, S. Xia, and X. Jiang, "Spiders: Low-cost wireless glasses for continuous in-situ bio-signal acquisition and emotion recognition," in *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 2020, pp. 27–39.

[3] Y. Gao, Y. Jin, S. Choi, J. Li, J. Pan, L. Shu, C. Zhou, and Z. Jin, "Sonicface: Tracking facial expressions using a commodity microphone array," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–33, 2021.

[4] T. L. Nwe, F. S. Wei, and L. C. De Silva, "Speech based emotion classification," in *Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology. TENCON 2001 (Cat. No. 01CH37239)*, vol. 1. IEEE, 2001, pp. 297–301.

[5] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd annual international conference on mobile computing and networking*, 2016, pp. 95–108.

[6] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[7] P. Ekmann, "Universal facial expressions in emotion," *Studia Psychologica*, vol. 15, no. 2, p. 140, 1973.

[8] E. A. Clark, J. Kessinger, S. E. Duncan, M. A. Bell, J. Lahne, D. L. Gallagher, and S. F. O'Keefe, "The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review," *Frontiers in psychology*, vol. 11, p. 920, 2020.

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[10] J. V. Patil and P. Bailke, "Real time facial expression recognition using realsense camera and ann," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 2. IEEE, 2016, pp. 1–6.

[11] T.-W. Shen, H. Fu, J. Chen, W. Yu, C. Lau, W. Lo, and Z. Chi, "Facial expression recognition using depth map estimation of light field camera," in *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE, 2016, pp. 1–4.

[12] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "Wife: Wifi and vision based intelligent facial-gesture emotion recognition," *arXiv preprint arXiv:2004.09889*, 2020.

[13] Y. Chen, R. Ou, Z. Li, and K. Wu, "Wiface: facial expression recognition using wi-fi signals," *IEEE Transactions on Mobile Computing*, vol. 21, no. 1, pp. 378–391, 2020.

[14] S. Choi, Y. Gao, Y. Jin, S. j. Kim, J. Li, W. Xu, and Z. Jin, "Ppgface: Like what you are watching? earphones can" feel" your facial expressions," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–32, 2022.

[15] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones, "Modeling stylized character expressions via deep learning," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 136–153.

[16] D. Rommel, J. Nandrino, M. Jeanne, R. Logier *et al.*, "Heart rate variability analysis as an index of emotion regulation processes: interest of the analgesia nociception index (ani)." in *2012 Annual international conference of the IEEE engineering in medicine and biology society*. IEEE, 2012, pp. 3432–3435.

[17] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 73–78.

[18] D. Verma, S. Bhalla, D. Sahnan, J. Shukla, and A. Parnami, "Expressear: Sensing fine-grained facial expressions with earables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–28, 2021.

[19] X. Song, K. Huang, and W. Gao, "Facelistener: Recognizing human facial expressions via acoustic sensing on commodity headphones," in *21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2022.

[20] G. Wang, L. Zhang, Z. Yang, and X.-Y. Li, "Socialite: Social activity mining and friend auto-labeling," in *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*. IEEE, 2018, pp. 1–8.

[21] F. Han, L. Zhang, X. You, G. Wang, and X.-Y. Li, "Shad: Privacy-friendly shared activity detection and data sharing," in *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2019, pp. 109–117.

[22] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 313–325.

[23] C. Li, M. Liu, and Z. Cao, "Wihf: enable user identified gesture recognition with wifi," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 586–595.

[24] C. Li, Z. Cao, and Y. Liu, "Deep ai enabled ubiquitous wireless sensing: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–35, 2021.

[25] E. Hof, A. Sanderovich, M. Salama, and E. Hemo, "Face verification using mmwave radar sensor," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2020, pp. 320–324.

[26] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, "Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 54–66.

[27] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 321–336.

[28] Z. Gao, A. Li, D. Li, J. Liu, J. Xiong, Y. Wang, B. Li, and Y. Chen, "Mom: Microphone based 3d orientation measurement," in *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 2022, pp. 132–144.

[29] Y. Xie, F. Li, Y. Wu, H. Chen, Z. Zhao, and Y. Wang, "Teethpass: Dental occlusion-based user authentication via in-ear acoustic sensing," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1789–1798.

[30] D. Li, J. Liu, S. I. Lee, and J. Xiong, "Lasense: Pushing the limits of fine-grained activity sensing using acoustic signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–27, 2022.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[33] W. Mao, M. Wang, and L. Qiu, "Aim: Acoustic imaging on a mobile," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018, pp. 468–481.

[34] Z. Zhou, Y. Zhang, X. Yu, P. Yang, X.-Y. Li, J. Zhao, and H. Zhou, "Xhar: Deep domain adaptation for human activity recognition with smart devices," in *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2020, pp. 1–9.

[35] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.

[36] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *arXiv preprint arXiv:2105.02358*, 2021.

[37] Y.-C. Tung, D. Bui, and K. G. Shin, "Cross-platform support for rapid development of mobile acoustic sensing applications," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018, pp. 455–467.

[38] J. Chen, U. Hengartner, H. Khan, and M. Mannan, "Chaperone: Real-time locking and loss prevention for smartphones," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 325–342.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] R. Nakano, "Scikit-plot," https://github.com/reiinakano/scikit-plot, 2017.