UltraCLR: Contrastive Representation Learning Framework for Ultrasound-based Sensing

XUN WANG, State Key Lab. for Novel Software Technology, Nanjing University, China
ZHIZHENG YANG, State Key Lab. for Novel Software Technology, Nanjing University, China
WEI WANG*, State Key Lab. for Novel Software Technology, Nanjing University, China
HAIPENG DAI, State Key Lab. for Novel Software Technology, Nanjing University, China
SHUYU SHI, State Key Lab. for Novel Software Technology, Nanjing University, China
QING GU, State Key Lab. for Novel Software Technology, Nanjing University, China

We propose UltraCLR, a new contrastive learning framework that fuses dual modulation ultrasonic sensing signals to enhance gesture representation. Most existing ultrasound-based gesture recognition tasks rely on a large amount of manually labeled samples to learn task-specific representations via end-to-end training. However, they cannot exploit unlabeled continuous gesture signals that are easy to collect. Inspired by recent self-supervised learning techniques, UltraCLR aims to autonomously learn a ubiquitous gesture signal representation that can benefit all tasks from low-cost unlabeled signals. We use the STFT heatmap as a secondary input and leverage the contrastive learning framework to improve the high-quality Channel Impulsive Response (CIR) heatmap input representations. The learned representations can better represent the spatial-position information and intermediate states of gesture movement. With the representation learned by UltraCLR, we can greatly reduce the complexity of downstream gesture recognition tasks so that they can be completed using a simple classifier trained with a small training set and a lower computational cost. Our experimental results show that UltraCLR outperforms state-of-the-art gesture recognition systems with only a few labeled samples, and achieves more than 85% reduction in computational complexity and over 9x improvement in inference speed.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Ultrasound-based sensing, contrastive learning, gesture recognition

1 INTRODUCTION

Device-free gesture recognition is one of the novel human-computer interaction (HCI) solutions typically empowered by machine learning techniques [20, 36, 38, 46]. By intelligently analyzing sound or Radio Frequency (RF) signals reflected by the human body, device-free sensing systems allow users to interact with their devices

*Wei Wang is the corresponding author.

This work is supported in part by National Natural Science Foundation of China under Numbers 62272213, 62272223, and supported in part by the program B for Outstanding PhD candidate of Nanjing University.

Authors' addresses: Xun Wang, xunwang@smail.nju.edu.cn, State Key Lab. for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023; Zhizheng Yang, State Key Lab. for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023, yzz@smail.nju.edu.cn; Wei Wang, State Key Lab. for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023, wangwei.ww@gmail.com; Haipeng Dai, State Key Lab. for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023, 274175744@qq.com; ShuYu Shi, State Key Lab. for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023, ssy@nju.edu.cn; Qing Gu, State Key Lab. for Novel Software Technology, Nanjing University, 163 Xianlin Avenue, Nanjing, Jiangsu, China, 210023, guq@nju.edu.cn.

1550-4859/2023/5-ART $15.00
https://doi.org/10.1145/3597498

without wearing any sensors. To enable intelligent and accurate gesture recognition, the machine learning model often needs a large number of gesture samples for training. Therefore, how to reduce the efforts involved in collecting training samples is one of the key challenges for device-free gesture recognition.

Traditional task-specific gesture classification frameworks need a large amount of *labeled* gesture samples for training. These supervised training approaches face two challenges in the training and deployment phase. First, the supervised training dataset has to cover different scenarios, users, and perspectives to improve the generalization ability of the model. Collecting and labeling gesture samples would take a formidable amount of human effort. Sometimes even the end-users are involved in the dataset collection process. Second, each existing supervised dataset tends to serve only one specific task and is difficult to be reused by other gesture classification tasks. Coincidentally, there is no widely-accepted supervised dataset in device-free gesture recognition as influential as ImageNet in computer vision.

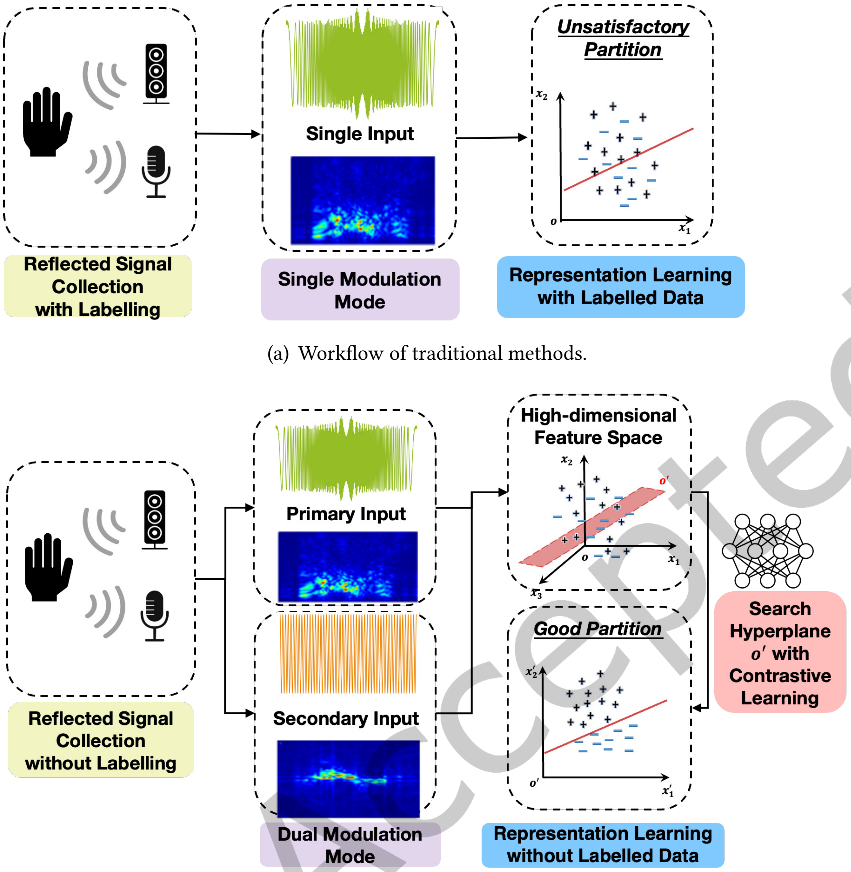
Self-supervised learning schemes can exploit *unlabeled* data to help various downstream classification tasks, especially for computer vision [5, 11, 15, 32, 33, 41]. As the representative of self-supervised learning methods, contrastive learning schemes learn efficient representations by reducing the distance between samples that are similar to each other. In computer vision, contrastive learning mainly utilizes data augmentation techniques such as rotation or translation to generate samples of the same object. However, such data augmentation schemes cannot be applied to gesture signals, as these transformations distort the critical information of gesture speed or movement distance. To the best of our knowledge, there is no contrastive-learning-based representation learning scheme for gesture signals.

This paper presents UltraCLR, a simple self-supervised framework for contrastive learning of gesture representations based on unlabeled ultrasonic signals. Unlike traditional methods that use only one modulation mode as input, we propose to use two different types of inputs that can be independently gathered from the same ultrasound signal. One is the *primary* input, which contains high-quality information about the gesture movements. The other is the *secondary* input, which is weakly correlated to the gesture. As shown in Figure 1, a combination of primary and secondary inputs increases the dimension of the feature space compared with the traditional methods. However, we find that the self-supervised framework based on contrastive learning can find a suitable hyperplane in the feature space for representation projection. Interestingly, such a low-dimensional hyperplane is more suitable as a feature representation space for the primary inputs. So we can gather unlabeled samples without user intervention to train the encoder for the primary input, and this process does not require adding additional sensors. As the representation only relies on the spatial-temporal relationship of movements, users do not need to perform specific gestures when gathering unlabeled samples. Instead, they can move freely and continuously as they want, and the unlabeled samples can cover a more extensive sample space that includes different gestures, users, perspectives, and environments. Downstream models can use the extracted powerful representations to reduce the complexity of the classification tasks and the inference time so that the models can be efficiently deployed on mobile devices.

We face two critical technical challenges when designing the UltraCLR framework.

First, *how can we design the sensing signal to extract both the primary and the secondary inputs?* To address this problem, we design an ultrasound signal that combines two signal modes: Orthogonal Frequency-Division Multiplexing (OFDM) and single-frequency Continuous Wave (CW). We add guard bands between them to ensure that they do not interfere with each other. By separately demodulating different signals, we extract high-quality heatmap inputs, Channel Impulsive Response (CIR), from the OFDM signal [20, 30] and coarse heatmap inputs, Short-Term Fourier Transform (STFT), from the CW signal [36, 37], using the same gesture signal frame.

Second, *how does the framework we design make good use of the above inputs and learn the valuable features from unsupervised data?* Given that the same gesture instance is correlated to both heatmaps, there is a correspondence between the CIR and STFT heatmaps for the same signal frame. Therefore, we can perform a contrastive learning task to match the STFT and CIR representations of the same frame in the unsupervised dataset. Although we


Reflected Signal Collection with Labelling
Single Input
Single Modulation Mode
Unsatisfactory Partition
Representation Learning with Labelled Data
Reflected Signal Collection without Labelling
Primary Input
Secondary Input
Dual Modulation Mode
High-dimensional Feature Space
Good Partition
Search Hyperplane o' with Contrastive Learning
Representation Learning without Labelled Data

(a) Workflow of traditional methods.

(b) Workflow of UltraCLR.

Fig. 1. Difference of workflows for current schemes on solving ultrasound-based sensing tasks.

have experimentally demonstrated that using both inputs at the same time leads to further performance gains, this requires the end task to collect inputs with dual modulation modes. From the perspective of practicality and reusability, the end-tasks do not need to gather secondary inputs or even be aware of the secondary inputs to enjoy the performance improvement from the enhanced representation of the primary inputs.

To demonstrate the effectiveness and generalizability of the learned representations, we conducted a series of extensive comparative experiments, and the results show that UltraCLR only requires a simple classifier and fewer labeled data to be fine-tuned to achieve the same recognition accuracy as the fully supervised model. This indicates that the UltraCLR representations are more efficient than the original CIR heatmaps.

The main contributions of our work are as follows:

- We design a dual-modulated signal to extract both the high-quality and low-quality heatmap inputs from the same signal.
- Our new self-supervised learning framework, UltraCLR, utilizes the unlabeled dataset based on dual-modulated signals by exploiting the secondary STFT heatmaps to enhance the representation of the primary CIR heatmaps.

• We conduct extensive experiments on learned representations using an unlabeled dataset with more than four hours of gesture movements from different types of users. Our experimental results show that, with the extracted representations, task-specific models can be trained with few labeled samples to reduce over supervised training overhead. Simultaneously, these models have lower computational complexity and allow for faster inference on resource-limited devices.

2 RELATED WORK

Existing works closely related to our approach can be classified into the following four categories.

Ultrasound-based gesture recognition tasks. Ultrasound signals have fine-grained environmental sensing capabilities and are privacy-preserving so that they gradually become alternative solutions to optical sensors in human activity recognition (HAR) tasks [9, 13, 22, 23, 30, 36, 43]. Existing deep learning models have achieved high-precision gesture recognition for specific tasks using extensive labeled training datasets. UltraGesture [20] extracts Channel Impulsive Response (CIR) heatmaps from ultrasonic gesture signals modulated and uses CNN for supervised training. However, a large number of gesture samples need to be collected and labeled to achieve acceptable performance. To reduce the cost of acquiring labeled data, RobuCIR [38] proposes data augmentation strategies and combines CNN and LSTM for the classification model. Both works use specific datasets for the given classification task for specific gesture sets, which increases the difficulties for other researchers to reuse their datasets for comparison.

Representation learning frameworks for sensing signals. Learning signal representations has always been the focus of the research in HAR tasks. Early works attempted to migrate supervised models from the source domain to the target domain or fine-tune the model directly using small datasets in the new domain to get domain-correlated features [10, 14, 18, 39, 44, 47]. However, these transfer learning methods still require a certain amount of labeled data in new scenarios, which undoubtedly damages the universality and convenience of the system. Recent works have recognized the value of using massive amounts of unlabeled data and proposed self-supervised learning frameworks for Inertial Measurement Units (IMU) based HAR [4, 24, 31, 40]. Among them, SelfHAR [31] trains the teacher network with a small amount of supervised data to guide the pre-labeling of large amounts of unlabeled data. LIMU-BERT [40] designs lightweight BERT self-supervised models for IMU data, and trains them on unlabeled datasets to extract global features. However, IMU signals have different physical properties than sound and RF signals, and we cannot apply these methods to device-free gesture signals that are noisy and have complex multipath interference. Some recent work on self-supervised human activity recognition based on RF signals has also made progress. Among them, RadarAE [26] uses the MAE-like mask method to train the self-encoder, but it is still doubtful whether the pixel-based regression method can obtain advanced semantic information such as the spatial motion state. RF-URL[27] uses the advantages of the antenna array to obtain angle and distance information to construct a contrastive model. However, this single-mode method cannot be applied to ultrasonic signals, because the number of ultrasonic channels that common commercial equipment can obtain is not enough to support the acquisition of accurate angles .

Multi-modal supervised learning frameworks. It has been recently proved that when the data size remains constant, the richer the modalities, the higher the performance upper bound of the network model obtained by data fusion training [17]. The diversity of sensing devices provides us with rich multi-modal data. Therefore, it has become a trend to use multi-modal data for supervised learning [1, 8, 12, 21, 28]. Existing works [19, 45, 46] design teacher-student networks that use visual information to guide models to obtain user's location information and body's reflection distributions from RF signals. Other studies use homogenous multi-modalities (ultrasonic and acoustic) for speech enhancement [29] and user authentication [8]. However, multi-modal supervised learning means more overhead for manual processing of labels, so we explore the possibility of combining multi-modal data with unsupervised learning.

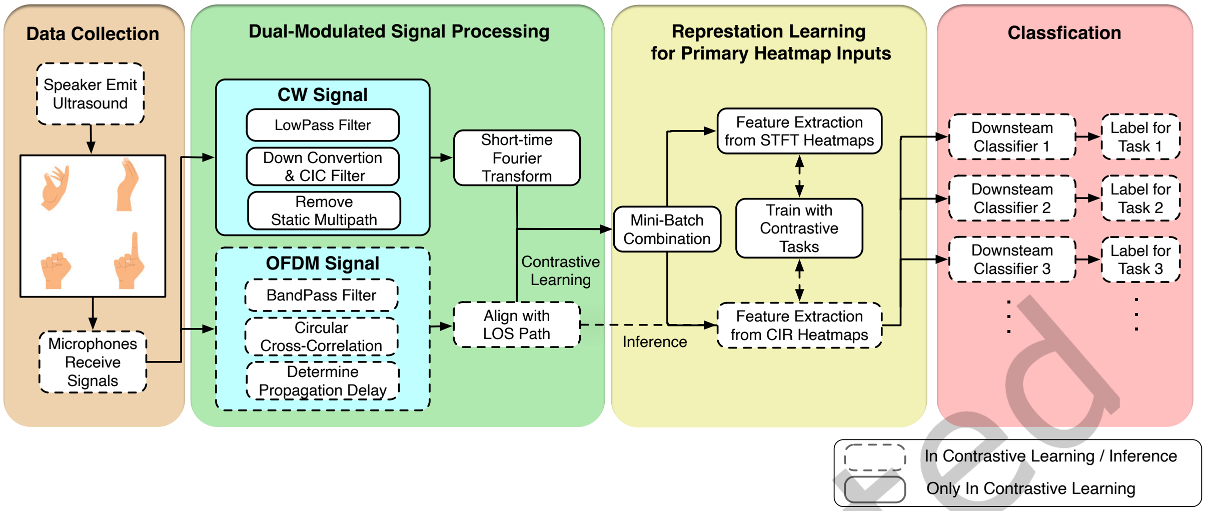

Data Collection
Speaker Emit Ultrasound
Microphones Receive Signals
Dual-Modulated Signal Processing
CW Signal
LowPass Filter
Down Convertion & CIC Filter
Remove Static Multipath
OFDM Signal
BandPass Filter
Circular Cross-Correlation
Determine Propagation Delay
Short-time Fourier Transform
Align with LOS Path
Contrastive Learning
Inference
Represtation Learning for Primary Heatmap Inputs
Mini-Batch Combination
Feature Extraction from STFT Heatmaps
Train with Contrastive Tasks
Feature Extraction from CIR Heatmaps
Classfication
Downsteam Classifier 1
Downsteam Classifier 2
Downsteam Classifier 3
Label for Task 1
Label for Task 2
Label for Task 3
In Contrastive Learning / Inference
Only In Contrastive Learning

Fig. 2. System framework for UltraCLR.

Contrastive learning methods in computer vision. The core idea of contrastive learning is to learn the representation of samples by constructing practical positive and negative examples for discrimination. Compared with the generative method, contrastive learning does not need to reconstruct pixel-level features, and it focuses on learning high-level semantic information, which is simpler and more efficient. Well-known contrastive learning frameworks [3, 5–7, 11, 15], construct different views of the original image through data augmentation. However, unlike images, data augmentation on ultrasonic signals often destroys their physical characteristics of the signal and causes severe distortion.

3 FRAMEWORK OVERVIEW

UltraCLR is a self-supervised learning framework for device-free ultrasound-based sensing tasks. The system collects continuous unlabeled acoustic signals, extracts efficient high-level feature representations from these signals, and then uses the learned encoders for data processing in various downstream sensing tasks. The system architecture consists of the following components, as shown in Fig. 2.

Data collection and dual-modulated signal processing(Section 4). First, we choose an acoustic signal that combines two modulation modes, a single-frequency continuous wave signal and an OFDM signal, as the transmitted signal of the system. We then use different preprocessing methods to obtain the acoustic features we are interested in and try to remove the noise interference caused by static reflection paths in the environment. We take the CIR heatmaps obtained from OFDM signals as the primary input and the STFT heatmaps from CW signals as the secondary input.

Representation learning for primary signal inputs(Section 5). We design a simple binary classification task for feature encoders that process continuous unlabeled signal inputs. The positive sample pairs refer to the primary and secondary signal inputs with the same timestamp, and the negative sample pairs correspond to the primary and secondary input samples at different times. We evaluate the performance of UltraCLR in experiments in Section 6, demonstrating that the trained encoder can be directly used downstream on different gesture datasets and simplify the task-specific model-building process.

4 HOMOLOGOUS DUAL-MODULATED SIGNALS

In this section, we first introduce the motivation and advantages of homologous dual-modulated signals. We then introduce our ultrasound signal design and data preprocessing methods to obtain primary and secondary data inputs that can characterize the motion state of gestures.

4.1 Motivation of Homologous Dual-Modulated Signals

Our design of the homologous dual-modulated signal is motivated by the following two aspects. First, multi-modal data's latent representation capability has been proven to increase with data modality [17, 32]. It has been shown that the richer the modalities of the data, the stronger the latent representation performance that the model can be trained from the combined signal. Second, the collection and processing of homologous signals require fewer instruments and are naturally synchronized. Multi-modal signals from different sources, such as visual signals and radio frequency signals, often require multiple sensors, which increases hardware costs and is troubled with synchronization issues. In contrast, homogenous multi-modal signals, such as voice and ultrasound, can be collected with the same device or sensor. Therefore, we choose to fuse different modulation modes of the ultrasound signal, which can share the same set of speakers and microphones. In this way, the signal collection system can be easily deployed, reducing sample acquisition cost and providing the appropriate latent representation ability of multi-modal data.

4.2 Selection of Ultrasound Sensing Signals

We consider the following two types of gesture information captured by ultrasound signals. The first type of gesture information characterizes the change of the position of the hand in space via fine-grained distance measurements. Because the high-level semantic information of gestures comes from the trajectories of hand movements in space [37]. A common practice is to calculate the direct reflection distance of the gesture signal by recording the Time of Arrival (TOA) and further improve the frequency band utilization of the signal with the help of OFDM technology [30, 35]. One way to calculate the propagation delay is to calculate the received signal's Channel Impulse Response (CIR). The propagation delay of different signal components from the transmitter to the receiver is obtained by calculating the correlation between the received signal $R(t)$ and the transmitted sequence $S(t)$.

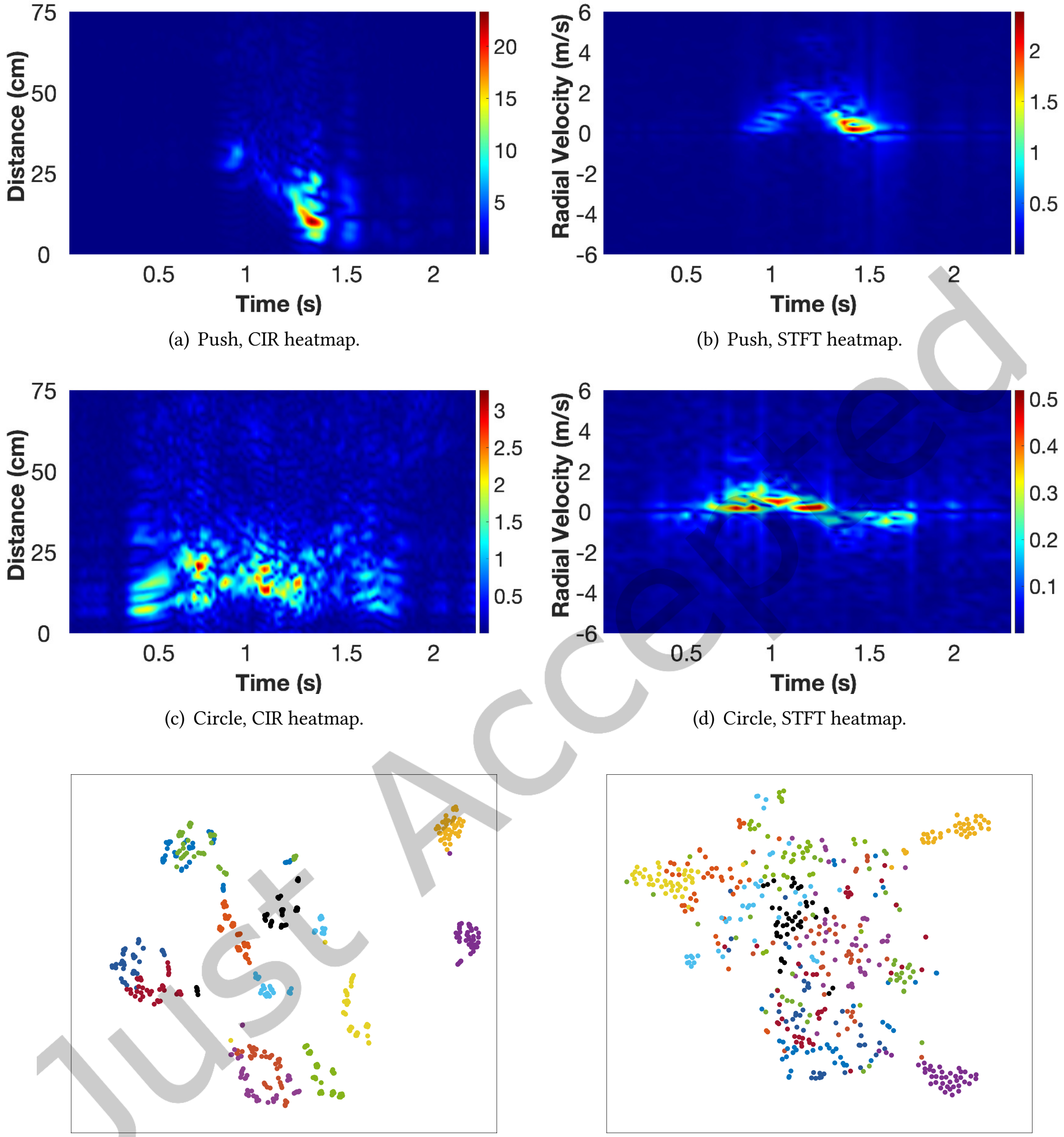
It should be noted that the FMCW signal and some chirp signals also contain the range information, but these signals have periodic hopping, which will cause some speakers to emit irrelevant noise, and the processing complexity of these signals is higher than that of OFDM signal. Therefore, we select the OFDM signal as the primary input.

In the multipath propagation model, the CIR function can be described as

$$h(t) = R(t) * \overline{S(t)} = \sum_{p \in P} A_p e^{-j2\pi d_p/\lambda} \delta(t - \tau_p), \tag{1}$$

where τ_p is the delay of path p and $\delta(t)$ is the Dirichlet function. Our data collection platform can adjust the impulse response length so that the CIR has a millimeter-level range resolution, sufficient to describe the centimeter-level finger movement. The rich features embedded in CIRs need to be first consolidated via encoding schemes, so that nearby points in the feature space represent similar movements. Therefore, we introduce a second signal to help learn representations.

The second type of gesture information characterizes the movement speed of the hand via fine-grained Time-Frequency measurements of the Doppler shift. In general, the user's gesture speed ranges from −65 to 65 cm/s, introducing Doppler frequency shifts of −117 to 117 Hz when the center frequency of the transmitted signal is 15 kHz. Therefore, an alternative solution for describing the gesture change is to use the short-term



(a) Push, CIR heatmap.

(b) Push, STFT heatmap.

(c) Circle, CIR heatmap.

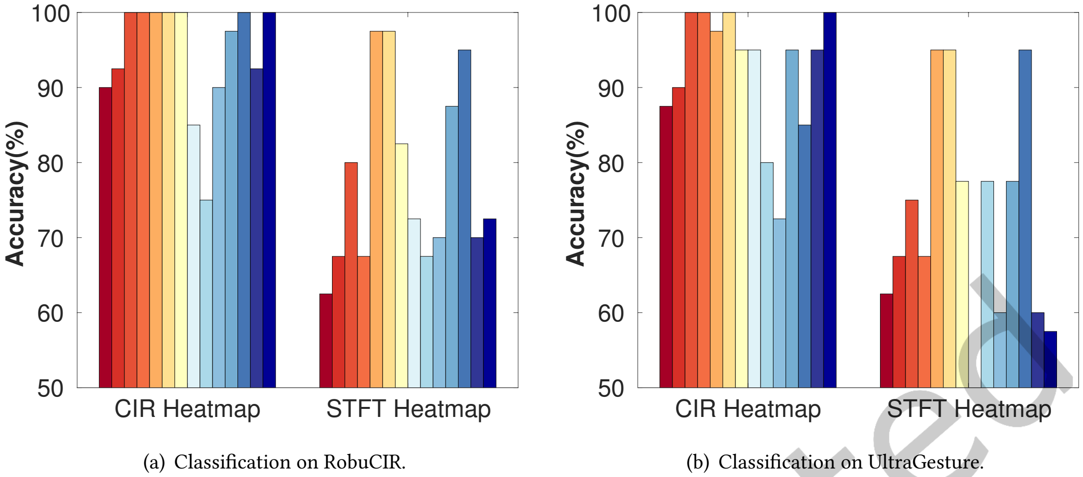
(d) Circle, STFT heatmap.

(e) Feature visualization from CIR.

(f) Feature visualization from STFT.

Fig. 3. The heatmaps from the two different modulated signals are diverse.

high-resolution Doppler frequency shift. However, the OFDM signal chooses to sacrifice the high sampling rate of the signal to obtain the precise spatial absolute position (propagation delay). Therefore, we introduce another



(a) Classification on RobuCIR.

(b) Classification on UltraGesture.

Fig. 4. Performance under state-of-the-art classification methods with different heatmaps as input.

standard ultrasonic sampling signal, the Continuous Wave signal (CW). The CW signal is the simplest device-free sensing signal, a sinusoid function of a single frequency. When transmitting a CW signal, the received signal is a linear combination of different reflection paths in set P:

$$B(t) = \sum_{p \in P}^{P} A_p e^{-j2\pi d_p(t)/\lambda}, \tag{2}$$

where $d_p(t)$ is the distance of the path p, A_p is the semi-constant complex-valued attenuation along with path p, and λ is the wavelength. Note that when there is a gesture movement, the correspondent path p will have a time-dependent distance $d_p(t)$, which incurs a Doppler shift in the received signal.

4.3 Primary and Secondary Signal Input

As discussed earlier, we can compute the Short-Time Fourier Transform (STFT) of the CW baseband signal to obtain a high-resolution Time-Frequency heatmap, which is called the STFT heatmap in later parts of this paper. The STFT heatmap reflects how fast the hand moves so that we can use it to classify different gestures [37]. At the same time, each frame of the CIR signal is a Dirichlet function, and its peak point corresponds to the propagation delay of ultrasonic waves through solids or gases. Therefore, we can align the CIR frames in sequence according to the position of peaks and calculate the difference between the aligned signal frames. In this way, we can get the energy change distribution map that reflects the movement at different reflection distances, which is called the CIR heatmap in later parts.

We compare CIR heatmaps and STFT heatmaps of several common gestures in Figure 3. We observe that CIR heatmaps contain the position information of the gesture, *e.g.*, hand moving closer or circling in Figure 3(a) and 3(c), and can show the intermediate state information of different stages of the motion over time. In contrast, STFT heatmaps in Figure 3(b) can only reflect the relative speed changes of gestures with a coarse resolution. Furthermore, the STFT heatmaps combine Doppler shift at different distances so that it does not show the hand-part variations in complex gestures such as circling the hands around in Figure 3(d). To demonstrate the difference in representation capability of the two heatmaps, Figure 3(e) and 3(f) show the embedding visualization

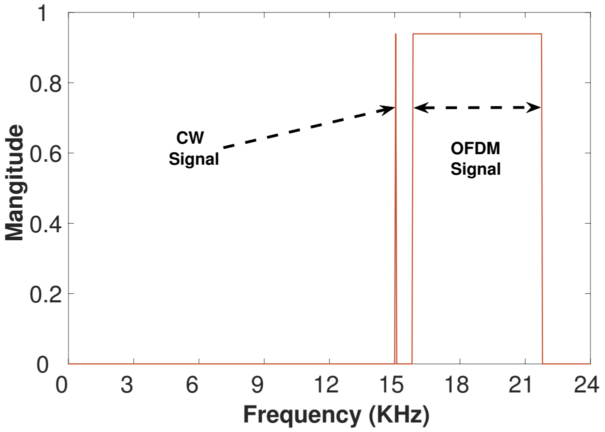


Fig. 5. The transmitted signal in frequency domain.

of t-SNE [34] of different gestures after encoded by UltraGesture. We observe that features extracted from CIR heatmaps better distinguish between gestures than STFT heatmaps. We further study their performance on the same supervised gesture dataset using models in UltraGesture [20] and RobuCIR [38]. The bars of the same color in the Figure 4(a) and 4(b) represent the gesture classification results of the same category, and the categories of gestures are in Table 1. From these figures, we observe that under SOTA supervised learning models, the representation capability of STFT heatmaps is far inferior to that of CIR heatmaps. Therefore, we call the CIR heatmaps the *primary* inputs and the STFT heatmaps the *secondary* inputs.

4.4 Dual-modulated Signal Design

As shown in Figure 5, we interpolate CW and OFDM signals into the same frame in a frequency-divided manner to simultaneously obtain STFT heatmaps and CIR heatmaps. Our equipment supports up to a 48 kHz sampling rate. We allocate a frequency of 15 kHz to the CW signal and 15.75 $\sim$ 21.75 kHz to the OFDM signal. We add a 750 Hz guard band between the CW and OFDM signal to not interfere with each other under high Doppler shifts caused by high-speed movements. since our target scene is a close range perception, we can minimize the impact of the detection signal on the user by adjusting the speaker's volume. With the further iteration of sensing devices, there are more and more 96 KHz or 192 KHz acoustic sensing devices on the market, which means that the available frequency band will be greatly expanded. Therefore, this problem will be solved entirely with sensor hardware development.

During demodulation, we first mix the received signal with the 15 kHz CW signal and then extract the 6 kHz baseband signal using the CIC filter [16]. We then estimate the static component in the baseband signal using a window of the length of 48 samples and remove it from the baseband signal. Only the LOS and multipath signals related to the gesture movement are left in the baseband signal of the CW. We then obtain a single frame via FFT with a window size of 1024 samples. According to the estimation of the range of gesture velocity, we retain 40 frequency points from -117 Hz to 117 Hz, corresponding to the round-trip gesture speed from -1.3 m/s to 1.3 m/s with a velocity resolution of 6.5 cm/s. Further, we use linear interpolation to expand each frame to 200 frequency points to enrich the frequency details of the STFT heatmap in low-frequency bands.

We also use a bandpass filter to separate the OFDM signal and perform cross-correlation to extract the CIR heatmaps. We align the cross-correlation result frames according to the distance bin of the highest peak and select the values of 256 distance bins after the highest peak as the OFDM feature. We align the OFDM frame with the

STFT window so that both heatmaps have the same frame rate. Note that the processing cost for dual-modulation extraction is considerably higher than for single modulation. To solve this problem, on the unlabeled dataset, we use the secondary STFT heatmaps as an auxiliary to improve the encoder's ability to extract features from the primary CIR heatmaps. In the training/inference stage of the classifier, we only need to use the single modulation CIR heatmaps so that the computational cost of the end-system is not increased. Furthermore, the potentially audible CW signal at 15 kHz can also be removed in the end-system.

5 CONTRASTIVE LEARNING FRAMEWORK FOR ULTRASONIC SENSING SIGNALS

In this section, we first introduce the theoretical advantages of contrastive learning methods. Then we elaborate on the contrastive learning framework used by UltraCLR, and finally we further explain the implementation details and parameters.

5.1 Motivation of Using Contrastive Learning

For contrastive learning, we need to choose the function f in the encoding function space $\mathcal{F}$ that minimizes the loss function of the downstream classification task. Let C denote the set of all latent classes in the unsupervised dataset, then any element x in either the unsupervised or the downstream supervised dataset should belong to a subclass of C. We denote positive samples of the sample x, e.g., from the same signal frame, as x^+ and negative samples, e.g., from different frames, as x^-. As shown in [25], there is a specific upper bound for the supervised loss $\mathcal{L}_{sup}(\hat{f})$ in the contrastive learning task. With probability at least $1-\delta$, $\forall f \in \mathcal{F}$,

$$\mathcal{L}_{sup}(\hat{f}) \leq \mathcal{L}^{\neq}_{un}(f) + \beta s(f) + \eta\, Gen_m, \tag{3}$$

where $\mathcal{L}_{un}(f)$ is the loss of contrastive learning that x^+ and x^- are not in the same latent class. $s(f)$ is the intraclass deviation of function f, β and η are constant values. Note that Gen_m approaches 0 when the number of samples m increases. The upper bound shows that when the encoding function space $\mathcal{F}$ is rich enough, the unsupervised contrastive learning task combined with a large number of unsupervised samples can help us find a suitable encoding function f to achieve excellent performance in downstream tasks.

A commonly used self-supervised loss function for contrastive learning is InfoNCE Loss $\mathcal{L}_m$ [33]:

$$\mathcal{L}_m = -\mathbb{E}_X\left[\frac{d(f(x), f(x^+))}{\sum_{x' \in X} d(f(x), f(x'))}\right], \tag{4}$$

where $d(x, y)$ is the similarity between x and y. It can be easily proved that minimizing the InfoNCE loss is equivalent to maximizing the mutual information of the positive sample pairs. This guides us to carefully select suitable positive and negative sample pairs to achieve effective feature extraction.

5.2 Contrastive Learning Framework

UltraCLR seeks the consistency of the characteristics of the two types of heatmaps while minimizing the contrastive loss in the latent space. As shown in Figure 6, the learning framework mainly contains the following three components:

Input: CIR and STFT heatmaps. As an alternative, we use two different heatmaps extracted from the same signal instead of data augmentation methods. We use the STFT heatmap $S(f, T)$ and the CIR heatmap $H(\tau, T)$ corresponding to the same gesture with a duration of T as input. The new features can be expressed as $h_c(f)$ and $h_o(\tau)$ after passing through two specific encoders f_o and f_c, respectively. The consistency between the STFT and CIR features is that the propagation paths for the gesture signal are fixed and do not change with different perspectives. In the view of the STFT feature $h_c(f)$, the signal components with the same Doppler frequency f are classified into the same class. Whereas, in the view of the CIR feature $h_o(\tau)$, signal components with the same propagation distance $\tau \times c$ are grouped together. In UltraCLR, the nonlinear projection heads reorganize

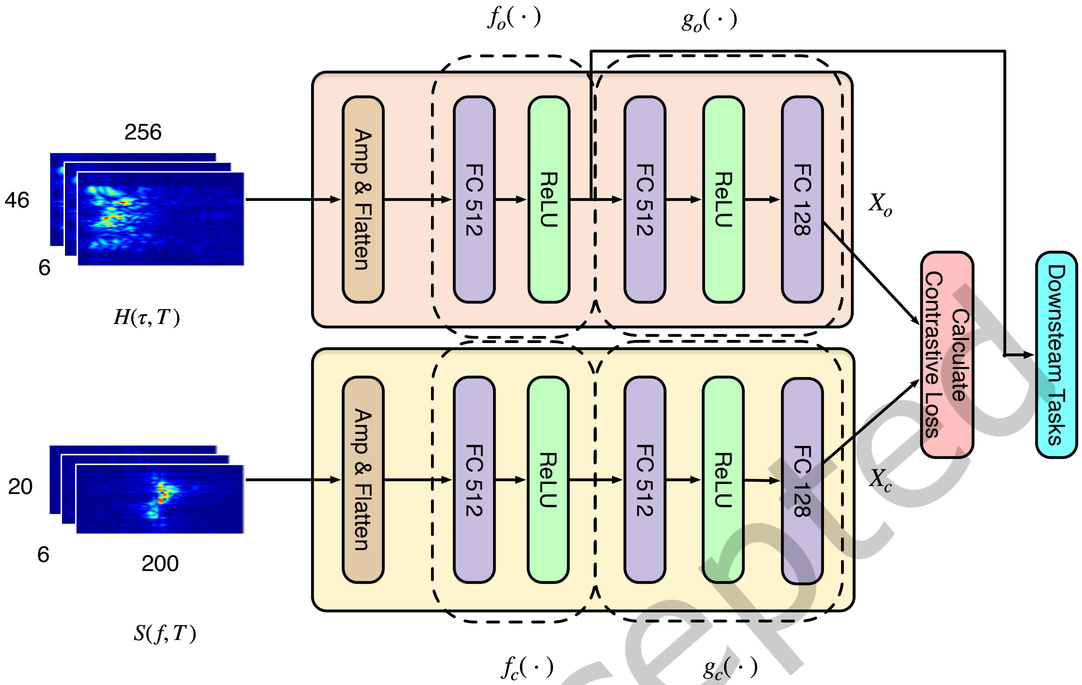
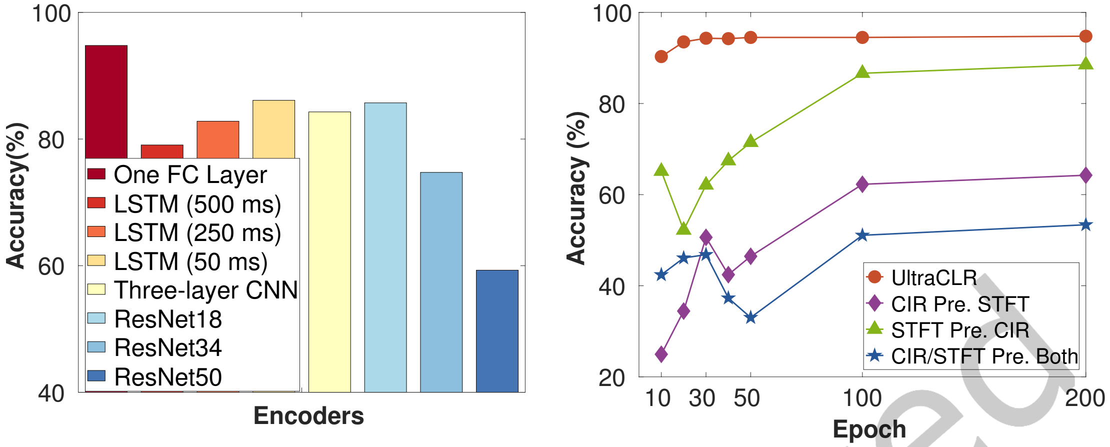


Fig. 6. Network for contrastive learning framework.

the signal features with consistency across different perspectives to enable the encoder to extract key elements that reflect the physical process of gesture movement.

Encoders: two encoders for dual-modality. Two neural network encoders, $f_c(\cdot)$ and $f_o(\cdot)$, to learn representation vectors from the STFT heatmaps extracted from CW signals and the CIR heatmaps extracted from OFDM signals. We evaluate different candidate encoders and find that using a single fully connected network as the encoder for both $f_c(\cdot)$ and $f_o(\cdot)$ gives the best performance. We first try to use some classic CNN models as encoders, including three-layer CNN, ResNet18, ResNet34, and ResNet50. However, we find that ResNet and CNN always converge quickly on the pre-training task after a few epochs. This indicates that the convolution operation allows these encoders to find the shortcut, such as movement time alignments, to correctly match the features extracted from STFT and CIR heatmaps. When the receptive field of the encoder scans the STFT heatmap and the CIR heatmap, it is equivalent to performing sliding filtering on the STFT heatmap and the CIR heatmap. The "smart" encoder will quickly discover that the energy distribution of the STFT heatmap and the CIR heatmap, which are positive sample pairs, have a strong positive correlation along the time axis. Moreover, it can easily separate them from other negative sample pairs in this way.

The above shortcut does not help the encoder extract invariant features related to gesture movement, and the terrible classification performances on the SG dataset are shown in Figure 7(a). Due to GPU memory resource limitation, we can only set smaller batch-size for deeper ResNet34 and ResNet50 networks, which leads to further performance degradation. We also consider using LSTM-based encoders. We fix the duration of a single sample and only change the length of the sample segment corresponding to each time step in the LSTM-based encoder. The experimental results show that as the sample duration in each time step decreases, the LSTM-based encoder



(a) Classification on SG dataset with different encoders.

(b) UltraCLR vs BYOL-like models.

Fig. 7. Discussions on the network network of UltraCLR.

improves in its ability to extract features. However, it is still not as effective as a single fully connected layer, because LSTM also utilizes the same shortcut, the strong temporal correlation of the two heatmaps, to perform self-supervised tasks. Therefore, our solution is to choose a less smart encoder. Besides, without using data augmentation methods, we hide the strong correlation of the energy distribution of positive samples. We end up choosing a single fully-connected layer as the encoder. After the STFT heatmaps and CIR heatmaps are vectorized, their strong correlations in energy distribution along the time axis are hidden. In this case, the encoder must extract the valid invariance of the gesture motion state to find out the positive sample pair composed of the corresponding STFT heatmap and CIR heatmap. Our simple experiments on supervised dataset SG shown in Figure 7(a) also support our judgment on different types of encoders.

Projection head and Loss. Two specific projection heads, $g_c(\cdot)$ and $g_o(\cdot)$, are responsible for further mapping the learned representation to the space for calculating the contrastive loss. As indicated by recent contrastive learning research, we also find that comparison based on the results from projection heads is better than directly using representations extracted from encoders. Given a set of positive pairs $\{(X_o, X_c)\}$, the contrastive prediction task aims to identify $x_{c,i}$ in $\{(X_o, X_c)\}$ for a given $x_{o,i}$. We choose InfoNCE loss in Eq. (4) , which is typical for contrastive learning tasks. After training, the encoder can convert raw CIR heatmaps into a better representation as feature vectors with the size of 512.

5.3 Discussions

Inappropriate data augmentation methods for ultrasound. For computer vision tasks, multiple data augmentation operations (rotating the image and adding Gaussian white noise) to the unlabeled image data can increase the difficulty of the contrastive learning task, making the model focus on the invariant characteristics of the image. However, we cannot use similar data augmentation methods to process ultrasonic sensing signals. The propagation model of the ultrasound signal is essentially a series of wave equations, and the characteristics related to movement are often the relevant parameters of these equations. Any inappropriate transformation may cause abrupt changes in these unknown parameters, and we cannot guarantee whether the transformed signal represents the same physical movement process as the original.

Finding suitable sample pairs and mini-batch training. The only preliminary information we can obtain from the unlabeled data set is the different heat map representations corresponding to the same gesture trajectory. We do not know anything about the specific meaning of the gesture expression. Therefore, we can only construct positive pairs based on temporal alignment. Similarly, we can only construct negative sample pairs based on the principle of time inconsistency because heat maps at different times corresponding to different spatial trajectories. It should be noted that the heatmap type does not limit our negative sample pairs; for example, two STFT heatmaps or two CIR heatmaps at a different time also constitute a pair of negative samples.

We randomly select N samples of (X_o, X_c) to form a mini-batch, and all pairs of positive and negative samples come from combining these samples with $2N$ heatmap sets. There are $2(N-1)$ pairs of negative samples corresponding to each (X_o, X_c) positive sample pair within a mini-batch. This way, we reduce the storage overhead for storing negative sample pairs. In order to choose a suitable batch size, we design the following experiment. First, we change the input batch size (256, 512, 1024, 2048, limited by GPU memory) and train the self-supervised model to the best state. Then, we verify the classification accuracy on three different downstream supervised datasets, respectively. We find that with a batch size of 512, UltraCLR has the best overall classification performance downstream. However, when the batch size is further increased, the accuracy does not improve further as other contrastive learning works, which may be limited by the unsupervised dataset we provide. In addition, during the experiment, we have found it easier to distinguish samples composed of features from the same modulation mode than to distinguish sample pairs composed of dual modalities. It indicates that the dual-modulation mode plays a crucial role in extracting the invariance of gesture features.

Comparison with other contrastive learning frameworks. We test the mainstream contrastive learning frameworks in computer vision methods, such as SimCLR [5], MoCo [15], and BYOL [11]. There are significant differences between UltraCLR and these existing frameworks. As mentioned above, the data augmentation methods widely used in imaging cannot be applied to ultrasound signals. Therefore, we use multi-modal instead of data augmentation. At the same time, because the feature distributions of the CIR and STFT heatmaps are different, it is not suitable to use a single encoder for dual-modal feature learning. Therefore, in UltraCLR, we replace the momentum encoder, which is standard in the image contrast learning framework, with two independent encoders.

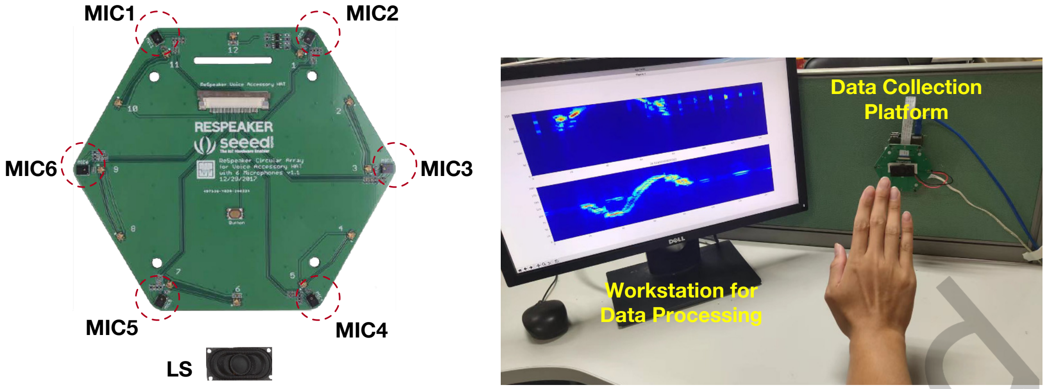
We have also tried other contrastive learning framework, e.g., the BYOL-like models, which adds a new prediction module behind the projection head. We design three types of self-supervised tasks: STFT features predict CIR features, CIR features predict STFT features, and STFT and CIR features predict each other. The results are shown in Figure 7(b). In the above three types of tasks, the efficiency of CIR predicting STFT features is nearly 20% higher than that of STFT predicting CIR features, which also confirms that the CIR heatmap has a more vital gesture feature expression ability than the STFT heatmap. However, the performance of these three types of tasks is far inferior to maximizing the mutual information between pairs of positive samples directly.

6 EXPERIMENTAL RESULTS

In this section, we first introduce the datasets for the experiments and various models for comparison. Then we compare the performance of UltraCLR with other baseline models on the complete dataset/few-shot datasets and explore the classification impact of adding unlabeled data. Finally, we illustrate the overhead advantages of UltraCLR at deployment.

6.1 Platform setup and data collection

We have implemented a data collection system for ultrasonic sensing signals on Raspberry Pi. We deploy and collect data in a laboratory scenario, as shown in Figure 8. The platform consists of a single loudspeaker and Respeaker's ring-shaped 6-channel microphone array. We fix the loudspeaker in the center of the microphone



(a) Hardware platform.

(b) Environment for data collection.

Fig. 8. The experiment setup.

array, and the direct distance of sound propagation is 5cm. The platform is placed vertically, and the user performs gestures at a distance of 20-40 cm from the platform. Collected data is saved through the Raspberry Pi and transmitted to the workstation for further analysis and processing. During the collection of unlabeled data, we ensure that the user's hand is within the sensor's detection range, regardless of whether the user's hand is stationary or moving. Therefore, there are currently no unlabeled samples without hands. If the hand is allowed to leave the sensor's detection range, we could obtain the CIR of the pure environment at the beginning of a collection. Then, we could calculate the similarity between the current CIR and the environment CIR in real-time to judge whether the current sample is NULL. The design of our ultrasonic sensing signal is introduced in Section 4.

Unlabeled gesture dataset USG. We collect unlabeled gesture signals from 8 volunteers of various ages, weights, and hand shapes. The total time duration of the gesture dataset exceeds 4 hours. We provide them with a set of common gestures to improve efficiency, but users can freely choose their favorite gestures to perform. After gesture signals are collected, we do not label or filter them. Instead, we directly divide the continuous gesture signals into unlabeled samples using a sliding window with an equal duration of one second.

Our secondary STFT heatmap has a three dimension of 6 × 200 × 20 (channels × FFT samples × frames) for each sample. When calculating the primary CIR heatmaps from the OFDM signals, we select 256 time-delay feature points after the correlation peak as the gestures feature, covering a distance of 0.9 meters. The primary CIR heatmap has a dimension of 6 × 256 × 46 (channels × time delay × frames).

Downstream labeled gesture datasets SG, SG-L, and SG-G. We have designed and collected three labeled datasets, SG, SG-L, and SG-G, respectively, to test the feature extraction capability of UltraCLR. These three supervised datasets are collected after different users perform the 14 types of gestures in Table 1. Meanwhile, these 14 categories are also included in the random gesture set provided to volunteers when collecting unsupervised data. The first supervised dataset, SG, contains 2800 samples (8 × 14 × 25) obtained by eight users performing each type of gesture 25 times with their right hand. The other two datasets, SG-L and SG-G, contain 1750 labeled gesture samples (5 × 14 × 25) for five users, respectively. When collecting dataset SG-L, we ask volunteers to perform gestures with their left hand. All volunteers use their right hand and wear white flannel gloves that we prepared in advance to collect the data in SG-G. Note that none of the volunteers are involved in collecting the unsupervised dataset. Taking the SG dataset as an example, Figure 9(d) presents the distribution of peak speeds

Table 1. List of labeled gestures.

Gesture	Label	Gesture	Label
Flick	A	Anti-Flick	B
Clockwise	C	Anti-Clockwise	D
Push	E	Pull	F
Page Up	G	Page Down	H
Page Left	I	Page Right	J
Click	K	Hover	L
Tick	M	Cross	N

Table 2. Baseline performance with other supervised and self-supervised Learning methods.

Dataset	**Testers**	**Features**	**Model**	**Precision** (mean±std)	**Recall** (mean±std)	**F1-Score** (mean±std)
SG	8	CIR	UltraGesture [20]	94.24 ± 0.51	94.11 ± 0.52	94.14 ± 0.53
SG	8	CIR	RobuCIR [38]	93.34 ± 1.31	93.17 ± 1.35	93.17 ± 1.34
USG & SG	8	CIR	DNN Model	91.42 ± 0.22	91.02 ± 0.23	90.97 ± 0.22
USG & SG	8	CIR	SimCLR [5]	91.19 ± 0.27	90.76 ± 0.46	90.78 ± 0.44
USG & SG	8	CIR	**UltraCLR**	**94.88 ± 0.41**	**94.78 ± 0.41**	**94.80 ± 0.41**
USG & SG	8	CIR & STFT	**UltraCLR**	**95.82 ± 0.31**	**95.76 ± 0.34**	**95.73 ± 0.34**
USG & SG-L	5	CIR	DNN Model	92.43 ± 2.17	91.71 ± 2.67	91.62 ± 2.64
USG & SG-L	5	CIR	SimCLR [5]	82.46 ± 2.00	81.50 ± 2.04	81.10 ± 2.09
USG & SG-L	5	CIR	**UltraCLR**	**95.74 ± 0.75**	**95.50 ± 0.89**	**95.45 ± 0.91**
USG & SG-L	5	CIR & STFT	**UltraCLR**	**96.72 ± 0.68**	**96.50 ± 0.79**	**96.46 ± 0.82**
USG & SG-G	5	CIR	DNN Model	89.08 ± 1.33	88.21 ± 1.87	88.14 ± 1.90
USG & SG-G	5	CIR	SimCLR [5]	76.56 ± 1.40	74.21 ± 1.25	73.82 ± 1.29
USG & SG-G	5	CIR	**UltraCLR**	**96.93 ± 0.89**	**96.64 ± 1.00**	**96.61 ± 1.01**
USG & SG-G	5	CIR & STFT	**UltraCLR**	**98.55 ± 0.40**	**98.43 ± 0.43**	**98.43 ± 0.42**

for gesture samples from different categories. We find that the peak speed difference of the same gesture can reach up to 90 cm/s (ranging from 40 cm/s to 130 cm/s), due to different users' habits.

6.2 Frameworks in Comparison

UltraCLR. The input is the features learned by UltraCLR proposed in this paper, and the classifier is the simple 1-layer LSTM neural network. We divide the labeled data into four segments with a duration of 1s and arrange the learned representation output in sequence as the input of LSTM. We use a single fully-connected layer as the base encoder and a 2-layer MLP as the projection head. We train InfoNCE loss with a learning rate 5×10^{-5} and weight decay of 5×10^{-4}. We set the mini-batch size as 512 and train the model for 200 epochs.

DNN Model. The DNN Model has the same network architecture as UltraCLR, which consists of a single fully-connected layer and an LSTM network. The difference is that the DNN Model is an end-to-end supervised model that is trained and tested directly on downstream supervised datasets.

UltraGesture [20]. UltraGesture uses a 5-layer CNN concatenated with a single fully-connected layer for supervised gesture classification. Initially, the input is the single-channel CIR heatmap calculated by a signal based on the Barker-code sequence, and we change the input into our multi-channel CIR heatmap for comparison.

RobuCIR [38]. RobuCIR uses the magnitude and phase heatmaps of the CIR as input and uses two identical neural networks to handle them separately. Each neural network contains a 3-layer CNN feature extractor and a single-layer LSTM network classifier. RobuCIR takes the weighted summation of the probability vectors output

by the two LSTMs as the final prediction result . To be fair, we use the multi-channel CIR amplitude heatmap as the input of RobuCIR in our experiments, which is consistent with UltraCLR.
SimCLR [5]. SimCLR is a self-supervised learning framework of contrastive learning for images. We apply this framework directly to raw CIR heatmaps and use the data augmentation method proposed in RobuCIR (translation and scaling operations on heatmaps) to construct positive sample pairs. For a fair comparison, we keep the data processing method and classifier setup consistent with UltraCLR.

6.3 Baseline Performance

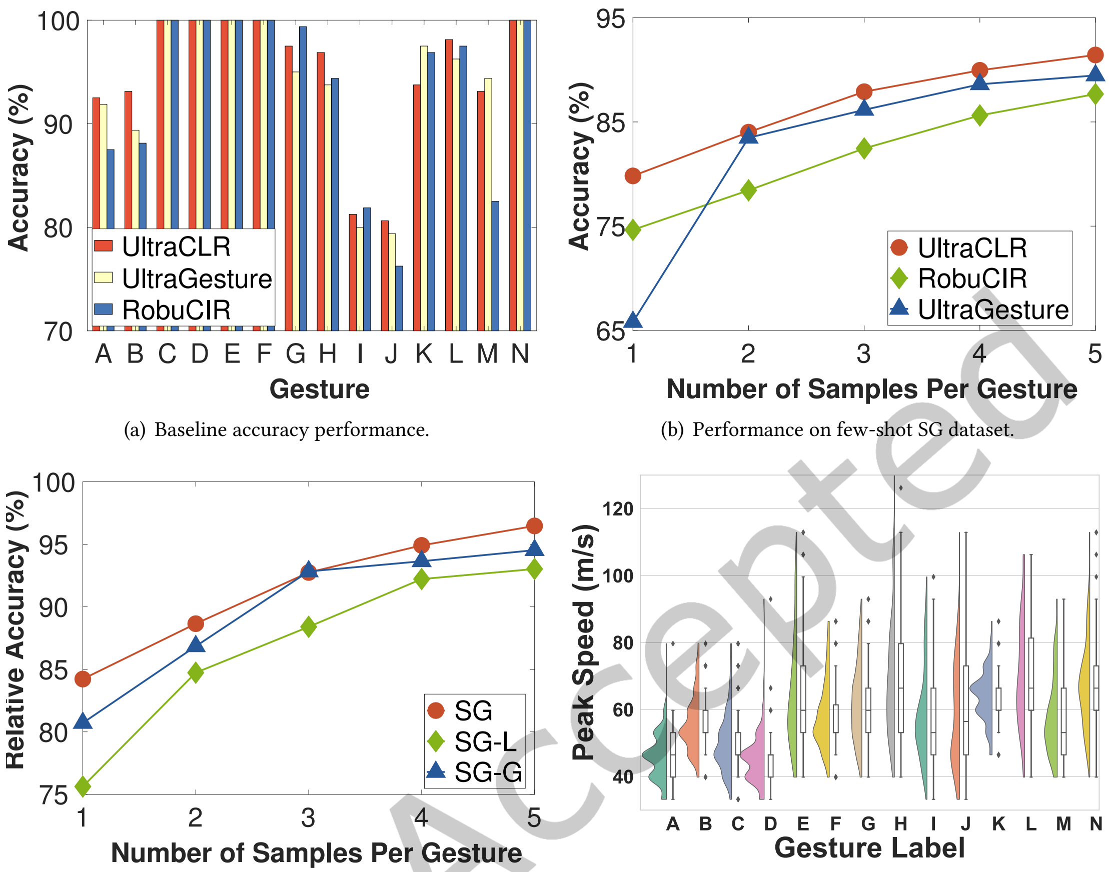
We first compare the performance of UltraCLR and other baseline models for gesture classification on the dataset SG. We divide the data into training and test sets according to the 80%: 20% rule and perform four-fold cross-validation on the training set. As shown in Table 2 and Figure 9(a), UltraCIR outperforms other baseline models. Note that for self-supervised learning models, the size of the labeled training data is only about 8.1% of the training data samples. Furthermore, the 512-bit input vector extracted by UltraCLR is compressed by 99.2% compared to the original CIR heatmap, demonstrating the effectiveness of the CIR representation extracted by UltraCIR. SimCLR compresses and extracts features on the same scale for heatmaps as another self-supervised model. However, there is a significant performance loss compared with supervised learning methods. The result validates our judgment that we cannot directly apply conventional image data augmentation strategies to gesture signals.

To further verify the effectiveness of UltraCLR based on the contrastive learning between the primary and secondary signal inputs, we verify the classification performance on two other supervised datasets, SG-L and SG-G, and compare the results with SimCLR. It should be emphasized that the unlabeled dataset does not contain any data collected by the volunteers with their left hand or wearing gloves, so the unlabeled dataset provides little prior knowledge. The training set, test set, and cross-validation settings are consistent with previous experiments. Compared with SimCLR, the experimental results clearly show that UltraCLR performs equally well on SG-L and SG-G. It shows that the contrastive learning framework based on the primary and secondary signal inputs can focus on preserving the spatial state information of gesture motion while realizing data compression. In contrast, SimCLR again caused a deep accuracy drop due to inappropriate data augmentation. When we use both new encoded features (from CIR and STFT heatmaps) as the input of downstream tasks, the combined inputs will further improve the classification performance of UltraCLR on the above three datasets. It is in line with our perception that the upper performance of the multi-modal model is higher than that of the single-modal model. In most cases, downstream tasks usually consist of only one modulation mode, such as OFDM, while UltraCLR has been shown to outperform existing supervised models.

It should also be noted that the supervised learning model, DNN Model, which has the same architecture as UltraCLR, shows a significant performance degradation on the above three supervised datasets. While the supervised learning model has certain advantages when trained and tested on the same source datasets, UltraCLR performs well in the absence of prior knowledge of the downstream environment, demonstrating the advantage of self-supervised learning in performance. This further supports the effectiveness of self-supervised learning.

6.4 Improvements on Few-Shot Supervised Samples

In this section, we compare the effectiveness of our model in the few-shot supervised learning scenario, which is important for continuous learning tasks on mobile devices [2]. We randomly divide the labeled dataset SG into the training and testing sets, where the testing set contains 20% samples for each user. We then use 4-fold cross-validation to further select the validation set from the training set, and randomly select a few samples from the remaining training data for training. To show how UltraCLR helps models learn from only a few gesture samples, we gradually increase the training sample size from one to five samples per gesture.



(a) Baseline accuracy performance.

(b) Performance on few-shot SG dataset.

(c) UltraCLR performance on few-shot datasets.

(d) Peak speed distribution for samples in dataset SG.

Fig. 9. Performances of UltraCLR and other baseline models on downstream tasks.

Figure 9(b) shows that UltraCLR significantly outperforms UltraGesture and RobuCIR because the model is pre-trained on unlabeled data. Each user only needs to provide two samples per gesture to ensure a classification accuracy of more than 85% for UltraCLR. At the same time, as the number of samples increases, UltraCLR converges to the maximum accuracy faster. This indicates that the performance improvement of UltraCLR comes from feature learning on unlabeled datasets. UltraCLR can effectively learn the motion intermediate state information of sensing signals through contrastive learning and further improve the representation ability of CIR heatmaps. This capability allows UltraCLR to be quickly applied to different supervised datasets of a similar type, as shown in Figure 9(c). When using the entire training set(15 training samples for each gesture type), UltraCLR achieves the highest gesture classification accuracy on SG, SG-L, and SG-G with 94.78%, 97.29%, and 96.64%, respectively. It should be noted that the unlabeled dataset does not contain prior knowledge of left-hand gestures and gloved gestures. With very few training samples, e.g., no more than two samples per gesture, the performance of UltraCLR on the SG-L and SG-G datasets is not as good as the performance on the SG dataset. However, with just a little more data, UltraCLR's excellent learning ability of intermediate state motion

Table 3. Improvement of adding unlabeled data.

Data (User Number)	1	3	5	7	8
Accuracy (%)	92.72	93.21	93.97	94.24	94.78
Increase (%)	-	0.49	0.76	0.27	0.54

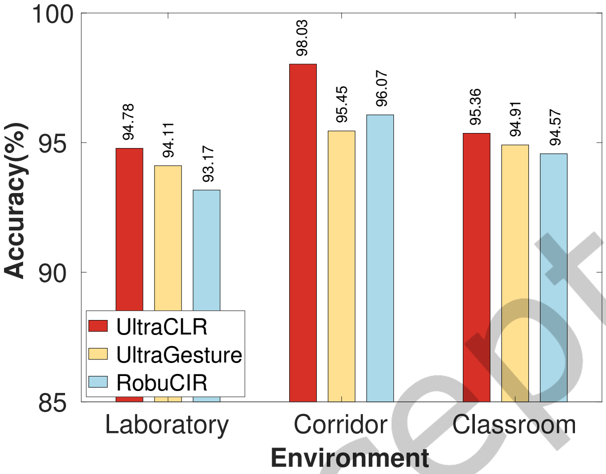


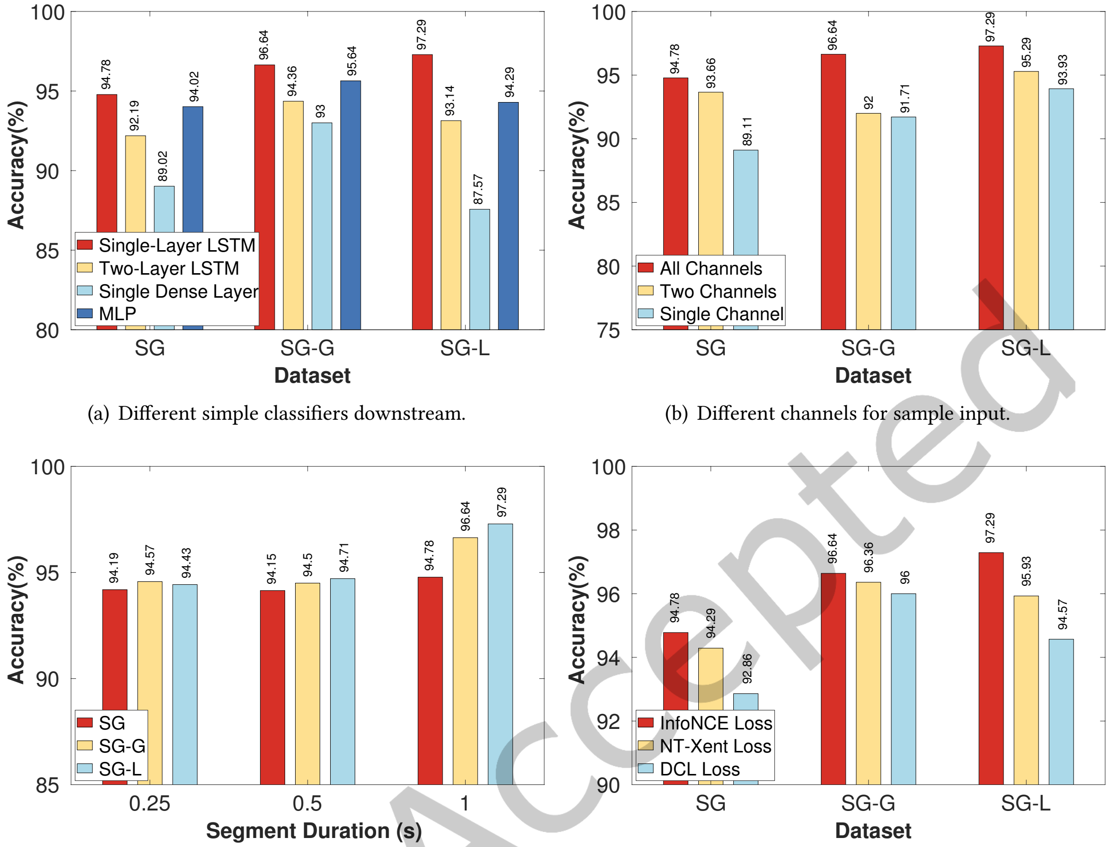
Fig. 10. Comparison on different downstream environments.

information can significantly improve its classification performance. Figure 9(c) further shows that UltraCLR can achieve more than 90% of the optimal accuracy on each supervised dataset when using four samples per gesture. This shows that UltraCLR's contrastive learning framework can significantly reduce the use of supervised data while ensuring minimizing the performance loss.

6.5 Impact of Adding More Unlabeled Samples

We design the following experiments to verify that adding more unlabeled data is helpful for UltraCLR to extract the invariant features of sensing signals. We first randomly select a volunteer's unlabeled data as the training set and gradually add new volunteer data until all the unlabeled data are used. We perform the classification task on the SG dataset to evaluate the feature extraction ability of UltraCLR. In the classification task, the training set and the test set are divided by a ratio of 4:1. Meanwhile, we employ 4-fold cross-validation on the training set to calculate the final classification performance.

Table 3 shows the performance of the CIR representation learned by adding unlabeled data from different numbers of new users. As we add new unlabeled samples, the accuracy of the classification model is improved. It shows that UltraCLR can continuously optimize the extraction of CIR heatmaps on more unlabeled data. We also find that the increased accuracy rate slows down with more unlabeled data, mainly because of the limited supervised dataset and the performance cap on the CIR heatmaps. Overall, it gives us a new perspective on fine-tuning specific recognition models for new users. In addition to requiring limited supervised samples from new users, we can also collect long-term, unlabeled data from users to improve the classification performance while not disturbing the users.



(a) Different simple classifiers downstream.

(b) Different channels for sample input.

(c) Different unlabeled sample duration for self-supervised learning.

(d) Different self-supervised loss on UltraCLR.

Fig. 11. Discussions on other important model setting.

6.6 Model adaptation in different scenarios.

To verify the applicability of UltraCLR to supervised datasets in different environments, we collect new gesture datasets respectively in the corridor and the classroom. Each dataset contains gesture samples from four different volunteers, and it should be noted that new volunteers does not participate in the collection of the unlabeled dataset. Each volunteer repeats each gesture in Table 1 25 times, resulting in a total of 1400 samples for each sample set.

We divide the training set and the test set in a ratio of 4:1 and performed four-fold cross-validation to select the best classification model for testing. We also compare UltraGesture and RobuCIR in the two new scenarios, and the result is shown in Figure 10. The results show that UltraCLR can achieve better classification performance than single-dataset supervised models in new scenarioa. This fully demonstrates that UltraCLR can be applied to new classification tasks in different scenarios.

6.7 Discussions

Selection of simple classifiers. With the powerful feature extraction capability of UltraCLR, we only need a simple classifier to accurately classify downstream tasks. We compare four types of classifiers: single-layer LSTM network, double-layer LSTM network, single-layer fully connected network, and double-layer fully connected network (multi-layer perceptron). We compare the performance of the above classifiers on the three downstream supervised datasets SG, SG-L, and SG-G. The method of dividing the dataset and calculating the average accuracy was consistent with previous experiments. Figure 11(a) shows that the single-layer LSTM network performs best on all datasets compared to other classifiers. Increasing the number of layers of LSTM does not lead to further performance improvement. At the same time, the performance of using MLP as a classifier is second only to that of using a single-layer LSTM network. Considering the number of parameters, we finally select the one-layer LSTM network as the benchmark classifier for downstream tasks.

Impact of multiple channels. The multi-channel acoustic signal provided by the circular array can significantly enhance the ability to describe complex gestural movements. However, only two microphones are usually deployed on the most common mobile smart devices. Therefore, we design the following experiments to demonstrate the feasibility of deploying UltraCLR to smartphones. We consider two cases, one microphone or two microphones as the receiver. We randomly select 1 or 2 channels from the existing six microphone channels for the experiment. We ensure that the selected two channels are on opposite sides of the microphone array (with a distance of 10cm from each other). We change UltraCLR to accommodate the new channel count and keep other settings unchanged. We test the performance of UltraCLR on three downstream datasets SG, SG-G, and SG-L. Figure 11(b) shows the performance of single- and dual-channel UltraCLR on different downstream datasets. As the number of channels decreases, the accuracy of the classifier decreases. This is because the signal components with different one-dimensional distance information are also reduced along with the reduction of the number of channels. Therefore, complex 3D gesture movements cause more significant errors in the low-channel UltraCLR model. Nevertheless, for dual-channel UltraCLR, its accuracy has been maintained above 90%, which is enough to prove the feasibility of deploying UltraCLR on smart mobile devices.

Impact of sample duration. As mentioned earlier, UltraCLR does not need to ensure that the length of the unlabeled samples is the same as that in the downstream supervised dataset. However, our concern is whether the different duration of unlabeled samples will affect downstream tasks. Therefore, we conduct the following experiments: We adopt three unlabeled sample segmentation schemes with duration of 0.25 s, 0.5 s, and 1 s. Different segmentation schemes will result in a multiplied difference in the size of the final unlabeled data set, leading to an exponential increase in the size of positive and negative sample pairs. Hence, the difficulty of self-supervised learning varies. Therefore, we ensure that the size of the unlabeled dataset is the same under different segmentation schemes through uniform sampling. We test the performance of different segmentation schemes on three downstream supervised datasets, and the other parameter settings in this experiment are the same as the benchmark experiments. The experimental results are shown in the Figure 11(c). Under the premise of the same unlabeled data size, a longer-time segmentation scheme can achieve better performance. At the same time, the model trained by shorter samples is less effective. The two heatmaps we use have a low SNR locally, which makes the self-supervised model unable to learn effectively. Therefore, we fix the duration of unlabeled samples to 1 s.

Impact of different self-supervised loss. Apart from the most commonly used InfoNCE loss function, we have also tried using the NT-Xent loss used in SimCLR and the DCL loss used in [42]. The NT-Xent loss and InfoNCE loss are similar, both being the variants of cross-entropy loss, differing only in whether positive sample pairs are included in the overall consideration. The DCL loss is proposed to address the negative-positive coupling issue caused by InfoNCE loss, by reducing the batch size, but its downside is that it may not capture similarities between complex targets. When choosing different loss functions, it is essential to use them appropriately based

Table 4. Model efficiency comparison.

Model	Parameters	FLOPs	Inference Time
UltraGesture	2.27M	72G	22.0ms
RobuCIR	7.39M	64G	31.5ms
UltraCLR	20.2M	10G	2.3ms

on the characteristics of the dataset and the downstream task. We also evaluated the classification performance of UltraCLR trained with different loss functions on three supervised datasets separately. As shown in the Figure 11(d), InfoNCE loss performs the most stably in the ultrasound gesture recognition task and is the most appropriate choice.

Efficiency comparison. Table 4 compares UltraCLR with baseline models regarding the number of parameters, computational complexity and inference time. The inference processes are performed on the same server with a single NVIDIA GeForce RTX 3070 GPU. UltraCLR contains 18.1M encoder parameters, which are frozen and do not need to be trained for classification tasks on dataset SG. The inference time is the execution time of both the encoder and the classifier for inferring a gesture sample. For each experiment, the final result is the average of 2000 repetitions of execution. We observe that UltraCLR reduces the computational complexity by more than six times compared with the other two CNN-based models. Because the quantity of parameters for UltraCLR mainly comes from the encoder, which is a single-layer fully connected network. Despite increasing the number of parameters, the fully connected network uses a matrix multiplication operation, which requires fewer operating instructions than that of the convolution operation. This difference in FLOPs is reflected in the time overhead consumed by UltraCLR, which is much smaller than that of CNN-based supervised models. In conclusion, the time required for inference using UltraCLR is significantly less due to the use of a fully connected layer encoder. At the same time, since the encoder parameters are frozen when training UltraCLR, we only need to train the LSTM layer, which further reduces the training time. Therefore, the inference speeds of UltraCLR are nearly 9 times faster than the other two systems. Our new self-supervised UltraCLR model takes a step further by enhancing the performance efficiency of various downstream supervised datasets. As demonstrated in our comparison with existing methods, UltraCLR has shown advantages in achieving lightweight model overhead, real-time model operation performance, and high accuracy operation results.

7 CONCLUSION

This paper presents UltraCLR, a contrastive learning framework for extracting effective gesture representations from unlabeled ultrasonic signals. With the dual-modulated signal input paradigm, we only need to collect secondary signal inputs when training the encoder, while the end-task only utilizes the enhanced representation of the primary signal inputs. We envision our approach can be applied to a broader range of sensing tasks, as the dual-input extraction can be used in other scenarios, such as WiFi-based and IMU-based sensing.

REFERENCES

[1] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2022. IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-Based Activity Recognition Using IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2022), 1–20.

[2] Romil Bhardwaj, Zhengxu Xia, Ganesh Ananthanarayanan, Junchen Jiang, Yuanchao Shu, Nikolaos Karianakis, Kevin Hsieh, Paramvir Bahl, and Ion Stoica. 2022. Ekya: Continuous Learning of Video Analytics Models on Edge Compute Servers. In *Proceedings of USENIX NSDI*. 119–135.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Proceedings of NeurlPS*. 9912–9924.

[4] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A Systematic Study of Unsupervised Domain Adaptation for Robust Human-Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*

4, 1 (2020), 1–30.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of ICML*. 1597–1607.

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020. Big Self-Supervised Models Are Strong Semi-Supervised Learners. In *Proceedings of NeurlPS*. 22243–22255 pages.

[7] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. *CoRR* abs/2003.04297 (2020), 1–20.

[8] Yanjiao Chen, Meng Xue, Jian Zhang, Qianyun Guan, Zhiyuan Wang, Qian Zhang, and Wei Wang. 2021. ChestLive: Fortifying Voice-Based Authentication with Chest Motion Biometric on Smart Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–25.

[9] Haiming Cheng and Wei Lou. 2021. Push the Limit of Device-free Acoustic Sensing on Commercial Mobile Devices. In *Proceedings of IEEE INFOCOM*. 1–10.

[10] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. Metasense: Few-shot Adaptation to Untrained Conditions in Deep Mobile Sensing. In *Proceedings of ACM SenSys*. 110–123.

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Proceedings of NeurlPS*. 21271–21284.

[12] Kaiwen Guo, Hao Zhou, Ye Tian, Wangqiu Zhou, Yusheng Ji, and Xiang-Yang Li. 2022. Mudra: A Multi-Modal Smartwatch Interactive System with Hand Gesture Recognition and User Identification. In *Proceedings of IEEE INFOCOM*. 100–109.

[13] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: Using the Doppler Effect to Sense Gestures. In *Proceedings of SIGCHI*. 1911–1914.

[14] Zijun Han, Lingchao Guo, Zhaoming Lu, Xiangming Wen, and Wei Zheng. 2020. Deep Adaptation Networks based Gesture Recognition Using Commodity WiFi. In *Proceedings of IEEE WCNC*. 1–7.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of IEEE CVPR*. 9729–9738.

[16] Eugene Hogenauer. 1981. An Economical Class of Digital Filters for Decimation and Interpolation. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29, 2 (1981), 155–162.

[17] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What Makes Multi-Modal Learning Better than Single (Provably). In *Proceedings of NeurlPS*. 10944–10956.

[18] Hang Li, Xi Chen, Ju Wang, Di Wu, and Xue Liu. 2022. DAFI: WiFi-Based Device-Free Indoor Localization via Domain Adaptation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2022), 1–21.

[19] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. 2019. Making the Invisible Bisible: Action Recognition Through Walls and Occlusions. In *Proceedings of IEEE ICCV*. 872–881.

[20] Kang Ling, Haipeng Dai, Yuntang Liu, Alex X. Liu, Wei Wang, and Qing Gu. 2022. UltraGesture: Fine-Grained Gesture Sensing and Recognition. *IEEE Transactions on Mobile Computing* 21, 7 (2022), 2620–2636.

[21] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB De Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. MilliEgo: Single-chip MMWave Radar Aided Egomotion Estimation via Deep Sensor Fusion. In *Proceedings of ACM SenSys*. 109–122.

[22] Yang Qifan, Tang Hao, Zhao Xuebing, Li Yin, and Zhang Sanfeng. 2014. Dolphin: Ultrasonic-based Gesture Recognition on Smartphone Platform. In *Proceedings of IEEE CSE*. 1461–1468.

[23] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: Enabling Fine-Grained Hand Gesture Detection by Decoding Echo Signal. In *Proceedings of ACM UbiComp*. 474–485.

[24] Andrea Rosales Sanabria, Franco Zambonelli, and Juan Ye. 2021. Unsupervised Domain Adaptation in Activity Recognition: A GAN-based Approach. *IEEE Access* 9 (2021), 19421–19438.

[25] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A Theoretical Analysis of Contrastive Unsupervised Representation Learning. In *Proceedings of ICML*. 5628–5637.

[26] Zhiyao Sheng, Huatao Xu, Qian Zhang, and Dong Wang. 2022. Facilitating Radar-Based Gesture Recognition With Self-Supervised Learning. In *Proceedings of IEEE SECON*. 154–162.

[27] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. 2022. RF-URL: Unsupervised Representation Learning for RF Sensing. In *Proceedings of ACM MobiCom*. 282–295.

[28] Ke Sun, Chen Chen, and Xinyu Zhang. 2020. "Alexa, Stop Spying on Me!" Speech Privacy Protection Against Voice Assistants. In *Proceedings of ACM SenSys*. 298–311.

[29] Ke Sun and Xinyu Zhang. 2021. UltraSE: Single-Channel Speech Enhancement Using Ultrasound. In *Proceedings of ACM MobiCom*. 160–173.

[30] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing Touch Gestures on Surfaces of Mobile Devices using Acoustic Signals. In *Proceedings of ACM MobiCom*. 591–605.

[31] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. 2021. SelfHAR: Improving Human Activity Recognition through Self-Training with Unlabeled Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (2021), 1–30.

[32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Multiview Coding. In *Proceedings of ECCV*. 776–794.

[33] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018), 1–20. http://arxiv.org/abs/1807.03748

[34] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. *Journal of machine learning research* 9, 11 (2008), 2579–2605.

[35] Haoran Wan, Shuyu Shi, Wenyu Cao, Wei Wang, and Guihai Chen. 2021. RespTracker: Multi-user Room-scale Respiration Tracking with Commercial Acoustic Devices. In *Proceedings of IEEE INFOCOM*. 1–10.

[36] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-Free Gesture Tracking Using Acoustic Signals. In *Proceedings of ACM MobiCom*. 82–94.

[37] Xun Wang, Ke Sun, Ting Zhao, Wei Wang, and Qing Gu. 2020. Dynamic Speed Warping: Similarity-based One-shot Learning for Device-free Gesture Signals. In *Proceedings of IEEE INFOCOM*. 556–565.

[38] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2022. Push the Limit of Acoustic Gesture Recognition. *IEEE Transactions on Mobile Computing* 21, 5 (2022), 1798–1811.

[39] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. OneFi: One-Shot Recognition for Unseen Gesture via COTS WiFi. In *Proceedings of ACM SenSys*. 206–219.

[40] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *Proceedings of ACM SenSys*. 220–233.

[41] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. 2019. Unsupervised Embedding Learning via Invariant and Spreading Instance Feature. In *Proceedings of IEEE CVPR*. 6210–6219.

[42] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. 2022. Decoupled contrastive learning. In *Proceedings of ECCV*. 668–684.

[43] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-Grained Acoustic-Based Device-Free Tracking. In *Proceedings of ACM MobiSys*. 15–28.

[44] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing. In *Proceedings of ACM MobiCom*. 305–320.

[45] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-Wall Human Pose Estimation Using Radio Signals. In *Proceedings of IEEE CVPR*. 7356–7365.

[46] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D Skeletons. In *Proceedings of ACM SIGCOMM*. 267–281.

[47] Han Zou, Jianfei Yang, Yuxun Zhou, Lihua Xie, and Costas J Spanos. 2018. Robust WiFi-enabled Device-free Gesture Recognition via Unsupervised Adversarial Domain Adaptation. In *Proceedings of IEEE ICCCN*. 1–8.