Diffusion Model-based Contrastive Self-Supervised Learning for Human Activity Recognition

Chunjing Xiao, Yanhui Han, Yane Hou, Fangzhan Shi, Kevin Chetty

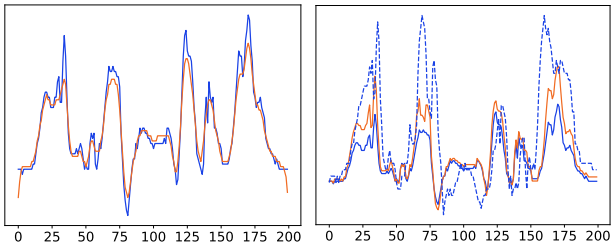
Abstract—WiFi Channel State Information (CSI)-based activity recognition plays a crucial role for vast Internet of Things applications. However, recognition models powered by supervised techniques are confronted with the difficulty of collecting numerous labeled data, which is time-consuming and labor-intensive. To address this issue, we design a diffusion model-based Contrastive self-supervised Learning framework for human Activity Recognition (CLAR) using WiFi CSI. Based on the contrastive learning framework, we mainly propose two components for CLAR to meet the scenarios for CSI-activity recognition. To effectively enlarge the distribution of training data, we propose a denoising diffusion probabilistic model (DDPM)-based time series-specific augmentation model, which can combine two samples to generate diverse augmented data. To efficiently capture the difference of the sample importance, we present an adaptive weight algorithm, which can adaptively adjust the weights of positive sample pairs for learning better data representations. The experiments suggest that CLAR achieves significant gains compared to state-of-the-art methods.

Index Terms—Contrastive learning, self-supervised learning, diffusion probabilistic models, WiFi CSI, activity recognition.

1 Introduction

Human activity recognition is considered a key aspect for a variety of real-world applications, such as health monitoring and smart home [1], [2]. Among a great many recognition techniques, WiFi Channel State Information (CSI)-based approaches have the potential to achieve device-free, non-intrusive and privacy-friendly activity sensing, when compared to camera-based or wearable sensor-based methods [3], [4]. Correspondingly, a great many studies have been initiated on WiFi CSI-based activity recognition.

While, most of the models are powered by supervised machine learning methods, where a large training dataset with annotations is needed to maintain an acceptable performance, makes the training phase time consuming, labor intensive, and expensive. Consequently, collecting numerous labeled data is one of the major hurdles in applying these methods for practical applications [1], [5]. Contrastive self-supervised learning can be a potential solution to overcome the limitations associated with the lack of labels, because it can effectively leverage an enormous number of unlabelled samples to train the model without using labels [6]. Contrastive self-supervised learning has shown superior performance in the image processing [7] [8] and natural language processing [9] [10]. However, directly applying contrastive learning to activity recognition tasks is confronted with two additional issues.



(a) Gaussian blur (b) DDPM-based augmentation

Fig. 1. Augmented data by different methods. (a) The augmented waveform (orange) by Gaussian blur is almost the same to the original one (blue). (2) The augmented waveform (orange) by our DDPM-based augmentation method can combine the characteristics of the two samples (solid and dotted blue).

First, prevailing augmentation approaches in contrastive learning, such as Gaussian blur and color distortion, hardly change the shape of the CSI waveform, leading to sub-optimal performance. General data augmentation methods are particularly designed for image data, which focus on manipulating pixels to generate augmented data. However, WiFi CSI is a kind of time-series data, and manipulating points in CSI data by these methods scarcely change its waveform. An example is presented in Figure 1(a), which suggests that the augmented waveform by Gaussian blur (orange) is quite similar to the original one (blue). However, if two augmented samples are the same in contrastive learning, few benefits can be provided for performance improvement [11]. Hence, these augmentation methods can only provide limited effectiveness for CSI data.

Second, typical contrastive learning models fail to con-

- *Chunjing Xiao is with the School of Computer and Information Engineering, Henan University, Kaifeng 475004, China, and also with University College London, London WC1E 6BT, UK (e-mail: Chunjing.Xiao@ucl.ac.uk).*
- *Yanhui Han and Yane Hou are with the School of Computer and Information Engineering, Henan University, and also with the Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475004, China (e-mail: hanyanhui@henu.edu.cn, houyane@henu.edu.cn).*
- *Fangzhan Shi and Kevin Chetty are with the Department of Security and Crime Science, University College London, London WC1E 6BT, UK (e-mail: Fangzhan.shi.17@ucl.ac.uk, k.chetty@ucl.ac.uk).*

Manuscript received April 19, 2005; revised August 26, 2015. (Corresponding author: Yane Hou.)

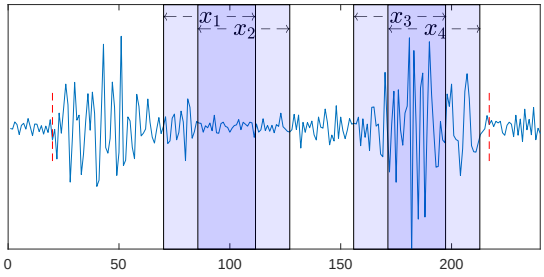


Fig. 2. Positive sample pairs extracted from an activity where the dotted red lines are the start and end points and there is pause near the center. Compared to positive pair (x_3, x_4), positive pair (x_1, x_2) should provide less clues for learning data representation because they contain more pause data.

sider the difference of the sample importance during model training. In contrastive learning, the same weights are generally assigned to all the positive sample pairs for model training. However, for CSI-based activity recognition, different positive sample pairs might provide various clues for learning data representation. For some activities consisting of multiple strokes, such as drawing X and lying down, there might be a pause between two strokes. If positive sample pairs extracted from CSI data contain more pause data, they will provide less clues for learning data representation, and should play a minor role for model training, and vice versa. An example of drawing X is presented in Figure 2, where the dotted red lines are the real start and end points of the activity. In this activity, there is a pause between the two strokes. Compared to positive pair (x_3, x_4), positive pair (x_1, x_2) contains more pause data and should provide minor clues for learning data representation.

To address these issues, we propose a diffusion model-based Contrastive self-supervised Learning framework for human Activity Recognition (CLAR) using WiFi CSI. On the basis of the contrastive learning model, we design two components for the scene of CSI-activity recognition: a denoising diffusion probabilistic model (DDPM)-based time series-specific data augmentation model and an adaptive weight algorithm. The designed augmentation model takes as inputs a source sample and a reference sample from users with different habits, and produces a new sample with the combined characteristics of them. These augmented data can effectively amplify training data and enhance generalization capacity of the model. The adaptive weight algorithm adaptively computes the weights of positive sample pairs, which are imposed on the contrastive loss to boost model performance.

Specifically, in the DDPM-based time series-specific data augmentation model, we feed a Gaussian noise into the reverse diffusion process of DDPM [12] to generae a clean CSI data by gradual denoising. During this denoising process, we regard the source sample and the reference sample as the conditions, and impose them into the reverse diffusion process of DDPM to generate a new sample with compromised characteristics of them. These generated samples not only differ from the input ones in CSI waveform, but also complement the limited training data to enhance model robustness. By combining both source and reference samples, the generated samples have different waveforms from the source and reference ones. Moreover, for CSI-based activity recognition, it is difficult to gather enough training data to cover all kinds of motions habits, since waveforms of CSI data collected from users with different motion habits can be different even they perform the same action [13], and different persons have various motion habits. Our designed augmentation method can generate augmented data with new characteristics, which can complement limited gathered data. A visual example of the generated sample is presented in Figure 1(b), where the source sample is from the user with the habit of tending to draw a small circle (solid blue), and the reference for a large circle (dotted blue). Correspondingly the generated sample is the one for a middle circle (orange).

In the adaptive weight algorithm, we try to adjust the weights of positive pairs in model training to capture the difference of the sample importance and enhance model performance. For CSI data, different positive sample pairs provide various clues for learning data representation, i.e., positive sample pairs with less activity data should play a minor role for model training since they contain less clues for learning data representation, and vice verse. Hence, for each positive sample pair, we first compute a response map to reflect the amount of activity data in the positive pair, and then calculate the weight based on the response map. This weight will be incorporated into the contrastive loss to enhance the model performance. By incorporating the DDPM-based augmentation model and the adaptive weight algorithm into the basic contrastive learning framework, our model can efficiently boost the recognition performance.

We summarize the main contributions of this paper as follows:

- We propose a diffusion model-based Contrastive self-supervised Learning framework for Activity Recognition using WiFi CSI, CLAR, which can address the problem of the shortage of labeled data.
- We design a DDPM-based time series-specific augmentation method to produce augmented samples with new characteristics, which can amplify training data to enhance generalization capacity of the model.
- We present an adaptive weight algorithm, which can adaptively adjust the weights of positive sample pairs in the contrastive loss to enhance model performance.
- Experiment results illustrate that our framework outperforms the state-of-the-art approaches.

2 Preliminaries

In this section, we give necessary background information of the contrastive learning framework and the denoising diffusion probabilistic model.

2.1 Contrastive Learning Framework

Contrastive learning learns a representation by maximizing similarity and dissimilarity over data samples which are organized into similar (positive) and dissimilar (negative) pairs, respectively. Typical contrastive learning methods adopt the noise contrastive estimation (NCE) objective for discriminating different instance in the dataset. Concretely, NCE objective encourages different augmentations of the same instance to be pulled closer in a latent space yet

pushes away different instances' augmentations. In general, the framework can be summarized as the following components: (i) A data augmentation module that transforms any given data example randomly resulting in two correlated views of the same example, such as random cropping, color jittering, and random flipping. (ii) An encoder network f which extracts representation vectors from augmented data examples by mapping it into a d-dimensional space $\mathbb{R}^d$. (iii) A projection head h which further maps extracted representations into a hyper-spherical (normalized) embedding space. This space is subsequently used for a specific pretext task, i.e., contrastive loss objective for a batch of positive/negative pairs. The InfoNCE [14] objective can be expressed as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(sim\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathrm{I}_{[k\neq i]} \exp\left(sim\left(z_i, z_k\right)/\tau\right)}, \quad (1)$$

where $sim(,)$ denotes cosine similarity, z denotes the outputs from the non-linear projection head as used in the original SimCLR work, and τ is a temperature hyper-parameter scaling the distribution of distances.

2.2 Denoising Diffusion Probabilistic Model

Denoising diffusion probabilistic models (DDPM) [12] is a class of generative models that show superior performance in unconditional image generation. It learns a Markov Chain which gradually converts a simple distribution (e.g., isotropic Gaussian) into a data distribution. Generative process learns the reverse of the DDPM forward (diffusion) process: a fixed Markov Chain that gradually adds noise to data. Here, each step in the forward process is a Gaussian translation:

$$q\left(z^t|z^{t-1}\right) := N\left(z^t; \sqrt{1-\beta_t}z^{t-1}, \beta_t \mathrm{I}\right), \quad (2)$$

where $\beta_1, \ldots, \beta_T$ is a fixed variance schedule rather than learned parameters [12]. Eq. (1) is a process finding z^t by adding a small Gaussian noise to the latent variable z^{t-1}. Given clean data z^0, sampling of z^t can be expressed in a closed form:

$$q\left(z^t|z^0\right) := N\left(z^t; \sqrt{\bar{\alpha}_t}z^0, \left(1-\bar{\alpha}_t\right)\mathrm{I}\right), \quad (3)$$

where $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Therefore, z^t is expressed as a linear combination of z^0 and ε:

$$z^t = \sqrt{\bar{\alpha}_t}z^0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad (4)$$

where $\varepsilon \sim N\left(0, \mathrm{I}\right)$ has the same dimensionality as data z^0 and latent variables $z^1, \ldots, z^T$.

Since the reverse of the forward process, $q\left(z^{t-1}|z^t\right)$, is intractable, DDPM learns parameterized Gaussian transitions $p_\theta\left(z^{t-1}|z^t\right)$. The generative (or reverse) process has the same functional form [15] as the forward process, and it is expressed as a Gaussian transition with learned mean and fixed variance [12]:

$$p_\theta\left(z^{t-1}|z^t\right) = N\left(z^{t-1}; \mu_\theta\left(z^t, t\right), \sigma_t^2 \mathrm{I}\right). \quad (5)$$

Further, by decomposing μ_θ into a linear combination of z^t and the noise approximator ε_θ, the generative process is expressed as:

$$z^{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(z^t, t\right)\right) + \sigma_t^2\varepsilon, \quad (6)$$

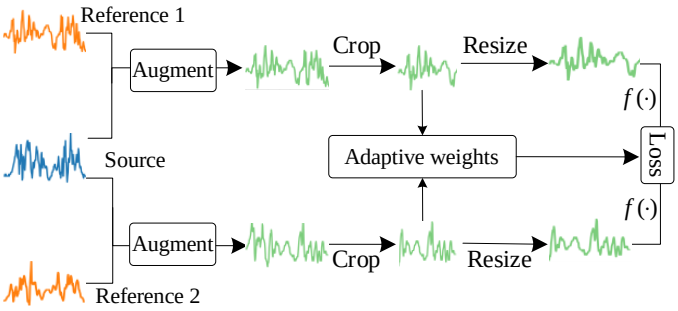


Fig. 3. CLAR framework. During the training process, the reference and source samples are fed into our designed DDPM-based augmentation model to generate augmented data with new characteristics. These augmented data are further processed by cropping and resizing to build the contrastive loss. Meanwhile, the weight of each sample pair is computed by our devised adaptive algorithm and further is incorporated into the contrastive loss to enhance model performance.

which ε is a noise suggesting that each generation step is stochastic. Here ε_θ represents a neural network with the same input and output dimensions and the noise predicted by the neural network ε_θ in each step is used for the denoising process in Eq. 6.

3 CLAR Framework

In this section, we present the diffusion model-based contrastive self-supervised learning framework for human activity recognition (CLAR). First, we illustrate an overview of the recognition framework. Next, we present the DDPM-based data augmentation method, which will generate augmented data with characteristics of both the source and reference samples. Finally, we illustrate the adaptive weight algorithm to compute weights of different positive sample pairs for the contrastive loss.

3.1 Overview of the Proposed Framework

To address the shortage of labeled training data, we design a new contrastive learning framework for human activity recognition, CLAR, whose overview is illustrated in Figure 3. The model takes a source sample and two reference samples as inputs. The source sample and the two reference samples are first combined by our designed DDPM-based time series-specific augmentation mothod to generate two augmented samples as a positive pair. Then, the augmented samples are processed by the cropping and resizing operations for building the contrastive loss. During these procedure, the weights of sample pairs are computed by our proposed adaptive weight algorithm, and these weights are incorporated into the contrastive loss to enhance the model robustness.

Compared with typical contrastive learning models such as SimCLR [8], we design a DDPM-based data augmentation model and an adaptive weight algorithm to satisfy the requirement of CSI-based activity recognition and enhance recognition performance. Since the prevailing augmentation approaches mainly focus on image and text processing, which are ineffective for CSI data. Besides, since limited training data cannot cover all the motion habits, augmentation models for CSI-based activity recognition should be able to generate augmented data with new motion habits.

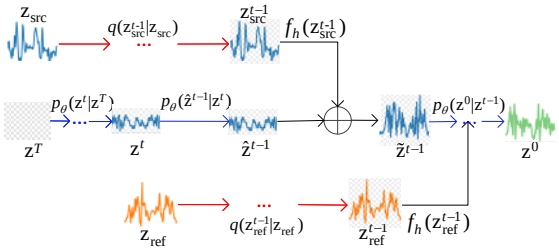


Fig. 4. DDPM-based data augmented model. Red arrows → indicate the forward diffusion, blue ones → refer to the reverse diffusion, and ⊕ is the concatenation operation.

Hence we design a DDPM-based time series-specific data augmentation model, which can combine two samples from users with different habits to generate a augmented sample with compromised ones. This augmentation method is conducive to improving the generalization ability and recognition performance of the model.

Further, for CSI data, various positive sample pairs provide different clues for learning data representation , i.e., positive sample pairs with less activity data should play a minor role for model training since they contain less clues for learning data representation, and vice verse. Therefore, we propose an adaptive weight algorithm to adjust the weights of positive samples in model training. These weights are incorporated into the contrastive loss to enhance the data representations, and futher improve recognition performance.

3.2 DDPM-based Data Augmentation Model

Contrastive learning algorithms try to learn representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. Hence, the data augmentation operation is crucial in learning data representations [8]. Various augmentation methods for contrastive learning, such as Gaussian blur and color distortion, have been designed to enhance performance of image processing [7] [8] and natural language processing [9] [10]. However, the prevailing augmentation approaches might be improper for WiFi CSI data. For example, the typical augmentation way, Gaussian blur, generally yields limited effectiveness for WiFi CSI-based activity recognition, because the augmented CSI has nearly the same waveform with the original one.

Moreover, for CSI-based activity recognition, the augmentation models should generate diverse augmented samples to enhance the coverage of training data. Due to the diversity of user habits, it is difficult to gather enough training data to cover all kinds of user habits. However, the limited training data might lead to inferior generalization capacity when the test samples are from users with different motion habits. For example, when asking two users to draw a circle, one might draw a big circle, while another may draw a small circle. The recognition model trained based on the data from these two users cannot accurately identify the data from a user who tending to draw a middle circle. Hence, the augmentation method should be able to generate augmented data with new characteristics.

Towards this purpose, we design a DDPM-based data augmentation model, which can combine two samples from users with different habits to generate a augmented sample with characteristics of both users. Specifically, inspired by superiority of DDPM on image and audio generation [16], [17], [18], [19], we introduce DDPM as the basic framework to build the augmentation model. This model takes a source sample and a reference sample as inputs, and outputs a compromised sample by combining both input samples. The main idea is that the source and reference samples are considered as the conditions, which are exerted on the reverse diffusion (denoising) process of DDPM to generate augmented samples. In this way, our designed augmentation model iteratively exert features of the source and reference samples on the generative process to manufacture augmented samples. This gradual refining process can produce more suitable data.

Figure 4 presents the framework of the designed DDPM-based augmentation model. In this model, we first feed a Gaussian noise into the reverse diffusion process to generate a clean CSI waveform through gradual denoising, i.e., $z^T \rightarrow z^t \rightarrow \hat{z}^{t-1} \rightarrow \tilde{z}^{t-1} \rightarrow z^{t-1} \rightarrow z^0$. During this denoising process, the features of the source and reference samples, z_{src} and z_{ref}, are extracted, and iteratively injected into the latent variable $\hat{z}^{t-1}$. As a result, the generated (augmented) data z_0 contains the characteristics of both source and reference data, which can be regarded as a compromise of them. Assuming the source sample refers to CSI data performed by the user tending to draw a big circle and the reference sample for a small circle, the augmented data can be considered as the one for a middle circle.

Concretely, on the basis of the reverse diffusion process $p_\theta\left(z^{t-1}|z^t\right)$, we exert the conditions c_{src} and c_{ref} on the reverse diffusion process. Hence, we approximate the Markov transition under the condition c_{src} and c_{ref} as follows:

$$\begin{aligned} &p_\theta\left(z^{t-1}|z^t, c_{\text{src}}, c_{\text{ref}}\right) \\ &\quad\approx p_\theta\Big(z^{t-1}|z^t, f_l\left(z^{t-1}\right) = \sigma\left(f_l(z_{\text{src}}^{t-1}), f_l(z_{\text{ref}}^{t-1})\right)\Big), \end{aligned} \tag{7}$$

where z_{src}^{t-1} and z_{ref}^{t-1} are sampled by Equation 4, $f_l(\cdot)$ is a low-pass filter, and σ is a aggregation function which concatenates z_{src}^{t-1} and z_{ref}^{t-1} based on the warping path. Here, the warping path, produced using Dynamic Time Warping (DTW) [20], [21], maps the elements of two data sequences to minimize the distance between them. Here, we adopt a warping path, instead of the default shortest path, to concatenate them because the warping path can more appropriately keep the shape of the waveforms [22].

Equation 7 tries to incorporate z_{src}^{t-1} and z_{ref}^{t-1} into the generated data. Hence, the generated data will has the compromised characteristics of both them. According to this equation, in each transition from z_t to z_{t-1}, the features of both source and reference samples are extracted and then injected into the latent variable. To this end, we first adopt the forward process (Euqation 4) to compute z_{src}^{t-1} and z_{ref}^{t-1} from z_{src} and z_{ref}, respectively:

$$\begin{aligned} \mathbf{z}_{\text{src}}^{t-1} &\sim q(\mathbf{z}_{\text{src}}^{t-1}|\mathbf{z}_{\text{src}}), \\ \mathbf{z}_{\text{ref}}^{t-1} &\sim q(\mathbf{z}_{\text{ref}}^{t-1}|\mathbf{z}_{\text{ref}}). \end{aligned} \tag{8}$$

Then, we adopt the reverse process (Equation 5) to compute latent variable $\hat{z}^{t-1}$ from z^t:

$$\hat{\mathbf{z}}^{t-1} \sim p_\theta(\hat{\mathbf{z}}^{t-1}|\mathbf{z}^t). \tag{9}$$

As a result, the augmented sample is refined by matching $f_l(\hat{\mathbf{z}}^{t-1})$ of $\hat{\mathbf{z}}^{t-1}$ with that of $\sigma(f_l(z_{\text{src}}^{t-1}), f_l(z_{\text{ref}}^{t-1})))$ as follows:

$$\begin{aligned}\tilde{z}^{t-1} =&(1-\gamma_1)\hat{z}^{t-1} + \gamma_1\left(f_l(\sigma(\hat{z}^{t-1}, z_{\text{src}}^{t-1})) - f_l(\hat{z}^{t-1})\right),\\ z^{t-1} =&(1-\gamma_2)\tilde{z}^{t-1} + \gamma_2\left(f_l(\sigma(\tilde{z}^{t-1}, z_{\text{ref}}^{t-1})) - f_l(\tilde{z}^{t-1})\right),\end{aligned} \tag{10}$$

where $\gamma_1, \gamma_2 \in [0, 1]$ denote the hyper-parameters to adjust the weights. The matching operation by Equation 10 ensures the conditions c_{src} and c_{ref} in Equation 7, which further enables the conditional generation based on DDPM. In this way, through injecting the features of both samples collected from users with various habits into the latent variable in the generative process, the generated (augmented) data can possess compromised characteristics of them. Hence, augmented data can be considered to be the one collected from another users with different habits. Both augmented samples and source samples will be used for model training.

3.3 Adaptive Weighting

In contrastive learning, cropping is a commonly used way to extract views for building positive sample pairs [23], [24]. For each activity data, we also adopt cropping operations to extract two views (samples) from the same activity data to form a positive pair. While, for CSI data, the clues provided by different positive pairs should be various in learning data representation. For some activities, there may be a pause among the action. For example, for drawing X, a pause can occur between the two strokes, and for lying down, it can occur between sitting and lying. Hence, some positive pairs extracted by cropping operations might contain more activity data, while others might include more pause data. Correspondingly, the positive pairs containing more activity data should provide more clues for learning data representation, and vice versa. An example of drawing X is presented in Figure 2, where the positive pair (x_3, x_4) should play a more important role for model training than (x_1, x_2), because the former contains more activity data.

Towards this goal, we propose an adaptive weight algorithm to adjust the importance of positive pairs for model training by assigning various weights to different positive pairs. This algorithm first computes a response map which can reflect the amount of activity data in the positive pair, and then computes the weights based on the response map for constructing the contrastive loss.

Concretely, to compute the response map, we first select a template w^T with length H from the the CSI sequence in the absence of activity data, called *static template*. To avoid the selection bias, we choose multiple static templates. Then, for each sample from positive pairs, we split it into overlapping windows using a sliding window, each with length H, where the sliding step is 1. For window l extracted from sample x_i, we adopt a response score to reflect the amount of containing activity data:

$$S_l = \frac{1}{M}\sum_{k=1}^{M}\text{DTW}(w_l, w_k^T), \tag{11}$$

where M is the number of selected static templates, and $\text{DTW}(w_l, w^T)$ denotes the DTW distance between the w_l and w^T. The bigger distance between w_l and w^T indicates that w_l is more different from static template w^T, i.e., w_l contains more activity data. Therefore, response score S_l reflects the amount of activity data in window l. The response scores of the windows in x_i are merged to form the response map of this sample.

After obtaining response maps, we calculate the weights of sample x_i for model training:

$$W_i = \left(\frac{1}{N_w}\sum_{k=1}^{N_w}\text{I}\left(S_k, \sigma_s\right)\right)^{\alpha}, \tag{12}$$

where α denotes the power which controls the scale of weights, N_w refers to the number of the windows extracted from x_i, and $\text{I}(,)$ is the indicator of the presence of activity data, and is defined as:

$$\text{I}(S_k, \sigma_s) = \begin{cases} 1, & \text{if} S_k > \sigma_s \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

Here σ_s is a threshold to determine whether this window is regarded as data in the presence of activities. σ_s can be set to the average of the response scores, i.e., $\sigma_s = \left(\sum_{k=1}^{N_w} S_k\right)\Big/ N_w$. Further, for a positive pair (x_i, x_j), its weight is the aggregation of the weights of both samples:

$$W_{(i,j)} = \text{Aggregate}(W_i, W_j), \tag{14}$$

where $\text{Aggregate}(,)$ sums the two items. This weigh suggests the amount of containing CSI data in the presence of activities. Hence, the positive paris with a larger weight contains more clues and should play a more significant role in the model training.

3.4 Overall Model

Taking the augmented data and adaptive weights into account, we formulate the loss function as follows:

$$\mathcal{L}_{i,j}^{aug} = -\log\frac{\exp\left(W_{(i,j)} * sim\left(\hat{z}_i, \hat{z}_j\right)\Big/\tau\right)}{\sum_{k=1}^{k=2N}\text{I}_{k\neq i}\exp\left(sim(\hat{z}_i, \hat{z}_k)/\tau\right)}, \tag{15}$$

where N is the length of the minibatch, τ is a temperature hyper-parameter scaling the distribution of distances. $\hat{z}_i$ and $\hat{z}_j$, which form a positive pair, are two embeddings which are extracted from the two augmented samples derived from the same source sample, and $\hat{z}_i$ and $\hat{z}_k$, which form a negative pair, are derived from the different source samples.

Further, we also adopt the original data without the process of our designed augmentation model to build the contrastive loss to capture the characteristics of original training data. This loss is defined as:

$$\mathcal{L}_{i,j}^{ori} = -\log\frac{\exp\left(W_{(i,j)} * sim\left(z_i, z_j\right)\Big/\tau\right)}{\sum_{k=1}^{k=2N}\text{I}_{k\neq i}\exp\left(sim(z_i, z_k)/\tau\right)}, \tag{16}$$

where z_i, z_j and z_k are the embeddings of the original samples without being processed by our augmentation model, and the other parameters are the same to Eqation 15. As a reult, the overall loss is the sum of them:

$$\mathcal{L}_{i,j}^{all} = \mathcal{L}_{i,j}^{aug} + \mathcal{L}_{i,j}^{ori}. \tag{17}$$

After obtaining the trained model, it is used to extract representations of activity samples. Further, a linear classifier is adopted to classify the representations into corresponding activity categories.

4 Experimental Evaluation

In this section, we evaluate the effectiveness of the proposed CLAR by comparing it with the several baselines with different techniques. Also, we conduct ablation studies and inspect the role of the augmentation model and labeled data size. The data and code are available online[1].

4.1 Experiment Setup

For the evaluation, we conduct the experiments on two WiFi CSI-based behavior recognition datasets. **SignFi data** [25] consists of 1,250 CSI sequences, each of which represents a sign language gesture. These activities are performed by 4 users with each activity repeated for 10 times. **DeepSeg data** [26] is composed of 1,500 human activities from 5 users with various shapes and ages. For these experiments, We will use 80% of all data as the training set and the rest as the test set. In the training set, we select 30% and 20% of data as labeled data for fine-tuning the classifier on SignFi and DeepSeg, respectively. For selection of source and reference samples, if they are labeled, we select two samples from the same activity category as the source and reference ones. If they are unlabeled, for a source sample, we randomly select one sample from its top 10 most similar samples as the reference one. Here, we adopt DTW to compute the similarity degree between samples. For the DDPM-based augmentation model, we use 1,000 diffusion steps considering both efficiency and effectiveness. During the optimization process of CLAR, the learning rate and batch size are set to 0.0001 and 50, respectively. The hyper-parameters α in Equation 12 and τ in Equation 15 are set 0.5 and 0.1, individually, for both datasets. For all the following experiments, the accuracy and F1-score are adopted as metrics for performance comparison.

4.2 Baselines

To prove the effectiveness and superiority of the proposed model, we choose activity recognition methods with different technologies as the baselines, including GAN-based [27], [13], Meta learning-based [28], [29] and self-supervised contrastive learning-based [30], [8], [31] appoaches:

- **ManiGAN** [27]: A GAN-based semi-supervised learning method incorporating manifold regularization. The method exhibits obvious merits on image classification compared to other GAN-based and non-GAN-based semi-supervised methods.
- **CsiGAN** [13]: A GAN-based activity recognition model using WiFi CSI. This model introduces a new complement generator and optimizes the outputs and loss functions of the discriminator to improve performance of activity recognition.
- **RF-Net** [29]: A unified meta-learning framework for RF-enabled one-shot activity recognition. It delivers the capability of being adaptive to new environments with very few labeled data.
- **MetaAct** [28]: A meta learning-based adaptable activity recognition model. This approach is specifically designed for recognizing activities across scenes and categories using WiFi CSI.

1. https://github.com/ChunjingXiao/CLAR

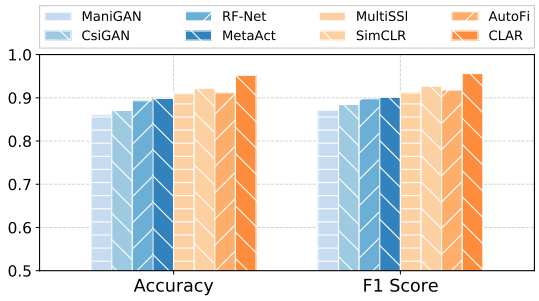


Fig. 5. The activity recognition performance for SignFi data

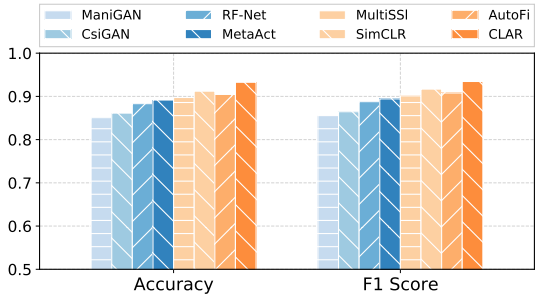


Fig. 6. The activity recognition performance for DeepSeg data

- **MultiSSL** [30]: A self-supervised learning method for human activity recognition. MultiSSL learns accelerometer representations by training a temporal convolutional neural network to recognize the transformations applied to the raw input signal.
- **SimCLR** [8]: A simple framework for contrastive learning of visual representations. SimCLR learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space.
- **AutoFi** [31]: A self-supervised learning activity recognition model using WiFi CSI. AutoFi fully utilizes unlabeled low-quality CSI samples to learn the knowledge, which is further transfered to specific tasks.

4.3 Recognition Performance Comparison

Figure 5 and 6 report the results of our model and the baseline models across the two datasets: SignFi and DeepSeg. From these results, we have following observations. First, ours model CLAR consistently yields better performance on the two datasets. For example, compared to SimCLR, CLAR exhibits improvements of more than 3% and 2% on the SignFi and DeepSeg datasets, respectively. CLAR achieves more distinct improvement on SingFi. The reason is that there are more activity categories on SingFi, meaning fewer labeled samples per category. The limited labeled samples lead to inferior performance for the baselines. However, by generating augmented data and taking advantage of unlabeled data, our method CLAR can efficiently address this issue and achieve higher performance.

Second, meta learning-based methods, RF-Net and MetaAct, exceed the two GAN-based semi-supervised baselines, CsiGAN and ManiGAN. Since meta learning-based methods are designed for the scenarios with a few labeled

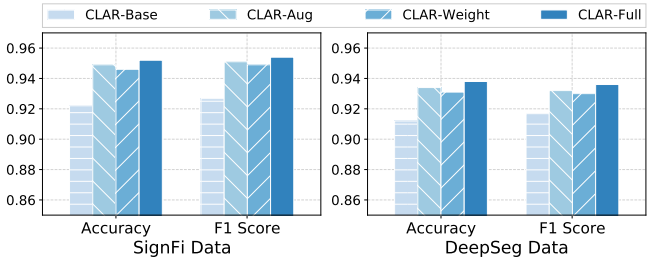


Fig. 7. The performance with different design choices for SignFi data and DeepSeg data.

samples. Therefore, under the environment with limited labeled samples, the meta learning-based models can obtain better performance than the semi-supervised models, which generally require a given number of labeled sample to obtain expected performance.

Third, self-supervised models outperform the others baselines. In particular, SimCLR, which is designed for image processing, also achieve relatively good performance, compared with the GAN-based models. This indicates that the self-supervised techniques can effectively benefit WiFi CSI-based human activity recognition especially for the scenarios with limited training data. However, by incorporating our designed augmentation model and adaptive weight algorithm, our model CLAR significantly outperforms these baselines.

4.4 Ablation Study

Here we investigate the contribution of the two important components in CLAR, i.e., the augmentation model and adaptive weight algorithm. Specifically, we investigate the role of different components by considering the following variants of our model: (1) *CLAR-Base* is the basic contrastive learning framework that removes the DDPM-based time series-specific augmentation model and the adaptive weight algorithm. (2) *CLAR-Aug* is the contrastive learning framework with the DDPM-based augmentation model but without the adaptive weight algorithm. (3) *CLAR-Weight* is the contrastive learning framework with the adaptive weight algorithm but without DDPM-based augmentation model. (4) *CLAR-Full* is our proposed model fully incorporating all the components.

The experimental results using SignFi data and DeepSeg data are presented in Figure 7. We summarize the observations from this figure as follows. First, CLAR-Full performs the best, while CLAR-Base is the worst model, which implies that the main components we proposed can significantly improve the recognition performance. Second, when incorporating the DDPM-based augmentation model, CLAR-Aug obtains better results than CLAR-Base. This is because that the limited samples are augmented by our designed method, which can benefit the model in improving generalization capacity on the test data. Third, CLAR-Weight outperforms CLAR-Base by a certain margin. The results prove the motivation of our model, i.e., introducing the adaptive wights can enable the model to capture more characteristics of activity data and further significantly enhance recognition performance.

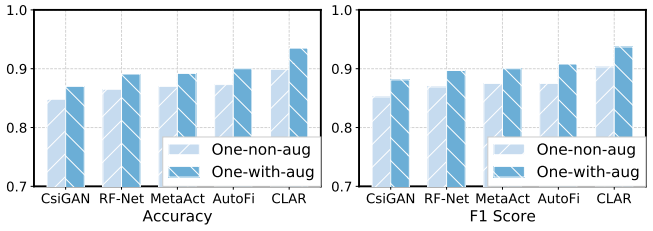


Fig. 8. The performance with/without the augmented data for SignFi data.

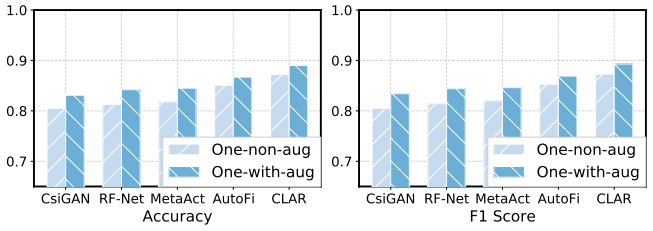


Fig. 9. The performance with/without the augmented data for DeepSeg data.

4.5 Role of the Augmentation Model

The analyses in the previous section suggest our designed augmentation model can effectively contribute to the performance improvement. Here, we futher inspect the efficacy of the augmented data when applying them to other activity recognition models. We select four baseline approaches which are specially designed for WiFi CSI-based activity recognition under cross scenes: *CsiGAN* [13], *RF-Net* [29], *MetaAct* [28] and *AutoFi* [31]. We evaluate the model performance with/without the augmented data for model training, named as *one-with-aug*/*one-non-aug*. To inspect the generalization capacity, we conduct these experiments under the left-out scene, i.e. the data of one user are extracted as the test data, and others as the training data.

Figure 8 and Figure 9 show the accuracy and F1 for these four baselines and our CLAR with/without the augmented data generated by our DDPM-based augmentation model. As shown in Figure 8, the performance of one-with-aug substantially exceeds that of one-non-aug for all the models on SignFi data. For example, the F1 of one-with-aug for AutoFi is about 3.2% higher than that of one-non-aug. The DeepSeg dataset, presented in Figure 9, also exhibit the similar trends. This results indicate that our DDPM-based augmentation model can generate effective augmented smaples by combining multiple samples. The generated data can enlarge the distribution of training data and further enhance generalization capacity. Also, our augmentation model can be applied to other similar recongition models.

4.6 Role of Labeled Data Size

Our model requires a number of labeled data to fine-tune the classifier. Here we investigate the role of labeled data size. For these experiments, we select p = [40, 60, 100]% of the training samples as unlabeled data, and select q% as the labeled data.

As shown in Figure 10 and Figure 11, for all the p values, our model achieves increasing accuracy and F1 score with

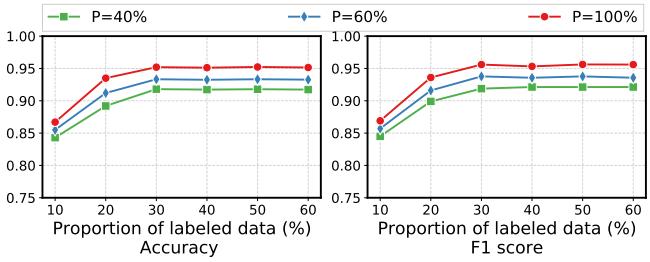


Fig. 10. The performance with different size of labeled data on SignFi data

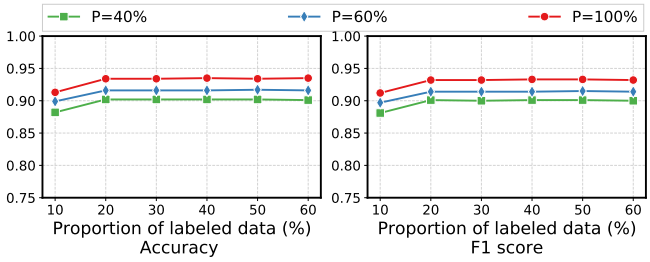


Fig. 11. The performance with different size of labeled data on DeepSeg data

the rise of the labeled data size on both datasets, which indicates that the labeled data size has an important impact on recognition performance for our model. While, when selecting 30% and 20% of the labeled data on SignFi and DeepSeg, individually, the performance becomes stable on both datasets, i.e., their accuracies are almost the same with that at 60%. This suggests that our model can efficiently take advantage of a few labeled samples to obtain expected performance.

Moreover, the growth rates of the two datasets are different. The accuracies on SignFi data increase sharply, while they are relatively stable on DeepSeg data. The reason behind is because the number of labeled samples per class is very different for the two datasets. In fact, there are 10 and 30 labeled samples per category for SignFi data and DeepSeg data, individually. Hence, the same ratio means the various number of labeled data for these two datasets, which further leads to different performance. These results suggest that our model needs a certain amount of training data to achieve better performance. However, the number of labeled data, such as 10 per category, is easily affordable by human labeling.

5 Related Work

This work is mainly related to two research areas: CSI-based activity recognition and contrastive self-supervised learning. Here, we will present an overview of the most closely related works in each area, and highlight the major differences between our study and these works.

5.1 CSI-based Activity Recognition

The studies on WiFi CSI-based activity recognition can be divided into two genres according to the number of available labeled data: supervised and semi-supervised approaches and few-shot learning-based methods.

Supervised methods mainly aim to adopt a number of labeled data to train classification models for activity identification. The researchers principally exploit different features or/and techniques to enhance recognition performance. For example, Chen *et al.* [32] extract information from both time and frequency domains for an end-to-end neural network model to conduct activity recognition, and they apply point-wise grouped convolution and depth-wise separable convolutions to confine the model scale and speed up the inference execution time. Zhang *et al.* [33] present a data augmentation method to transform and synthesize CSI data for alleviating the influence of motion inconsistency and subject-specific issue, and adopt a Dense-LSTM to classify activities. Sruthi *et al.* [34] propose a multi-stage deep learning model consisting of a convolutional neural network and other popular deep neural architectures, such as Alexnet, Googlenet and Squeezenet, for WiFi sensing-based human activity recognition. Xiao *et al.* [26] explore the interaction between the activity segmentation and classification to improve activity recognition performance. Chen *et al.* [35] design an attention based bi-directional long short-term memory model for passive human activity recognition using WiFi CSI signals. Shi *et al.* [36] propose an innovative scheme, which combines an activity-related feature extraction and enhancement method and matching network. The proposed scheme can be directly applied in new/unseen environments without retraining. *Semi-Supervised approaches* try to leverage unlabeled data to compensate for the shortage of labeled data for activity recognition. For instance, Xiao *et al.* [13] proposed a semi-supervised generative adversarial network to exploit unlabeled data for CSI-based activity recognition. Yuan *et al.* [37] propose a human continuity activity semi-supervised recognizing method in multi-view IoT network scenarios. They combine supervised activity feature extraction with unsupervised encoder-decoder modules, which can capture continuity activity features from sensor data streams.

Few-shot learning-based methods intend to recognize a set of target classes by learning with sufficient labeled samples from a set of source classes but only with a few labeled samples from the target classes [38]. Due to the difficulty of collecting numerous labeled data in the target domain, this technique is widely applied to the field of activity recognition. For instance, Zhang *et al.* [28] present an adaptable CSI activity recognition framework based on meta-learning, which can apply to new environments or new types of activities by fine-tuning the model with very little train effort. Zhang *et al.* [39] propose a graph-based few-shot learning framework with dual attention mechanisms for human activity recognition. The model uses a feature extraction layer, including the convolutional block attention module, to extract activity related information from CSI data. Wang *et al.* [40] propose a multimodal CSI-based activity recognition framework, which leverages existing WiFi infrastructures and monitors human activities from CSI measurements. Wang *et al.* [41] propose a few shot learning-based human activity recognition framework, which can achieve expected performance in recognizing new categories through a small amount of samples to fine-tune the model parameters and avoid retraining the network from scratch. Shi *et al.* [42] design a human activity recognition scheme using matching

network with enhanced CSI to perform one-short learning for recognizing human activities in a new environment. Feng *et al.* [43] propose a few-shot human activity recognition method, which leverages a deep learning model for feature extraction and classification and implements knowledge transfer in the manner of model parameter transfer.

Difference: These supervised and semi-supervised methods require a number of labeled samples to obtain considerable performance. Meanwhile, the few-shot learning-based methods need sufficient labeled data from the source domain for model training. Instead, we introduce contrastive learning for CSI-based activity recognition, which can exploit unlabeled data to derive reliable recognition for scenarios where only small amounts of labeled training samples can be collected.

5.2 Contrastive Self-Supervised Learning

Contrastive self-supervised learning is an improtant division of self-supervised learning [44]. This technique tries to transform one item into multiple views, minimizes the distance between views from the same item, and maximizes the distance between views from different items in a feature map [8]. Contrastive methods have been applied to multiple fields, such as image processing, voice and natural language processing, and activity recognition.

In the field of *image processing*, various methods have been initiated to augment data and build effective views. For example, SimCLR [8] proposes the composition of multiple data augmentations, e.g., Grayscale, Random Resized Cropping, Color Jittering, and Gaussian Blur, to make the model more robust. InfoMin [45] introduces an information maximization principle which suggests that a good augmentation strategy should reduce the mutual information between the positive pairs while keeping the downstream task-relevant information intact. To explore the use of negative samples, InstDisc [46] proposes a memory bank to store the representation of all the images in the dataset. Meanwhile, SwAV [47] proposes to compute cluster assignments online while enforcing consistency between cluster assignments obtained from views of the same image. MoCo [48] increases the number of negatives by using a momentum contrast mechanism that forces the query encoder to learn the representation from a slowly progressing key encoder and maintains a long queue to provide a large number of negative examples. SupCon [49] shows that the positive and negative instances created by SimCLR do not take into account the correlation of features between different pictures belonging to the same class. Clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes. WCL [50] proposes a k-nearest neighbor based multi-crops strategy. They store the feature for every batch and then use these features to find the K closest samples based on the cosine similarity at the end of each epoch. CLSA [51] proposes to build stronger augmentation by a random combination of different augmentations.

In addition, contrastive learning is also widely used for *voice and natural language processing*. For example, in the field of voice, Yakura *et al.* [52] introduces self-supervised contrastive learning to acquire feature representations of singing voices. Tang *et al.* [53] proposes a novel one-shot voice conversion framework based on vector quantization voice conversion and AutoVC. In the domain of natural language processing, Qin *et al.* [10] utilized contrastive learning to explicitly align similar representations across source language and target language. Han *et al.* [54] proposed a cross-lingual contrastive learning framework to learn FGET models for low-resource languages.

Recently, contrastive learning is adopted to enhance the performance of *sensor-based activity recognition*. For example, Jain *et al.* [55] present a collaborative self-supervised learning method for senor-based activity recognition, which leverages natural transformations in the sensor datasets collected from multiple devices to perform contrastive learning. Khaertdinov *et al.* [56] combine a transformer-based encoder into a contrastive self-supervised learning framework to learn effective feature representations for sensor-based human activity recognition. Haresamudram *et al.* [57] introduce masked reconstruction as a viable self-supervised pretraining objective for wearable sensing device-based human activity recognition. Saeed *et al.* [30] design a multi-task self-supervised approach, which presents a multi-task temporal convolutional network to learn generalizable features from sensory data. Xu *et al.* [58] design a dual-stream contrastive learning model that can process and learn the raw WiFi CSI data in a self-supervised manner. Liu *et al.* [59] introduce a short-time fourier neural network-based contrastive self-supervised representation learning framework, which takes both time-domain and frequency-domain features into consideration. Koo *et al.* [60] devise a self-supervised learning task that pairs the accelerometer and the gyroscope embeddings acquired from the same activity instance. Wang *et al.* [61] propose a sensor data augmentation method for contrastive learning, which introduces variable domain information and simulates realistic activity data by varying the sampling frequency to maximize the coverage of the sampling space. Wang *et al.* [62] present a new contrastive learning framework for sensor-based human activity recognition, which first clusters the instance representations, and for each instance, samples from different clusters are regarded as negative pairs.

Difference: Compared to these methods, we present a novel diffusion model-based augmentation way for contrastive learning, which can combine two samples from users with different habits into a new one with compromised characteristics. Different from these augmentation methods, our augmentation method can generate effective samples to fill the gap among limited training data, and further enhance the generalization capacity of the model. Also, we propose an adaptive weight algorithm to assign appropirate weights to different positive sample pairs, which is ignored in the aforementioned approaches.

6 Conclusions

In this paper, we presented a diffustion model-based contrastive self-supervised learning framework for human activity recognition using WiFi CSI, CLAR. In this framework, we designed a DDPM-based time series-specific augmentation model, which can merge two samples from users with different motion habits to generate augmented samples

with combined characteristics for amplifying training data and enhancing generalization capacity. Also, we presented an adaptive weight algorithm, which can adaptively adjust the weights of positive sample pairs for learning better data representations. Based on two datasets, experimental results illustrate that CLAR significantly outperforms the stateof-the-art baselines.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (62072077).

REFERENCES

[1] F. Gu, M. Chung, M. H. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1–34, 2021.

[2] Y. Yang, H. Wang, R. Jiang, X. Guo, J. Cheng, and Y. Chen, "A review of iot-enabled mobile healthcare: technologies, challenges, and future trends," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9478–9502, 2022.

[3] Y. Ma, G. Zhou, and S. Wang, "Wifi sensing with channel state information: A survey," *ACM Computing Surveys*, vol. 52, no. 3, pp. 1–36, 2019.

[4] R. Zhang, X. Jing, S. Wu, C. Jiang, J. Mu, and F. R. Yu, "Device-free wireless sensing for human detection: The deep learning perspective," *IEEE Internet Things Journal*, vol. 8, no. 4, pp. 2517–2539, 2021.

[5] C. Xiao, Y. Lei, C. Liu, and J. Wu, "Mean teacher-based cross-domain activity recognition using wifi signals," *IEEE Internet of Things Journal*, 2023.

[6] H. Haresamudram, I. Essa, and T. Plötz, "Assessing the state of self-supervised human activity recognition using wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–47, 2022.

[7] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 1747–1756.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 1597–1607.

[9] J. M. Giorgi, O. Nitski, B. Wang, and G. D. Bader, "Declutr: Deep contrastive learning for unsupervised textual representations," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 879–895.

[10] L. Qin, Q. Chen, T. Xie, Q. Li, J. Lou, W. Che, and M. Kan, "Gl-clef: A global-local contrastive learning framework for cross-lingual spoken language understanding," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 2677–2686.

[11] R. Zhu, B. Zhao, J. Liu, Z. Sun, and C. W. Chen, "Improving contrastive learning by visualizing feature transformation," in *International Conference on Computer Vision*, 2021, pp. 10 286–10 295.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[13] C. Xiao, D. Han, Y. Ma, and Z. Qin, "Csigan: Robust channel state information-based activity recognition with gans," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 191–10 204, 2019.

[14] C. Wu, F. Wu, and Y. Huang, "Rethinking infonce: How many negative samples do you need?" *The International Joint Conference on Artificial Intelligence*, pp. 2509–2515, 2022.

[15] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2256–2265.

[16] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2022.

[17] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

[18] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations*, 2021.

[19] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: bilateral denoising diffusion models for fast and high-quality speech synthesis," in *International Conference on Learning Representations*, 2022.

[20] Z. Zhang, R. Tavenard, A. Bailly, X. Tang, P. Tang, and T. Corpetti, "Dynamic time warping under limited warping path length," *Information Sciences An International Journal*, vol. 393, pp. 91–107, 2017.

[21] T. Phan, É. P. Caillault, A. Lefebvre, and A. Bigand, "Dynamic time warping-based imputation for univariate time series data," *Pattern recognition letters*, vol. 139, pp. 139–147, 2020.

[22] C. Xiao, S. Chen, F. Zhou, and J. Wu, "Self-supervised few-shot time-series segmentation for activity recognition," *IEEE Transactions on Mobile Computing*, pp. 1–14, 2022.

[23] X. Peng, K. Wang, Z. Zhu, M. Wang, and Y. You, "Crafting better contrastive views for siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 031–16 040.

[24] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Advances in Neural Information Processing Systems*, 2020, pp. 6827–6839.

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

[26] C. Xiao, Y. Lei, Y. Ma, F. Zhou, and Z. Qin, "Deepseg: Deep-learning-based activity segmentation framework for activity recognition using wifi," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5669–5681, 2021.

[27] B. Lecouat, C. S. Foo, H. Zenati, and V. R. Chandrasekhar, "Semi-supervised learning with gans: Revisiting manifold regularization," *International Conference on Learning Representations*, 2018.

[28] Y. Zhang, X. Wang, Y. Wang, and H. Chen, "Human activity recognition across scenes and categories based on csi," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2411–2420, 2022.

[29] S. Ding, Z. Chen, T. Zheng, and J. Luo, "Rf-net: a unified meta-learning framework for rf-enabled one-shot human activity recognition," *Proceedings of the Conference on Embedded Networked Sensor Systems*, pp. 517–530, 2021.

[30] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–30, 2019.

[31] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "Autofi: Towards automatic wifi human sensing via geometric self-supervised learning," *IEEE Internet of Things Journal*, 2022.

[32] Z. Chen, C. Cai, T. Zheng, J. Luo, J. Xiong, and X. Wang, "Rf-based human activity recognition using signal adapted convolutional neural network," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 487–499, 2021.

[33] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-lstm for human activity recognition using wifi signal," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4628–4641, 2021.

[34] P. Sruthi and S. K. Udgata, "An improved wi-fi sensing-based human activity recognition using multi-stage deep learning model," *Soft Computing*, vol. 26, no. 9, pp. 4509–4518, 2022.

[35] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "Wifi csi based passive human activity recognition using attention based blstm," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2714–2724, 2019.

[36] Z. Shi, Q. Cheng, J. A. Zhang, and R. Yi Da Xu, "Environment-robust wifi-based human activity recognition using enhanced csi and deep learning," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 24 643–24 654, 2022.

[37] R. Yuan and J. Wang, "The human continuity activity semi-supervised recognizing model for multi-view iot network," *IEEE Internet of Things Journal*, 2023.

[38] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys*, vol. 53, no. 3, pp. 1–34, 2020.

[39] Y. Zhang, Y. Chen, Y. Wang, Q. Liu, and A. Cheng, "Csi-based human activity recognition with graph few-shot learning," *IEEE Internet Things Journal*, vol. 9, no. 6, pp. 4139–4151, 2022.
[40] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Multimodal csi-based human activity recognition using gans," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17 345–17 355, 2021.
[41] Y. Wang, L. Yao, Y. Wang, and Y. Zhang, "Robust csi-based human activity recognition with augment few shot learning," *IEEE Sensors Journal*, vol. 21, no. 21, pp. 24 297–24 308, 2021.
[42] Z. Shi, J. A. Zhang, R. Y. Xu, and Q. Cheng, "Environment-robust device-free human activity recognition with channel-state-information enhancement and one-shot learning," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 540–554, 2022.
[43] S. Feng and M. F. Duarte, "Few-shot learning-based human activity recognition," *Expert Systems with Applications*, vol. 138, p. 112782, 2019.
[44] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2023.
[45] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision*, 2020, pp. 776–794.
[46] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.
[47] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.
[48] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9726–9735.
[49] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, 2020, pp. 18 661–18 673.
[50] M. Zheng, F. Wang, S. You, C. Qian, C. Zhang, X. Wang, and C. Xu, "Weakly supervised contrastive learning," in *International Conference on Computer Vision*, 2021, pp. 10 022–10 031.
[51] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2022.
[52] H. Yakura, K. Watanabe, and M. Goto, "Self-supervised contrastive learning for singing voices," *IEEE-ACM Transactions on Audio Speech and Language Processing*, vol. 30, pp. 1614–1623, 2022.
[53] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning," in *International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4613–4617.
[54] X. Han, Y. Luo, W. Chen, Z. Liu, M. Sun, B. Zhou, F. Hao, and S. Zheng, "Cross-lingual contrastive learning for fine-grained entity typing for low-resource languages," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 2241–2250.
[55] Y. Jain, C. I. Tang, C. Min, F. Kawsar, and A. Mathur, "Collossl: Collaborative self-supervised learning for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–28, 2022.
[56] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Contrastive self-supervised learning for sensor-based human activity recognition," in *IEEE International Joint Conference on Biometrics*, 2021, pp. 1–8.
[57] H. Haresamudram, A. Beedu, V. Agrawal, P. L. Grady, I. Essa, J. Hoffman, and T. Plötz, "Masked reconstruction based self-supervision for human activity recognition," in *Proceedings of the 2020 ACM International Symposium on Wearable Computers*, 2020, pp. 45–49.
[58] K. Xu, J. Wang, L. Zhang, H. Zhu, and D. Zheng, "Dual-stream contrastive learning for channel state information based human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 329–338, 2023.
[59] D. Liu, T. Wang, S. Liu, R. Wang, S. Yao, and T. Abdelzaher, "Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective," in *International Conference on Computer Communications and Networks*, 2021, pp. 1–10.
[60] I. Koo, Y. Park, M. Jeong, and C. Kim, "Contrastive accelerometer–gyroscope embedding model for human activity recognition," *IEEE Sensors Journal*, vol. 23, no. 1, pp. 506–513, 2022.
[61] J. Wang, T. Zhu, J. Gan, L. L. Chen, H. Ning, and Y. Wan, "Sensor data augmentation by resampling in contrastive learning for human activity recognition," *IEEE Sensors Journal*, vol. 22, no. 23, pp. 22 994–23 008, 2022.
[62] J. Wang, T. Zhu, L. Chen, H. Ning, and Y. Wan, "Negative selection by clustering for contrastive learning in human activity recognition," *IEEE Internet of Things Journal*, pp. 1–13, 2023.

Chunjing Xiao received the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China. He is currently an Associate Professor with the School of Computer and Information Engineering, Henan University, Kaifeng, China. He was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. His current research interests include recommender systems, representation learning, and Internet of Things.

Yanhui Han received the B.E degree from the School of Computer and Information Engineering, Henan University, China, in 2021. She is currently pursuing the M.S. degree in the School of Computer and Information Engineering, Henan University. Her current research interests include Internet of Things, wireless networks and data analytics.

Yan-e Hou received the Ph.D. degree from Henan University, Kaifeng, China. She is currently an Associate Professor with the School of Computer and Information Engineering, Henan University, Kaifeng, China. Her research interests currently include intelligent optimization algorithms, artificial intelligent and its relative applications.

Fangzhan Shi received the B.Eng. in Telecommunication Engineering in 2017 and M.Sc. in Robotics in 2018 at Hangzhou Dianzi University, China and University College London respectively. He worked as an artificial intelligence engineer at Supcon, China in 2019 and 2020. He is currently a PhD student in the department of security and crime science, University College London. His research interest is joint communication and sensing.

Kevin Chetty is an Associate Professor at University College London where he leads the Urban Wireless Sensing Lab. He has pioneered work in passive WiFi sensing; an area of radar research expected to drive advancements in ubiquitous sensing and smart environments. Dr. Chetty has developed patented techniques for high-throughput data processing in passive wireless systems to facilitate real-time operation, and demonstrated the first through-the-wall detections using the technology. He has over 100 conference and journal publications in the application of radar systems and signal processing techniques for situational awareness and human behaviour classification using micro-Doppler signatures, machine learning and software-defined sensors.