APSIPA Transactions on Signal and Information Processing, 2024, 13, e300

Original Paper

CRNet: Robust Millimeter Wave Gait Recognition Method Based on Contrastive Learning

Zhen Meng[1*], Anfu Zhou[1], Huadong Ma[1] and Qian Zhang[2]

[1] *Beijing University of Posts and Telecommunications, Beijing, China*
[2] *Hong Kong University of Science and Technology, Hong Kong, China*

ABSTRACT

Gait recognition based on millimeter waves (mmWave) has a wide range of applications, such as smart homes and health monitoring, and attracted extensive attention due to its non-contact, privacy protection, and light-independent characteristics. Existing gait recognition based on mmWave performs well on the datasets collected at a single time or environment. However, the recognition accuracy declines significantly with the time and environment domains shifting, which affects its practical applications. In this paper, we propose a novel mmWave gait recognition CRNet that is robust to both time and environment, which is realized through two-stage training. Specifically, the first stage designs a contrastive learning strategy to pre-train the encoder module, which aims at learning the general gait features across different seen time and environment domains. The second stage further trains the classification module based on specific recognition tasks. After the two-stage training, CRNet experiments on test sets with unseen time or environment domains. We collect a mmWave-based multi-person gait recognition dataset with multiple time and environment domains. Experiments show that CRNet still performs well in unfa-

*Corresponding author: Huadong Ma, mhd@bupt.edu.cn.

Received 30 September 2023; Revised 14 January 2024
ISSN 2161-1823; DOI 10.1561/116.00000067

miliar domains which increases the accuracy from 75.7% to 91.2% compared to the baseline.

Keywords: Gait Recognition, Millimeter Wave Sensing, and Contrastive Learning.

1 Introduction

Gait recognition refers to the recognition of a person's identity through their walking style. Compared to recognition technologies such as facial, iris, and fingerprint recognition, it has advantages such as long-distance, non-contact, and difficulty in disguise [18]. It has enormous research value and broad application prospects. Traditional visual-based gait recognition [16, 2] is line-of-sight and sensitive to light conditions. What's more, the video images obtained by the camera take the risk of privacy leakage.

mmWave has the characteristics of non-line-of-sight, privacy protection, and light-independent. Therefore, mmWave attracts more and more attention in smart homes [14, 13], health monitoring [15, 12, 29], and other fields. Gait recognition based on mmWave has been widely studied [17, 27, 10, 27, 28, 24], which greatly improves the accuracy of gait recognition and explores different scenarios in daily life, such as recognition under the coexistence of multiple people. However, existing mmWave-based recognition algorithms mainly focus on datasets with a single time or environment domain, neglecting the robustness of algorithms when the time and environment domain shift, which is crucial in practical applications.

In practical applications, we discover that the accuracy of existing mmWave-based gait recognition performance dropped from the original 96% to 39% with time and environment domains shifting, which will be detailed analyzed in Section 3. This is because the gait features learned have strong background dependence. There are two factors that bring the background noise. The first is that, over time, people's clothing, mood, and physical condition affect the gait itself. The second is that different environments can bring different noise impacts due to multipath effects [6, 4]. Therefore, we need to design a robust gait recognition algorithm that is insensitive to the time or environment domain shifting.

In this paper, we first collect a mmWave gait dataset from multiple people which is collected at multiple times and environments. Especially, the time span is greater than one year. What's more, we propose a novel mmWave gait recognition method CRNet which is robust to time and environment. The key idea of CRNet is to remove background noisy features of gait samples and obtain robust gait feature expression. To achieve this purpose, CRNet

is trained through two stages. Specifically, the first stage ignores specific recognition tasks and mainly focuses on pre-training a robust feature encoder module through contrastive learning strategies. Contrastive learning shortens the distance between gait features with the same labels in different seen time and environment domains, which reduces the background noise. After the pre-trained process, the second stage trains the classification module based on specific recognition tasks, which enables CRNet to perform the recognition tasks. We test the robust gait recognition performance of CRNet on test sets with unseen time or environment domains. Experimental results show that our method is more robust when time or environment domains shift. Specifically, CRNet improves the accuracy above 15% than the baselines.

The contributions of this paper are as follows:

- We propose a robust gait recognition method CRNet based on two-stage training strategies, which aims at utilizing contrastive learning to obtain robust gait feature expression with time and environment domains shifting.
- We collect a mmWave-based multi-person gait recognition dataset with multiple time and environment domains. Experimental results show that our method CRNet still performs well in the test sets that are collected in unseen environments and time. What's more, CRNet increases the accuracy from 75.7% to 91.2% compared to the baseline.

The rest of the paper is organized as follows: Related work is reviewed in Section 2. Motivation is reviewed in Section 3. The CRNet method is proposed in Section 4. Experimental results are shown in Section 5. Finally, the conclusion and future work are given in Section 6.

2 Related Work

2.1 Gait Recognition

Gait recognition has a wide range of applications in security checks, health monitoring, and new types of human-computer interaction. People try to solve this problem with many different methods, such as vision-based methods or wireless-based methods. The traditional vision-based gait recognition methods [2, 11, 16] perform very well, but there are several limitations. First, cameras capture real-life images, which can lead to personal information being leaked once the camera is hacked or hijacked. Second, the camera is easily affected by lighting conditions. For example, they cannot obtain effective images in dark environments.

To address the above issues, researchers attempt to use wireless signals to capture human gait data. Among these wireless sensing works, WiFi-based

work [8] holds a place. Some WiFi-based methods use WiFi channel state information (CSI) for personnel perception [26, 31, 25, 34, 31]. However, WiFi signals are difficult to segment to isolate the impact between people, so they struggle to identify multiple people at the same time. Another part of WiFi-based work uses frequency-modulated continuous wave (FMCW) in the WiFi frequency for sensing [9, 5, 32]. Due to the bandwidth limitations of the WiFi frequency, the spatial resolution of the perceived signal is limited.

mmWave has a high band, which provides high spatial resolution. There are two main mmWave data formats to extract gait features. One is the point cloud [17, 10, 33]. For example, literature [17] proposes a mmWave gait point cloud dataset and uses ResNet18 [7] to extract temporal and spatial characteristics of gait for multi-people. Another is the gait spectrogram [19, 27]. For example, literature [27] captures different gait patterns of the lower limbs in terms of step length, duration, instantaneous lower limb velocity, and distance between lower limbs through mmWave in the Range-Doppler domain, and then uses a convolutional neural network (CNN) to perform the classification. The existing mmWave-based gait recognition methods explore the gait characteristics in different situations such as single person and multiple persons on the collected single time or environment dataset. However, researchers ignore the issue of robustness of gait recognition over time and environment migration, which is crucial in real life.

2.2 Contrastive Learning

To improve the robustness of gait recognition, one promising direction is to learn the general gait features over time and environment migration, which means eliminating the background noise when extracting the gait features. Contrastive learning [3] focuses on learning the general features between samples of the same category under a self-supervised fashion, and removing the task-unrelated noise features. Specifically, it learns a feature encoder that makes the data features similar to the same category, and the data features different with the different category. Thus contrastive learning becomes a desired solution for extracting the general gait features. The literature [30] adopts contrastive learning for learning robust unsupervised representations of graph data. The literature [1] enforces consistency between cluster assignments produced for different augmentations (or views) of the same image, which aims to enhance the image feature representations. The literature [20] encourages two elements (corresponding patches) to map to a similar point in a learned feature space. The literature [22] integrates mature RF signal processing techniques with unsupervised representation learning frameworks by contrastive learning in which different signal representations construct positive and negative pairs. Although contrastive learning is very effective in extracting general features, in the millimeter wave scenario, there are still challenges in applying contrastive

learning to extracting general mmWave-based gait features over time and environment migration.

3 Motivation

The gait recognition based on mmWave has received widespread attention and achieved good results on the datasets collected at a single time or environment. In the practical application, we discover a very serious problem: the neural network mmGaitNet [17] trained on a single dataset performs well in a short period, but its recognition performance severely decreases over time and environment. The performance of mmGaitNet on different test sets is shown in Figure 1. The accuracy of the test sets of "Similar-Time", "Unseen-Time" and "Unseen-Environment" are 96%, 39%, 50.5%. The environment of the test sets "Similar-Time" and "Unseen-Time" is consistent with the train set, while "Unseen-Environment" is different. The collection time of "Similar-Time" is similar to the train set. The collection time of "Unseen-Time" is one month away from the train set. The collection time of "Unseen-Environment" is one year away from the train set. The neural network mmGaitNet [17] fails in the test sets "Unseen-Time" and "Unseen-Environment" even though it performs well on the test set "Similar-Time" which is similar to the train set on the time and environment.

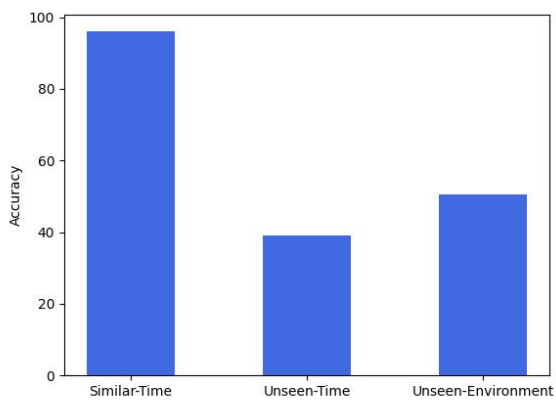


Figure 1: The performance of mmGaitNet which is trained by a single dataset. The environment of the test sets "Similar-Time" and "Unseen-Time" is consistent with the train set, while "Unseen-Environment" is different. The collection time of "Similar-Time", "Unseen-Time", and "Unseen-Environment" is about two days, a month, and a year away from the train set.

This is because the existing expression features learned under a single dataset not only include human gait features but also background noise. Two factors form the background noise. The first is that, over time, people's clothing, mood, and physical condition affect the gait itself. The second is that, in different environments, there are different environmental noise caused by multipath effects. In this paper, we design a gait recognition algorithm based on contrastive learning to remove background noise to obtain robust feature representation.

4 Method

4.1 Task Definition

We represent each gait sample as o_i. o_i is a $t \times n \times 5$ matrix, where t denotes time frames, and n denotes the number of points in a point cloud. 5 means each point cloud contains five attributes $\{X, Y, Z, V, S\}$, where X, Y, Z denote the spatial location, V denotes the redial speed and S denotes the signal strength of the points. The goal of this task is to construct a gait recognition model to predict the correct person with the sample o_i, i.e., $P(y_i|o_i)$, where y_i is the golden person label.

The main reason for the non-robustness of gait recognition is caused by background noise over time and environment migration. Therefore, we perform contrastive learning to reduce the noise impact and thus can improve the robustness of gait recognition. The framework of the proposed CRNet is shown in Figure 2. Specifically, CRNet is trained through two stages:

- The first stage mainly focuses on pre-training a robust feature encoder module through contrastive learning strategies. We create augmented gait samples according to different seen time and environment domains. Then input the gait sample and the augmented gait sample into the feature encoder module and the linear layer to extract the gait features respectively. Later we use contrastive learning to narrow the distance between the gait sample and the augmented gait sample under the same person label.

- The second stage mainly focuses on the specific recognition tasks. Based on the robust feature encoder module trained in stage one, we froze the parameters of the feature encoder module and used the labeled dataset to train the parameters of the classification module, which is composed of a fully connected layer. So the classification module learns recognition knowledge based on the supervised signals and retains it in the parameters.

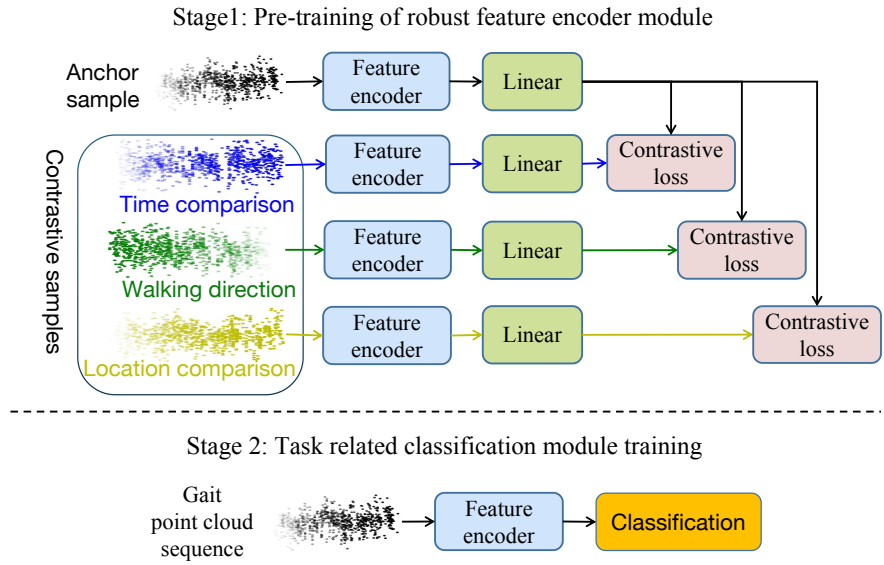


Figure 2: The framework of the proposed CRNet.

4.2 Robust Gait Recognition Framework

After the training, the test of the gait recognition model only needs to input the gait sample into the feature encoder and classification modules to obtain the person identification results.

4.3 Pre-trained Robust Feature Encoder

4.3.1 The Contrastive Augmented Gait Sample for Robust Training

We collect the augmented gait sample according to the gait sample's three factors: time, location, and walking direction. The augmented gait sample generated with the time factor can remove the time noise. Similarly, the location and direction factors can remove the environmental noise.

For each input sample o_i, we obtain the x-axis spatial location l_i, time t_i, and human walking direction d_i, respectively. Specifically, For the sample $o_i \in \mathbb{R}^t \times \mathbb{R}^p \times \mathbb{R}^5$, we extract the matrix $m_x \in \mathbb{R}^t \times \mathbb{R}^p$ corresponding to attribute X and then average matrix m_x to calculate the location l_i. When collecting the gait sample, we record the data collection batch and date and treat it as the corresponding time t_i. For the direction information, because each input has point clouds with sequence t time, we calculate the x-axis spatial location for the 1-th time and t-th time, corresponding to the 1-th

matrix $c_1 \in \mathbb{R}^p$ and t-th matrix $c_t \in \mathbb{R}^p$ of the matrix m_x, respectively. Then we subtract the two x-axis spatial location values to determine the direction $d_i \in \{0, 1\}$.

For each input sample o_i, we construct three positive samples from the above three factors, respectively. For the location factor, we randomly select a sample that meets the following description and treat it as the positive sample o_j, where $l_i \neq l_j, t_i = t_j$, and $d_i = d_j$. Similarly, for the time factor, we randomly select a positive sample o_z, where $l_i = l_z, t_i \neq t_z$, and $d_i = d_z$. For the direction factor, we randomly select a positive sample o_k, where $l_i = l_k, t_i = t_k$, and $d_i \neq d_k$.

4.3.2 Robust Feature Encoder Network Structure

We follow [17] and utilize the residual block to perform the gait recognition task. The structure is shown in Figure 3. Specifically, we divide the input sample or the augmented gait sample $o_i \in \mathbb{R}^t \times \mathbb{R}^p \times \mathbb{R}^5$ into five attributes matrices $m_j \in \mathbb{R}^t \times \mathbb{R}^p$, where $\forall j \in \{X, Y, Z, V, S\}$. X, Y, Z denote the spatial location, V denotes the redial speed and S denotes the signal strength of the points. Due to the significant difference in the values of the five attributes, we first utilize the batch normalization algorithm to normalize the values of the five attributes separately. For the feature encoder module, firstly, the combination of the convolution layer and residual layer is used to encode the five attribute matrices, respectively. Secondly, the feature fusion layer is utilized to fusion the five attribute features to obtain the overall feature of the input gait sample.The above process can be summarized as follows:

$$\begin{aligned} o_i &= \{m_X, m_Y, m_Z, m_V, m_S\}, \\ a_i &= \text{residual}(\text{convolution}(m_i)), \forall i \in \{X, Y, Z, V, S\} \\ f_i &= \text{fusion}(a_X, a_Y, a_Z, a_V, a_S) \end{aligned} \tag{1}$$

4.3.3 Robust Contrastive Learning

After introducing the feature encoder structure, we utilize the contrastive learning strategy to perform task-unrelated pre-training on the feature encoder, enabling it to extract general gait features and remove the background noise. For the input sample o_i and the corresponding feature f_i, we can obtain the augmented gait samples $\{o_j, o_z, o_k\}$ and their corresponding features $\{f_j, f_z, f_k\}$. We transform these samples' features through a linear layer, and then utilize contrastive learning to shorten the feature distances of the input sample and each augmented gait sample, respectively. The above process can be summarized as follows:

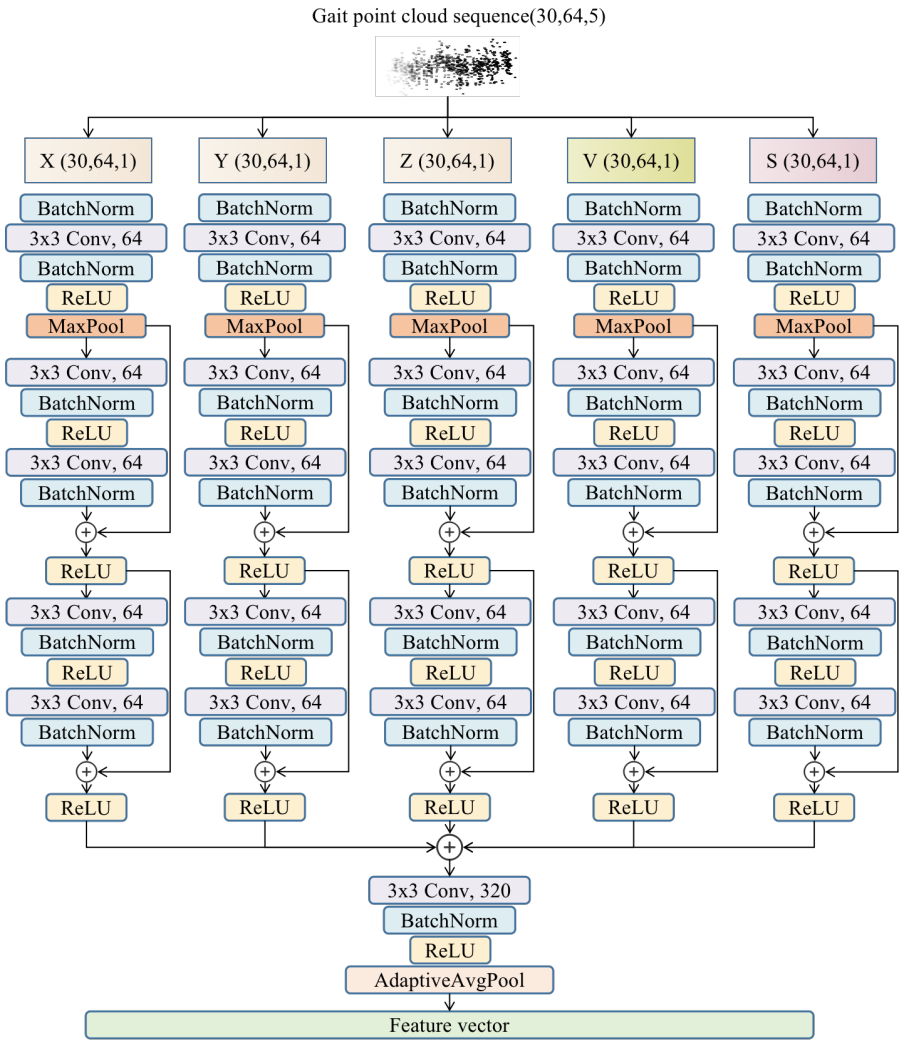


Figure 3: Feature encoder network structure of CRNet.

$$\begin{aligned}
\hat{f}_i =& \text{linear}(f_i), \hat{f}_j = \text{linear}(f_j), \hat{f}_z = \text{linear}(f_z), \hat{f}_k = \text{linear}(f_k), \\
\mathcal{L}^{Con}_{(i,j)}(\theta_1) &= -\log \frac{\exp(\hat{f}_i, \hat{f}_j)}{\sum_{e=1}^{4N} \exp(\hat{f}_i, \hat{f}_e)}, \\
\mathcal{L}^{Con}_{i}(\theta_1) &= \mathcal{L}^{Con}_{(i,j)}(\theta_1) + \mathcal{L}^{Con}_{(i,z)}(\theta_1) + \mathcal{L}^{Con}_{(i,k)}(\theta_1),
\end{aligned} \tag{2}$$

where N is the batch size. $4N$ means the combination of N input gait samples and $3N$ augmented gait samples. θ_1 means the trainable parameters of the feature encoder module and the linear layer.

4.4 The Classification Module

In the training classification module, we freeze the parameters of the feature encoder module and use the labeled dataset to train the parameters of the classification module, which is composed of a fully connected layer. So the classification module learns recognition knowledge based on the supervised signals and retains it in the parameters. Specifically, given the input sample o_i, we first obtain its gait features through the pre-trained robust feature encoder, which can be represented as f_i. Then the classification module outputs the gait recognition probability distribution. Later we utilize the label information to train the classification parameters. The loss function of the above process can be summarized as follows:

$$\begin{aligned} \mathcal{L}_i^{Class}(\theta_2) &= -\log P(y_i|o_i), \\ &= -\log\big(\frac{\exp(score_{y_i})}{\sum_{j=1}\exp(score_j)}\big), \end{aligned} \tag{3}$$

where θ_2 means the trainable parameters of the classification module. y_i is considered as the golden label here. $score_{y_i}$ means the probability score of that sample o_i belongs to y_i.

4.5 Loss Function

After the two-stage training strategies, we obtain a robust gait recognition model CRNet, which is insensitive to the time or environment domains shifting. Specifically, the feature encoder module focuses on learning the general gait features across different time and environment domains, and the classification module focuses on learning the task-related features based on the supervised signals. The total loss for the proposed CRNet is shown as follows:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum(\mathcal{L}_i^{Con}(\theta_1) + \alpha\mathcal{L}_i^{Class}(\theta_2)) + \lambda||\theta||_2^2, \tag{4}$$

where θ means the trainable parameters of CRNet, which is the combination of θ_1 and θ_2. α and λ are hyperparameters.

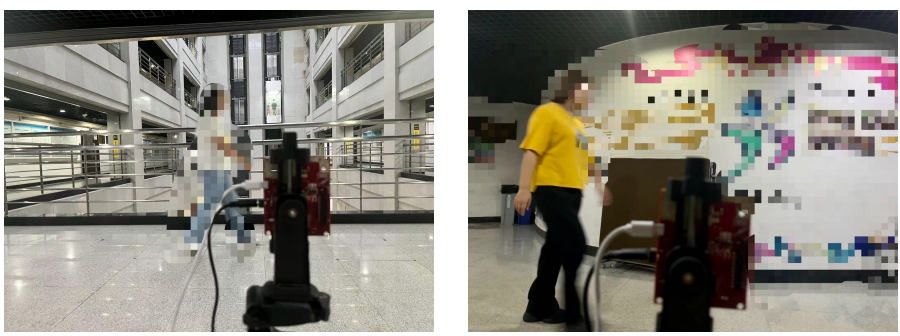
5 Experiment

5.1 Dataset

We first collect a mmWave gait dataset from multiple people which contains two environments and the period for data collection is greater than one year. The data structure of the dataset is shown in Table 1. The two environments for data collection are shown in Figure 4. We use a commercial mmWave device TI IWR1443 to sense pedestrian gait data.

Table 1: The structure of the dataset.

name	environment	time span
Training set	scene 1	/
Test-1	scene 2	16 months
Test-2	scene 2	18 months
Test-3	scene 1	15 days



(a) Scene 1. (b) Scene 2.

Figure 4: Environment for data collection

Training set: The training set is collected in scene 1 of the teaching building. The period for training set collection is approximately 12 days, with multiple collections. When collecting the gait sample using a mmWave device, we record the data collection batch and date. Gait samples collected in the same batch and on the same date are considered to be at the same time.

Test set: The collect environment of "Test-1" and "Test-2" is scene 2 of the teaching building. The collect environment of "Test-3" is consistent with the training set. The collection time of "Test-3" is similar to the training set. The collection time of "Test-1" is 16 months away from the training set. The

collection time of "Test-2" is 18 months away from the training set. When collecting data sets, to better reproduce the gait in real scenes, volunteers are asked to imagine that they are in a situation where they want to pick up something on the opposite side. Instead of mechanically passing through the board for data collection.

Volunteer: Since the period of the dataset is relatively long, we have initially explored the situation of 3 volunteers, there are two male and one female volunteer. However, the three volunteers were very similar in body shape, which brings challenges to gait recognition. The age of the volunteers is between 25 and 28. The height of the volunteers is between 158cm and 170 cm. The weight of the volunteers is nearly 120kg.

5.2 Implementation Details

Hardware: TI IWR1443 has 3 transmit antennas and 4 receive antennas, the antennas are all used when collecting data. The elevation angle of the antenna is 30 degrees. The horizontal angle is 120 degrees. The device transmits a signal with Frequency modulated continuous wave (FMCW). The specific parameters of the signal are shown in Table 2. Finally, the device directly transmits the perceived mmWave point cloud data of a person to a computer through a data cable.

Table 2: Millimeter wave radar parameters.

Chirp Parameter (Units)	Value
Start Frequency(GHz)	77
Frequency Slope (MHz/us)	70
Bandwidth (GHz)	4
Idle Time(us)	81
Ramp End Time(us)	57.14
Samples Per Chirp	224
ADC Sampling Frequency (Ksps)	4558
Chirps Per Frame	16*3
Frame Duration (ms)	100
Range Resolution (m)	0.044
Max Unambiguous Range (m)	8
Max Radial Velocity (m/s)	2.35
Radial Velocity Resolution (m/s)	0.3
Azimuth Resolution (deg)	15
Elevation Resolution (deg)	58

Data Format: We use the sliding window method to intercept gait sample data on the walking path. The step size of the sliding window is 1, the length of the intercepted gait sequence is 30 frames, and the points of each frame are copied to 64. There are 5 attributes of each point, which are 3D coordinates, radial velocity, and confidence. The gait sample data is deleted once the number of points in a certain frame of a gait sample is less than 15. This is because the point cloud of millimeter wave sensing is mainly concentrated on the torso, and a point cloud that is too sparse cannot describe gait information.
Training Setting: First of all, for the training data, we normalize each attribute of the point cloud separately in the encoder. That is because the five attributes of a point cloud contain three physical meanings which are spatial coordinates, speed, and confidence. Second, for the network structure, we use batch-normalize to the features after the convolution operation and use ReLU as the activation function. Finally, for the training process, we implemented our network in PyTorch, and the optimizer is SGD [21]. We pre-train the feature encoder module with an initial learning rate of 0.5. The learning rate has decreased to one-tenth of the original values for the 40th, 60th, and 80th epochs respectively. We select the result of the 100th iteration as the parameters of the pre-trained feature encoder. The batch size used in both stages of network training is 128. The number of training samples for each category is 1500.

5.3 Experiment Result

5.3.1 The Baseline Performances

We regard the model mmGaitNet [17] as the baseline. We train our model and the baseline on the collected training set and test on the three test sets which are collected at different times and environments.

Specifically, the effect of mmGaitNet trained on this training set is better than the result mentioned in Section 3. The accuracy of mmGaitNet trained on this training set is 64%. However, the accuracy of mmGaitNet mentioned in Section 3 is 39%. That is because the training set is collected multiple times over 12 days. However, mmGaitNet mentioned in Section 3 is trained by a dataset that is collected one time as usual. This result shows that the neural network trained by multiple collected data can essentially help the neural network remove background noise and obtain more stable and robust feature expressions.

5.3.2 Robust Contrastive Learning

CRNet pre-trains feature encoder by contrastive learning strategies in the first stage, and the second stage trains the classification module based on specific

recognition tasks. We demonstrate the effectiveness of CRNet by comparing it with the baseline. The results in Table 3 illustrate two aspects. Firstly, our method has a large degree of improvement compared to the baseline in all indicators. Specifically, the average accuracy of our method CRNet in "Test-1", "Test-2" and "Test-3" is 92.5%, 86%, 94% respectively. However, the average accuracy of mmGaitNet is only 78.5%, 64%, 84%. Our method CRNet has above a 15% improvement in accuracy on the three test sets, which demonstrates the effectiveness of our proposed model CRNet.

Table 3: Experimental results on three datasets.

Test sets	Methods	precision	recall	f1-score	accuracy
Test-1	mmGaitNet	81.5%	78%	71.5%	78.5%
	CRNet	**96%**	**93%**	**94%**	**93.5%**
Test-2	mmGaitNet	68%	59%	57%	64%
	CRNet	**89%**	**83%**	**85%**	**86%**
Test-3	mmGaitNet	85%	86.5%	84%	84.5%
	CRNet	**93.5%**	**94%**	**93.5%**	**94%**
Mean	mmGaitNet	78.2%	74.5%	70.8%	75.7%
	CRNet	**92.8%**	**90%**	**90.8%**	**91.2%**

Secondly, our method still performs well even when time and environment domains shift. Specifically, "Test-2" is quite different from the training set in the environment and time domain. The baseline method has an accuracy of only 64% on "Test-2" which is almost unusable. However, our method performs well achieving an accuracy of 86%. CRNet pre-trains the feature encoder to extract general gait features. Experimental results prove the robustness of our proposed model in removing background noise through contrastive learning strategies.

5.4 Ablation Experiment

To fine-grained analysis of the effectiveness of the proposed model CRNet, we perform the ablation experiments, mainly analyzing the impact of batch-normal, three factors for augmented gait samples, such as walking direction, position, time, etc. The experimental results are shown in the Table 4. The item "w/o BN" means not considering data batch-normal. The item "w/o time" means the augmented gait samples do not consider the time factor. Similarly, the item "w/o position" means not considering the position factor, and the item "w/o direction" means not considering the direction factor.

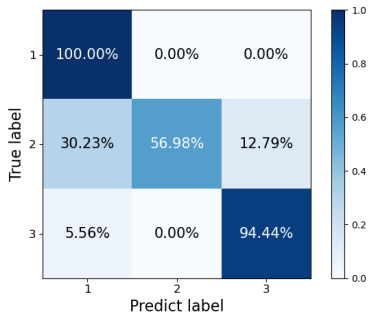
Table 4: The results of ablation experiments.

	Method	precision	recall	f1-score	accuracy
Test-1	CRNet	**96%**	93%	**94%**	93.5%
	– w/o BN	93%	**93.5%**	92.5%	**94%**
	– w/o time	87%	88.5%	85%	88%
	– w/o position	90.5%	90%	89%	90%
	– w/o direction	92.5%	88%	85.5%	87.5%
Test-2	CRNet	**89%**	**83%**	**85%**	**86%**
	– w/o BN	80%	79%	78%	78%
	– w/o time	66%	61%	59%	67%
	– w/o position	80%	68%	68%	71%
	– w/o direction	77%	61%	58%	65%
Test-3	CRNet	**93.5%**	**94%**	**93.5%**	**94%**
	– w/o BN	93.5%	94%	93.5%	94%
	– w/o time	90%	91%	90%	90.5%
	– w/o position	92.5%	93.5%	92.5%	92.5%
	– w/o direction	92%	92.5%	92%	92%

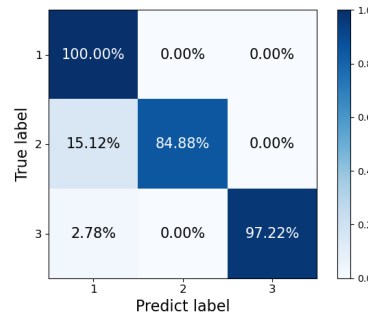
From the results, we observe that the three augmented factors and data batch-normal operation can significantly improve the performance of the three test sets under different time and environment domains. Meanwhile, we find that compared to "Test-3", adding each ablation element on "Test-1" and "Test-2" resulted in a significant improvement in CRNet performance. Specifically, after adding position factors, the f1-score on "Test-1" increased by 9%. while it only increased by 1% on "Test-3". One possible reason is that "Test-3" and the training set have similar time and the same environment domains, while "Test-1" and "Test-2" have different time and environment domains with the training set. The above analysis further proves the effectiveness of CRNet in obtaining robust gait feature expression with time and environment domains shifting.

5.5 Analysis of Fine-grained Recognition Accuracy

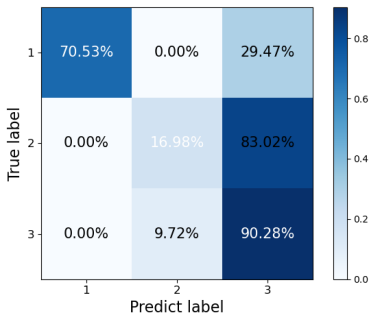
In this section, we present a fine-grained display of the accuracy changing on different labels before and after utilizing the proposed CRNet, which is shown in Figure 5. We utilize the confusion matrix to demonstrate the variation of the accuracy performance across three test sets. From the results, we can observe two interesting phenomena. Firstly, CRNet can improve the recognition accuracy of each label on the three test sets. Specifically, on the test set "Test-1", CRNet improves the accuracy by 28%, and 3% on categories



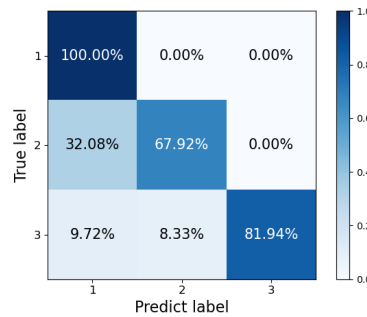
(a) The confusion matrix of "Test-1" before utilizing CRNet.



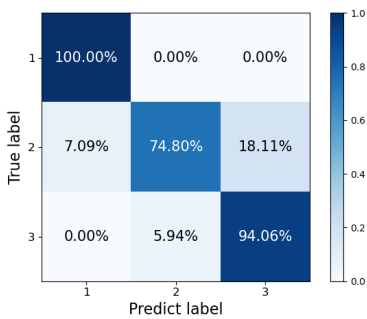
(b) The confusion matrix of "Test-1" after utilizing CRNet.



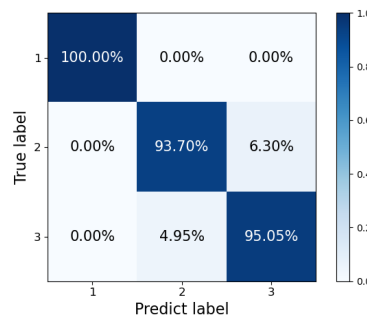
(c) The confusion matrix of "Test-2" before utilizing CRNet.



(d) The confusion matrix of "Test-2" after utilizing CRNet.



(e) The confusion matrix of "Test-3" before utilizing CRNet.



(f) The confusion matrix of "Test-3" after utilizing CRNet.

Figure 5: The display of confusion matrices related to recognition accuracy on three datasets.

2 and 3 and respectively. The above analysis further proves the generality of CRNet in removing background noise from gait features. Secondly, compared with other categories, the second category is more affected by the environment.

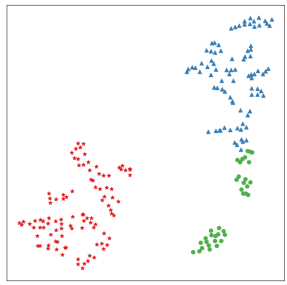
This is because compared to other categories, the second category is more casual when walking, so the variance of the second category in intuitive walking speed, stride length, and other factors is larger. Otherwise, the second category is very similar in body shape to the third category.

5.6 Gait Features Aggregation Analysis

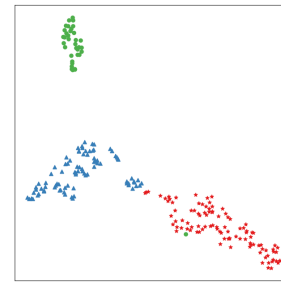
To better display the distribution of gait data under different labels in the feature space, we use the T-SNE algorithm [23] to convert non-evaluable high-order gait features into observable low-order gait features, and display them in the feature space. The results are shown in Figure 6. From the results, we observe that before applying CRNet, gait features of different labels tend to mix, especially on the "Test-2" and "Test-3" datasets. Besides this, after applying CRNet, gait features belonging to the same labels can be better aggregated among the three datasets, which verifies the ability of CRNet to extract gait features under background noise interference. In addition, after utilizing CRNet, we find that the feature distribution of "Test-1" and "Test-2" tends to be consistent, but their feature distribution is significantly different from "Test-3". Compared to "Test-1" and "Test-2", "Test-3" exhibits significant changes in both environmental and time domains, resulting in a more pronounced transfer of gait features. However, after using contrastive learning in CRNet, gait features belonging to different labels can still be well distinguished.

6 Conclusion and Future Work

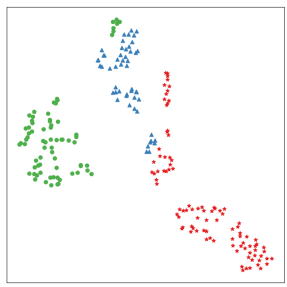
In this paper, we focus on the robust mmWave gait recognition task, based on the discovery that the recognition accuracy declines significantly with the time and environment domains shifting, which affects the practical applications of mmWave gait recognition. To address the aforementioned challenges, we propose CRNet that is robust to both time and environment, which is realized through two-stage training with contrastive learning strategies. In addition, in order to verify the effectiveness of CRNet, we collect a mmWave gait dataset from multiple people under multiple times and environments. Experiments prove CRNet performs well in test sets with unseen time or environment domains. In the future, we will increase the number of people in the dataset to verify the effectiveness of our proposed model on a larger dataset with more people's gait samples.



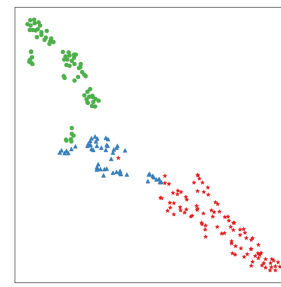
(a) The sample features distribution before utilizing CRNet on "Test-1".



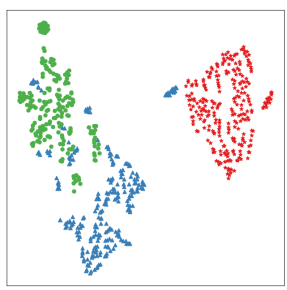
(b) The sample features distribution after utilizing CRNet on "Test-1".



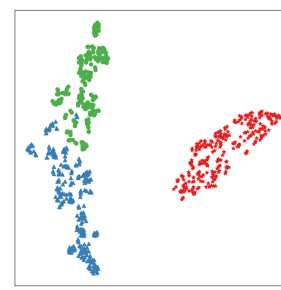
(c) The sample features distribution before utilizing CRNet on "Test-2".



(d) The sample features distribution after utilizing CRNet on "Test-2".



(e) The sample features distribution before utilizing CRNet on "Test-3".



(f) The sample features distribution after utilizing CRNet on "Test-3".

Figure 6: The distribution of gait data in feature space on three datasets.

7 Acknowledgments

This work was supported in part by Beijing Natural Science Foundation (L223002), in part by the Innovation Research Group Project of NSFC under Grant 61921003, in part by the Youth Top Talent Support Program, in part by NSFC Project under Grant 62302058.

References

[1] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[2] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition", in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, No. 01, 2019, 8126–33.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations", in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Vol. 119, *Proceedings of Machine Learning Research*, PMLR, 2020, 1597–607.

[4] Y. Chen, H. Deng, D. Zhang, and Y. Hu, "Speednet: Indoor speed estimation with radio signals", *IEEE Internet of Things Journal*, 8(4), 2762–74.

[5] L. Fan, T. Li, R. Fang, R. Hristov, Y. Yuan, and D. Katabi, "Learning longterm representations for person re-identification using radio signals", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 10699–709.

[6] Z. Hao, H. Yan, X. Dang, Z. Ma, P. Jin, and W. Ke, "Millimeter-Wave Radar Localization Using Indoor Multipath Effect", *Sensors*, 22(15), 5671.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2016, 770–8.

[8] Y. He, Y. Chen, Y. Hu, and B. Zeng, "WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi", *IEEE Internet of Things Journal*, 7(9), 8296–317.

[9] C.-Y. Hsu, R. Hristov, G.-H. Lee, M. Zhao, and D. Katabi, "Enabling identification and behavioral sensing in homes using radio reflections", in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, 1–13.

[10] Y. Huang, Y. Wang, K. Shi, C. Gu, Y. Fu, C. Zhuo, and Z. Shi, "HDNet: Hierarchical Dynamic Network for Gait Recognition using Millimeter-wave radar", in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, 1–5.

[11] S. Li, W. Liu, and H. Ma, "Attentive spatial–temporal summary networks for feature learning in irregular gait recognition", *IEEE Transactions on Multimedia*, 21(9), 2361–75.

[12] K. Liang, A. Zhou, Z. Zhang, H. Zhou, H. Ma, and C. Wu, "mmStress: Distilling Human Stress from Daily Activities via Contact-less Millimeter-wave Sensing", *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(3), 110:1–110:36, DOI: 10.1145/3610926, https://doi.org/10.1145/3610926.

[13] H. Liu, K. Cui, K. Hu, Y. Wang, A. Zhou, L. Liu, and H. Ma, "mTransSee: Enabling Environment-Independent mmWave Sensing Based Gesture Recognition via Transfer Learning", *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(1), 23:1–23:28, DOI: 10.1145/3517231, https://doi.org/10.1145/3517231.

[14] H. LIU, X. LIU, X. XIE, X. TONG, and K. LI, "PmTrack: Enabling Personalized mmWave-based Human Tracking".

[15] X. Liu, W. Jiang, S. Chen, X. Xie, H. Liu, Q. Cai, X. Tong, T. Shi, and W. Qu, "PosMonitor: Fine-Grained Sleep Posture Recognition With mmWave Radar", *IEEE Internet of Things Journal*.

[16] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, "Joint intensity and spatial metric learning for robust gait recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 5705–15.

[17] Z. Meng, S. Fu, J. Yan, H. Liang, A. Zhou, S. Zhu, H. Ma, J. Liu, and N. Yang, "Gait recognition for co-existing multiple people using millimeter wave sensing", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 01, 2020, 849–56.

[18] B. B. Mjaaland, P. Bours, and D. Gligoroski, "Walk the walk: Attacking gait biometrics by imitation", in *Information Security: 13th International Conference, ISC 2010, Boca Raton, FL, USA, October 25-28, 2010, Revised Selected Papers 13*, Springer, 2011, 361–80.

[19] M. Z. Ozturk, C. Wu, B. Wang, and K. R. Liu, "GaitCube: Deep data cube learning for human recognition with millimeter-wave radio", *IEEE Internet of Things Journal*, 9(1), 546–57.

[20] T. Park, A. A. Efros, R. Zhang, and J. Zhu, "Contrastive Learning for Unpaired Image-to-Image Translation", in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, Vol. 12354, *Lecture Notes in Computer Science*, Springer, 2020, 319–45.

[21] S. Ruder, "An overview of gradient descent optimization algorithms", *CoRR*, abs/1609.04747.

[22] R. Song, D. Zhang, Z. Wu, C. Yu, C. Xie, S. Yang, Y. Hu, and Y. Chen, "Rf-url: unsupervised representation learning for rf sensing", in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, 282–95.

[23] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.", *Journal of machine learning research*, 9(11).

[24] C. Wang, P. Gong, and L. Zhang, "Stpointgcn: Spatial temporal graph convolutional network for multiple people recognition using millimeter-wave radar", in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 3433–7.

[25] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using wifi signals", in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, 363–73.

[26] Q. Xu, Y. Chen, B. Wang, and K. R. Liu, "Radio biometrics: Human recognition through a wall", *IEEE Transactions on Information Forensics and Security*, 12(5), 1141–55.

[27] X. Yang, J. Liu, Y. Chen, X. Guo, and Y. Xie, "MU-ID: Multi-user identification through gaits using millimeter wave radios", in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, IEEE, 2020, 2589–98.

[28] Y. Yao, H. Zhang, P. Xia, C. Liu, F. Geng, Z. Bai, L. Du, X. Chen, P. Wang, B. Han, *et al.*, "mmSignature: Semi-supervised human identification system based on millimeter wave radar", *Engineering Applications of Artificial Intelligence*, 126, 106939.

[29] H. Yin, S. Yu, Y. Zhang, A. Zhou, X. Wang, L. Liu, H. Ma, J. Liu, and N. Yang, "Let IoT Know You Better: User Identification and Emotion Recognition Through Millimeter-Wave Sensing", *IEEE Internet Things J.*, 10(2), 1149–61, DOI: 10.1109/JIOT.2022.3204779, https://doi.org/10.1109/JIOT.2022.3204779.

[30] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations", *Advances in neural information processing systems*, 33, 5812–23.

[31] Y. Zeng, P. H. Pathak, and P. Mohapatra, "WiWho: WiFi-based person identification in smart spaces", in *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, IEEE, 2016, 1–12.

[32] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 7356–65.

[33] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mid: Tracking and identifying people with millimeter wave radar", in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, IEEE, 2019, 33–40.

[34] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. Spanos, "Wifi-based human identification via convex tensor shapelet learning", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.