# See Your Emotion from Gait Using Unlabeled Skeleton Data

**Haifeng Lu**[1], **Xiping Hu**[1,2*], **Bin Hu**[1,2*]

[1] School of Information Science and Engineering, Lanzhou University
[2] School of Medical Technology, Beijing Institute of Technology
luhf18@lzu.edu.cn, huxp@bit.edu.cn, bh@bit.edu.cn

## Abstract

This paper focuses on contrastive learning for gait-based emotion recognition. The existing contrastive learning approaches are rarely suitable for learning skeleton-based gait representations, which suffer from limited gait diversity and inconsistent semantics. In this paper, we propose a Cross-coordinate contrastive learning framework utilizing Ambiguity samples for self-supervised Gait-based Emotion representation (CAGE). First, we propose ambiguity transform to push positive samples into ambiguous semantic space. By learning similarities between ambiguity samples and positive samples, our model can learn higher-level semantics of the gait sequences and maintain semantic diversity. Second, to encourage learning the semantic invariance, we uniquely propose cross-coordinate contrastive learning between the Cartesian coordinate and the Spherical coordinate, which brings rich supervisory signals to learn the intrinsic semantic consistency information. Exhaustive experiments show that CAGE improves existing self-supervised methods by 5%–10% accuracy, and it achieves comparable or even superior performance to supervised methods.

## Introduction

Human emotion recognition is a vital task in the field of affective computing. Most prior works detect human emotion through physiological signals such as electroencephalogram (EEG), electrocardiography (ECG) (Jia et al. 2021), and non-physiological signals such as facial expressions, tone of voice, walking styles, *etc.* (Ringeval et al. 2019; Sun, Su, and Fan 2022). However, fast, effective, and reliable emotion recognition is a challenge, especially when self-reported emotions are unreliable or people are disturbed by environmental factors (Fernández-Dols and Ruiz-Belda 1995; Quigley, Lindquist, and Barrett 2014).

Since gait can be observed at a distance and without cooperative subjects during acquisition, it is an emerging research topic in these years (Lu et al. 2022). On the one hand, existing physiological and psychological studies point out that there is internal link between gait and emotions (Kleinsmith and Bianchi-Berthouze 2012; Deligianni, Guo, and Yang 2019). On the other hand, due to the development of human pose estimation algorithms and depth sensors, 3D skeleton has gradually become a significant feature representation to study gait (Cao et al. 2017; Zhang 2012). Therefore, we focus on using skeleton-based gait for emotion perception.

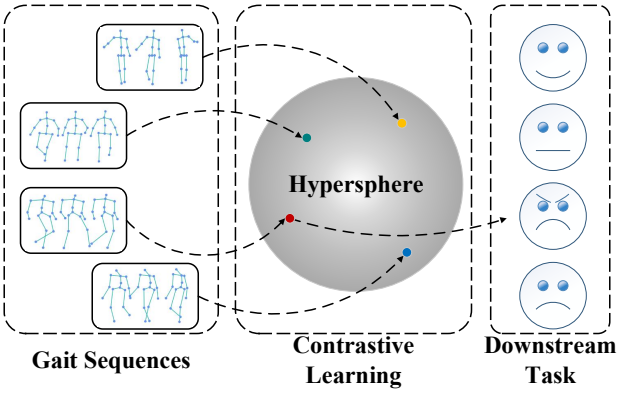



Figure 1: We propose a self-supervised emotion recognition framework that named CAGE, which maps the gait sequences into features residing on a hypersphere space for different downstream tasks.

Most existing skeleton-based methods rely on handcrafted features (Li et al. 2016; Crenn et al. 2016; Daoudi et al. 2017). Moreover, deep learning approaches like STEP (Bhattacharya et al. 2020a), ProxEmo (Narayanan et al. 2020), TEW (Bhattacharya et al. 2020b), and TNTC (Hu et al. 2022), are also important method for learning gait representations. However, all of them rely on supervised paradigms, in which a significant number of annotations for gait sequences are indispensable. In particular, annotating dataset requires tremendous human workforce, which is expensive and time-consuming. Furthermore, high inter-class similarity between emotions usually leads to an inconsistent annotating or even mislabeling (Bhattacharya et al. 2020a; Wang et al. 2006). Under this circumstance, how to devise an effective method, which can learn gait representations from unlabeled skeleton data, is a significant task.

**Motivation.** Recently, self-supervised learning (SSL) has been proved effective in the field of computer vision. However, the existing contrastive learning methods usually rely

*Corresponding authors

on data augmentation, which could distort data' structures, causing the dimensional collapse in contrastive learning (Wang and Qi 2022; Li et al. 2022). For gait sequences, the semantics of skeleton-based gait sequence is more likely to change after data augmentation. In other word, data augmentation provides the diverse semantic information while bringing the problem of semantic inconsistency. Therefore, balancing the semantic consistency and diversity is a critical issue for learning gait representations.

To this end, we propose a **C**ross-coordinate contrastive learning framework utilizing **A**mbiguity samples for self-supervised **G**ait-based **E**motion representation (CAGE) that can map gait sequences into on a hypersphere for downstream task, as shown in Figure 1. First, we propose a simple Contrastive Learning framework using skeleton-based Gait for self-supervised Emotion Representation (GE-CLR), which has a similar structure to MoCo (He et al. 2020). Second, by changing speed and joint angle, positive examples are pushed into ambiguous semantic space, which is called ambiguity samples. Then, we propose Ambiguity Contrastive Learning (ACL), which adds ambiguity samples into the negative set and can achieve the same effectiveness as traditional contrast learning without using negative samples. Furthermore, we transform the original data to the Spherical coordinate, and propose Cross-Coordinate Contrastive Learning ($C^3L$), which refers to the samples in the Cartesian coordinate to guide the learning process in the Spherical coordinate for pursuing cross-coordinate consistency information. To the best of our knowledge, CAGE is the first to explore SSL for gait-based emotion recognition. Experimental evaluations on public dataset demonstrate CAGE achieves comparable or even superior performance to supervised methods.

In summary, our main contributions include:

- We explore the effectiveness of ambiguity samples and introduce ACL, which bring diverse semantics to improve the generalizability of the gait representations.
- We propose $C^3L$, which leverages different coordinate-representations of the gait sequences to pursue the semantic invariance.
- We evaluate our model on public datasets, and achieve state-of-the-art results under the different evaluation protocol.

## Related Work

### Emotion Recognition Using Skeleton-based Gait

Some early work identified emotion from gait by using the geometric, morphological, or anthropometric attributes of skeleton data. (Crenn et al. 2016) extract 136 skeleton descriptors from motion, geometric, and frequency-domain, respectively, for body expression recognition. In (Li et al. 2016), the authors use the time-frequency analysis in signal processing to extract 44 kinematic features for emotion recognition. Meanwhile, a few works exploit deep learning to obtain gait representations. (Bhattacharya et al. 2020a) propose a classifier network based on Spatial Temporal Graph Convolutional Network (ST-GCN) (Yan, Xiong, and Lin 2018) to detect human emotion from gait. In (Zhuang et al. 2020), the authors propose a global link and shrinkage block that is effective in improving the performance of the classifier. (Narayanan et al. 2020) propose ProxEmo, which maps 3D skeleton data into an image for emotion recognition. The authors of (Hu et al. 2022) use two-stream network with Transformer-based to classify the image that is obtained by encoding skeleton joint and affective features. However, SSL used on gait-based emotion recognition is rarely explored.

### Self-Supervised Learning (SSL)

SSL aims to learn feature representations from unlabeled data, which usually generates supervision by carefully designing various pretext tasks, *e.g.* predicting rotation (Zhai et al. 2019), jigsaw puzzles (Noroozi and Favaro 2016; Noroozi et al. 2018), image inpainting (Pathak et al. 2016). With the proposal of MoCo (He et al. 2020), contrastive learning has gradually become an important branch of SSL. Contrastive learning essentially maps samples onto a hypersphere through alignment and uniformity (Wang and Isola 2020; Wang and Liu 2021). MoCo (He et al. 2020) and MoCov2 (Chen et al. 2020b) introduce momentum update mechanism and a queue-based memory bank. SimCLR (Chen et al. 2020a) promotes contrastive self-supervised learning through larger batch size and multi-layer perceptron. (Grill et al. 2020) propose BYOL that predicts previous versions of its outputs to avoid the use of negative pairs. NNCLR (Dwibedi et al. 2021) consider that there are potential positive samples in latent space, and using the nearest-neighbor to find positive samples in the memory bank for improving performance on ImageNet classification. Recently, (Zhang et al. 2022) propose SimMoCo, which focuses on hardness-aware property and removes dictionary as well as momentum. The above approaches have laid a solid theoretical foundation for CAGE.

### Unsupervised Skeleton Representation

According to network architecture, the most unsupervised skeleton representation approaches can be divided into two classes: (1) Single-batch. (Zheng et al. 2018) propose that using an auto-encoder-based GAN to explore the long-term global motion dynamics and learn skeleton representations. P&C (Su, Liu, and Shlizerman 2020) obtains more discriminative features by predicting and clustering skeleton sequences. (Lin et al. 2020) introduce an integrate multiple tasks framework called $MS^2L$, which solves multiple pretext tasks simultaneously (*e.g.*, jigsaw puzzle recognition and motion prediction). (Su, Lin, and Wu 2021) focus on motion consistency and continuity to learn the intrinsic dynamic motion consistency features. (2) Dual-batches. AS-CAL (Rao et al. 2021) introduces contrastive learning into unsupervised action recognition and proposes a generic paradigm based on momentum encoder. In (Thoker, Doughty, and Snoek 2021), the authors propose ISC that uses different network architecture to encode skeleton sequences and learns similarities between different skeleton representations. 3s-CrosSCLR (Li et al. 2021) and AimCLR (Guo et al. 2022) use cross-stream skeleton sequences for

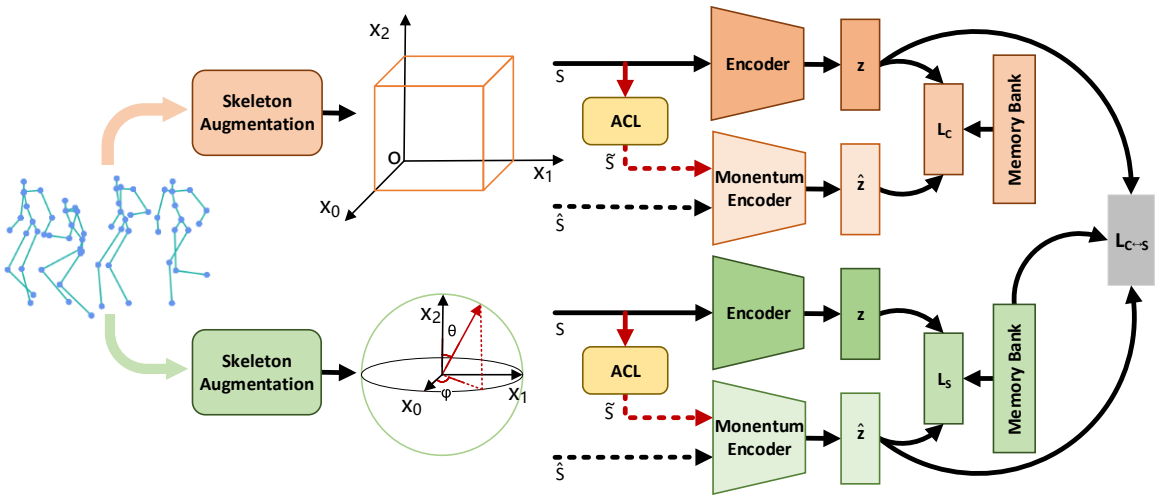



Figure 2: The overall pipeline of the proposed CAGE. We first augment the gait sequence into three different views ($\mathbf{S}$, $\hat{\mathbf{S}}$, and $\tilde{\mathbf{S}}$). Then, the encoder extracts $z$ and $\hat{z}$ while ACL is used to blend ambiguity samples $\tilde{\mathbf{S}}$ into the memory bank. Furthermore, our approach learns gait representations with cross-coordinate supervision, building contrastive loss cross the Cartesian coordinate and the Spherical coordinate. Note that the memory bank is maintained by $\hat{z}$.

contrastive learning, and improve the performance of their model through knowledge mining strategy. However, these methods highly depend on data augmentation, and they are rarely explore the semantic consistency and diversity of the augmented gait sequences. Thus, we introduce CAGE for unsupervised gait-based emotion representation.

## Method

**Problem definition.** Suppose $\mathbf{S} = (S_1, \ldots, S_T)$ is a 3D skeleton sequence that contains $T$ consecutive frames, where $S_i \in \mathbb{R}^{J*3}$ contains $J$ different body joints. Let $\Phi = \left\{\mathbf{S}^{(i)}\right\}_{i=1}^{N}$ be a gait set containing $N$ sequences. Our task is to train an encoder $f(\mathbf{S};\theta)$, which learns an effective gait representations from $\Phi$, without using any label. The overview of our approach is shown in Fig. 2, and the detail of each technical component are presented below.

### GE-CLR Overview

Contrastive learning has been widely used in the field of skeleton-based action recognition (Rao et al. 2021; Li et al. 2021). Inspired by this, we propose a simple **C**ontrastive **L**earning framework utilizing skeleton-based **G**ait for self-supervised **E**motion **R**epresentation (GE-CLR), which has a similar structure to MoCo (He et al. 2020).

GE-CLR has a data augmentation function $\mathcal{T}(\cdot)$, which randomly transforms the original gait sequences into two different views ($\mathbf{S}$ and $\hat{\mathbf{S}}$). Augmentation methods include *Crop* (Li et al. 2021) and *Joint Jittering* (Thoker, Doughty, and Snoek 2021). *Crop* is temporal augmentation strategy, which symmetrically pads $T/\gamma$ frames to the sequence and then randomly crops it to the original length. $\gamma$ is defined as padding ratio. *Joint Jittering* is spatial augmentation strategy, which randomly transforms the joint position. The transformation is defined as Eq. 1.

$$\mathbf{S}_{Jitter} = \mathbf{S}[:, j] \cdot \begin{bmatrix} r_{00} & r_{01} & r_{02} \\ r_{10} & r_{11} & r_{12} \\ r_{20} & r_{21} & r_{22} \end{bmatrix}, \tag{1}$$

where $j$ is a subset of the joints $J$, and $\{r_{00}, \cdots, r_{22}\}$ are the jitter factors sampled randomly from $[-1, 1]$.

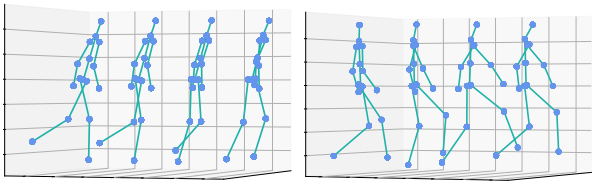
One gait sequence's different augments are positive samples, other gait sequences' different augments are considered as negative samples. Two encoders $f(\mathbf{S};\theta)$ and $\hat{f}(\hat{\mathbf{S}};\hat{\theta})$ project both views of the original gait sequence to an embedding feature space ($z$ and $\hat{z}$). We use the memory bank $\mathcal{N}_i$ that is a first-in-first-out queue to store the feature representation of negative samples. The memory bank is updated per iteration by $\hat{z}$. Specifically, the parameter of $\hat{f}$ is updated as: $\hat{\theta} \leftarrow m\hat{\theta} + (1-m)\theta$, where $m$ is a momentum coefficient. GE-CLR uses InfoNCE (van den Oord, Li, and Vinyals 2018) to calculate contrastive loss, formulated as Eq. 2.

$$\begin{aligned} \mathcal{L}_{Base} &= \mathcal{L}_{InfoNCE}(\hat{z}, z, \mathcal{N}_i) \\ &= -\log \frac{\exp(\hat{z} \cdot z/\tau)}{\exp(\hat{z} \cdot z/\tau) + \sum_{n \in \mathcal{N}_i} \exp\left(\hat{z} \cdot z_n/\tau\right)}, \end{aligned} \tag{2}$$

where $z_n \in \mathcal{N}_i$, and $\tau$ denotes the temperature hyper-parameter.

### Ambiguity Contrastive Learning (ACL)

Unlike images, the semantics of skeleton-based gait sequence is unstable. As shown in Figure 3, although the two gait sequences are sampled consecutively and alternately, respectively, the observer can easily tell they are dissimilar. So, why does the same gait sequence show different semantics? (Wang, Jiao, and Liu 2020) point out that human visual system is sensitive to video pace and speed play a crucial role in



(a) 1X-speed (b) 2X-speed

Figure 3: Visualization of the same sample at different playback speeds. (a) is raw sample that is labeled as "sad" without any pre-processing. (b) is sampled alternately (2X-speed), and adding a little noise.

video representation. The related works (Crenn et al. 2016; Wang et al. 2020) also prove that speed and joint angle are important features when using gait for mental state detection. Therefore, we think that changing speed and irregular distortions may lead to gait sequences with ambiguous semantic information, and thus the transformed gait sequences are pushed into other semantic space. Inspired by this, we propose ACL that adds ambiguity samples to the memory bank for withstanding semantic change.

**Ambiguity Transform.** Given a gait sequence $\mathbf{S}$, it is transformed by $\tilde{\mathcal{T}}(\cdot)$ into $\tilde{\mathbf{S}}$. We call $\tilde{\mathbf{S}}$ ambiguity sequence, meaning that the sequence has uncertain semantic information and no clear label. The ambiguity transform function $\tilde{\mathcal{T}}(\cdot)$ consists of two parts, which achieves the goal of pushing positive samples into ambiguous semantic space by changing speed and joint angle.

First, $\tilde{\mathbf{S}}$ is divided into $\tilde{\mathbf{S}}_{up}$ and $\tilde{\mathbf{S}}_{down}$ according to different playback speed. We suppose $r$ is playback speed, $\tilde{\mathbf{S}}_{up}$ is the upsampling sequence. For instance, when $r = 2$ and starting from the $i^{th}$ frame, we keep the frame set $\{i, i+r, i+2r, ...\}$. When $r < 1$, $\tilde{\mathbf{S}}_{down}$ is the downsampling sequence, we use linear interpolation to get new sequence, formulated as Eq. 3.

$$\tilde{\mathbf{S}}_{down} = \mathbf{S}[L : L + rT], \tag{3}$$

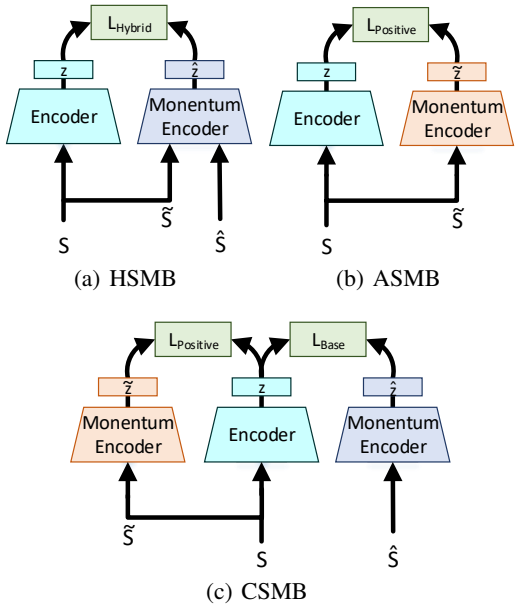
where $L$ is randomly selected starting frame, $L \in [0, T - rT]$, $r \in (0, 1)$.

Second, in order to make a gait sequence deviate from the original semantics as much as possible, we use the shear transformation that simulates the potential morphological changes of the gait sequence (Yang et al. 2022). The transformation formula is shown in Eq. 4.

$$\tilde{\mathbf{S}} = \tilde{\mathbf{S}}[:, j] \cdot \begin{bmatrix} 1 & s_{12} & s_{13} \\ s_{21} & 1 & s_{23} \\ s_{31} & s_{32} & 1 \end{bmatrix}, \tag{4}$$

where $j$ is a subset of the joints, $s_{12}, s_{13}, s_{21}, s_{23}, s_{31}, s_{32}$ are shear factors sampled randomly from $[-0.5, 0.5]$.

In prototypical contrastive learning paradigm, the memory bank is maintained by the negative samples. Since the ambiguity samples could have deviated from the original semantics, we propose using the ambiguity samples $\tilde{\mathbf{S}}$ to obtain




(a) HSMB (b) ASMB

(c) CSMB

Figure 4: Conceptual comparison of three memory bank updating mechanisms.

more diverse negative samples and mine latent representation. In other words, the memory bank is updated by both ambiguity samples and negative samples. In this paper, we propose three strategies for updating the memory bank, as shown in Figure 4.

**Hybrid Samples Memory Bank (HSMB).** Since the semantics of ambiguity samples $\tilde{\mathbf{S}}$ could have changed, we consider $\tilde{\mathbf{S}}$ as negative example and blend it directly into the memory bank, as shown in Eq. 5. When calculating the similarity between the positive samples and the negative samples, we use Eq. 6 as loss function. Figure 4.(a) gives an overview of HSMB.

$$\tilde{\mathcal{N}}_i = \left\{ \hat{f}\left(\mathcal{T}(\hat{S}_n); \hat{\theta}\right), \tilde{z} \mid \forall n \neq i \right\}, \tag{5}$$

where $\tilde{z}$ is obtained by $\hat{f}\left(\tilde{\mathbf{S}}; \hat{\theta}\right)$.

$$\mathcal{L}_{Hybrid} = \mathcal{L}_{InfoNCE}\left(\hat{z}, z, \tilde{\mathcal{N}}_i\right). \tag{6}$$

**Ambiguity Samples Memory Bank (ASMB).** GE-CLR needs to maintain the memory bank that contains thousands of negative samples, which brings redundant computation during contrastive training. In order to get rid of the dependence on negative samples, we focus on the similarity between positive samples and ambiguous samples, the contrastive loss function is written as Eq. 7. On the one hand, we are unconcerned about whether different samples have different representations, like BYOL (Grill et al. 2020). On the other hand, we keep the main architecture of MoCo, and the samples in the memory bank is replaced by $\tilde{z}$. As shown in Figure 4.(b), we transform raw sample into $\mathbf{S}$, $\hat{\mathbf{S}}$, and $\tilde{\mathbf{S}}$ for contrastive learning without negative samples.

$$\mathcal{L}_{Positive} = \mathcal{L}_{InfoNCE}\left(\hat{z}, z, \tilde{z}\right). \tag{7}$$

**Cooperation Samples Memory Bank (CSMB).** Previous work prove that multi-level supervision can enhance instance discrimination (Yao et al. 2021; Xie et al. 2021; Tian, Krishnan, and Isola 2020; Han, Xie, and Zisserman 2020). Inspired by this, we facilitate GE-CLR via the ambiguity samples discrimination task. The pre-trained gait encoder is learned to not only differentiate the normal positive samples from the memory bank, but also increase the similarity between ambiguous sample and positive samples. The overview architecture of cooperation samples memory bank is illustrated in Figure 4.(c). The complete loss function is defined as Eq. 8.

$$\mathcal{L}_{Cooperation} = \mathcal{L}_{\text{Base}} + \mathcal{L}_{Positive}. \tag{8}$$

## Cross-Coordinate Contrastive Learning ($C^3L$)

Most existing skeleton-based gait recognition methods are performed in the Cartesian coordinate, and have not yet explored the rich cross-coordinate supervision information. The Cartesian coordinate mainly reflects the position changes of joints, while the Spherical coordinates can better reflect the angle changes of joints. Moreover, we have already discussed that speed and joint angles are important indicators for studying the relationship between emotion and gait (Wang et al. 2020). However, joint angle is a hidden variable in the Cartesian coordinate, which is not easy to be captured by neural network. Considering that the two coordinates can be converted to each other, complementary information preserved in different coordinates can assist the operation to explore the inherent semantic information in different coordinate. More specifically, we propose $C^3L$, which trains both the Cartesian coordinate and the Spherical coordinate data stream together via their respective encoder simultaneously.

First, we convert the data to the Spherical coordinate according to Eq. 9. Then, we feed two group data under different coordinates to GE-CLR and get gait representations, respectively, as shown in Figure 2.

$$\begin{cases} r = \sqrt{x_0^2 + x_1^2 + x_2^2} \\ \theta = \arccos \frac{x_2}{r} \\ \varphi = \arctan \frac{x_1}{x_0} \end{cases}, \tag{9}$$

where $[x_0, x_1, x_2]$ denotes the 3D coordinates of body joint, $r$ is is the radial distance from the origin of the coordinates, $\varphi \in [0, 2\pi]$ is azimuth, $\theta \in [0, \pi]$ is elevation.

In order to pursue cross-coordinate consistency in contrastive learning, we refer to the gait sequences in the Cartesian coordinate to guide the learning process in the Spherical coordinate. The Cartesian↔Spherical contrastive loss can be written as Eq. 10.

$$\mathcal{L}_{C \leftrightarrow S} = \mathcal{L}_{InfoNCE}\left(\hat{z}, z, \mathcal{N}_i^S\right), \tag{10}$$

where $\mathcal{N}_i^S$ is the memory bank obtained in the Spherical coordinate.

As a summary, the complete loss function of CAGE can be defined as follows:

$$\mathcal{L}_{CAGE} = \lambda_1 \mathcal{L}_C + \lambda_2 \mathcal{L}_S + \mathcal{L}_{C \leftrightarrow S}, \tag{11}$$

where $\mathcal{L}_C$ and $\mathcal{L}_S$ represent the input data from the Cartesian coordinate and the Spherical coordinates, respectively. In this paper, $\mathcal{L}_C$ and $\mathcal{L}_S$ are calculated the same method as $\mathcal{L}_{Cooperation}$. $\lambda_1$ and $\lambda_2$ are the weight coefficients to trade off the importance of different loss.

# Experiments

## Datasets

In this paper, we use Emotion-Gait (E-Gait) datasets (Bhattacharya et al. 2020a) to evaluate our approach. The dataset contains 2,177 gait samples covering four emotions (happiness, sadness, neutral, angry). It consists of two subsets. Part $I$ contains 342 gait sequences, in which 90 participants were asked to imagine different emotions during walking, and then the camera recorded their gait. In part $II$, 1835 samples from ELMD dataset (Habibie et al. 2017) were annotated by the 10 annotators. All samples are skeleton data and contain $J = 16$ body joints. We use a split of 4:1 for training and testing sets and fix it, and this split is used for all experiments.

## Experimental Settings

All the experiments are conducted on with the PyTorch framework (Paszke et al. 2019). We train our model on one NVIDIA Titan V GPU. For data pre-processing, we process each gait sequence according to ISC (Thoker, Doughty, and Snoek 2021) and resize them to the same length (250 frames) by padding zeros. We take the top-1 accuracy (Acc) as evaluation criteria.

**Self-supervised Pretext Training.** For data augmentation, we set the joint jittering $j = 10$ and the padding ratio $\gamma = 6$. For the encoder, A-GCN (Shi et al. 2019) is adopted as the backbone. For contrastive settings, we follow that in AimCLR (Guo et al. 2022), but the size of the negatives set $\mathcal{N}$ is 2,048. In particular, the size of ASMB is the same as the mini-batch size. For GE-CLR, we use SGD ($momentum = 0.9$ and $weight - decay = 0.0001$) optimizer and train for 200 epochs with a learning rate of 0.1, which is multiplied by 0.1 at epoch 100 and 160. The mini-batch size is set to 32. CAGE follows that in GE-CLR, but the model trains for 300 epochs with a learning rate of 0.1 (multiplied by 0.1 at epoch 150 and 250).

**Linear Evaluation Protocol.** We append a linear classifier to the frozen encoder, and then training the classifier supervisedly. The model is trained for 100 epochs with the learning rate 30 (multiplied by 0.1 at epoch 50 and epoch 70).

**Finetuned Evaluation Protocol.** Attaching the trained encoder to a linear classifier, and then training the whole model for emotion recognition task. The model is trained for 20 epochs with the learning rate 0.001 (multiplied by 0.1 at epoch 10).

**Semi-supervised Evaluation Protocol.** Following finetune protocol, we fine-tune the pre-trained encoder and the final classification layer, but only 5%, 10%, 20%, and 50% randomly selected labeled data are used.

| GE-CLR | ACL | $C^3L$ | Acc(%) |
|---|---|---|---|
| ✓ | | | 76.38 |
| ✓ | ✓ | | 78.67 |
| ✓ | | ✓ | 76.83 |
| ✓ | ✓ | ✓ | **79.59** |

Table 1: Performance comparison of different components of our method. "✓" indicates that the corresponding model component is used.

| Sampling Method | Playback Speed ($r$) | Acc(%) |
|---|---|---|
| Up | 2 | 77.06 |
| Up | 4 | **78.67** |
| Up | 8 | **78.67** |
| Down | 0.1 | 77.29 |
| Down | 0.3 | **78.44** |
| Down | 0.5 | 77.06 |

Table 2: Performance comparison of different playback speed.

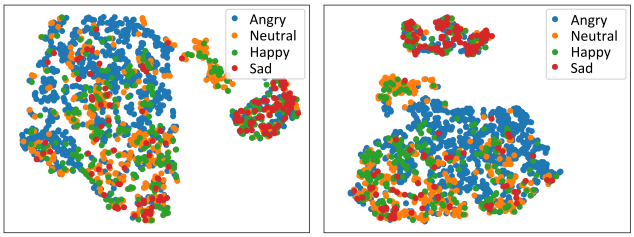
## Ablation Study

We conduct ablation study to demonstrate the effectiveness of our method. All the experiments in this section follow the unsupervised pre-training and linear evaluation protocol.

**Effectiveness of ACL and $C^3L$.** We perform ablation study to provide solid validation of each model component. As report in Table 1, we first observe that when introducing ACL, the accuracy is improved by 2.29%, which fully verify that ambiguous samples can make the encoder learn more movement features for gait-based emotion recognition. However, it is worth noting that compared with ACL, the gait representations built by $C^3L$ show limited effectiveness (only 0.45% accuracy improvement). That's because, $C^3L$ with two inputs (the Cartesian coordinate and the Spherical coordinate) is difficult to converge during pre-training. Therefore, we increase the number of epoch to achieve stable gait representations. The experimental results show that CAGE achieves the highest accuracy when $C^3L$ is further introduced.

**Qualitative Results.** We apply t-SNE (Van der Maaten and Hinton 2008) visualization of the embedding distribution of GE-CLR and CAGE in Figure 5. From the visual results, we have the following conclusions: (1) Our approach can roughly cluster the embeddings of the same class and map them uniformly on a hypersphere. (2) Features of CAGE presents a more discriminative distribution than GE-CLR, which makes the emotion classes that overlapped seriously more discriminative distribution.

**Effectiveness of Playback Speed.** As hyper-parameter $r$ determines the speed of samples, influencing the similarity between positive samples and ambiguity samples, we study how $r$ impacts the performance in ACL. In the subsection, we conduct ablation study using HSMB. From Table 2, we can see that the smaller $r$ leads to the similar augmented gait sequences rather than bringing rich semantic information. However, ACL is always better than GE-CLR under the same protocol no matter using $\tilde{\mathbf{S}}_{up}$ or $\tilde{\mathbf{S}}_{down}$. Finally, we consider two playback speed candidates according to different sampling method, where the corresponding speeds $r$ are 4, 0.3, respectively.




(a) GE-CLR (b) CAGE

Figure 5: t-SNE visualization of embeddings at different model during pre-training. Embeddings from four emotions are sampled and visualized with different colors. More visual results are shown in Appendix.

**Effectiveness of Ambiguity Samples.** In this subsection, we evaluate the effectiveness of different memory bank (HSMB, ASMB, and CSMB). From Figure 6, we have several observations: (1) ACL performs mostly better than GE-CLR, which verifies the necessity of ambiguity samples. (2) Using the upsampling sequences $S_{up}$ produces evident performance gain (1.37%-2.29% accuracy) when compared with normal memory bank without ambiguity samples. (3) Although the performance of ASMB is not stable, the result is obtained without negative samples. It proves our claim that ambiguous samples that have deviated from the original semantics can help the model learn novel movement patterns and improve the gait representations.

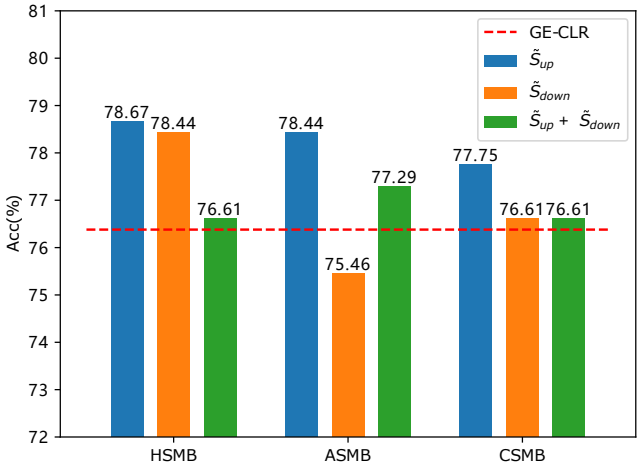



Figure 6: Comparison of linear evaluation results of the different memory bank.

| Input | ACL | Acc(%) |
|---|---|---|
| Cartesian coordinate | - | 76.38 |
| Cartesian coordinate | ✓ | 78.67 |
| Spherical coordinate | - | 72.02 |
| Spherical coordinate | ✓ | 74.08 |
| Cross-coordinate | - | 76.83 |
| Cross-coordinate | ✓ | **79.59** |

Table 3: Performance comparison of different coordinate.

| Method | Acc(%) |
|---|---|
| *Supervised* | |
| Hand-crafted Features(IC3D 2016) | 66.22 |
| ST-GCN(AAAI 2018) | 65.62 |
| Base-STEP(AAAI 2020a) | 78.24 |
| TEW(ECCV 2020b) | 81.89 |
| STEP(AAAI 2020a) | 82.15 |
| TNTC(ICASSP 2022) | 85.97 |
| *Unsupervised* | |
| ISC(ACM MM 2021) | 72.93 |
| 3s-CrosSCLR(CVPR 2021) | 76.83 |
| AimCLR(AAAI 2022) | 74.31 |
| **CAGE(Ours)** | **79.59** |

Table 4: Linear evaluation results.

**Effectiveness of Different Coordinate.** We conduct experiments on different inputs to verify the performance of our approach. From Table 3, one can see that the accuracy on the Cartesian coordinate input and the Spherical coordinate input are 76.38% and 72.02%, respectively. Especially, when introducing ACL, the accuracy is improved by 2.29% and 2.06%. We also observe that pre-training with two inputs side by side is considerably better than only learning with a single input. It also proves that our approach can pursue cross-coordinate semantic consistency for learning suitable features.

## Comparison with State-of-the-art

As few unsupervised results are reported on gait-based emotion recognition, we compare GE-CLR and CAGE with some skeleton-based self-supervised methods that are similar to gait-based emotion recognition task.

**Linear Evaluation Results.** We conduct an extensive comparison with existing supervised methods and recent paper in the field of skeleton-based self-supervised action recognition. As shown by Table 4, compared with unsupervised method, the accuracy of CAGE is improved by 2.76%-6.66%. What's more, our approach not only defeats unsupervised methods, but also achieves comparable or even superior performance to some supervised methods. From these results, we can observe that our approach enjoys obvious advantages over existing skeleton-based methods in terms of gait-based emotion performance.

| Method | Acc(%) |
|---|---|
| ISC(ACM MM 2021) | 71.56 |
| 3s-CrosSCLR(CVPR 2021) | 79.36 |
| AimCLR(AAAI 2022) | 76.15 |
| **GE-CLR(Ours)** | 81.88 |
| **CAGE(Ours)** | **82.57** |

Table 5: Finetuned evaluation results.

| Method | 5% | 10% | 20% | 50% |
|---|---|---|---|---|
| ISC(ACM MM 2021) | 60.78 | 61.69 | 62.16 | 67.89 |
| 3s-CrosSCLR(CVPR 2021) | 62.61 | 70.87 | 72.48 | 70.41 |
| AimCLR(AAAI 2022) | 49.54 | 61.93 | 61.01 | 65.82 |
| **GE-CLR(Ours)** | 66.51 | 74.31 | 77.75 | 81.42 |
| **CAGE(Ours)** | **70.64** | **78.90** | **79.13** | **81.65** |

Table 6: Semi-supervised evaluation results.

**Finetuned Evaluation Results.** For fair comparisons, we follow the same protocol as 3s-CrosSCLR (Li et al. 2021), but the number of epoch is 20. As shown in Table 5, our approach surpasses ISC (Su, Lin, and Wu 2021), 3s-CrosSCLR (Li et al. 2021) and AimCLR (Guo et al. 2022) by 3.21%-11.01% accuracy on finetuned evaluation protocol. It shows that CAGE can more easily capture gait features and learn better gait representations.

**Semi-supervised Evaluation Results.** The results in Table 6 reveal that our approach is always better than other unsupervised methods under the same proportion of labeled data no matter. In particular, when there is only 5% labeled subset, our approach performs better than other model by a large margin (8.03%-21.10% accuracy improvement). It indicates that our approach is especially suited to learn from a small amount of labeled data.

## Conclusion

In this paper, we propose a cross-coordinate contrastive learning framework named CAGE, which mainly focuses on obtaining effective gait representations from unlabeled skeleton-based data for emotion recognition. We focus on the semantic consistency and diversity, which are two critical factors for learning gait representations. We propose ACL to learn diverse semantics and $C^3L$ to pursue the intrinsic semantic consistency information of gait sequences. Experiments indicate that CAGE significantly outperforms existing skeleton-based unsupervised methods under a variety of evaluation protocols, and its performance is comparable or even superior to some supervised learning methods.

## Ethics Statement

Gait-based emotion recognition as an important emerging research topic possesses great value. However, illegal or improper use of gait-based emotion recognition technologies could pose serious threat to the public privacy and society security. Especially, in the field of human-computer interaction, telling a participant that his emotional state is negative may trigger further emotions. Therefore, the researcher needs to carefully evaluate the application of gait-based emotion recognition to prevent accidental injury. Our models must only be used for the purpose of research.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), in part by the National Natural Science Foundation of China (Grant No.61632014, No.61627808).

## References

Bhattacharya, U.; Mittal, T.; Chandra, R.; Randhavane, T.; Bera, A.; and Manocha, D. 2020a. STEP: Spatial Temporal Graph Convolutional Networks for Emotion Perception from Gaits. In *AAAI*, 1342–1350.

Bhattacharya, U.; Roncal, C.; Mittal, T.; Chandra, R.; Kapsaskis, K.; Gray, K.; Bera, A.; and Manocha, D. 2020b. Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping. In *ECCV*, 145–163. Springer.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 7291–7299.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, X.; Fan, H.; Girshick, R. B.; and He, K. 2020b. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297.

Crenn, A.; Khan, R. A.; Meyer, A.; and Bouakaz, S. 2016. Body expression recognition from animated 3D skeleton. In *2016 International Conference on 3D Imaging (IC3D)*, 1–7.

Daoudi, M.; Berretti, S.; Pala, P.; Delevoye, Y.; and Bimbo, A. D. 2017. Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices. In *International Conference on Image Analysis and Processing*, 550–560. Springer.

Deligianni, F.; Guo, Y.; and Yang, G.-Z. 2019. From emotions to mood disorders: A survey on gait analysis methodology. *IEEE journal of biomedical and health informatics*, 23(6): 2302–2316.

Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *CVPR*, 9588–9597.

Fernández-Dols, J.-M.; and Ruiz-Belda, M.-A. 1995. Expression of emotion versus expressions of emotions. In *Everyday conceptions of emotion*, 505–522. Springer.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 21271–21284.

Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; and Ding, R. 2022. Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-supervised Action Recognition. In *AAAI*, volume 36, 762–770.

Habibie, I.; Holden, D.; Schwarz, J.; Yearsley, J.; and Komura, T. 2017. A recurrent variational autoencoder for human motion synthesis. In *BMVC*.

Han, T.; Xie, W.; and Zisserman, A. 2020. Self-supervised co-training for video representation learning. In *NeurIPS*, 5679–5690.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Hu, C.; Sheng, W.; Dong, B.; and Li, X. 2022. TNTC: two-stream network with transformer-based complementarity for gait-based emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3229–3233. IEEE.

Jia, Z.; Lin, Y.; Wang, J.; Feng, Z.; Xie, X.; and Chen, C. 2021. HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *ACM MM*, 1047–1056.

Kleinsmith, A.; and Bianchi-Berthouze, N. 2012. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1): 15–33.

Li, B.; Zhu, C.; Li, S.; and Zhu, T. 2016. Identifying emotions from non-contact gaits information based on microsoft kinects. *IEEE Transactions on Affective Computing*, 9(4): 585–591.

Li, J.; Pascal, V.; Yann, L.; and Yuandong, T. 2022. Understanding dimensional collapse in contrastive self-supervised learning. *ICLR*.

Li, L.; Wang, M.; Ni, B.; Wang, H.; Yang, J.; and Zhang, W. 2021. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In *CVPR*, 4741–4750.

Lin, L.; Song, S.; Yang, W.; and Liu, J. 2020. $Ms^2l$: Multi-task self-supervised learning for skeleton based action recognition. In *ACM MM*, 2490–2498.

Lu, H.; Xu, S.; Hu, X.; Ngai, E.; Guo, Y.; Wang, W.; and Hu, B. 2022. Postgraduate Student Depression Assessment by Multimedia Gait Analysis. *IEEE MultiMedia*, 29(2): 56–65.

Narayanan, V.; Manoghar, B. M.; Dorbala, V. S.; Manocha, D.; and Bera, A. 2020. Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In *IROS*, 8200–8207. IEEE.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 69–84. Springer.

Noroozi, M.; Vinjimoor, A.; Favaro, P.; and Pirsiavash, H. 2018. Boosting self-supervised learning via knowledge transfer. In *CVPR*, 9359–9367.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *CVPR*, 2536–2544.

Quigley, K. S.; Lindquist, K. A.; and Barrett, L. F. 2014. Inducing and measuring emotion and affect: Tips, tricks, and secrets. *Handbook of research methods in social and personality psychology*, 220–252.

Rao, H.; Xu, S.; Hu, X.; Cheng, J.; and Hu, B. 2021. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569: 90–109.

Ringeval, F.; Schuller, B.; Valstar, M.; Cummins, N.; Cowie, R.; Tavabi, L.; Schmitt, M.; Alisamir, S.; Amiriparian, S.; Messner, E.-M.; et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 3–12.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 12026–12035.

Su, K.; Liu, X.; and Shlizerman, E. 2020. Predict & cluster: Unsupervised skeleton based action recognition. In *CVPR*, 9631–9640.

Su, Y.; Lin, G.; and Wu, Q. 2021. Self-Supervised 3D Skeleton Action Representation Learning With Motion Consistency and Continuity. In *ICCV*, 13328–13338.

Sun, X.; Su, K.; and Fan, C. 2022. VFL—A deep learning-based framework for classifying walking gaits into emotions. *Neurocomputing*, 473: 1–13.

Thoker, F. M.; Doughty, H.; and Snoek, C. G. 2021. Skeleton-Contrastive 3D Action Representation Learning. In *ACM MM*, 1655–1663.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *ECCV*, 776–794. Springer.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, F.; and Liu, H. 2021. Understanding the behaviour of contrastive loss. In *CVPR*, 2495–2504.

Wang, J.; Jiao, J.; and Liu, Y.-H. 2020. Self-supervised video representation learning by pace prediction. In *ECCV*, 504–521.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 9929–9939.

Wang, T.; Li, C.; Wu, C.; Zhao, C.; Sun, J.; Peng, H.; Hu, X.; and Hu, B. 2020. A gait assessment framework for depression detection using kinect sensors. *IEEE Sensors Journal*, 21(3): 3260–3270.

Wang, X.; and Qi, G.-J. 2022. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, X.-J.; Zhang, L.; Jing, F.; and Ma, W.-Y. 2006. Annosearch: Image auto-annotation by search. In *CVPR*, volume 2, 1483–1490. IEEE.

Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; and Luo, P. 2021. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, 8392–8401.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Yang, J.; Lu, H.; Li, C.; Hu, X.; and Hu, B. 2022. Data augmentation for depression detection using skeleton-based gait information. *Medical & Biological Engineering & Computing*, 2665–2679.

Yao, T.; Zhang, Y.; Qiu, Z.; Pan, Y.; and Mei, T. 2021. Seco: Exploring sequence supervision for unsupervised representation learning. In *AAAI*, volume 2, 7.

Zhai, X.; Oliver, A.; Kolesnikov, A.; and Beyer, L. 2019. S4l: Self-supervised semi-supervised learning. In *ICCV*, 1476–1485.

Zhang, C.; Zhang, K.; Pham, T. X.; Niu, A.; Qiao, Z.; Yoo, C. D.; and Kweon, I. S. 2022. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *CVPR*, 14441–14450.

Zhang, Z. 2012. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2): 4–10.

Zheng, N.; Wen, J.; Liu, R.; Long, L.; Dai, J.; and Gong, Z. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *AAAI*, volume 32.

Zhuang, Y.; Lin, L.; Tong, R.; Liu, J.; Iwamot, Y.; and Chen, Y.-W. 2020. G-GCSN: Global Graph Convolution Shrinkage Network for Emotion Perception from Gait. In *Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops*.