# Skeleton-Based Pre-Training with Discrete Labels for Emotion Recognition in IoT Environments

Zhen Zhang, Feng Liang, Wei Wang *Member, IEEE*, Runhao Zeng* *Member, IEEE*, Xiping Hu* *Member, IEEE*, Victor C.M. Leung *Life Fellow, IEEE*,

***Abstract*—Self-supervised emotion recognition leveraging skeleton-based data offers a promising approach for classifying emotional expressions within the extensive amount of unlabeled data gathered by sensors in the Internet of Things (IoT). Recent advancements in this field have been driven by contrastive learning-based or generative learning-based self-supervised methods, which effectively tackle the issue of sparsely labeled data. In emotion recognition tasks, the emotional high-level semantics embedded in the skeleton data are more important than the subtle joint movements. Compared to existing methods, discrete label prediction can encourage SSL models to abstract high-level semantics in a manner similar to human perception. However, it is challenging to comprehensively capture emotional information expressed in skeleton data solely from joint-based features. Moreover, emotional information conveyed through body movements may include redundant details that hinder the understanding of emotional expression. To overcome these challenges, we propose a novel discrete-label-based emotion recognition framework named the Appendage-Informed Redundancy-ignoring (AIR) discrete label framework. First, we introduce the Appendage-Skeleton Partitioning (ASP) module, which leverages limb movement data from the original skeleton to explore emotional expression. Next, we propose the Appendage-refined Multi-scale Discrete Label (AMDL) module, which transforms traditional self-supervised tasks into classification tasks. This design continuously extracts emotional semantics from skeleton data during pre-training, functioning similarly to predicting categories and subsequently classifying samples. To further reduce the nonessential information in skeleton data that may negatively impact the generation of accurate emotional categories, we propose the Appendage Label Refinement (ALR) module. It refines the generated categories by using the relationships between the skeleton and the various appendages obtained via ASP module. Finally, to maintain consistency across multiple scales, we introduce the Multi-Granularity Appendage Alignment (MGAA) method. By incorporating features from both coarse and fine scales, MGAA mitigates the encoder's sensitivity to noise and enhances its overall robustness. We evaluate our approach on the Emilya, EGBM, and KDAE datasets, where it consistently outperforms state-of-the-art methods under various evaluation protocols.**

***Index Terms*—Internet of Things (IoT), Body Skeleton-based Analysis, Self-supervised, Affective Computing**

(*Corresponding author: Xingping Hu; Runhao Zeng)

Zhen Zhang is with the Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China, the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, China, the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Gansu, China (e-mail:zhangzhen19@lzu.edu.cn)

Feng Liang and Runhao Zeng are with the Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China and the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, China (e-mail:fliang@smbu.edu.cn, zengrh@smbu.edu.cn)

Wei Wang and Xiping Hu are with the Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China, the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, China, and the School of Medical Technology, Beijing Institute of Technology, China (e-mail:ehomewang@ieee.org, huxp@smbu.edu.cn)

Victor C.M. Leung is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, China, and the Department of Electrical and Computer Engineering, The University of British Columbia, Canada

## I. Introduction

With the rapid development of the Internet of Things (IoT), data interactions are occurring constantly [1], [2]. To better understand humans' thoughts and feelings, emotion recognition is crucial for comprehending their behavior and decision-making processes. Within the IoT ecosystem, emotion recognition enables the creation of smarter, context-aware environments where interconnected devices can dynamically adapt to users' emotional states, thereby enhancing interactions and overall user experiences. Current emotion recognition methods primarily focus on analyzing facial expressions [3]–[6], speech [7], [8], text [9], and physiological signals such as electroencephalography (EEG) [10], [11] or electrocardiography (ECGs) [12]. However, facial expression-based approaches can be unreliable in cases of 'mock expressions' or misleading self-reports of emotional responses [13], [14]. Additionally, the resolution of the collected facial data significantly affects the accuracy of recognition. Speech- or text-based methods may be less suitable in public settings or for large-scale crowds [15]. Physiological signal-based approaches, on the other hand, involve demanding data acquisition processes that limit their practicality in everyday IoT applications [16], [17]. These challenges underscore the necessity for more powerful emotion recognition techniques.

The existing research shows certain differences in people's body movements with different emotions [18]. With the increasing maturity of depth sensors in the IoT environments [19] and human pose estimation algorithms [20], [21], the cost of acquiring skeleton data has gradually decreased. As a result, emotion recognition based on body skeleton data has garnered increasing attention. Early body skeleton-based approaches to emotion recognition primarily depended on handcrafted features [22]–[24]. The above approach relies heavily on the set feature extraction rules that depend on the domain knowledge setting, which limits the generalization ability of the method [25]. With the rise of deep learning, the limitations of manual feature extraction methods have been broken. STEP [26], ProxEmo [27], and TNTC [28] respectively attempted to automatically extract the high-level

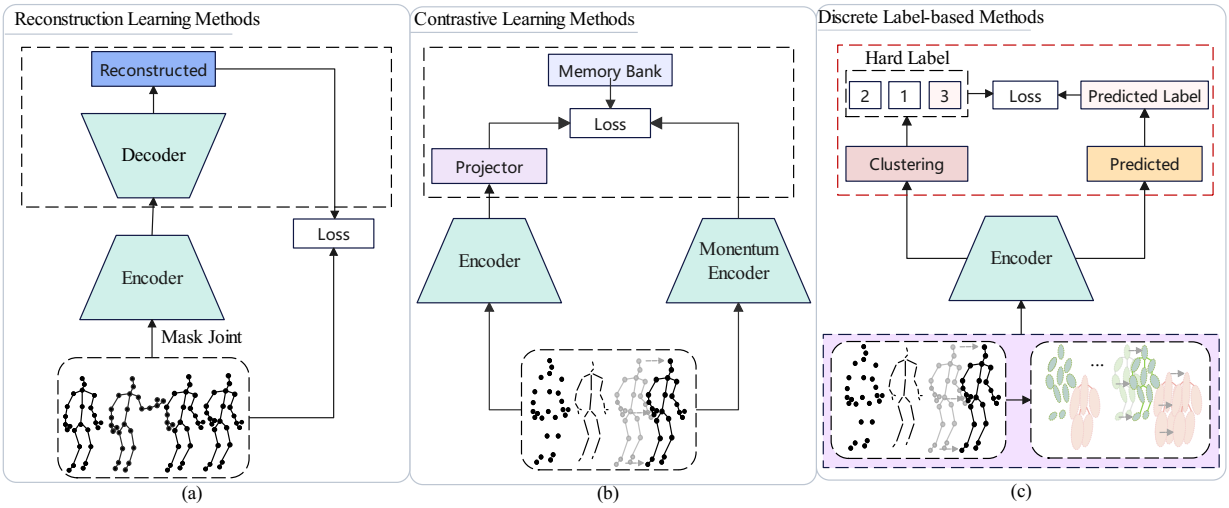


Fig. 1. Three commonly used self-supervised learning frameworks. Among these, (a) is the reconstruction learning method, (b) is the contrastive learning method, and (c) is the Discrete Label-based method.

emotional categories information contained in skeleton data by means of spatiotemporal graph convolution and group convolution. It is important to note that the aforementioned methods, as supervised learning approaches, rely heavily on a substantial amount of labeled body skeleton-based emotional data. However, the data annotation process is notably labour-intensive and time-consuming. Moreover, the labeling process imposes stringent requirements on annotators to prevent topic deviation, which could lead to incorrect labeling.

To effectively utilize unlabeled emotional data, researchers have employed self-supervised learning (SSL) methods to learn diverse emotion representations from these unlabeled datasets. In recent years, SSL has achieved significant success using contrastive-based methods [15], [29], as depicted in Fig. 1(a), and reconstruction-based methods [30], [31], as shown in Fig.1(c), for skeleton emotion recognition tasks. However, it is generally believed that these SSL frameworks primarily ensure the accuracy of low-level features while neglecting high-level semantic abstractions [32]–[35]. In a constrained IoT environment, it is critical to improve the encoder's ability to extract the high-level emotional-semantic representation contained in skeleton data. Compared with other SSL frameworks, SSL method based on discrete labels makes encoder pay more attention to extracting high-level semantic information in emotional expression during pre-training [34].

Although discrete label prediction offers these advantages and has achieved significant success across various domains [33], [36]–[39], its application in general skeleton-based emotion recognition tasks remains challenging for two reasons. Firstly, in the dynamic emotional expression process based on skeleton data, in addition to the micro-level information (i.e., the changes in the original 24 joints included in the skeletal structure), the coarse-grained motion variations of the appendage segments also convey significant emotional information [40]. To leverage the more comprehensive emotional representations contained within skeleton data for generating discrete labels, it is essential to consider how to enable methods to extract micro-level information from the skeleton while simultaneously capturing the valuable structural information of the auxiliary limb segments. Secondly, when humans understand bodily emotional expressions, they do so by extracting and clustering the macroscopic relationships between different body parts. For example, when expressing happiness, individuals involuntarily swing their upper and lower limbs with varying amplitudes. Even when the same person's limb movement habits do not significantly differ across different environments, identifying these commonalities can convey unique and recognizable high-level emotion information [41]. Humans can classify emotions by capturing the prominent changes in the upper and lower limb regions while eliminating redundant details that exhibit minimal variation. To achieve understanding and generalized discriminative capabilities comparable to humans, it is essential to consider how to eliminate redundant information contained in the skeleton data to obtain more refined discrete labels.

To address the aforementioned challenges, we propose a novel Appendage-Informed Redundancy-ignoring (AIR) discrete label framework, which refines discrete labels to mitigate these issues. To capture the additional motion variations contained in skeleton data, we introduce an Appendage-Skeleton Partitioning (ASP) module. This module segments human limbs based on nodes in the skeleton data and inputs this information into the encoder along with fine-grained details (joint, bone, and motion data from the original skeleton). This process ensures that the encoder learns comprehensive skeletal information during pre-training. Additionally, to reduce the redundant details present when extracting and clustering data of the same emotion, we propose the Appendage-refined Multi-scale Discrete Label (AMDL) module. Its simple structure is shown in Fig. 1(b). In this framework, we leverage the original skeleton data and the coarse-scale appendage information obtained by the ASP module to label the unlabeled data through clustering. In each iteration, we use the AMDL module to generate emotion labels for the unlabeled skeleton data and use them to optimize the SSL model. In this process, we also propose an Appendage Label Refinement (ALR) Module,

which utilizes the coarse-grained appendage information from ASP to amplify features that help distinguish between categories and reduce the noise caused by highly similar regions, such as the human torso, during clustering. Finally, to ensure the consistency of all information contained in the skeleton data, we introduce Multi-Granularity Appendage Alignment (MGAA). By leveraging features from both macro and fine-grained scales, we reduce the encoder's sensitivity to noise and enhance its robustness.

In summary, our new self-supervised learning framework for body movement emotion recognition provides three key contributions.

- We propose the ASP module to uncover latent macro appendage information in skeleton data. This module underscores the significance of appendage-based movements in emotion recognition, enabling the encoder to capture more comprehensive skeletal information. Consequently, it enhances the overall performance of emotion recognition.
- We propose the AIR framework, which transforms SSL training into a simple process of iterative pretraining using discrete labels. This framework leverages the more comprehensive information obtained from the ASP module to assign discrete labels to the corresponding data samples. In this process, we introduce the ALR module to eliminate redundant information during emotion category recognition and enhance the encoder's ability to capture high-level emotional representations.
- Extensive experiments conducted on three datasets (i.e., Emilya, EGBM, KDAE) validate the effectiveness and transferability of the proposed framework. The results demonstrate that our approach surpasses state-of-the-art self-supervised techniques across multiple evaluation protocols.

## II. Related Work

### A. Emotion Recognition from Skeleton-based Movements

Early approaches to identifying emotions through body skeleton-based movement and posture primarily relied on extracting hand-crafted features. [42] extracted features at multiple levels using three-dimensional motion data of full-body movement. These feature vectors were then input into a support vector machine for classification. [43] use covariance descriptors derived from 3D skeleton joint sequences, representing them within the non-linear riemannian manifold of symmetric positive definite matrices. This allowed them to leverage geodesic distances and geometric means on the manifold to perform emotion classification. In another approach, [44] explores the contribution of different body movement representations to the classification of emotions expressed in various movement tasks. Their findings suggest that submotion characteristics of action-related joints (e.g., temporal features of foot motion during walking) capture additional emotional properties. These features, when combined with multi-level descriptions that include multidirectional representations of full-body posture and a discrete analysis of movement dynamics, enhance the recognition of emotional body expressions.

In recent years, several studies [45], [46] have utilized deep learning models to learn emotion representations from body movements, typically processed as skeleton data. Given that skeleton data is a non-Euclidean form of data, Graph Convolutional Network (GCN)-based methods have garnered significant attention [26], [47]. For instance, STEP [26] is one of the first attempts to classify perceived human emotions from skeleton-based data using GCN. [48] introduces a novel AT-GCN network for skeleton sequences, which effectively captures discriminative spatiotemporal features. The AT-GCN simultaneously learns representations for multiple tasks, including emotion recognition, identity recognition, and auxiliary prediction. Additionally, [49] propose a self-attention enhanced spatial-temporal GCN for skeleton-based emotion recognition. In this model, the spatial convolution component models the body's skeletal structure as a static graph, while the self-attention mechanism dynamically constructs additional connections between joints to provide supplementary information. More recently, EPIC [47] introduces a joint reconstruction method to uncover latent connections between body joints, followed by an ST-GCN to identify emotions. While these approaches rely on supervised learning to extract emotional features from GCNs, the limited availability of labeled affective skeleton data, coupled with the potential for mislabeling, can negatively impact model performance and generalizability. To address this, we pre-train encoders on unlabeled skeleton sequences using a self-supervised learning framework, enabling the extraction of more robust and efficient emotional representations.

### B. Body Skeleton-based Self-supervised Methods

Self-supervised learning aims to learn effective feature embedding functions from unlabeled data.

Previous work has primarily focused on tasks such as predicting rotations [50], jigsaw puzzles [51], [52], and image inpainting [53]. With the advent of methods like MoCo [54] and SimCLR [55] methods, contrastive learning techniques have shown remarkable performance. The purpose of these techniques is to bring the features of homologous samples closer and samples from different sources further apart [15]. As for the reconstruction pre-training objective, Audio2Vec [56] proposed the CBoW task to reconstruct the acoustic feature of an audio clip of pre-determined duration based on past and future clips. MAE [57] proposed masking random patches of the input image and reconstruct the missing pixels.

There have been numerous efforts to apply these SSL methods for skeleton-based tasks, such as skeleton-based action recognition. For example, [58] introduces comparative learning based on momentum updating and proposes a series of skeleton data augmentation strategies, which laid the groundwork for subsequent research. [59] proposes a cross-view contrastive learning framework for unsupervised 3D skeleton-based action recognition by leveraging multiview complementary supervision signals. This method integrates both single-view contrastive learning and cross-view consistent knowledge

mining modules in a collaborative learning framework. [60] proposes a novel cross-modal mutual distillation framework, formulating the cross-modal interaction as a bidirectional knowledge distillation problem. Recently, some researchers have explored contrastive learning techniques to address self-supervised emotion recognition tasks based on the body skeleton. [29] designs a cross-coordinate contrastive learning framework for self-supervised emotion representation from body skeletons. The method uses an uprising transformation to push positive samples into an ambiguous semantic space, enabling the model to capture high-level skeleton semantics while maintaining semantic diversity. Additionally, [15] proposes an SSA method for gait skeleton emotion recognition, incorporating upper body jitter and random spatiotemporal masking to generate diverse positive samples, helping the model learn more distinctive features. As for the reconstruction pre-training objective, SkeletonMAE [61] applied the Masked Autoencoder (MAE) approach to 3D skeleton action representation learning, which employs a skeleton-based encoder-decoder transformer for spatial coordinate reconstruction. Skeleton2Vec [62] used a transformer-based teacher encoder taking unmasked training samples as input to create latent contextualized representations as prediction targets.

While other skeleton-based tasks (e.g., action recognition) typically focus on distinguishing fine-grained differences among various classes, emotion recognition instead places greater emphasis on the high-level semantic information conveyed by the entire skeleton [29]. Compared with existing self-supervised frameworks, the SSL approach based on discrete labels is more adept at capturing these higher-level semantics, thereby conveying the essential meaning of each class [34]. Although SSL methods based on discrete labels have already been broadly applied to tasks in audio [36], [37], vision [33], [39], and language [63], they remain largely unexplored in the context of skeleton-based emotion recognition. To uncover the emotional semantics concealed within substantial amounts of unlabeled skeleton data, we propose leveraging discrete label-based methods to capture high-level affective information for emotion recognition.

### C. Discrete Labels-based self-supervised methods

We effectively explore a self-monitoring strategy grounded in discrete labels. For example, SpCL [64] partitions unlabeled data into cluster-level and unclustered instance-level categories using clustering as a feature representation in supervised signal learning. [65] partitions unlabeled data into cluster-level and unclustered instance-level categories using clustering as a feature representation in supervised signal learning. Similarly, [66] proposes the CSTCN method, which constructs a supervised signal of action sequences via an online clustering mechanism, complemented by data augmentation and triplet contrastive sample construction strategies.

Nevertheless, when applying discrete label-based SSL methods to skeleton-based emotion recognition, two major challenges must be addressed. First, compared to other data types (e.g., images), skeleton data inherently carries less information. To encourage the encoder to extract a broader range of emotional signals from skeleton data during discrete label-based pretraining, we propose the Appendage-Skeleton Partitioning (ASP) Module. This module enables the encoder to capture a more comprehensive representation of emotions. The second challenge arises from the presence of redundant skeletal information during emotional expressions. For instance, in a display of happiness, limb movements typically exhibit a larger amplitude than the torso. However, the torso information contained in the skeleton may compromise the accuracy of the final discrete label generation, thereby hindering the encoder's capacity to learn high-level emotional semantics during pretraining. To address this issue, we introduce a discrete label-based pretraining paradigm, termed Appendage-refined Multi-scale Discrete Label (AMDL), which allows the encoder to acquire high-level emotional semantics during the pretraining phase. In addition, we propose the Appendage Label Refinement (ALR) module to mitigate the adverse effects of redundant skeletal information on discrete label generation, thereby further enhancing the encoder's capability to extract emotional semantics.

## III. METHOD

**Problem Definition.** Given a skeleton sequence $X$ including $T$ frames of $J$ body joints under the global spatial coordinate, the body skeleton sequence of a specific subject can be represented as $\{v_j^t | t \in (1, \ldots, T), j \in (1, \ldots, J)\}$, where $v_j^t$ represents the spatial coordinate of joint $j$ at frame $t$ of that subject and can be denoted by a three-dimensional vector $(x_j^t, y_j^t, z_j^t)$. Hence, $X \in \mathbb{R}^{3\times T\times J}$. Each skeleton sequence $X$ implicitly corresponds to an emotion label $Y \in \{1, \ldots, C\}$, where $C$ is the number of emotion categories. The goal of unsupervised body skeleton-based emotion recognition is to train an encoder using a set of unlabeled skeleton data $X$ that contains emotional information, enabling the encoder to effectively extract emotional representations.

**Overview.** We propose a self-supervised emotion recognition framework called the Appendage-Informed Redundancy-ignoring (AIR) discrete label framework, as illustrated in Fig. 2. To tackle the challenge of extracting multi-scale appendage information from body skeleton-based data and to ensure that the encoder captures more comprehensive emotional representations, we propose the Appendage-Skeleton Partitioning (ASP) module (Section III-A). The ASP module extracts multi-scale appendage information from the original body skeleton data. These appendage information input into the encoder in both Joint, Bone, or Motion ($J|B|M$) and Joint, Bone, and Motion ($J\&B\&M$) skeleton formats. To abstract high-level emotion semantics, we propose the Appendage-refined Multi-scale Discrete Label (AMDL) module (Section III-B). In this framework, the encoder extracts feature map $\mathbb{M}$ clusters from the multi-skeletons to generate and assign hard discrete labels $Y_h$ to the corresponding skeletons. This approach transforms encoder training into a classification task, simplifying the overall process. To further mitigate the noise introduced during clustering and reduce the effect of redundant skeletal information, we propose the Appendage Label Refinement (ALR) module (Section III-B2). The ALR

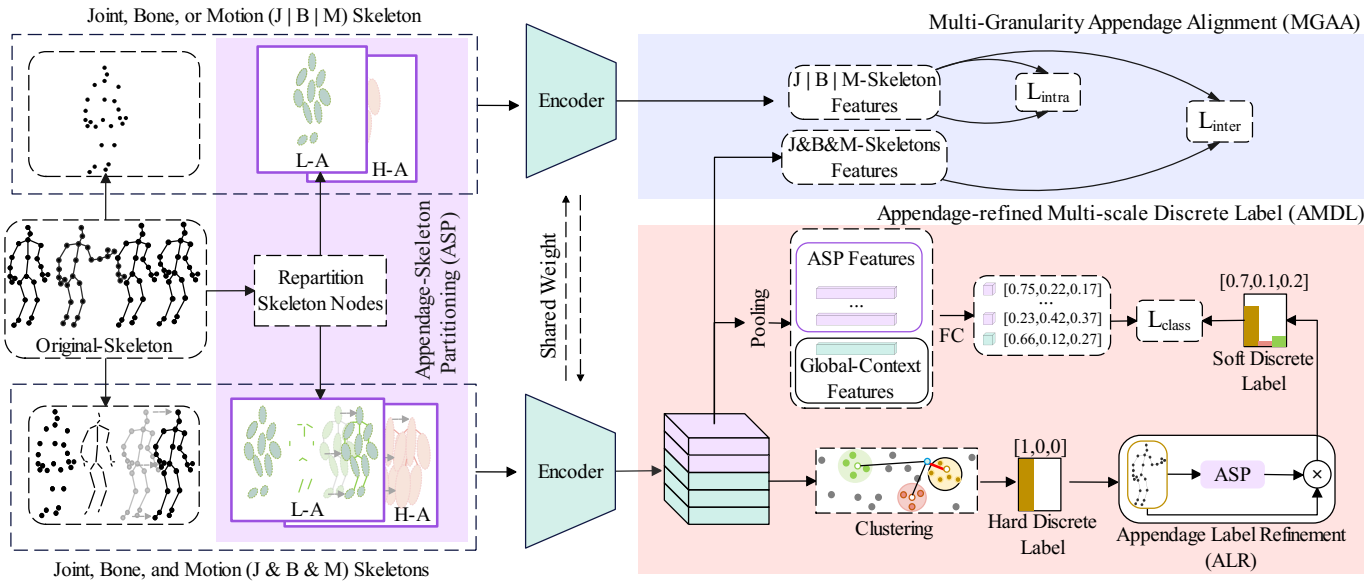


Fig. 2. The overall framework of the proposed AIR is illustrated as follows. Given an input sequence $X$, the Appendage-Skeleton Partitioning (ASP) module is applied to obtain $X_{L-A}$ and $X_{H-A}$. Subsequently, both $J|B|M$ and $J\&B\&M$ skeleton representations of this data are fed into the encoder. Using the multi-skeleton data generates a feature map, which is used for clustering to produce hard labels (with the brown category in the figure as an example). Then, the Appendage Label Refinement (ALR) module is employed to refine these hard labels. Simultaneously, to enhance the encoder's robustness, the Multi-Granularity Appendage Alignment is utilized to align the macro-scale appendage data with the original skeleton data.

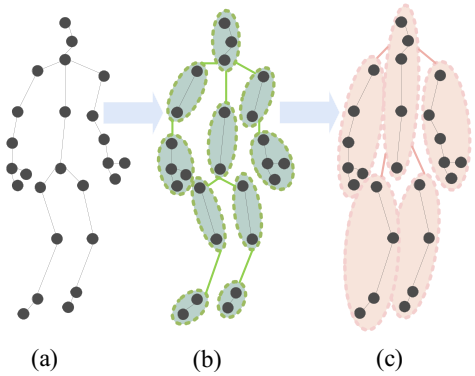


Fig. 3. Macro-scale appendage change process. (a) is the original skeleton data, (b) is the Low-Appendage (LA) connection, and (c) is the High-Appendage (HA) connection.

refines soft labels $Y_s$ by calculating the correlation between the skeleton and its corresponding High-Appendage (HA) data $X_{HA}$. Finally, to ensure alignment and consistency between the $J|B|M$ and $J\&B\&M$ skeleton data, we propose Multi-Granularity Appendage Alignment (MGAA) (Section III-C). This step enhances the encoder's robustness and stability. After the encoder is pre-trained, we use some evaluation protocols (i.e., the linear evaluation protocol, transfer evaluation protocol, and so on.) to verify the emotion representation extraction ability of the encoder.

### A. Appendage-Skeleton Partitioning Module

In the field of body skeleton-based emotion recognition, many studies tend to restrict the use of original skeleton ( i.e., joints, bones, or motions) information. However, expressing emotions through body language typically involves a holistic process that engages multiple appendages of the body. Beyond the discriminative information provided by the coordinates of each joint, latent movement information among different appendages of the human body also provides abundant useful cues for emotion recognition [67], [68]. For instance, performing a "happy" emotion involves concurrent interactions among arms, legs, and the torso [69]. Motivated by [70], the Appendage-Skeleton Partitioning (ASP) module addresses this challenge and enables the framework to systematically unearth information on the connections among appendages as a supplement to existing skeleton-based information.

To facilitate access to the connection relationships between parts and limbs, we reclassify the constituent nodes within the skeletal data, as illustrated in Fig. 3. It sequentially processes connections within and across body parts and limbs, obtaining more comprehensive connection information from a global perspective. Specifically, Fig. 3(a) showcases typical skeletal data that are composed of $J$ nodes ($X \in \mathbb{R}^{3\times T\times J}$), whereas Fig. 3(b) and Fig. 3(c) depict skeletal data constituted by Low-Appendage (LA) ($X_{LA} \in \mathbb{R}^{3\times T\times 10}$) and High-Appendage (HA)($X_{HA} \in \mathbb{R}^{3\times T\times 5}$), respectively. Using this module, we can get the corresponding Multi-Scale Appendage information.

### B. Appendage-refined Multi-scale Discrete Label Framework

To more effectively capture the high-dimensional affective semantic information embedded in a large number of unlabeled skeletons, we propose a self-monitoring approach based on discrete labels. The details of our method are presented below.

*1) Discrete Label Generation:* The discrete labels-based method enables direct training of the encoder by generating discrete hard labels $Y_h$ from unlabeled data, allowing the encoder to effectively extract emotional representations from

the skeleton data. Given $N$ unlabeled skeleton-based data $X$, a two-stage training scheme is alternately adopted in each training generation: (1) generating hard labels $Y_h$ via clustering the features of the unlabeled training skeleton-based instances, (2) training the encoder $F(\cdot)$ with the discrete hard labels. Before training the encoder with the clustering-based framework, we initialize the generation discrete labels using the DBSCAN [71] algorithm. In skeleton-based paper such as [66] follows the above clustering process, but some issues need to be addressed. The density-based clustering algorithm may produce outliers during the clustering process. These outliers can impact the final classification results, leading to significant deviations between the generated labels and the actual distribution [72].

*2) Suppress the effects of skeleton redundancy information:* As illustrated in Fig. 2, the $X$, $X_{LA}$, and $X_{HA}$ input encoder $F(\cdot)$ to get the feature map $\mathbb{M}$, and using the $\mathbb{M}$ to generate the initialization hard label $Y_h$. However, the feature map $\mathbb{M}$, focusing mainly on the overall context, can occasionally neglect specifics related to appendage features, and certain features may contain information irrelevant to body skeleton-based emotion. To address the above issue, inspired by [65], we propose an Appendage Label Refinement (ALR) module for clustering, as shown in Fig. 4. The module aims to minimize the distance between the generated labels and the actual distribution. We first perform K-Nearest Neighbor (KNN) classification to select the top-1 samples independently for the skeleton-based input data $X$ and the High-Appendage information $X_{HA}$, which is generated by the ASP module and then calculate the consistency score $S$.

Given $R(X,k)$ and $R(X_{H-A},k)$ as the sets of indices for the top-1 samples in the ranked list of KNN classification for $X$ and $X_{HA}$, respectively. This process is shown as follows:

$$S(X,X_{HA}) = \frac{|R(X,k) \cap R(X_{HA},k)|}{|R(X,k) \cup R(X_{HA},k)|}, S \in (0,1). \quad (1)$$

Utilizing the obtained consistency score $S$ mitigates noise in the hard label $Y_h$ generates the final soft label $Y_s$ and enhances the final recognition accuracy.

Learning all features $\mathbb{M}$ under identical pseudo-labels invariably leads to confusion among features, which adversely affects the final recognition accuracy [73]. To address this problem, the AMDL employs the consistency score to refine the pseudo-label, thereby improving recognition performance. Throughout this process, our proposed method applies cross-entropy loss to assess prediction accuracy and utilizes Kullback-Leibler (KL) divergence to ensure stability during the pre-training phase. The loss function for $X_{HA}$ features are defined as:

$$\begin{aligned} L_{HA} = \frac{1}{N_{HA}} \sum_{n=1}^{N_{HA}} (S \times CE(Y_s, y_{HA})) \\ +(1-S) \times D_{KL}(u||y_{HA}), \end{aligned} \quad (2)$$

where $N_{HA}$ represents the number of appendage-based skeletons, the $u$ denotes a uniform vector, and $CE$ and $D_{KL}$ denote the cross-entropy loss and KL divergence, respectively.

Meanwhile, we enhance the labels with more accurate information by aggregating predictions of High-Appendage skeleton features, with different weights assigned to each cross-agreement score. The loss function for this process is defined as:

$$L = -\sum_{i=1}^{N}((S \times y_{HA}) + (1-S) \times y) \cdot log(Y_s), \quad (3)$$

where $N$ represents the number of skeletons in the dataset. Based on Eq.2 and 3, the classification loss is:

$$L_{class} = L_{HA} + L. \quad (4)$$

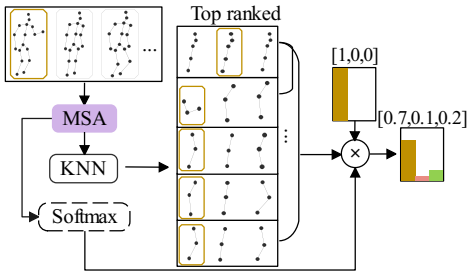


Fig. 4. Refined hard label. Utilizing the ASP module to extract High-Appendage (HA) information enables us to further minimize deviations from the true labels during the clustering process.

### C. Multi-Granularity Appendage Alignment

To ensure the harmonization between macro-scale appendage information and original skeleton data, we propose the multi-granularity appendage alignment to enhance the robustness of the encoder. Inspired by [74], [75], we utilize intra- and inter-scale semantic alignment regularization to learn more representative and discriminative features. The intra-scale semantic loss is the mean squared error between projected features of skeleton-based data samples, which is defined as:

$$L_{intra} = MSE(y^{single}, y^{multi}), y \in \{y_J, y_B, y_M\}, \quad (5)$$

where $MSE(\cdot)$ is the mean squared error and $y^{single}$ and $y^{multi}$ are the projected features of $X^{single}$ and $X^{multi}$. Notice, the $X^{single}$ (i.e., $X$, $X_{L-A}$, and $X_{H-A}$ contain $J|B|M$ information) and $X^{multi}$ (i.e., $X$, $X_{L-A}$, and $X_{H-A}$ contain $J\&B\&M$ information) include input skeleton-based data. The intra-scale semantic loss function enables the identification of more subtle differences between $J|B|M$ and $J\&B\&M$ information, effectively reducing the distance between samples of the same category.

The objective of the inter-scale semantic loss is to maximize the separation between data of different categories, thereby clarifying boundaries between these categories. The inter-scale semantic loss is defined as:

$$L_{inter} = MSE(y^{single}, \mathbb{N}), y \in (y_J, y_B, y_M), \quad (6)$$

where $\mathbb{N}$ represents the average value of the combined information of $y_J, y_B, and y_M$. Based on Eqs.4, 5, and 6, the overall loss of the AIR is:

$$L = L_{class} + L_{inra} + L_{inter}. \quad (7)$$

**Algorithm 1:** Training and Evaluation Process of AIR framework.

**Require:** Set the parameters required for the AIR framework $\alpha$
  **while** i $\leq$ Epochs **do**
    **if** i $\leq$ warmup **then**
      1. Update the encoder using Eq.5 and Eq.6 to ensure that it has the fundamental capability to extract features.
    **else if** i $>$ warmup **then**
      2. Utilize the ASP module to obtain comprehensive skeleton features, as shown in Section III-A.
      3. The AMDL module is employed to generate discrete labels for the data, while employing Eq.1 removes redundant details contained in the data.
      4. Update the encoder simultaneously using Eq.7.
    **end if**
  **end while**
  Validate the effectiveness and robustness of the pre--trained encoder through downstream tasks.

## IV. Experiments

### A. Datasets

**Emilya** [76] dataset contained 8260 samples of body movements expressing emotions. Eleven actors were asked to express eight distinct emotions in the context of seven daily actions, including Joy, Anger, Panic Fear, Anxiety, Sadness, Shame, Pride and Neutral. All data in this dataset is recorded using the Xsens MVN system. The system can capture 28 3D joints at a frame rate of 120 Hz.

**EGBM** [77] contains 560 samples captured by the Kinect V2 camera, each with a frame rate of 30 Hz. The data included 16 professional actors performing seven different emotions, including Happiness, Sadness, Neutral, Anger, Disgust, Fear and Surprise. Each emotion category contains 80 samples, each consisting of 3D coordinates provided by 25 joints.

**KDAE** [78] dataset is recorded by a portable wireless motion-capture system that can capture 72 joint node data at a frame rate of 125 Hz. The dataset consists of 1402 samples performed by 22 actors expressing seven emotions, namely Happiness, Sadness, Neutral, Anger, Disgust, Fear, and Surprise. Since nearly half of the 72 joint node species included in the sample are hand nodes. We pay attention to the whole body movement process, and hand nodes are directly excluded, and only 24 nodes are retained for analysis in our manuscript.

### B. Evaluation Protocol

**Linear evaluation protocol.** Initially, the encoder's weights are learned through self-supervised training. Subsequently, the encoder's weights are frozen, and a linear classifier is attached. Finally, the entire model is trained with labelled data to achieve the ultimate recognition accuracy.

**Transfer learning evaluation protocol.** The encoder is initially trained on a source dataset. Then, the pre-trained encoder, after being attached to a linear classifier, undergoes fine-tuning on a target dataset.

**Semi-supervised evaluation protocol.** First, the encoder uses the entire dataset for pre-training. Then, after being attached to a linear classifier, it is fine-tuned with a specific portion of labelled data.

TABLE I
Top-1 accuracy (%) comparisons with state-of-the-art methods using the linear evaluation protocol for body skeleton-based ($J\&B\&M$ skeleton) emotion recognition on datasets Emilya, EGBM, and KDAE. The $Emilya^*$ is a subset of the original dataset, which contains only four categories (i.e., anger, neutral, joy, and sadness)

| Method | Emilya | EGBM | KDAE | $Emilya^*$ |
|---|---|---|---|---|
| *supervised* | | | | |
| ST-GCN | 68.70 | 28.44 | 24.20 | 65.98 |
| AGCN | 79.78 | 28.15 | 37.27 | - |
| *self-supervised* | | | | |
| CrosSCLR | 20.52 | 45.87 | 44.84 | 66.5 |
| SCD Net | 61.57 | 37.62 | 39.86 | - |
| Skeleton Contrast | 60.23 | 34.86 | 33.45 | - |
| CMD | 57.98 | 38.53 | 35.94 | - |
| UmURL | 67.05 | 45.87 | 40.93 | - |
| SSAL | - | - | - | 77.34 |
| AIR | **69.67** | **56.88** | **55.51** | **81.55** |

### C. Implementation Details

All experiments are carried out using two RTX 4090 GPUs. In our setup, we select a simple Transformer, which comprises 1 layer and 1 head, as the feature extractor for the encoder. The training of the AIR uses the Adam optimizer with a weight decay of $5\times10^{-4}$. The mini-batch size is 128, and the learning rate is initially $5 \times 10^{-4}$ and subsequently decreased to $5 \times 10^{-5}$ at epoch 350 for Emilya, EGBM, and KDAE. The model undergoes pre-training for 450 epochs on these three emotion datasets. We randomly split the dataset into training and testing sets at a $4 : 1$ ratio. For DBSCAN [71], the $eps$ distance threshold is set to 0.6, and the hyperparameters for Jaccard distance, $k_1$ and $k_2$, are set to 30 and 6, respectively. It is worth noting that the hyperparameters of DBSCAN follow the settings specified in [65].

### D. Experimental Results

In this section, we evaluate the proposed AIR against the latest state-of-the-art methods across three key downstream tasks, as listed in Section 4.2. Through comprehensive comparative analyses, demonstrates superior performance compared to existing methods.

*1) Linear Evaluation Results:* To explore the classification performance of the AIR, we compare the top-1 accuracy of various methods using the linear evaluation protocol on different datasets, whose results are shown in Table I. Given the relatively limited scope of body skeleton movement-based emotion recognition (i.e., SSAL [15]), we have also included skeleton-based action recognition methods (i.e., ST-GCN [79], AGCN [80], CrosSCLR [59], SCD Net [81], Skeleton Contrast [82], CMD [60], and UmURL [75]) for comparison. Our method outperforms other methods in almost all cases in the linear evaluation protocol on Emilya, EGBM, and KDAE.

TABLE II
TOP-1 ACCURACY (%) COMPARISONS WITH STATE-OF-THE-ART METHODS USING THE TRANSFER-LEARNING EVALUATION PROTOCOL FOR BODY SKELETON MOVEMENT-BASED EMOTION RECOGNITION. THE TARGET DATASET IS KDAE.

| Method | Transfer to KDAE | |
|---|---|---|
| | Emilya | EGBM |
| CMD | 18.86 | 19.57 |
| SCD Net | 34.52 | 27.76 |
| Skeleton Contrast | 32.74 | 36.65 |
| UmURL | 38.43 | 30.25 |
| AIR | **43.77** | **38.08** |

TABLE III
TOP-1 ACCURACY (%) COMPARISONS WITH STATE-OF-THE-ART METHODS USING THE SEMI-SUPERVISED EVALUATION PROTOCOL FOR BODY SKELETON MOVEMENT-BASED EMOTION RECOGNITION ON THE EGBM DATASET.

| Method | 1% | 5% | 10% | 50% |
|---|---|---|---|---|
| CrosSCLR | **29.36** | 31.19 | 38.53 | 43.12 |
| SCD Net | 17.43 | 18.35 | 20.19 | 28.44 |
| CMD | 22.02 | 27.52 | 25.69 | 33.95 |
| Skeleton Contrast | 19.27 | 20.18 | 21.10 | 32.11 |
| UmURL | 25.69 | 29.36 | 33.03 | 39.45 |
| AIR | 26.61 | **35.78** | **39.45** | **52.29** |

For example, the results of our method surpass those of UmURL [75] and CrosSCLR [59] by about 2.62, 11, and 10 percentage points, respectively on the three data sets. It is worth noting that the sample size of EGBM and KDAE is smaller than that of Emilya. Still, our method performs better on the two datasets, indicating that our method has more advantages in processing small datasets. Meanwhile, AIR improved by 4.2 percentage points on the $Emilya^*$ dataset. Compared with the SSAL method, indicating that our method is more accurate in capturing body skeleton movement-based emotion features. This significant improvement demonstrates that the AIR framework effectively addresses self-supervised emotion recognition tasks via category classification. The notable increase in accuracy indicates that AIR can more efficiently extract high-level emotional semantic features from skeleton data, thereby enhancing the encoder's capability for emotional feature extraction.

### *E. Transfer Learning Evaluation Results*

To investigate the generalization capability of the AIR in the context of transfer learning, we compare its performance with those of other state-of-the-art methods using the transfer-learning evaluation protocol, whose results are presented in Table II. The source dataset for pre-training is either Emilya or EGBM, and the target dataset for fine-tuning is KDAE. When Emilya and EGBM as source datasets, the performance of the AIR surpasses other methods. This suggests that AIR, as a self-supervised method, can be utilized more effectively for pre-training using data that differ from the downstream tasks. The final recognition accuracy demonstrates that, compared with other methods, our approach can extract emotional features at a higher level of abstraction, thereby enhancing the robustness of the encoder during the self-supervised process.

### *F. Semi-supervised Evaluation Results*

Finally, using the semi-supervised evaluation protocol, we compare the AIR with other state-of-the-art methods to explore its ability to learn with few labels. Table III shows results on the EGBM dataset, with 1%, 5%, 10%, and 50% of the labeled data randomly sampled for fine-tuning. In the case of the semi-supervised evaluation protocol with 1% labeled data used for fine-tuning, the result of the AIR is significantly higher than those of others except for being slightly lower than that of CrosSCLR. This may be because labels generated from too little data during the clustering process introduce more noise, leading to significant discrepancies from the true categories and ultimately impairing the encoder's ability to extract accurate representations. Nonetheless, a comparative analysis reveals that our method substantially outperforms other approaches in nearly all scenarios. This finding indicates that, even under semi-supervised conditions, our approach maintains a strong ability to capture high-level affective semantics, thereby delivering superior classification results.

TABLE IV
THE ABLATION STUDY OF TOP-1 ACCURACY (%) RESULTS ON THE EGBM DATASET WHEN DIFFERENT COMBINATIONS OF FUNCTIONAL MODULES ARE ACTIVATED.

| Baseline | ASP | AMDL | AMDL + ALR | Accuracy |
|---|---|---|---|---|
| ✓ | | | | 45.87 |
| ✓ | ✓ | | | 47.71 |
| ✓ | | ✓ | | 46.79 |
| ✓ | | | ✓ | 52.29 |
| ✓ | ✓ | | ✓ | 56.88 |

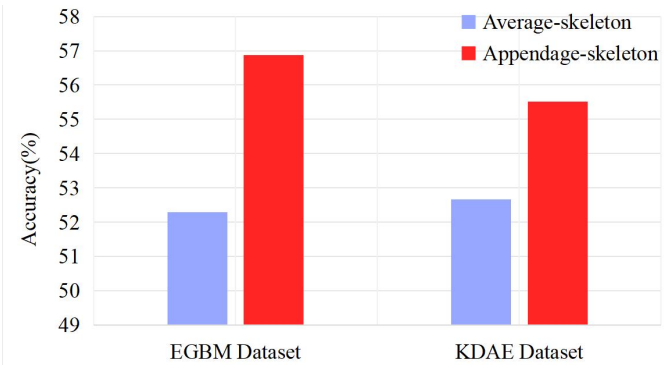


Fig. 5. Effects of appendage-based partitioning on recognition accuracy. In the ALR module, the influence of labels generated by different skeleton data partitioning on the ability of the final encoder to extract emotional representation.

## V. ABLATION STUDIES AND ANALYSIS

The ablation study evaluates the effectiveness of the ASP, AMDL, and ALR modules of the AIR. The results when different combinations of these modules are activated shown in Table IV. The baseline model is a variant of the proposed AIR when all these modules are deactivated.

### *A. Effectiveness of the appendage connection information*

We first investigate the effect of the ASP module to explore the effectiveness of appendage connection information. The ASP module enhances the baseline model by providing

TABLE V
THE ABLATION STUDY OF THE DIFFERENT GRANULARITY OF APPENDAGE LEVELS FEATURE ON THE EGBM DATASET. THE $Baseline^*$ INDICATES THE BASELINE MODEL OF ADDING THE ASP MODULE. DIFFERENT SCALE APPENDAGE INFORMATION IS USED, WITH "LA", "HA", AND "LH" REPRESENTING THE LOW-APPENDAGE SKELETON INFORMATION, HIGH-APPENDAGE SKELETON INFORMATION, AND A COMBINATION OF 'LA' AND 'HA' SKELETON INFORMATION.

| $Baseline^*$ | LA | HA | LH | Accuracy (%) |
|---|---|---|---|---|
| ✓ | | | | 47.71 |
| ✓ | ✓ | | | 48.90 |
| ✓ | | ✓ | | 53.46 |
| ✓ | | | ✓ | 56.88 |

additional information on the connections among human appendages besides the original skeleton-based information, resulting in a 2 percentage points increase in the final recognition accuracy compared to cases when the ASP module is deactivated. This demonstrates that in body skeleton-based emotion recognition, human appendage connection information plays a significant role in accurately identifying corresponding emotion categories.

### B. Validity of Multiple Appendage Scales

We further examine the effectiveness of different appendage scales by the ASP module, whose results are shown in Table V. Compared to the $baseline^*$ result with only original skeleton data, the LA and HA skeleton data increase the emotion recognition accuracy by $1.1$ and $4.5$ percentage points, respectively. The encoder's ability to extract emotional representations is further enhanced by incorporating various appendage information. Our findings indicate that incorporating both LA and HA information simultaneously is more effective than processing them separately. This underscores the crucial role of body information in addressing emotion recognition tasks based on skeletal data.

### C. Effectiveness of the Discrete label-based AMDL

Next, we examine the effectiveness of the AMDL. This module offers a novel perspective on presentation learning by transforming the conventional SSL framework into a process that directly predicts the emotional categories of skeleton data. Through this transformation, the emotional semantics inherent in skeleton data can be captured more effectively, thereby enhancing the encoder's ability to recognize emotions. As indicated in Table IV, the use of the clustering-based framework contributes a $1$ percentage points improvement in recognition performance compared to the baseline model. Additionally, compared to other methods in Table I, our approach achieves state-of-the-art performance using only the discrete label-based module. These findings suggest that the SSL method based on discrete tags can more effectively extract the high-level affective semantics embedded in the skeleton data, thereby improving overall emotion recognition performance.

Meanwhile, for the ALR module, we discussed whether to use the appendage information provided by the ASP module, as shown in Fig. 5. We use both equalized local skeleton

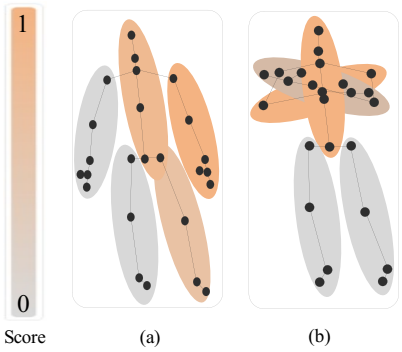


Fig. 6. Visualization of the score of the High-Appendage (HA) skeleton features. The gray limbs are the useless body information.

information and macro-scale appendage information to refine the categories generated by the clustering process. Comparisons show that the refinement based on appendage information more closely aligns with the true category distribution, significantly improving the final emotion recognition accuracy (e.g., by $4.5$ percentage points on the EGBM dataset). Fig. 6 also proves that different appendages contribute differently to the final clustering category, and some appendage information may even interfere with the final clustering category.

## VI. CONCLUSION

In this article, we propose an Appendage-Informed Redundancy-ignoring (AIR) discrete label framework to extract high-level emotional information embedded in large volumes of unlabeled skeleton data collected by smart cameras in IoT environments. The framework introduces three new submodules in addition to the base structure. First, we introduce the Macro-Scale Appendage (ASP) module, which incorporates appendage-scale information from the original skeleton data, enabling the encoder to learn more comprehensive emotional representations. Next, we propose an Appendage-refined Multi-scale Discrete Label (AMDL) module. Because high-level emotional semantics embedded in skeleton data are more significant than subtle joint movements in emotion recognition tasks, unlike previous approaches based on contrastive or generative learning for pre-training, the discrete label-based method encourages SSL models to capture high-level semantics in a manner akin to human perception. To further reduce nonessential information in skeleton data that may negatively affect the generation of accurate emotional categories, an Appendage Label Refinement (ALR) module is introduced, which exploits the complementary relationship between the appendage hierarchy features extracted by the ASP module and the original skeleton features, thereby mitigating label noise. Lastly, we have proposed a Multi-Granularity Appendage Alignment (MGAA) to align different scale appendage information during pre-training, enhancing the robustness of the encoder. Extensive experiments conducted on three datasets demonstrate the effectiveness of the proposed AIR framework, particularly under linear evaluation protocols. Additionally, transfer evaluation protocols confirm that the AIR method can further enhance the encoder's transferability,

offering a novel approach for pre-training encoders with large-scale skeleton-based unlabeled emotion data collected by smart cameras in IoT environments.

## REFERENCES


[1] O. Aouedi, T.-H. Vu, A. Sacco, D. C. Nguyen, K. Piamrat, G. Marchetto, and Q.-V. Pham, "A survey on intelligent internet of things: Applications, security, privacy, and future directions," *IEEE communications surveys & tutorials*, 2024.

[2] A. Rejeb, K. Rejeb, A. Appolloni, S. Jagtap, M. Iranmanesh, S. Alghamdi, Y. Alhasawi, and Y. Kayikci, "Unleashing the power of internet of things and blockchain: A comprehensive analysis and future directions," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 1–18, 2024.

[3] M. Mohammadpour, H. Khaliliardali, S. M. R. Hashemi, and M. M. AlyanNezhadi, "Facial emotion recognition using deep convolutional networks," in *2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI)*. IEEE, 2017, pp. 0017–0021.

[4] F. W. Smith and M. L. Smith, "Decoding the dynamic representation of facial expressions of emotion in explicit and incidental tasks," *Neuroimage*, vol. 195, pp. 261–271, 2019.

[5] H. Zhang and M. Xu, "Multiscale emotion representation learning for affective image recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 2203–2212, 2023.

[6] M. Yang, Y. Gao, L. Tang, J. Hou, and B. Hu, "Wearable eye-tracking system for synchronized multimodal data acquisition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 5146–5159, 2024.

[7] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-gcn: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 3793–3804, 2020.

[8] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *2011 IEEE workshop on applications of signal processing to audio and acoustics (Waspaa)*. IEEE, 2011, pp. 65–68.

[9] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," *arXiv preprint arXiv:1906.01267*, 2019.

[10] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[11] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2018.

[12] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1541–1554, 2022.

[13] R. E. Nisbett and T. D. Wilson, "Telling more than we can know: Verbal reports on mental processes." *Psychological review*, vol. 84, no. 3, p. 231, 1977.

[14] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[15] C. Song, L. Lu, Z. Ke, L. Gao, and S. Ding, "Self-supervised gait-based emotion representation learning from selective strongly augmented skeleton sequences," *arXiv preprint arXiv:2405.04900*, 2024.

[16] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[17] S. Xu, J. Fang, X. Hu, E.-H. Ngai, Y. Guo, V. Leung, J. Cheng, and B. Hu, "Emotion recognition from gait analyses: Current research and future directions," *Cornell University - arXiv,Cornell University - arXiv*, Mar 2020.

[18] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 505–523, 2021.

[19] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, 2012.

[20] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[21] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8877–8886.

[22] B. Li, C. Zhu, S. Li, and T. Zhu, "Identifying emotions from non-contact gaits information based on microsoft kinects," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 585–591, 2018.

[23] A. Crenn, K. Ahmed, A. Meyer, and S. Bouakaz, "Body expression recognition from animated 3d skeleton," *IEEE Conference Proceedings,IEEE Conference Proceedings*, Jan 2016.

[24] T. Randhavane, A. Bera, K. Kapsaskis, U. Bhattacharya, K. Gray, and D. Manocha, "Identifying emotions from walking using affective and deep features," *Cornell University - arXiv,Cornell University - arXiv*, Jun 2019.

[25] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[26] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "Step: Spatial temporal graph convolutional networks for emotion perception from gaits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 02, 2020, pp. 1342–1350.

[27] V. Narayanan, B. M. Manoghar, V. Sashank Dorbala, D. Manocha, and A. Bera, "Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8200–8207.

[28] C. Hu, W. Sheng, B. Dong, and X. Li, "Tntc: Two-stream network with transformer-based complementarity for gait-based emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3229–3233.

[29] H. Lu, X. Hu, and B. Hu, "See your emotion from gait using unlabeled skeleton data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1826–1834.

[30] G. Paoletti, C. Beyan, and A. Del Bue, "Graph laplacian-improved convolutional residual autoencoder for unsupervised human action and emotion recognition," *IEEE Access*, vol. 10, pp. 131 128–131 143, 2022.

[31] W. Sheng, X. Lu, and X. Li, "Mldt: Multi-task learning with denoising transformer for gait identity and emotion recognition," in *Proceedings of the 2021 4th Artificial Intelligence and Cloud Computing Conference*, 2021, pp. 47–52.

[32] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[33] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[34] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[35] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, "A comprehensive survey on contrastive learning," *Neurocomputing*, p. 128645, 2024.

[36] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[37] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[38] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," *arXiv preprint arXiv:2208.06366*, 2022.

[39] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *arXiv preprint arXiv:2208.10442*, 2022.

[40] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification," *arXiv preprint arXiv:2106.03069*, 2021.

[41] M. P. Murray, A. B. Drought, and R. C. Kory, "Walking patterns of normal men," *JBJS*, vol. 46, no. 2, pp. 335–360, 1964.

[42] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 1, pp. 1–31, 2016.

[43] M. Daoudi, S. Berretti, P. Pala, Y. Delevoye, and A. Bimbo, "Emotion recognition by body movement representation on the manifold of

symmetric positive definite matrices," *Cornell University - arXiv,Cornell University - arXiv*, Jul 2017.

[44] N. Fourati, C. Pelachaud, and P. Darmon, "Contribution of temporal and multi-level body cues to emotion classification," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 116–122.

[45] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Skeleton-based gait index estimation with lstms," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/icis.2018.8466522

[46] X. Sun, K. Su, and C. Fan, "Vfl—a deep learning-based framework for classifying walking gaits into emotions," *Neurocomputing*, p. 1–13, Feb 2022. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2021.12.007

[47] H. Lu, S. Xu, S. Zhao, X. Hu, R. Ma, and B. Hu, "Epic: Emotion perception by spatio-temporal interaction context of gait," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 5, p. 2592–2601, May 2024. [Online]. Available: http://dx.doi.org/10.1109/jbhi.2022.3233597

[48] W. Sheng and X. Li, "Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network," *Pattern Recognition*, p. 107868, Jun 2021. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2021.107868

[49] J. Shi, C. Liu, C. T. Ishi, and H. Ishiguro, "Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network," *Sensors*, p. 205, Dec 2020. [Online]. Available: http://dx.doi.org/10.3390/s21010205

[50] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. [Online]. Available: http://dx.doi.org/10.1109/iccv.2019.00156

[51] M. Noroozi and P. Favaro, *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles*, Jan 2016, p. 69–84. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46466-4_5

[52] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, May 2018.

[53] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2016.278

[54] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. [Online]. Available: http://dx.doi.org/10.1109/cvpr42600.2020.00975

[55] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *Cornell University - arXiv,Cornell University - arXiv*, Feb 2020.

[56] M. Tagliasacchi, B. Gfeller, F. de Chaumont Quitry, and D. Roblek, "Pre-training audio representations with self-supervision," *IEEE Signal Processing Letters*, vol. 27, pp. 600–604, 2020.

[57] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[58] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition," *Information Sciences*, p. 90–109, Aug 2021. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2021.04.023

[59] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. [Online]. Available: http://dx.doi.org/10.1109/cvpr46437.2021.00471

[60] Y. Mao, W. Zhou, Z. Lu, J. Deng, and H. Li, "Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation," Aug 2022.

[61] W. Wu, Y. Hua, C. Zheng, S. Wu, C. Chen, and A. Lu, "Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition," in *2023 IEEE international conference on multimedia and expo workshops (ICMEW)*. IEEE, 2023, pp. 224–229.

[62] R. Xu, L. Huang, M. Wang, J. Hu, and W. Deng, "Skeleton2vec: A self-supervised learning framework with contextualized target representations for skeleton sequence," *arXiv preprint arXiv:2401.00921*, 2024.

[63] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1. Minneapolis, Minnesota, 2019, p. 2.

[64] G. Yixiao, D. Chen, F. Zhu, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id." *Cornell University - arXiv,Cornell University - arXiv*, Jun 2020.

[65] Y. Cho, W. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification."

[66] M. Wang, X. Li, S. Chen, X. Zhang, L. Ma, and Y. Zhang, "Learning representations by contrastive spatio-temporal clustering for skeleton-based action recognition," *IEEE Transactions on Multimedia*, 2023.

[67] H. Rao, C. Leung, and C. Miao, "Hierarchical skeleton meta-prototype contrastive learning with hard skeleton mining for unsupervised person re-identification," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 238–260, 2024.

[68] S. Yang, J. Liu, S. Lu, E. M. Hwa, and A. C. Kot, "One-shot action recognition via multi-scale spatial-temporal skeleton matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[69] D. A. Winter, *Biomechanics and motor control of human movement*. John wiley & sons, 2009.

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

[71] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[72] X. Zhang, Y. Ge, Y. Qiao, and H. Li, "Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3436–3445.

[73] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7308–7318.

[74] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[75] S. Sun, D. Liu, J. Dong, X. Qu, J. Gao, X. Yang, X. Wang, and M. Wang, "Unified multi-modal unsupervised representation learning for skeleton-based action understanding," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2973–2984.

[76] N. Fourati and C. Pelachaud, "Perception of emotions and body movement in the emilya database," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, p. 90–101, Jan 2018. [Online]. Available: http://dx.doi.org/10.1109/taffc.2016.2591039

[77] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, and G. Anbarjafari, "Multimodal database of emotional speech, video and gestures," in *Pattern Recognition and Information Forensics: ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, August 20-24, 2018, Revised Selected Papers 24*. Springer, 2019, pp. 153–163.

[78] M. Zhang, L. Yu, K. Zhang, B. Du, B. Zhan, S. Chen, X. Jiang, S. Guo, J. Zhao, Y. Wang, B. Wang, S. Liu, and W. Luo, "Kinematic dataset of actors expressing emotions," *Scientific Data*, Sep 2020. [Online]. Available: http://dx.doi.org/10.1038/s41597-020-00635-7

[79] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[80] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[81] C. Wu, X.-J. Wu, J. Kittler, T. Xu, S. Ahmed, M. Awais, and Z. Feng, "Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5949–5957.

[82] F. M. Thoker, H. Doughty, and C. G. Snoek, "Skeleton-contrastive 3d action representation learning," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 1655–1663.