frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 05 July 2023
DOI 10.3389/fnins.2023.1225312

OPEN ACCESS

EDITED BY
Fudong Nian,
Hefei University, China

REVIEWED BY
Jianjun Sun,
Northwestern Polytechnical University, China
Shimeng Yang,
Anhui University, China

*CORRESPONDENCE
Ziliang Ren
✉ renzl@dgut.edu.cn

RECEIVED 19 May 2023
ACCEPTED 12 June 2023
PUBLISHED 05 July 2023

CITATION
Yang H, Ren Z, Yuan H, Xu Z and Zhou J (2023) Contrastive self-supervised representation learning without negative samples for multimodal human action recognition. *Front. Neurosci.* 17:1225312. doi: 10.3389/fnins.2023.1225312

Contrastive self-supervised representation learning without negative samples for multimodal human action recognition

Huaigang Yang[1], Ziliang Ren[1,2]*, Huaqiang Yuan[1], Zhenyu Xu[2] and Jun Zhou[1]

[1]School of Computer Science and Technology, Dongguan University of Technology, Dongguan, China, [2]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Action recognition is an important component of human-computer interaction, and multimodal feature representation and learning methods can be used to improve recognition performance due to the interrelation and complementarity between different modalities. However, due to the lack of large-scale labeled samples, the performance of existing ConvNets-based methods are severely constrained. In this paper, a novel and effective multi-modal feature representation and contrastive self-supervised learning framework is proposed to improve the action recognition performance of models and the generalization ability of application scenarios. The proposed recognition framework employs weight sharing between two branches and does not require negative samples, which could effectively learn useful feature representations by using multimodal unlabeled data, e.g., skeleton sequence and inertial measurement unit signal (IMU). The extensive experiments are conducted on two benchmarks: UTD-MHAD and MMAct, and the results show that our proposed recognition framework outperforms both unimodal and multimodal baselines in action retrieval, semi-supervised learning, and zero-shot learning scenarios.

KEYWORDS

human action recognition, multimodal representation, feature encoder, contrastive self-supervised learning, Transformer

1. Introduction

Automatic recognition framework is a research field that aims to develop systems capable of identifying and classifying human actions or behaviors, which is to enable machines to understand and interpret human behavior, with applications in areas including video surveillance, healthcare, sports analysis, and human-computer interaction (Li et al., 2016a,b; He et al., 2023). Different techniques in real life adopt different types of data inputs, but each modality has its own advantages and limitations (Sun et al., 2023). To achieve more robust and accurate feature extraction, some approaches improve the performance of models by aggregating the advantages of various modalities in a reasonable manner. Due to the success of deep learning in the past decades, a large number of ConvNets-based frameworks have made impressive achievements in the field of multimodal visual tasks (Grillini et al., 2021; Mughal et al., 2022; Li et al., 2023). However, most of them require many large amounts of labeled data, especially for multimodal data (Zhang et al., 2019, 2020), and labeling the data requires exponentially more time and effort (Li et al., 2009).

Recently, self-supervised representation learning has made significant progress on visual tasks, which is mainly divided into the pre-training and fine-tuning stages (Chen et al., 2020; Grill et al., 2020). In the pre-training stage, it focuses on constructing feature representations of different views by unlabeled samples. In the fine-tuning stage, these representations are used as inputs and fed into a small-scale linear classifier, which requires only a small amount of labeled data. Moreover, contrastive learning is one of the self-supervised learning, where the core concept is to pull the representation distance between positive samples closer and push the distance away from other negative samples. For example, the CMC framework (Tian et al., 2020) is mainly to form positive samples between different data modalities, and consider other different samples as negative sample pairs. Due to the problem of relying too much on negative sample pairs, it is necessary to set a large batch size or a queue for storing negative samples in the learning process, therefore leads to a complex model and is vulnerable to information collapse.

In order to overcome the above shortcomings, inspired by Barlow Twins and VICReg (Zbontar et al., 2021; Bardes et al., 2022), we propose a contrastive self-supervised learning framework for unimodal and multimodal without relying on negative samples. Our proposed method employs multimodal samples as input data, e.g., skeleton sequence and inertial measurement unit signal (IMU). The main contributions of this paper are as follows:

- A unimodal contrastive self-supervised framework is proposed to encode and learn feature representations for multimodal action recognition with skeleton sequence and IMU data.
- The proposed recognition framework is extended to multimodal contrastive self-supervised learning. The model is designed to obtain simple and efficient feature representations without negative samples.

The remainder of this paper is organized as follows. Section 2 presents an overview of related works. In Section 3, we provided a detailed introduction to the proposed method. Section 4 provides experimental results for benchmark datasets and comparisons with state-of-the-art. Section 5 concludes this paper and look forward to future work.

2. Related works

In this section, we discuss unimodal, multimodal, and contrastive learning methods for human action recognition from the perspective of input data modality.

2.1. Unimodal human action recognition

Unimodal human action recognition primarily focuses on classifying and recognizing actions by using a single modality, including RGB videos, depth and skeleton sequences, IMU data, etc. This field encompasses tasks such as feature extraction, feature representation, and the construction of deep learning models, including convolution neural networks (CNNs) (Andrade-Ambriz et al., 2022; Islam et al., 2022; Xu et al., 2022), recurrent neural networks (RNNs) (Shu et al., 2021; Shen and Ding, 2022; Wang et al., 2022), graph convolution networks (GCNs) (Cheng et al., 2020; Chi et al., 2022; Feng et al., 2022; Tu et al., 2022) and Transformer models (Chen and Ho, 2022; Mazzia et al., 2022; Ahn et al., 2023).

Since the skeleton sequence would not be sensitive to viewpoint variation and circumstance disturbance, there are numerous skeleton-based methods is developed for human action recognition. In CNN-based methods, Li et al. (2018) proposed an end-to-end convolutional co-occurrence feature learning framework from the perspectives of intra-frame representation and inter-frame representation of skeleton temporal evolutions, which introduced a global spatial aggregation method and discarded the local aggregation approach. In RNN-based methods, Xie et al. (2018) aimed to address the issue of skeleton variations in 3D spatiotemporal space, which proposed a spatiotemporal memory attention network based on RNN and CNN to perform frame recalibration of skeleton data in the temporal domain. Regarding GNN-based methods, Yan et al. (2018) emerged as a classic approach based on spatial-temporal graph convolution networks. The core idea was to model human body joints as graph nodes and the connections between joints as graph edges, and the multiple graph convolutional layers were stacked to extract high-level spatial-temporal features. In Transformer-based methods, Plizzari et al. (2021) model employed a spatial self-attention module to capture intra-frame interactions among different body parts and a temporal self-attention module to model inter-frame correlations.

For IMU data, due to its ability to provide good complementary features and better privacy protection, it is gradually being used for human action recognition tasks. Through convolutional layers and pooling layers, CNN (Yi et al., 2023) were able to capture local and global features in IMU data, extract relationships between skeleton body parts, and achieve accurate classification of different actions. In IMU-based human action recognition, RNN (Al-qaness et al., 2022) utilized their memory units (e.g., Long Short-Term Memory Units or Gated Recurrent Units) to capture the temporal evolution of skeleton sequence, extracting crucial motion patterns and action features from it. Additionally, there have been research efforts that combined the strengths of CNNs and RNNs to comprehensively utilize the spatiotemporal information in IMU data for human activity recognition (Challa et al., 2022; Dua et al., 2023). It is worth noting that, with the progress of research, other IMU-based human action recognition methods have emerged, such as those based on Transformers (Shavit and Klein, 2021; Suh et al., 2023).

2.2. Multimodel human action recognition

Due to the limitation of single modal, it is difficult to further improve the performance of recognition model.

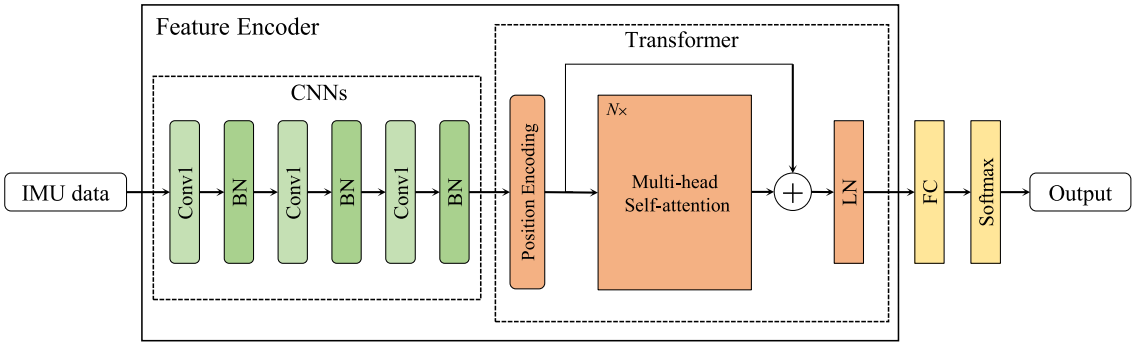

FIGURE 1
The feature encoder for IMU data. "BN" denotes batch normalization, "LN" indicates layer normalization, and $N\times$ represents that there are multiple multi-head self-attention modules.

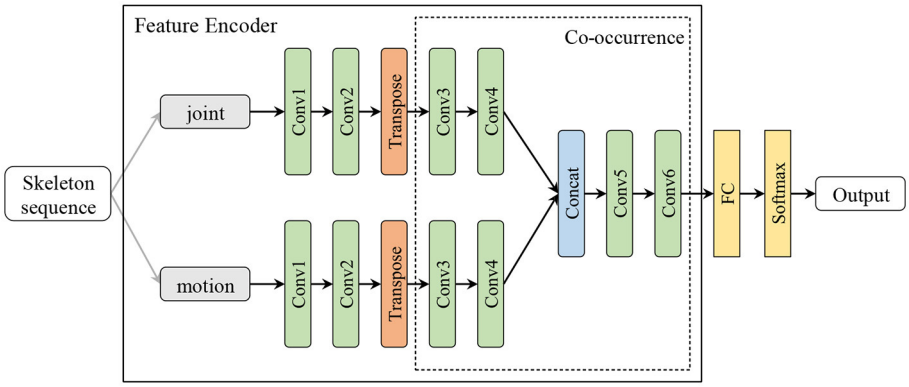

FIGURE 2
The feature encoder for skeleton sequence. The output channels of the 6 blocks 2D convolution layer are [64, 32, 32, 64, 128, 256]. The transpose layer transposes the dimensions of the input tensor according to the sequential parameters.

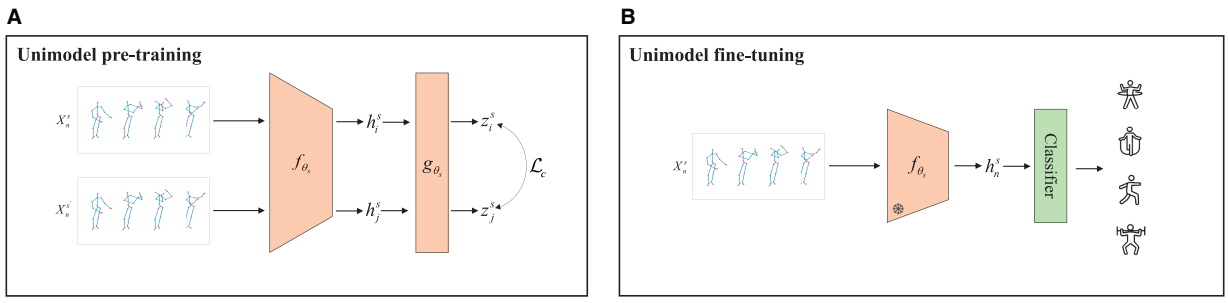

FIGURE 3
Contrastive learning framework for unimodal recognition. **(A)** Pre-training stage: for a skeleton sequence, the embedding representation z is generated by the same encoder f and projection head g after data augmentation using contrast loss L_c, respectively. **(B)** Fine-tuning stage: the labeled skeleton sequence is passed through the frozen encoder f, and then processed through the classifier to obtain the action recognition label.

Since the complementary information provided by different modalities, researchers have become interested in combining multimodal features to improve recognition performance, such as skeleton and IMU data (Das et al., 2020; Khaertdinov and Asteriadis, 2022). There are many excellent recognition models are developed to leverage the strengths of different modalities and achieve more robust and accurate action recognition. However, the main challenge in executing multimodal recognition

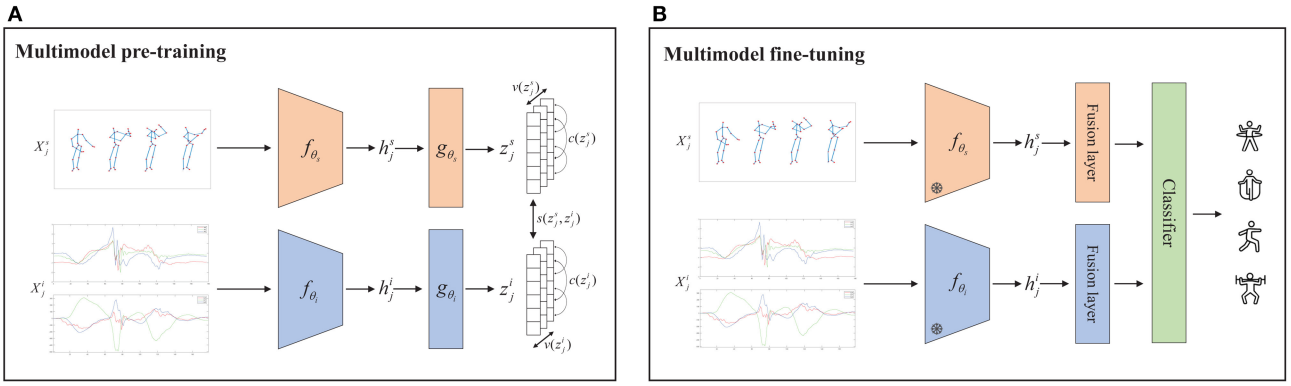

FIGURE 4
Contrastive learning framework for multimodal recognition. **(A)** Pre-training stage: the non-labeled skeleton sequences and IMU data were passed through modality-specific encoders *f* and projection heads *g* to generate embeddings representing *z* using contrasting loss. **(B)** Fine-tuning phase: the skeleton sequences with labels and the IMU data with labels were passed through frozen modality-specific encoders *f*, respectively, and then obtained action recognition label via fusion layer and classifier.

lies in effectively fuse the feature information from different modalities. Based on the above statement, the related work in multi-modal human action recognition can be roughly categorized into modality fusion and feature fusion, and we focus on the fusion method of skeleton sequence and IMU signal features.

Skeleton data provides precise positional information of human joints, while IMU data provides measurements from sensors such as accelerometers and gyroscopes (Das et al., 2020). By fusing skeleton and IMU data, more comprehensive and rich action features can be obtained. From the perspective of modality fusion, Fusion-GCN (Duhme et al., 2022) directly integrates IMU data into existing skeletons in the channel dimension during data preprocessing. Furthermore, RGB modality is processed to extract high-level semantic features, which are then fed into the GCNs as new nodes for fusion with other modalities. From the perspective of feature fusion (Khaertdinov and Asteriadis, 2022), features from different modalities are combined and integrated to achieve more representative and discriminative representations. In addition, cross-modal contrastive learning networks through knowledge distillation are also an effective identification method. Liu et al. (2021) proposed a Semantics-aware Adaptive Knowledge Distillation Network (SAKDN) that utilizes IMU data and RGB videos as inputs for the teacher and student model, respectively. The SAKDN adaptively fuses knowledge from different teacher networks and transfers the trained knowledge from the teacher network to the student network. The CMC (Tian et al., 2020) framework proposed a multi-modal learning architecture based on contrastive representation learning, which extended the representation learning to multiple modalities for improving the quality of the learned features with the number of modalities increased. It demonstrated the subtle relationship between mutual information across multiple modalities and multiple viewpoints. Similarly, CMC-CMKM (Brinzea et al., 2022) employed cross-modal knowledge distillation to perform feature-level fusion of IMU data and Skeleton information, which has achieved good recognition performance.

2.3. Contrastive learning for human action recognition

Recently, several advanced self-supervised learning methods have been proposed with excellent results in image and video tasks. Self-supervised contrast learning focuses on the variation between different views of the same or different samples, and better robust and transferable feature representations can be learned through contrast loss. SimCLR (Chen et al., 2020) incorporated a new contrastive loss function called Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent) into the network, which is a simple and effective contrastive learning framework. In contrast, BYOL (Grill et al., 2020) designed a more scalable and easily trainable self-supervised learning approach by contrasting the hidden representations in the network. Furthermore, to obtain more distinctive representations without requiring negative samples, Barlow Twins (Zbontar et al., 2021) minimized the correlation between features by employing the Barlow Twins loss. In addition, the biggest advantage of VICReg (Bardes et al., 2022) is its simplicity and effectiveness, which only necessary to compare along the batch dimension by invariance, variance and covariance, and does not require the weights of two branches to be shared.

In the case of action recognition tasks, most of the self-supervised contrastive learning is mainly applied to individual modalities, such as sensor data, skeleton sequence, or RGB video. To date, there has been a large number of works on fully supervised learning for multimodal human action recognition, and the disadvantage of these methods is that they require a large number of labeled samples for training. In contrary, to our knowledge, self-supervised contrastive learning frameworks

```
# f: encoder network
# lambda, mu, nu: coefficients of the
invariance, variance and covariance
 losses
# N: batch size
# D: dimension of the representations
# mse_loss: Mean square error loss function
# off_diagonal: off-diagonal elements
 of a matrix
# relu: ReLU activation function
for x_j^s, x_j^i in loader: # load a batch with
N samples
  # obtain augmented skeleton and
  # IMU samples
  x̃_j^s = 𝒯(x_j^s)
  x̃_j^i = 𝒯(x_j^i)
  # compute representations
  h_j^s = (f_θs(x̃_j^s)) # hidden layer feature
  h_j^i = (f_θs(x̃_j^i)) # hidden layer feature
  z_j^s = g_θs(h_θi)
  # embeddings for skeleton [N × D]
  z_j^i = g_θi(h_θi) # embeddings for IMU [N × D]
  # variance loss
  z_j^s = z_j^s − z_j^s.mean(dim = 0)
  z_j^i = z_j^i − z_j^i.mean(dim = 0)
  std_z_j^s = torch.sqrt(z_j^s.var(dim = 0) + 1e − 04)
  std_z_j^i = torch.sqrt(z_j^i.var(dim = 0) + 1e − 04)
  std_loss = torch.mean(relu(1 − std_z_j^s))
  +torch.mean(relu(1 − std_z_j^i))
  # invariance loss
  sim_loss = mse_loss(z_j^s, z_j^i)
  # covariance loss
  cov_z_j^s = (z_j^s.T @ z_j^s)/(N − 1)
  cov_z_j^i = (z_j^i.T @ z_j^i)/(N − 1)
  cov_loss = off_diagonal(cov_z_j^s).pow_(2).sum()/D
  +off_diagonal(cov_z_j^i).pow_(2).sum()/D
  # total loss
  loss = lambda ∗ sim_loss + mu ∗ std_loss
  +nu ∗ cov_loss
  # optimization step
  loss.backward()
  optimizer.step()
```

Algorithm 1. Multimodal pre-training pytorch pseudocode.

are rarely used in the field of multimodal human action recognition. Akbari et al. (2021) adopted a convolution-free Transformer architecture to train unlabeled video, audio, and text data end-to-end, and evaluated the model performance through downstream tasks such as video action recognition, audio event classification, image classification, and text-to-video retrieval. Inspired by VicReg (Bardes et al., 2022) and multimodal framework CMC, we propose a simple and effective self-supervised contrastive learning framework based on VICReg to address the multimodal human action recognition problem of IMU and skeleton data.

3. Methodology

3.1. Problem definition

Multimodal-based action recognition is defined as the fusion of different data modalities to obtain more comprehensive human pose and more precise action information. Specifically, for a given input $\{X^m | m \in M\}$ from a multimodal set M, the goal is to predict the label $y \in Y$ with the associated input X. In our work, we focus on IMU signal data and Skeleton sequences. IMUs could be used to measure the pose and acceleration of the human body with multivariate time series on the x, y and z axes for human motion recognition and analysis. Specifically, for S wearable sensors with S signal channels acquired at any t time stamp, we can define the input signal as $\boldsymbol{x_t} = \left[x_t^1, x_t^2, \ldots, x_t^S\right] \in \mathbb{R}^S$. Therefore, the IMU modal inputs are represented in matrix form as $X^i = [x_1, x_2, ..., x_T] \in \mathbb{R}^{T \times S}$ for any T time stamp. Furthermore, skeleton sequences can be collected by a pose estimation algorithm or a depth camera, which contain several joints of a human body, and each joint has multiple position coordinates. For a given skeleton sequence $X^s \in \mathbb{R}^{C \times T \times V}$, as 2D coordinates are used, the input channel $C = 2$, T denotes the number of frames in a sequence, and V means that the number of joints with respect to the dataset collection method.

3.2. Feature encoder

In order to obtain more effective features, we designed two feature encoders to handle IMU data and skeleton sequence, respectively, as shown in Figures 1, 2. In IMU data feature encoder, inspired by CSSHAR (Khaertdinov et al., 2021), we first employ a 1D convolution layer with 3 blocks for modeling in the temporal dimension, which includes a convolution kernel size of 3 and a feature map with channels of [32,64,128]. Furthermore, we employ a Transformer with a Multi-head self attention (heads $N = 2$) as the backbone to capture long-range dependencies from IMU data. Besides, inspired by hierarchical co-occurrence feature learning strategy, a two-stream framework is designed to learn and fuse the "joint" and "motion" features of skeleton sequences. Specifically, a skeleton sequence is divide into spatial joints and temporal motions. Then, they are fed into each of the four 2D CNN modules and assembled into semantic representations in both spatial and temporal domains, and point-level information of each joint is encoded independently.

3.3. Contrastive learning for unimodal recognition

As shown in Figure 3, given a skeleton sample in the pre-training, a positive sample pair X_n^s and $X_n^{s'}$ could be obtained in a small batch by normal data augmentation. Then, they are fed into an encoder f_{θ_s} with HCN to yield the hidden layer features as

$$h_i^s = f_{\theta_s}(X_n^s) \tag{1}$$

$$h_j^s = f_{\theta_s}(X_n^{s'}) \tag{2}$$

TABLE 1 Pre-training hyperparameter settings.

Modality	UTD-MHAD			MMAct		
	Learning rate	Training scale	Batch size	Learning rate	Training scale	Batch size
IMU	1e-2	100 epochs	128	1e-3	100 epochs	96
Skeleton	1e-2	100 epochs	128	1e-3	100 epochs	96
IMU+Skeleton	1e-3	200 epochs	256	1e-4	200 epochs	128

TABLE 2 The performance of action recognition for accuracy (%) and F1 score (%) is compared with the baseline methods.

Method	Modality	UTD-MHAD		MMAct cross-subject		MMAct cross-scene	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Supervised_transformer	IMU	79.77	79.59	62.15	62.32	78.27	71.86
Supervised_cooccurrence	Skeleton	93.49	93.43	80.53	81.93	78.61	74.30
SimCLR	IMU	64.65	64.64	52.32	51.94	66.16	60.28
SimCLR	Skeleton	92.09	91.87	75.97	76.75	72.62	62.04
Barlow Twins	IMU	58.60	57.69	45.17	44.11	59.96	51.77
Barlow Twins	Skeleton	88.84	88.24	67.86	69.24	60.68	52.34
Barlow Twins	IMU+Skeleton	91.63	91.72	82.17	81.98	82.70	80.05
CMC	IMU+Skeleton	95.12	95.08	82.05	<u>83.06</u>	84.01	82.41
CMC-CMKM§	IMU+Skeleton	95.81	95.74	<u>82.34</u>	82.69	85.24	83.60
Ours	IMU	75.58	75.93	49.04	47.08	60.81	53.80
Ours	Skeleton	86.05	86.23	73.78	75.66	74.94	73.29
Ours	IMU+Skeleton	<u>96.06</u>	**96.96**	**82.95**	**83.62**	<u>87.06</u>	<u>85.78</u>
Supervised	IMU+Skeleton	**96.51**	<u>96.36</u>	81.78	82.86	**89.47**	**87.94**

Bolded data indicate the best results, underlined data the second best. § represents the reproduced results.

Inspired by the Barlow Twins, the feature representations z_i^s and z_j^s are obtained by an MLP projection layer, which are denoted as

$$z_i^s = g_{\theta_s}(h_i^s) \tag{3}$$

$$z_j^s = g_{\theta_s}(h_j^s) \tag{4}$$

Finally, to explore the relationship between the two views X_n^s and $X_n^{s'}$, the cross-correlation matrix $\mathscr{C}$ between embedding z_i^s and z_j^s can be computed as follows

$$\mathscr{C}_{ij} = \frac{\sum_b z_{b,i} z'_{b,j}}{\sqrt{\sum_b \left(z_{b,i}\right)^2} \sqrt{\sum_b \left(z'_{b,j}\right)^2}}, \tag{5}$$

where b denotes the batch dimension, i and j represent the embedding dimension. Finally, by enforcing the empirical cross-correlation matrix between the embeddings Z^s of variations to be an identity matrix, the encoder could be used to capture the relationship between the two-stream siamese networks. The contrastive loss function is formulated as follows

$$L_c(Z^s) = \sum_i \left(1 - \mathscr{C}_{ii}\right)^2 + \beta \sum_i \sum_{j \neq i} \mathscr{C}^2{}_{ij} \tag{6}$$

Intuitively, the first term encourages the diagonal elements of $\mathscr{C}$ to converge to 1, so that the embedding is not subject to variation. The second term is intended to drive the different embedding components to be independent of each other, minimizing the redundancy of the output units and avoiding becoming a constant. β is a positive constant used to weigh the first term and against the second term.

3.4. Contrastive learning for multimodal recognition

Our proposed VICReg-based multimodal recognition framework focuses on generating and contrasting embeddings from the IMU data and skeleton sequence branches, which eventually form a joint embedding architecture with variance, invariance and covariance regularization. It is a self-supervised learning method that incorporates two different modality training architectures based on the principle of preserving the content of the embedding information.

As shown in Figure 4, given a multimodal training sample $\{x_j^s, x_j^i\}$, where s and i refer to skeleton and IMU data modalities respectively. The augmented inputs are

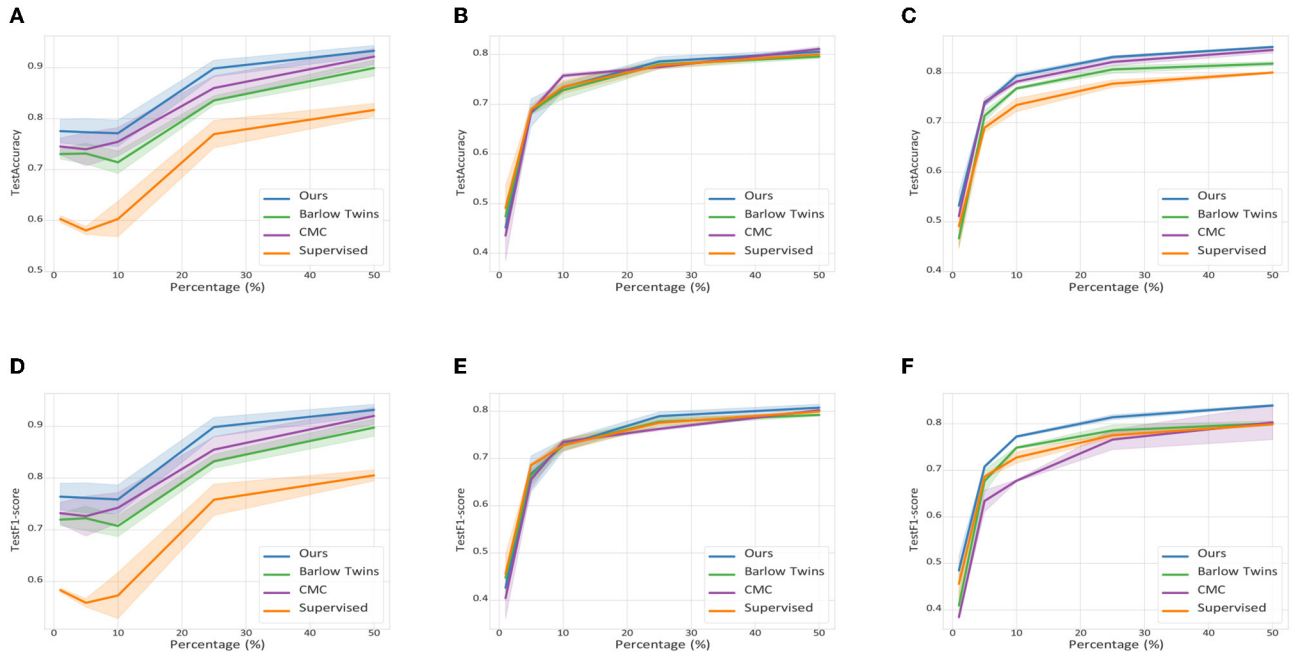

FIGURE 5
Average accuracy and F1 score with 95% confidence intervals for the semi-supervised learning scenario. **(A)** UTD-MHAD. **(B)** MMAct (cross-subject). **(C)** MMAct (cross-scene). **(D)** UTD-MHAD. **(E)** MMAct (cross-subject). **(F)** MMAct (cross-scene).

generated by modality-specific data augmentation in accordance with

$$x_j^s = \mathscr{T}(x_j^s) \tag{7}$$

$$\tilde{x}_j^i = \mathscr{T}(x_j^i) \tag{8}$$

In details, for the skeleton sequence augmentation methods are jittering, scaling, rotation, shearing, cropping and resizing, whereas the IMU data augmentation methods are jittering, scaling, rotation, permutation, shuffle of channel. Then, the feature representation of the two modalities are computed. Specifically, two modality-specific encoders f_{θ_s} and f_{θ_i} perform feature extraction to obtain the high-dimensional hidden layer features.

$$h_j^s = \left(f_{\theta_s}\left(\tilde{x}_j^s\right)\right) \tag{9}$$

$$h_j^i = \left(f_{\theta_s}\left(\tilde{x}_j^i\right)\right) \tag{10}$$

Both of these are passed through projection heads g_{θ_s} and g_{θ_i}, implemented by a multilayer perceptron, and finally generate mode-specific embeddings representations of the two modalities which are $z_j^s = g_{\theta_s}\left(h_{\theta_i}\right)$ and $z_j^i = g_{\theta_i}\left(h_{\theta_i}\right)$. The loss function is calculated at the embedding level with respect to z_j^s and z_j^i. We describe the three components of variance, invariance and covariance that constitute our loss function in the pre-training process.

Firstly, we define the variance regularization term v to adopt the form of a hinge function that represents the standard deviation of the embeddings along the batch dimension.

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max\left(0, \gamma - Std\left(z^j, \epsilon\right)\right), \tag{11}$$

where Std denotes the regularization standard deviation formula as:

$$Std(x, \epsilon) = \sqrt{\mathrm{Var}(x) + \epsilon}, \tag{12}$$

where we define $Z = [z_1, ..., z_n]$ consisting of n vectors of dimension d with embeddings z_j from the feature encoding network of two modalities. z^j is represented as the value of each vectors in Z in dimension j, γ denotes a fixed value of the standard deviation and defaults to 1 in our experiments. ϵ is a small scalar to guarantee data stability, which is set to 0.0001. The objective of this regularization term $v(Z)$ is to ensure that the variance of all embeddings Z^s and Z^i are close to γ in the current batch (s indicates the skeleton modality and i indicates the IMU modality), preventing all inputs from mapping on the same vector.

Secondly, we define the invariance regularization term s by using the mean square Euclidean distance between two positive sample pairs Z^s and Z^i. The formulation is as follows:

$$s(Z^s, Z^i) = \frac{1}{N}\sum_{j}^{N} \left\| z_j^s - z_j^i \right\|_2^2, \tag{13}$$

where N denotes the batch size, both embeddings Z^s and Z^i come from the siamese architecture of the two branches.

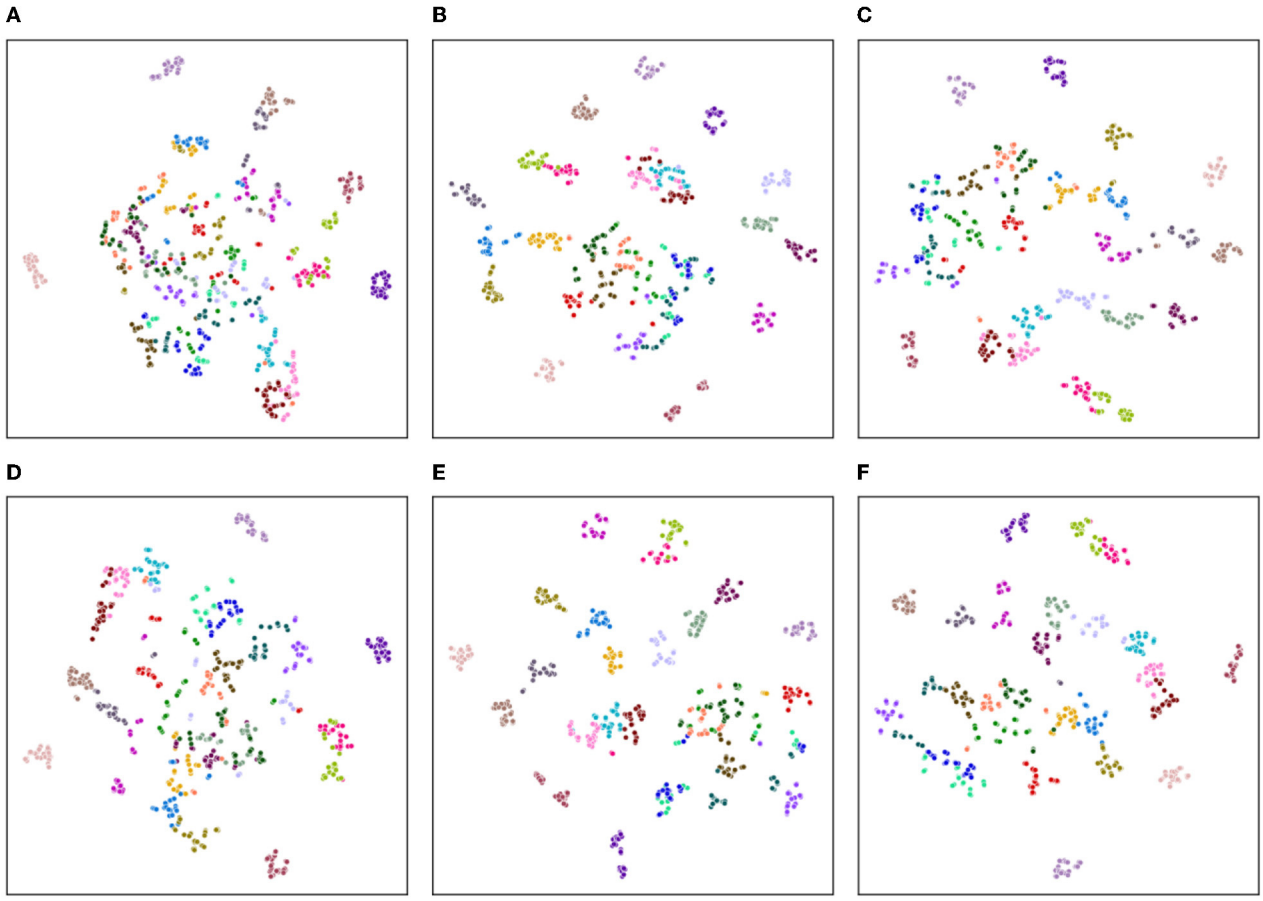

FIGURE 6
Visualization of representations learned using t-SNE for the UTD-MHAD benchmark. **(A)** Barlow Twins (IMU). **(B)** Barlow Twins (Skeleton). **(C)** Barlow Twins (IMU+Skeleton). **(D)** Ours (IMU). **(E)** Ours (Skeleton). **(F)** Ours (IMU+Skeleton).

Finally, the most critical component of the loss function, this term approximates the covariance between each pair of embedding variables to zero. Generally, it is the embeddings of the model that are decorrelated to each embedding variable to ensure the independence of the variables and prevent the model from learning similar or identical feature information. Inspired by Barlow Twins, we define the variance regularization term c as:

$$c(Z) = \frac{1}{d}\sum_{i\neq j}\left[C(Z)\right]_{i,j}^{2}, \tag{14}$$

where the $1/d$ scales this function at the dimensional level and $C(Z)$ denotes the covariance matrix of the embeddings Z. The formula is expressed as follows:

$$C(Z) = \frac{1}{N-1}\sum_{j=1}^{n}\left(z_j - \bar{z}\right)\left(z_j - \bar{z}\right)^{T}, \bar{z} = \frac{1}{N}\sum_{j}^{N} z_j. \tag{15}$$

Therefore, the overall loss function with weighted average of the invariance, variance and covariance terms could be expressed as follows:

$$L(Z^s, Z^i) = \lambda * s(Z^s, Z^i) + \mu * \left[v(Z^s) + v(Z^i)\right] + \varphi * \left[c(Z^s) + c(Z^i)\right], \tag{16}$$

where λ, μ, and φ are hyperparameters that measure the importance of each loss component. In our experiment, φ is set to 1 and a grid search is performed for the values of λ and φ with the basic condition $\lambda = \varphi > 1$.

The pseudo-code algorithm implementation is illustrated in Algorithm 1.

4. Experiments

4.1. Datasets

UTD-MHAD (Chen et al., 2015). The dataset is a multimodal dataset widely used for human action recognition, which includes RGB video, depth sequences, skeleton and IMU data. During the capturing process, 8 subjects perform 27 categories of actions, each individual repeating each action 4 times, for a total of

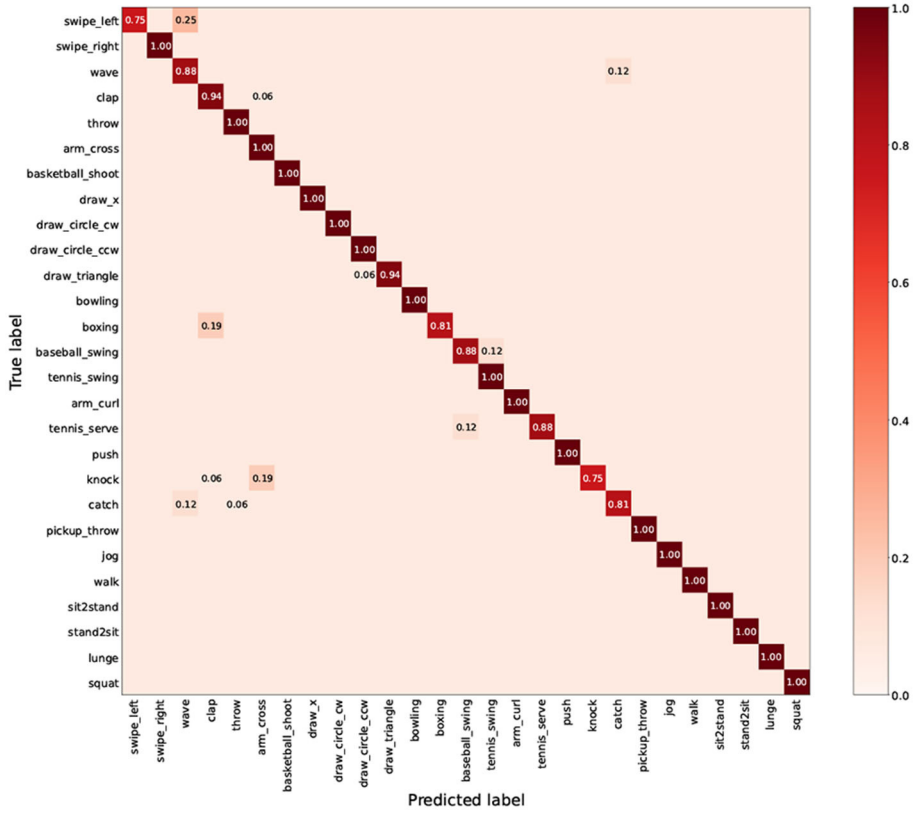

FIGURE 7
The normalized confusion matrix for UTD-MHAD.

861 samples. For the skeleton sequences, the Kinect camera would capture information regarding the subject's posture and movements. For the IMU data, the subjects were required to wear gloves, shoes and belts with IMU sensors attached, which recorded motion information on the subject's body parts, including accelerations, angular velocities and gyroscope data. Similar to the evaluation protocol in the original paper, we use data from odd-numbered subjects: 1, 3, 5, 7 as the training and validation sets, and data from even-numbered subjects: 2, 4, 6, 8 as the testing set, and report the accuracy and F1 score on the testing set.

MMACT (Kong et al., 2019). The dataset is a multimodal dataset consisting of 20 subjects performing 36 classes of actions, including skeleton sequences and IMU data. In this work, a challenge version of the dataset with 2D keypoints is adopted for the skeleton data. The IMU data is derived from smartphones including accelerometers, gyroscopes and orientation sensors. We verify our proposed recognition framework against the evaluation protocol from the previous study: cross-subject and cross-scene. For the cross-subject setting, the first 16 subject samples are used for training and validation, while the remaining ones are used for testing. For the cross-scene setting, the numbered 2 samples from the occlusion scene were used for testing and the rest for training, numbered 1, 3, 4. We report the accuracy and F1 score on the testing set.

4.2. Implementations details

Our experimental environment is implemented on the A5000 GPU platform using the Pytorch framework. Subsequently, we detailed three aspects: data pre-processing, pre-training and fine-tuning.

Data pre-processing. In order to normalize the IMU data and skeleton sequences, we employed a resampling method to uniformly represent all sequences with 50 time steps. Furthermore, to ensure consistency and comparability, we applied a standard normalization procedure to normalize the joints in all skeleton sequences. This normalization process involved scaling the joint positions based on the reference frame established by the first frame of each sequence. For data augmentation of skeleton sequences, we employ {jittering, random resized crops, scaling, rotation, shearing} for two benchmarks. For data augmentation of IMU data, we employ {jittering, scaling, permutation, rotation, channel shuffle}.

Pre-training. For the UTD-MHAD dataset, in unimodal pretraining, our proposed method uses a batch size of 100 and sets the random seed for both skeleton and IMU modalities to 28. The training is performed for 100 epochs with a learning rate of 1e-2 and Adam optimizer. In the case of multimodal pretraining, our proposed method increases the batch size to 200 epochs, adjusts the learning rate to 1e-3, and sets the training scale to 200 epochs. The optimizer remains Adam. For the MMAct dataset, we maintain

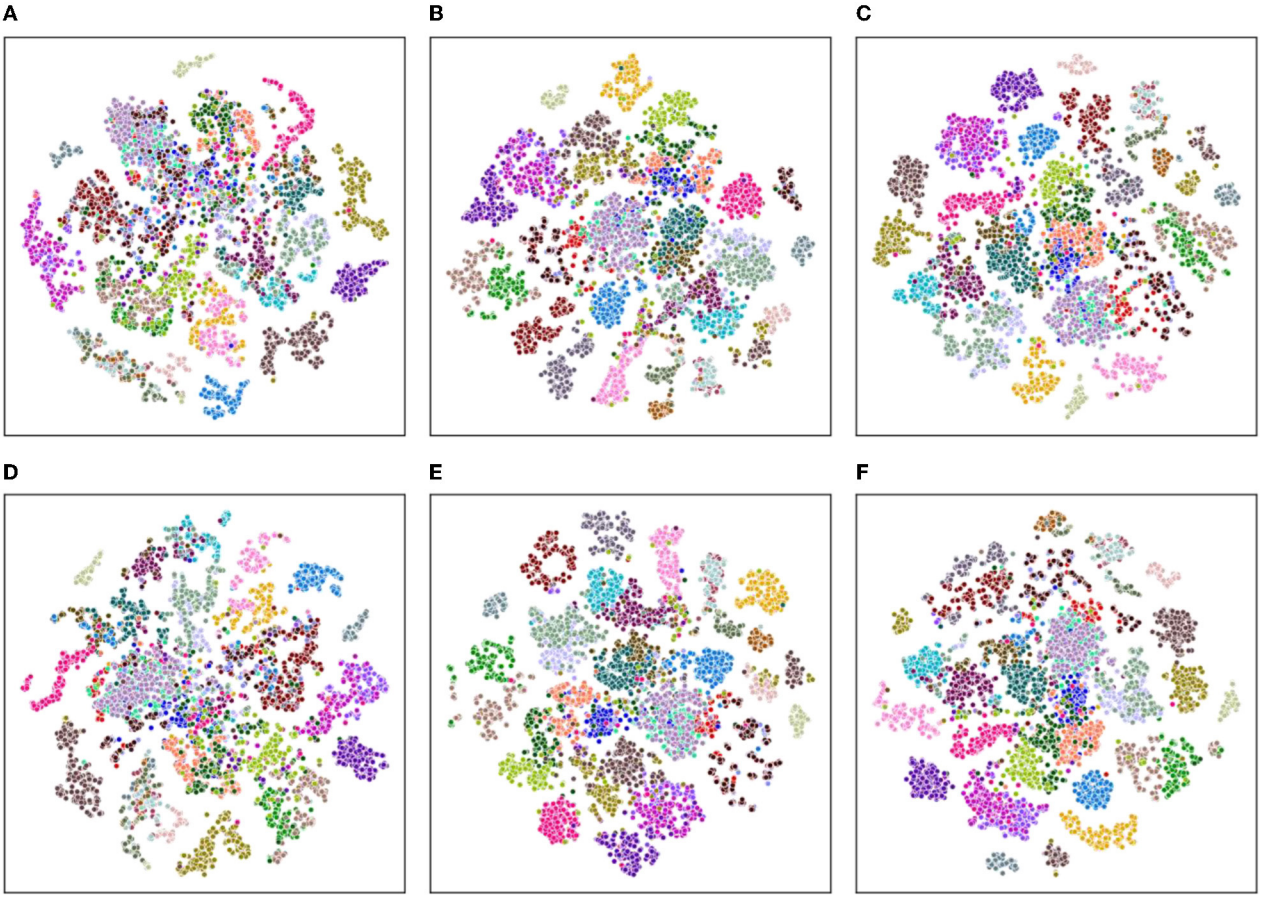

FIGURE 8
Visualization of representations learned using t-SNE for the MMAct benchmark. **(A)** Barlow Twins (IMU). **(B)** Barlow Twins (Skeleton). **(C)** Barlow Twins (IMU+Skeleton). **(D)** Ours (IMU). **(E)** Ours (Skeleton). **(F)** Ours (IMU+Skeleton).

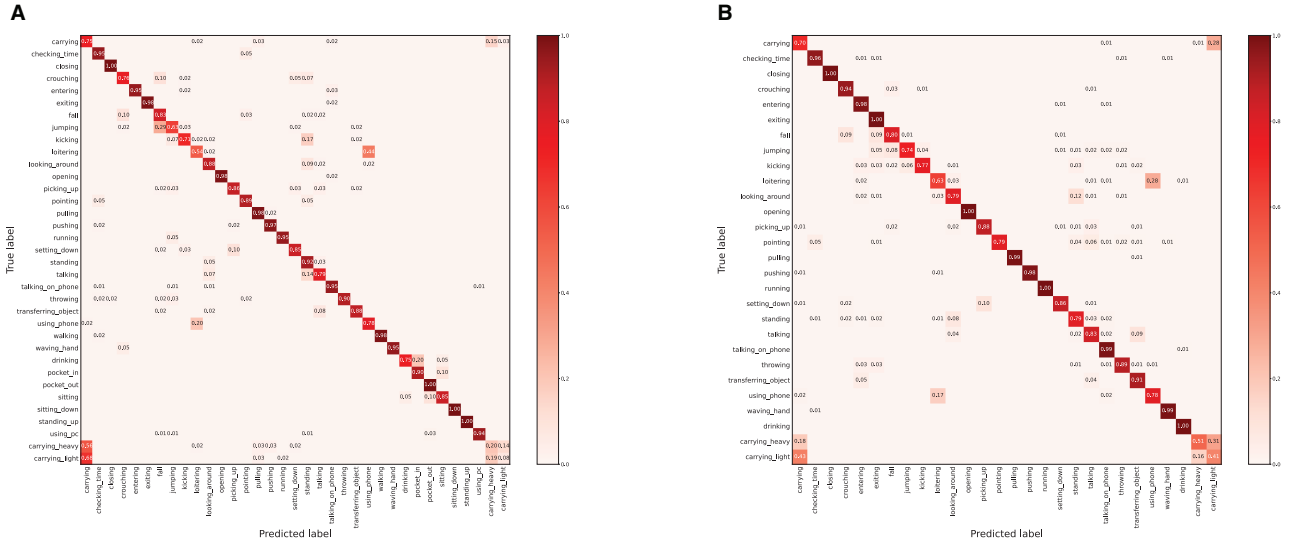

FIGURE 9
The normalized confusion matrix for MMAct benchmarks. **(A)** MMAct (cross-subject). **(B)** MMAct (cross-scene).

TABLE 3 Zero shot performance (%) on UTD-MHAD benchmark.

Modality	num_classes=1		num_classes=2		num_classes=5	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
IMU	73.95	73.95	73.49	73.79	75.35	75.54
Skeleton	88.84	88.43	87.91	87.84	89.77	89.51
IMU+Skeleton	95.58	95.59	93.95	93.85	96.05	96.00

TABLE 4 Zero shot performance (%) on MMAct benchmark.

Modality	num_classes=1		num_classes=2		num_classes=5	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
IMU	48.39	48.11	48.31	47.34	48.81	48.63
Skeleton	73.67	75.74	72.35	73.97	73.68	75.83
IMU+Skeleton	81.39	81.19	81.73	82.35	82.45	83.02

the same training settings as before, regardless of single or multi-modality. In unimodal pretraining, the learning rate is set to 1e-3, and the batch size is 96. In multimodal pretraining, we increase the batch size to 128 and adjust the learning rate to 1e-4. Similarly, the parameter initialization random seed is set to 28. All settings are shown in Table 1.

Fine-tuning. Following prior fine-tuning routines, we implemented modality-specific feature fusion layers for the multimodal fine-tuning process, including batch normalization and non-linear ReLU, mapping the embeddings of IMU data and skeleton sequence to the same size of 256. And then concatenated them up by a linear classifier with Softmax function. We train the samples with labels by fine-tuning the model both to 100 epochs either unimodal or multimodal for our action recognition task.

4.3. Evaluations

4.3.1. Learning feature representation

To evaluate the multimodal learned feature representation, we perform linear evaluation of the features extracted from a specific encoder and then input the labeled samples into the fine-tuned training encoder and linear classifier. The performance of our model is compared with existing state of the art methods, and the results as shown in Table 2.

From the accuracy and F1 score terms obtained from the linear evaluation, our method significantly outperforms unimodal (more than 20% for IMU and almost 10% for Skeleon) for two benchmarks when multimodal contrastive learning is implemented. When comparing the self-supervised learning baseline models, our method is superior to other contrastive learning methods in terms of the multimodal learning approach. However, for the unimodal learning approach, our method has relatively no advantage. It is possible that our method undergoes a certain degree of embeddings collapse when calculating the standard deviation and variance. Meanwhile, the accuracy and F1 score of our method are also slightly lower when comparing fully supervised learning, which may be due to the fact that the supervised learning approach can perform end-to-end feature extraction for specific modalities. It is worth noting that our proposed method achieves 82.95% accuracy and 83.62% F1 score for MMAct (cross-subject), which exceeds the supervised learning method by 1.17 and 0.76%, indicating that our method has a better learned feature representation for multimodal training.

4.3.2. Semi-supervised learning

In the experiments, we adopt proportional unlabeled IMU and Skeleton data to perform contrastive learning in the pre-training phase. In particular, we set a random percentage $p \in \{1\%, 5\%, 10\%, 25\%, 50\%\}$ to conduct the experiment. To obtain a reasonable fine-tuning result, we calculate the average accuracy under the evaluation protocol corresponding to that presented in the colored interval by repeating the training 10 times on each p. In addition, we train a supervised learning multimodal model using the same encoders (Transformer for IMU and Co-occurrence for Skeleton). Similarly, fine-tuning the two-stream siamese networks and performing feature fusion, the final recognition results are obtained by a linear classifier, especially noting that the weights of the encoders are randomly initialized.

As shown in Figure 5, despite training only a small number of labeled samples, the contrastive learning methods all exhibit excellent robustness and performance. Specifically, the contrastive learning based approach outperforms the supervised learning based approach when the labeled samples are less than 25%, regardless of the dataset. Besides, our proposed method is superior to both Barlow Twins and CMC contrastive learning based multimodal methods with arbitrary p values, which further validate the effectiveness and generalization ability of our proposed method.

4.3.3. Qualitative analysis

In order to evaluate the clustering effect of the model from a qualitative perspective, we employ a t-Distributed Stochastic

Neighbor Embedding (t-SNE, van der Maaten and Hinton, 2008) method to visualize the high-dimensional embeddings into a two-dimensional plane.

As shown in Figures 6, 7, we explore the IMU-based, Skeleton-based and multimodal approaches on the UTD-MHAD and MMAct datasets, respectively. Compared to the Barlow Twins, from an intuitive point of view, our proposed method is obviously effective in separating action class. Moreover, it is discovered that the multimodal data clustering is better than the unimodal clustering by fusing the features of IMU and Skeleton modalities. Furthermore, to measure the classification performance of our proposed method after fine-tuning, we performed accuracy evaluation by normalizing the confusion matrix. As shown in Figures 8, 9, we plot the normalized confusion matrices on UTD-MHAD, MMAct (cross-subject) and MMAct (cross-scene) to intuitively evaluate the performance of the classifier.

4.4. Zero shot setting

In the zero shot setting, we further explore the proposed method on the IMU and skeleton modalities through hiding certain action groups during the pre-training process. Specifically, we ensured that the action categories index [1, 2, 5] were not leaked during the training process by masking them.

As shown in Tables 3, 4, the performance of our model is compared with existing state of the art methods. Regarding UTD-MHAD benchmark for the unimodal evaluation, we could observe that the difference of the model is not significant after fine-tuning, but the skeleton sequence-based is much higher 15% than the IMU-based method. This is probably due to the fact that the skeleton sequences are modeled in both spatial and temporal dimensions, whereas IMU is only considered in the temporal dimension. For the multimodal evaluation, the model achieved 96.05% for accuracy and 96.00% for F1 score with class_id = $<$ 5 $>$ hidden, which is very close to the results achieved without the zero shot approach. Furthermore, regardless of the action class hidden, it is noted that the multimodal-based achieves much higher accuracy than the unimodal-based approach, exceeding the IMU-based approach by approximately 20% and the skeleton-based approach by approximately 6%. This validates that our proposed method achieves superior results with multimodal data inputs, which demonstrate the ability of the proposed method to learn complementary information.

5. Conclusion

In this paper, we propose a simple and effective contrastive self-supervised learning framework for human action recognition. Specifically, we construct a multimodal dataset by combining skeleton sequences and IMU signal data, and feed them into pretrained modality-specific two-stream networks for feature encoding. During the fine-tuning stage, labeled data is fed into the frozen encoders with weight initialization, and a linear classifier is applied to predict actions. Extensive experiments demonstrate that our proposed method outperforms unimodal approaches. It is worth noting that our model achieves comparable performance to pure supervised multimodal learning in certain metrics. In the future, we plan to further investigate other modalities, such as depth maps and RGB videos, to enhance multimodal human action recognition methods. Additionally, by incorporating knowledge distillation and unsupervised learning techniques, we aim to explore different ways of feature fusion between modalities to improve its performance in complex scenarios.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ZR and HYu: conceptualization and writing—review and editing. HYa: methodology and validation. ZR and ZX: software. JZ: formal analysis. HYu: resources. ZX and JZ: data curation and visualization. HYa and ZR: writing—original draft preparation. ZR: supervision and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of Guangdong Province (Nos. 2022A1515140119 and 2023A1515011307), Dongguan Science and Technology Special Commissioner Project (No. 20221800500362), Dongguan Science and Technology of Social Development Program (No. 20231800936242), and the National Natural Science Foundation of China (Nos. 61972090, U21A20487, and U1913202).

Acknowledgments

The authors thank everyone who contributed to this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ahn, D., Kim, S., Hong, H., and Ko, B. C. (2023). "Star-transformer: a spatio-temporal cross attention transformer for human action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3330–3339. doi: 10.1109/WACV56688.2023.00333

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., et al. (2021). "VATT: transformers for multimodal self-supervised learning from raw video, audio and text," in *Advances in Neural Information Processing Systems, Vol. 34,* eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Curran Associates, Inc), 24206–24221.

Al-qaness, M. A., Dahou, A., Abd Elaziz, M., and Helmi, A. (2022). Multi-resAtt: multilevel residual network with attention for human activity recognition using wearable sensors. *IEEE Trans. Indus. Inform.* 19, 144–152. doi: 10.1109/TII.2022.3165875

Andrade-Ambriz, Y. A., Ledesma, S., Ibarra-Manzano, M.-A., Oros-Flores, M. I., and Almanza-Ojeda, D.-L. (2022). Human activity recognition using temporal convolutional neural network architecture. *Expert Syst. Appl.* 191, 116287. doi: 10.1016/j.eswa.2021.116287

Bardes, A., Ponce, J., and Lecun, Y. (2022). "VICReg: variance-invariance-covariance regularization for self-supervised learning," in *ICLR 2022 - International Conference on Learning Representations*.

Brinzea, R., Khaertdinov, B., and Asteriadis, S. (2022). "Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition," in *2022 International Joint Conference on Neural Networks (IJCNN)* (Padua: IEEE), 1–8. doi: 10.1109/IJCNN55064.2022.9892522

Challa, S. K., Kumar, A., and Semwal, V. B. (2022). A multibranch cnn-bilstm model for human activity recognition using wearable sensor data. *Visual Comput.* 38, 4095–4109. doi: 10.1007/s00371-021-02283-3

Chen, C., Jafari, R., and Kehtarnavaz, N. (2015). "UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International Conference on Image Processing (ICIP)* (Quebec City, QC), 168–172.

Chen, J., and Ho, C. M. (2022). "MM-VIT: multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI), 1910–1921. doi: 10.1109/WACV51458.2022.00086

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (PMLR)*, 1597–1607.

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., and Lu, H. (2020). "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 183–192. doi: 10.1109/CVPR42600.2020.00026

Chi, H.-G., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., and Ramani, K. (2022). "InfoGCN: representation learning for human skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 20186–20196. doi: 10.1109/CVPR52688.2022.01955

Das, A., Sil, P., Singh, P. K., Bhateja, V., and Sarkar, R. (2020). MMHAR-ensemNet: a multi-modal human activity recognition model. *IEEE Sens. J.* 21, 11569–11576. doi: 10.1109/JSEN.2020.3034614

Dua, N., Singh, S. N., Semwal, V. B., and Challa, S. K. (2023). Inception inspired CNN-GRU hybrid network for human activity recognition. *Multimedia Tools Appl.* 82, 5369–5403. doi: 10.1007/s11042-021-11885-x

Duhme, M., Memmesheimer, R., and Paulus, D. (2022). "Fusion-GCN: multimodal action recognition using graph convolutional networks," in *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021* (Bonn: Springer), 265–281. doi: 10.1007/978-3-030-92659-5_17

Feng, L., Zhao, Y., Zhao, W., and Tang, J. (2022). A comparative review of graph convolutional networks for human skeleton-based action recognition. *Artif. Intell. Rev.* 55, 4275–4305. doi: 10.1007/s10462-021-10107-y

Grill, J.-B., Strub, F., Altche, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). "Bootstrap your own latent-a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems Vol. 33*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc), 21271–21284.

Grillini, A., Hernández-García, A., Renken, R. J., Demaria, G., and Cornelissen, F. W. (2021). Computational methods for continuous eye-tracking perimetry based on spatio-temporal integration and a deep recurrent neural network. *Front. Neurosci.* 15, 650540. doi: 10.3389/fnins.2021.650540

He, M., Hou, X., Ge, E., Wang, Z., Kang, Z., Qiang, N., et al. (2023). Multi-head attention-based masked sequence model for mapping functional brain networks. *Front. Neurosci.* 17, 1183145. doi: 10.3389/fnins.2023.1183145

Islam, M. M., Nooruddin, S., Karray, F., and Muhammad, G. (2022). Human activity recognition using tools of convolutional neural networks: a state of the art review, data sets, challenges, and future prospects. *Comput. Biol. Med.* 2022, 106060. doi: 10.1016/j.compbiomed.2022.106060

Khaertdinov, B., and Asteriadis, S. (2022). "Temporal feature alignment in contrastive self-supervised learning for human activity recognition," in *2022 IEEE International Joint Conference on Biometrics (IJCB)* (Abu Dhabi), 1–9. doi: 10.1109/IJCB54206.2022.10007984

Khaertdinov, B., Ghaleb, E., and Asteriadis, S. (2021). "Contrastive self-supervised learning for sensor based human activity recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, (Shenzhen: IEEE), 1–8. doi: 10.1109/IJCB52358.2021.9484410

Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., and Murakami, T. (2019). "MMACT: a large-scale dataset for cross modal human action understanding," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul), 8657–8666. doi: 10.1109/ICCV.2019.00875

Li, C., Zhong, Q., Xie, D., and Pu, S. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*. doi: 10.24963/ijcai.2018/109

Li, H., Liu, M., Yu, X., Zhu, J., Wang, C., Chen, X., et al. (2023). Coherence based graph convolution network for motor imagery-induced EEG after spinal cord injury. *Front. Neurosci.* 16, 1097660. doi: 10.3389/fnins.2022.1097660

Li, T., Cheng, B., Ni, B., Liu, G., and Yan, S. (2016a). Multitask low-rank affinity graph for image segmentation and image annotation. *ACM Trans. Intell. Syst. Technol.* 7, 1–18. doi: 10.1145/2856058

Li, T., Mei, T., Yan, S., Kweon, I.-S., and Lee, C. (2009). "Contextual decomposition of multi-label images," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 2270–2277. doi: 10.1109/CVPR.2009.5206706

Li, T., Meng, Z., Ni, B., Shen, J., and Wang, M. (2016b). Robust geometric p-norm feature pooling for image classification and action recognition. *Image Vision Comput.* 55, 64–76. doi: 10.1016/j.imavis.2016.04.002

Liu, Y., Wang, K., Li, G., and Lin, L. (2021). Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. *IEEE Trans. Image Process.* 30, 5573–5588. doi: 10.1109/TIP.2021.3086590

Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., and Chiaberge, M. (2022). Action transformer: a self-attention model for short-time pose-based human action recognition. *Pattern Recogn.* 124, 108487. doi: 10.1016/j.patcog.2021.108487

Mughal, N. E., Khan, M. J., Khalil, K., Javed, K., Sajid, H., Naseer, N., et al. (2022). EEG-fNIRS based hybrid image construction and classification using CNN-LSTM. *Front. Neurorobot.* 16, 873239. doi: 10.3389/fnbot.2022.873239

Plizzari, C., Cannici, M., and Matteucci, M. (2021). "Spatial temporal transformer network for skeleton based action recognition," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event* (Springer), 694–701. doi: 10.1007/978-3-030-68796-0_50

Shavit, Y., and Klein, I. (2021). Boosting inertial-based human activity recognition with transformers. *IEEE Access* 9, 53540–53547. doi: 10.1109/ACCESS.2021.3070646

Shen, X., and Ding, Y. (2022). Human skeleton representation for 3d action recognition based on complex network coding and LSTM. *J. Vis. Commun. Image Represent.* 82, 103386. doi: 10.1016/j.jvcir.2021.103386

Shu, X., Zhang, L., Sun, Y., and Tang, J. (2021). Host-parasite: graph LSTM-in-LSTM for group activity recognition. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 663–674. doi: 10.1109/TNNLS.2020.2978942

Suh, S., Rey, V. F., and Lukowicz, P. (2023). Tasked: transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation. *Knowledge Based Syst.* 260, 110143. doi: 10.1016/j.knosys.2022.110143

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., and Liu, J. (2023). Human action recognition from various data modalities: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3200–3225. doi: 10.1109/TPAMI.2022.3183112

Tian, Y., Krishnan, D., and Isola, P. (2020). "Contrastive multiview coding," in *Computer Vision-ECCV 2020: 16th European Conference* (Glasgow), 776–794. doi: 10.1007/978-3-030-58621-8_45

Tu, Z., Zhang, J., Li, H., Chen, Y., and Yuan, J. (2022). Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. *IEEE Trans. Multimedia* 25, 1819–1831. doi: 10.1109/TMM.2022.3168137

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, T., Li, 479 J., Wu, H.-N., Li, C., Snoussi, H., and Wu, Y. (2022). Reslnet: deep residual lstm network with longer input for action recognition. *Front. Comput. Sci.* 16, 166334. doi: 10.1007/s11704-021-0236-9

Xie, C., Li, C., Zhang, B., Chen, C., Han, J., Zou, C., et al. (2018). "Memory attention networks for skeleton-based action recognition," in *International Joint Conference on Artificial Intelligence* (Stockholm), 1639–1645. doi: 10.24963/ijcai.2018/227

Xu, K., Ye, F., Zhong, Q., and Xie, D. (2022). "Topology-aware convolutional neural network for efficient skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2866–2874. doi: 10.1609/aaai.v36i3.20191

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence* (Phoenix, Arizona). doi: 10.1609/aaai.v32i1.12328

Yi, M.-K., Lee, W.-K., and Hwang, S. O. (2023). A human activity recognition method based on lightweight feature extraction combined with pruned and quantized CNN for wearable device. *IEEE Trans. Cons. Electron.* 1. doi: 10.1109/TCE.2023.3266506

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). "Barlow twins: self-supervised learning via redundancy reduction," in *International Conference on Machine Learning (PMLR)*, 12310–12320.

Zhang, J., Wu, F., Hu, W., Zhang, Q., Xu, W., and Cheng, J. (2019). "Wienhance: towards data augmentation in human activity recognition using wifi signal," in *MSN* (Shenzhen), 309–314. doi: 10.1109/MSN48538.2019.00065

Zhang, J., Wu, F., Wei, B., Zhang, Q., Huang, H., Shah, S. W., et al. (2020). Data augmentation and dense-LSTM for human activity recognition using wifi signal. *IEEE Internet Things J.* 8, 4628–4641. doi: 10.1109/JIOT.2020.3026732