Multimedia Systems (2024) 30:267
https://doi.org/10.1007/s00530-024-01463-0

REGULAR PAPER

Multi-scale motion contrastive learning for self-supervised skeleton-based action recognition

Yushan Wu[1,2,6] · Zengmin Xu[1,2,5] · Mengwei Yuan[1,2] · Tianchi Tang[3] · Ruxing Meng[5] · Zhongyuan Wang[4]

Received: 12 April 2024 / Accepted: 19 August 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
People process things and express feelings through actions, action recognition has been able to be widely studied, yet under-explored. Traditional self-supervised skeleton-based action recognition focus on joint point features, ignoring the inherent semantic information of body structures at different scales. To address this problem, we propose a multi-scale Motion Contrastive Learning of Visual Representations (MsMCLR) model. The model utilizes the Multi-scale Motion Attention (MsM Attention) module to divide the skeletal features into three scale levels, extracting cross-frame and cross-node motion features from them. To obtain more motion patterns, a combination of strong data augmentation is used in the proposed model, which motivates the model to utilize more motion features. However, the feature sequences generated by strong data augmentation make it difficult to maintain identity of the original sequence. Hence, we introduce a dual distributional divergence minimization method, proposing a multi-scale motion loss function. It utilizes the embedding distribution of the ordinary augmentation branch to supervise the loss computation of the strong augmentation branch. Finally, the proposed method is evaluated on NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD datasets. The accuracy of our method is 1.4–3.0% higher than the frontier models.

Keywords Contrastive learning · Multi-scale · Attention mechanism · Data augmentation · Human skeleton

1 Introduction

Human action recognition is a fundamental problem in computer vision, with rich applications such as human-computer interaction, patient monitoring, and sports analysis [1–6]. The main objective of action recognition is to classify human actions from motion data, which mainly include RGB videos [7], depth maps [8], infrared images [9], and 3D skeleton sequences [10]. Most of the existing studies [11–13] utilize visual features in RGB or depth maps to recognize human movements. However, these methods are susceptible to background, light, and dark variations in practice. Recently, with the development of pose estimation algorithms [14] and sensors, body key points can be easily acquired to accurately estimate the human skeleton. Due to the advantages of computational efficiency and storage, as well as robustness to dynamic environmental changes and camera viewpoints [14, 15], there has been widespread interest in skeleton-based action recognition [16, 17].

In the past few years, most skeleton-based action recognition methods have taken advantage of supervised learning frameworks [18–22]. Fully supervised action recognition methods inevitably require large amounts of labeled data to drive them, but manual labeling large-scale datasets is particularly costly. As a result, researchers are increasingly using unlabeled skeleton data to learn human action representations. Self-supervised learning aims to obtain robust representations of samples from raw data without the usage

✉ Zengmin Xu
xzm@guet.edu.cn

1 School of Mathematics and Computing Science, Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China

2 Center for Applied Mathematics of Guangxi, Guilin University of Electronic Technology, Guilin 541002, Guangxi, China

3 Shuyeit.ai, Guangzhou 510670, Guangdong, China

4 NERCMS, School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China

5 Anview.ai, Guilin 541010, Guangxi, China

6 Guangxi Technological College of Machinery and Electricity, Nanning 530007, Guangxi, China

Published online: 10 September 2024

of expensive labels or annotations. Based on the learning paradigm, self-supervised learning methods can be categorized into generative [23–25] and contrastive methods. Generative methods capture spatial-temporal correlations by predicting the skeleton data of a mask. Zheng et al. [26] proposed reconstructing a mask's skeleton for long-term global motion dynamics. Wu et al. [27] following MAE's masking and reconstruction pipeline, we utilize a skeleton based encoder-decoder transformer architecture to reconstruct the masked skeleton sequences. Contrastive methods, also known as contrastive learning, use data augmentation to generate positive and negative samples to learn common features between similar instances, distinguishing differences between non-similar instances. Rao et al. [28] with Shear and Crop as data augmentation. Guo et al. [29] further proposed the use of more augmentations, such as Rotate, Axis Mask, and Flip, to improve the consistency of contrastive learning. This study also applies data augmentation to increase the model's performance.

Motivation Previous contrastive learning action recognition based on skeleton data priority node features at the joint level, ignoring semantic content at the coarse level (e.g., left hand, left foot, right upper limb, right lower limb, etc.). The "hand" at the part level and the "lower limb" at the body level can directly correspond to the semantic features, while joints at the joint level cannot. For example, as displayed in Fig. 1, the action "shake head" considers the head as the basic unit of movement, while the action "sitting down" emphasizes the bending of the lower limbs during the movement. The part level "head" and the body level "lower limbs" directly reflect the semantics of the action. Therefore, it is meaningful and reasonable to try to obtain rich semantic elements from different levels of skeleton scales. A few approaches have emphasized multi-scale skeleton

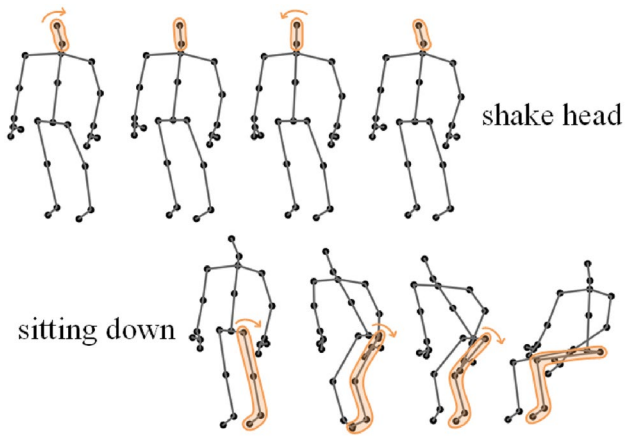

Fig. 1 The basic movement unit for the action "shake head" is the part level "head", while the basic movement unit for the action "sitting down" is the body level "lower limbs", both of which directly reflect the semantic of the actions

characterization [30–32], where these features are directly input into the model as multi-scale skeleton data, exploiting multi-scale or cross-scale information to optimize the model. However, these methods do not demonstrate attention to the semantic information embedded in different scale-level features. In purpose of better utilizing the large amount of semantic information contained in different scale-level skeleton features, we try to combine different scale information with the attention mechanism. Because the advantage that the attention mechanism can find the most effective information quickly can help us to utilize the large amount of semantic information embedded in different scale-level skeleton features, and further improve the model accuracy. As shown in Fig. 2, this study incorporated multi-scale skeleton features into the attention module to form semantic content that focuses on the joint, part, and body levels, rather than as input data at the beginning of the model.

Contribution Most current research on attention mechanisms focuses on spatial and temporal attention, which are difficult to encode the inherent motions in skeleton data. To mitigate this issue, as displayed in Fig. 2. We do inter-frame and inter-node differences for multi-scale features, explicitly combining human motion characteristics to form an attention module that concentrates on cross-frame and cross-node motion context. The operation of inter-frame and inter-node differences was inspired by the motion and bone views [33] of skeleton data. In cross-view or multi-view approaches to skeleton motion recognition [34, 35], additional view inputs, such as motion and bone, are offered to help the model capture high-quality motion representations. Motion views are obtained from the difference between front and back frames, bone views are obtained from the difference between pairs of joints. Both motion and bone are features during sports, since the motion is inter-frame information, the bone is intra-frame features [33], which are unified as motion information in this study. In the proposed attention module, the motion and bone features will be computed for each scale. Different from reference [33], we calculate bone features without considering joint pairs. Instead, similar to motion views, we take the difference between the previous and subsequent nodes.

The main contributions of this study can be summarized as follows:

- A multi-scale motion contrastive learning of visual representations (MsMCLR) framework was designed for self-supervised skeleton motion recognition. A strong augmentation branch is constructed in the framework, which combines multiple data augmentations to produce more motion patterns. A dual distributional divergence minimization method is introduced, and a multi-scale motion loss function is proposed, which takes into account the embedding distribution of ordinary augmented branches

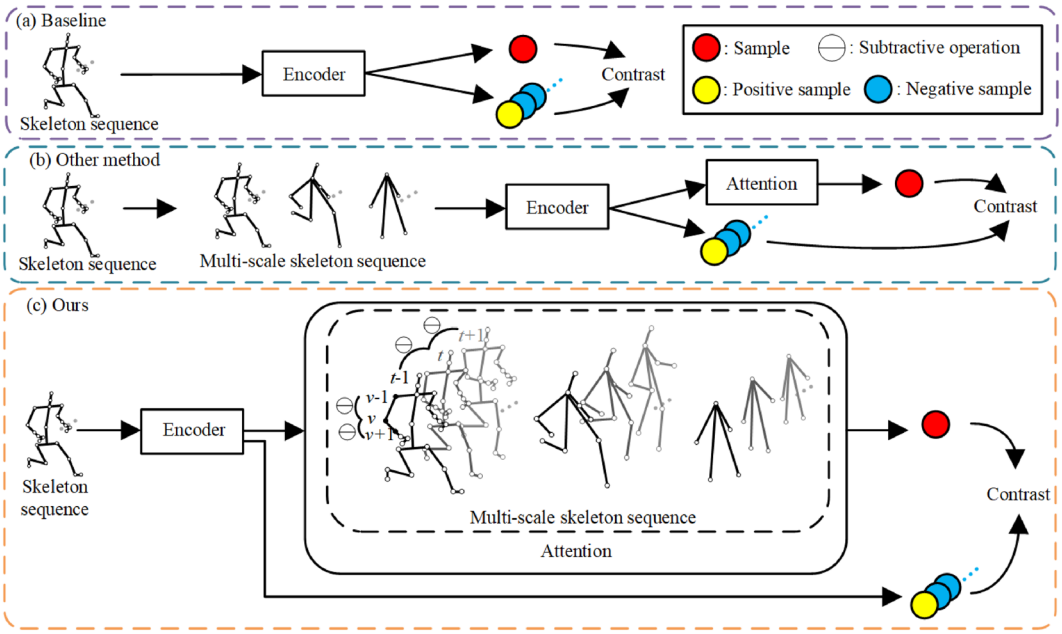

Fig. 2 Comparison of proposed and other methods. (**a**) The diagram is the underlying framework of SkeletonCLR [34], which does not effectively explore the inherent context of skeleton data. (**b**) Current methods consider multi-scale data, but only treat it as regular data input to the model, without fully utilizing the semantic content of multi-scale data. (**c**) This study combines different scale information with attention mechanisms to effectively explore the inherent features of skeletal data

to supervise the loss computation of strongly augmented branches.

- We propose a multi-scale motion attention (MsM Attention) mechanism, which incorporates the motion features of multi-scale skeletons into an attention mechanism, with priority attention on the contextual semantics obtained through cross-frame and cross-node.
- In this study, the proposed method was evaluated on the NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD datasets, achieving a consistent boost on their benchmarks.

2 Related work

2.1 Skeleton-based contrastive learning

To address skeleton-based action recognition, early methods were usually implemented based on manual features [36]. In recent years, deep learning methods, mainly represented by the SkeletonCLR [34] model, have followed the MoCov2 [37] framework for implementation. The SkeletonCLR method first augments a sample as a pair of positive samples. The historical positive samples are stored in the memory bank as negative samples. After encoding, the similarity between positive and positive samples, as well as between positive and negative samples, is calculated. Then the network is iteratively updated by InfoNCE loss [38]. CrosSCLR [34] is based on SkeletonCLR with the addition of cross-view (e.g., joint, bone, and motion) consistency knowledge mining in negative samples, exploiting complementarities between modalities. The AimCLR [29] model uses extreme augmentation to capture new motion patterns in sample features. In addition, the model performs nearest-neighbor mining on a single modality to expand the positive samples. However, none of these methods have explored valuable multi-scale body component relationships in skeleton structure or motion, without adequately capturing the relationships between physically connected nodes in the skeleton graph.

2.2 Multi-scale methods

Significant value exists in the human structure. For example, there is a strong motion correlation between the adjacent feet and thighs, as well as between the non-adjacent legs and arms when a person is walking. The different degrees of cooperation between the legs and arms at the body level, as well as between the feet and thighs at the part level, can capture unique motion patterns. Therefore, combining skeleton features at multi-scale is of great importance. Dang et al. [30] extracted features from fine to coarse scale and back again. Then they combined and decoded the characteristics extracted at each scale. Rao et al. [31] constructed a unified multi-level skeleton graph as a multi-head structural relationship layer, to comprehensively capture the relationships

between physically connected nodes in the skeleton graph. Xu et al. [32] designed a new pyramid aggregation attention mechanism that aggregates the attention map from the body, part, and joint levels step-by-step. All these methods choose multi-scale skeleton features as input or level-by-level aggregation, without considering inter-frame and inter-node motion information.

2.3 Self-attention mechanism

Integrate attention mechanisms to action recognition has become popular due to their excellent results in natural language processing tasks [39]. The self-attention mechanism is a variant of the attention mechanism that focal point dependencies within the data [39, 40]. Specifically, its input consists of Q (Query), K (Key), and V (Value) of dimension d. Q, K, and V are represented in the form of a matrix for fast computation. First the dot product of Q with all K is computed, each dot product result is divided by $\sqrt{d}$ for gradient stabilization. Then, the result obtained by applying the softmax activation function is multiplied by V to get the output.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{1}$$

Based on the traditional self-attention mechanism, this study proposes a multi-scale motion attention mechanism to combine local and global features of skeleton data by fusing skeleton features of different scales. Instead of obtaining the different scales of skeleton features at the early input stage, the division of joint, part, and body levels of skeleton features is performed only after the network encoder. Then inter-frame and inter-node differences are determined for each scale of skeleton features separately. The obtained features are fed into the self-attention mechanism to get the attention map that contains the explicit motion features.

3 Proposed approach

In this section, we introduce the proposed skeleton-based action recognition method as indicated in Fig. 3. Section 3.1 describes the overall framework of the model. Section 3.2

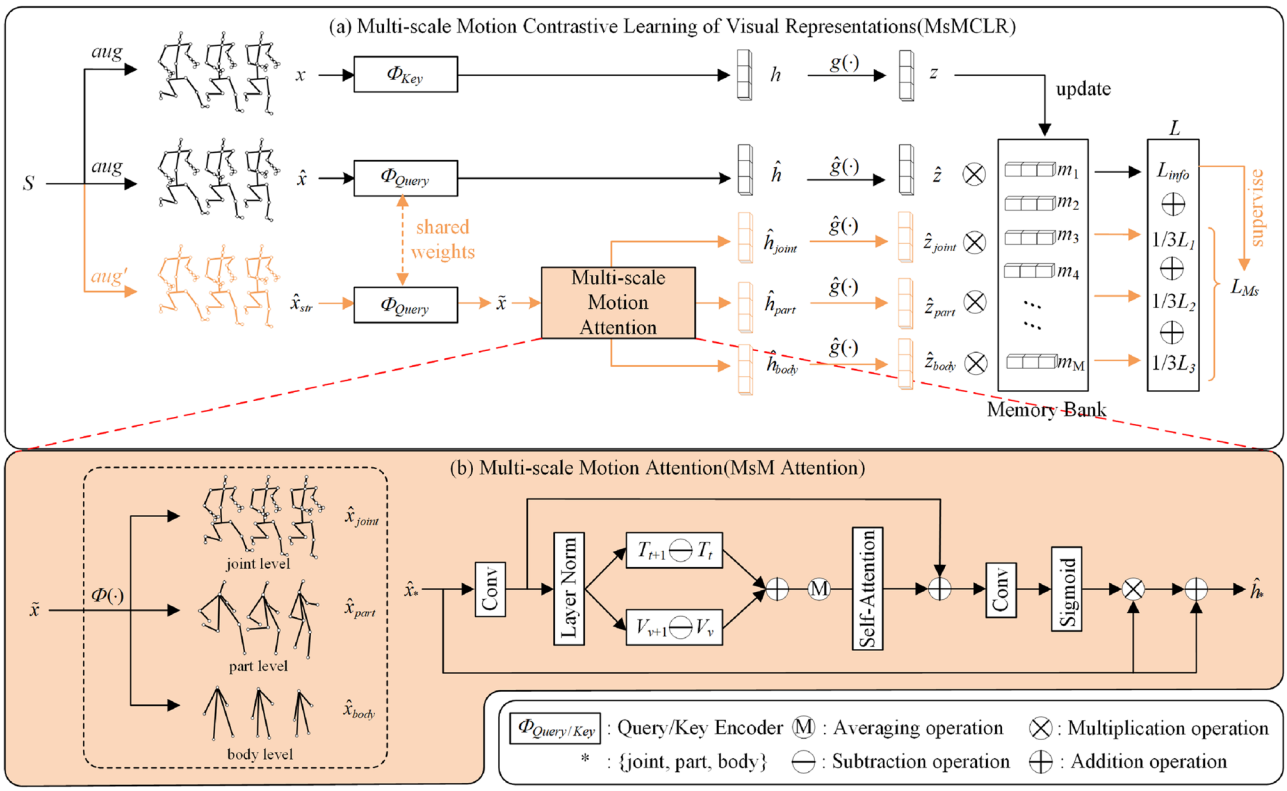

Fig. 3 MsMCLR overall structure diagram. In (a) *aug* stands for common augmentation and *aug′* stands for strong augmentation. MsM Attention is set on the strong augmentation branch. The loss function is computed by four feature embeddings on the Query branch (three from the strong augmentation branch and one from the normal augmentation branch) and one feature embedding on the Key branch are involved. The specific process of MsM Attention is shown in (b), where attention information at three skeleton scales is attended to. The proposal is displayed in color

describes the MSM Attention module in detail. Section 3.3 describes the combination of data augmentation in this study. Section 3.4 describes the loss functions used in the different branches.

3.1 MsMCLR method

The general framework of the proposed MsMCLR is revealed in Fig. 3. Different from previous models [34], we set up two Query branches separately for common augmented (aug) sequences and strong augmented (aug') sequences. The encoder $\boldsymbol{\Phi}_{Query}$ of the two Query branches shares the weights. The strong augmented sequence contains multiple data augmentations (see Sect. 3.3). The proposed MsM Attention is also set on this branch because strong augmentation provides more motion patterns. Given the input skeleton sequence S, different data augmentations are applied to obtain x, $\hat{x}$, $\hat{x}_{str} \in R^{C\times T\times V}$. Here, C, T, and V denote the numbers of channels, frames, and nodes, respectively. The features x, and $\hat{x}$ are encoded and mapped to obtain z and $\hat{z}$. Finally, the network is updated using InfoNCE losses.

When the human body performs a certain action, the basic human units at different levels contain inherent semantic characteristics. Inter-frame and inter-node motion information also contain potential semantics. Effective utilization of this context can enhance the action representation. Therefore, in this study, we propose a multi-scale motion attention mechanism that integrates skeleton hierarchy with inter-frame and inter-node motion features. Specifically, the input $\hat{x}_{str}$ passes through the encoder $\boldsymbol{\Phi}_{Query}$ to obtain $\widetilde{x} = \boldsymbol{\Phi}_{Query}(aug'(\hat{x}_{str}))$, where $\widetilde{x} \in R^{C\times T\times V}$. Then, $\widetilde{x}$ passes through the MsM Attention to obtain the feature $\hat{h}_{joint}$, $\hat{h}_{part}$, $\hat{h}_{body}$. Next, the multi-layer perception head $\hat{g}(\cdot)$ is applied to obtain the embeddings $\hat{z}_{joint} = \hat{g}(\hat{h}_{joint})$, $\hat{z}_{part} = \hat{g}(\hat{h}_{part})$, and $\hat{z}_{body} = \hat{g}(\hat{h}_{body})$. Finally, the network is updated according to L_{Ms} supervised by the ordinary augmented branching distribution(see Sect. 3.4).

In MsM Attention, the input $\widetilde{x}$ is first divided into three layers $\hat{x}_{joint}$, $\hat{x}_{part}$, and $\hat{x}_{body}$ by the function $\boldsymbol{\Phi}(\cdot)$, which is denoted by $\hat{x}_{*} =< \hat{x}_{joint}, \hat{x}_{part}, \hat{x}_{body} >$ in Fig. 3(b). Inter-frame and inter-node differences are computed respectively for each layer feature and averaged after merging the channel dimensions. Subsequently, the attention map corresponding to the three layers are computed then outputted as $\hat{h}_{joint}$, $\hat{h}_{part}$, and $\hat{h}_{body}$.

3.2 Multi-scale motion attention mechanisms

3.2.1 Multi-scale graph

In this study, we propose a multi-scale motion attention mechanism, as indicated in Fig. 3(b). The attention module converts the output of the encoder into three levels of features, involving the human body structure, i.e., joint, part, and body levels, allowing the hierarchical structure to be coded for human action recognition.

The human body posture is considered as a skeleton topology graph. The relationships between human joints can be flexibly learned using the ST-GCN encoder with contrastive learning. In the original skeletal graph, there are only physical connections between joints. However, there may be connections between body joints and between body parts that span beyond physical connections. Li et al. [41] revealed that a large amount of potential inherent semantic characteristics between limbs, parts, and joints is embedded in the abstract hierarchical structure of the skeleton, which crosses the physical connections when the human body performs a certain action. For this reason, in this study, we incorporated multi-scale skeleton features into the attention module to form the semantic information that spotlight the joint, part, and body levels, thereby enhancing action representation.

Taking the skeleton topology graph of the NTU RGB+D dataset as an example, the hierarchical structure in the human skeleton is shown in Fig. 4. The V joint-level nodes in (c) are spatially divided into P parts (e.g., hand, arm, and foot) in (e) and merged in (b). Each part's level node contains 2 to 3 joint-level nodes. Then, the V joint-level nodes in (c) are spatially divided into B body level parts (e.g., left upper limb, left lower limb, and torso) in (f) and merged in (d). Each body level node contains 4 to 6 joint-level nodes. To convert joint features $\widetilde{x} \in R^{C\times T\times V}$ into joint-level features $\hat{x}_{joint} \in R^{C\times T\times V}$, no additional operations are required. To convert joint features $\widetilde{x}$ into part-level features $\hat{x}_{part} \in R^{C\times T\times P}$, the node dimension V needs to be compressed to P through average operation. To convert joint features $\widetilde{x}$ into body-level features $\hat{x}_{body} \in R^{C\times T\times B}$, the node dimension V needs to be

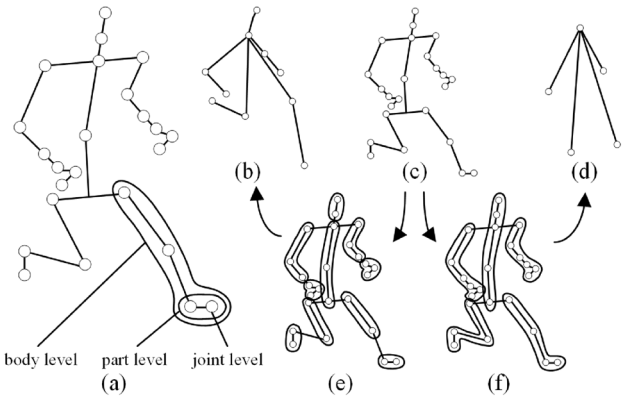

Fig. 4 Multi-scale skeleton merging process. (**c**) represents the joint level skeleton diagram, which is the original skeleton diagram. (**c**) is divided based on diagram (**e**), forming a part level skeleton diagram (**b**), and (**e**) is further aggregated into a body level skeleton diagram (**d**) based on diagram (**f**)

compressed to B through average operation. The above process is represented by the function $\boldsymbol{\Phi}(\cdot)$.

$$< \hat{x}_{joint}, \hat{x}_{part}, \hat{x}_{body} >= \boldsymbol{\Phi}(\widetilde{x}) \tag{2}$$

3.2.2 Motion attention

To integrate the motion context inherent in the different scale levels of the skeleton data, inspired by [42], motion elements between consecutive frames and between nodes are extracted, allowing attention to be focused on the motion patterns cross-frame and cross-node, which provides an important contribution to the model. It is cross-node because after dividing the skeleton into different scales (i.e., joint, part, and body levels), each node merges different joints of the original scale. Where the joint-level nodes merge a single joint point feature, part-level and body-level nodes merge multiple joint point features. Therefore, the inter-node semantic characteristics extracted from the different scale features contain semantic information between non-adjacent nodes.

As displayed in Fig. 3(b), taking $\hat{x}_{joint}$ in $\hat{x}_*$ as an example, performing convolution operation on $\hat{x}_{joint}$ to reduce the computational cost by decreasing the channel dimension. The normalization process is performed immediately after. Then, the computation of inter-frame and inter-node differences for joint-level features starts. The inter-frame difference is realized by subtracting the t th frame from the $t+1$ th frame in the time dimension T. The inter-node difference is obtained by subtracting node v and node $v+1$ in the joint dimension V. The two output features are summed and averaged. Next, the average result is flattened along the temporal dimension to calculate the attention map. The resulting attention map is summed with the output features of the first dimensionality reduction convolution. The result of the summation is convolved and passed through a sigmoid function, followed by multiplication and addition with the input feature $\hat{x}_{joint}$. The final obtained attention map activates the cross-frame and cross-node related features in the input feature $\hat{x}_{joint}$, while suppressing the unrelated features.

Except for the simultaneous conduct of inter-frame and inter-node differences, as indicated in Fig. 5, this study also explores applicable inter-frame or inter-node differences alone to capture semantic information. We compare the results of these three approaches in Sect. 4.4. According to the experimental results, this study ultimately determines both inter-frame difference and inter-node difference. This means that the Block in Fig. 5 is defaulted to subfigure (a).

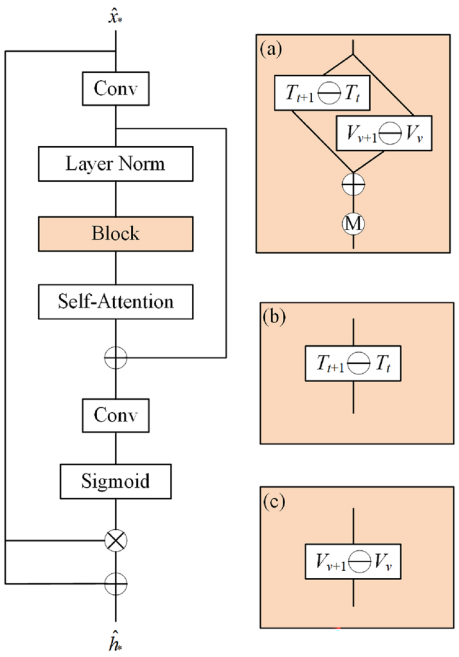

Fig. 5 Investigation on capturing motion attention. (**a**) Using both inter-frame and inter-node differences. (**b**) Using inter-frame differences. (**c**) Using inter-node difference

3.3 Data augmentation combination

Data augmentation plays a key role in contrastive learning [43]. Appropriately adding multiple data augmentations can enhance the accuracy of action recognition models, which has been proven [29]. Therefore, in this study, we designed a strong augmentation branch and combined eight data augmentations to provide more motion patterns for the model.

When inferring global structure, high-level information, namely the potential space, is more applicable than low-level information. For potential representations, global noise can be a serious noise. Strong data augmentation can produce more motion patterns to affect the potential space. Consequently, this study introduced the noise generation approach of Yoon et al. [44]. The method generates noisy skeleton features by adding some noise to randomly selected joints in the original skeleton sample. The randomly selected joints are fixed to 5 nodes. The noise is determined based on the range between the minimum and maximum values of the coordinates of the original skeleton sample. The Shear, Crop, Rotation, Gaussian Noise, Gaussian Blur, Joint Mask, and Channel Mask from [28] and [45] are also augmentation strategies for the strong augmentation branch. Only Shear and Crop were used for data augmentation in other branches.

However, compared with normal augmented sequences, strongly augmented sequences may not be able to maintain the identity of the original sequences due to the crucial changes in motion patterns, which leads to the degradation

of model performance. The loss function can handle this problem well.

3.4 Loss function

The MsMCLR framework has multiple Query branch outputs, requiring the calculation of multiple InfoNCE losses [38]. However, the feature sequences of the strong augmented branch change significantly, making it challenging to maintain identity with the original sequence. To address this issue, this study introduces the dual distribution divergence minimization (D^3M) method in AimCLR [29], which aims to maintain the identity of the changed sequence with the original sequence. The D^3M method exploits the embedding distribution of the normal augmented branch to supervise the loss computation of the strong augmented branch. Different from [29], we supervise the feature embeddings at all three scale levels.

For the common augmentation branch, the InfoNCE loss function is used directly, i.e., (3).

$$L_{Info} = -log\frac{\exp(\hat{z}\cdot z/\tau)}{\exp(\hat{z}\cdot z/\tau)+\sum_{i=1}^{M}\exp(\hat{z}\cdot m_i/\tau)} \tag{3}$$

Negative samples are stored in a queue-based memory bank, where the stored negative samples are denoted $m_i, i = 1,\ldots,M$. The conditional distribution of m_i with respect to $\hat{z}$ is given by (4), which encodes the likelihood that $\hat{z}$ will be assigned to an embedding of m_i. Similarly, the conditional distribution of z concerning $\hat{z}$ can be obtained as (5).

$$p(m_i|\hat{z}) = \frac{\exp(\hat{z}\cdot m_i/\tau)}{\exp(\hat{z}\cdot z/\tau)+\sum_{i=1}^{M}\exp(\hat{z}\cdot m_i/\tau)} \tag{4}$$

$$p(z|\hat{z}) = \frac{\exp(\hat{z}\cdot z/\tau)}{\exp(\hat{z}\cdot z/\tau)+\sum_{i=1}^{M}\exp(\hat{z}\cdot m_i/\tau)} \tag{5}$$

Similar to (4) and (5), the conditional distributions of $\hat{z}_{joint}$ are obtained based on positive and negative samples as $p(z|\hat{z}_{joint})$ and $p(m_i|\hat{z}_{joint})$, respectively. Similarly, the conditional distributions of $\hat{z}_{part}$ are $p(z|\hat{z}_{part})$ and $p(m_i|\hat{z}_{part})$. The conditional distributions of $\hat{z}_{body}$ are $p(z|\hat{z}_{body})$ and $p(m_i|\hat{z}_{body})$. Finally, minimizing a multi-scale motion loss function supervised by an ordinary augmented Query branching distribution, L_{Ms}, is revealed in (9). (10) is the total loss function of the MsMCLR model.

$$L_1 = -p(z|\hat{z})\log p(z|\hat{z}_{joint})\sum_{i=1}^{M}p(m_i|\hat{z})\log p(m_i|\hat{z}_{joint}) \tag{6}$$

$$L_2 = -p(z|\hat{z})\log p(z|\hat{z}_{part})\sum_{i=1}^{M}p(m_i|\hat{z})\log p(m_i|\hat{z}_{part}) \tag{7}$$

$$L_3 = -p(z|\hat{z})\log p(z|\hat{z}_{body})\sum_{i=1}^{M}p(m_i|\hat{z})\log p(m_i|\hat{z}_{body}) \tag{8}$$

$$L_{Ms} = (L_1 + L_2 + L_3)/3 \tag{9}$$

$$L = L_{Info} + L_{Ms} \tag{10}$$

4 Experiment

4.1 Datasets

NTU RGB+D 60 (NTU-60) [46] is one of the largest public datasets for skeleton action recognition. It contains 56,880 skeletal movement sequences. There are two evaluation benchmarks, including Cross-Subject (xsub) and Cross-View (xview) settings. For xsub, the training and test sets are from two non-intersecting sets of 20 subjects each. For xview, the training set contains 37,920 samples captured by cameras 2 and 3, while the test set contains 18,960 samples captured by camera 1. In addition, there are 302 erroneous samples to be ignored during training and evaluation.

NTU RGB+D 120 (NTU-120) [47] is an extension of the NTU-60 dataset with 57,367 skeleton sequences on an additional 60 action classes. Likewise, the authors present two evaluation benchmarks, including Cross-Subject (xsub) and Cross-Setup (xset) settings. For xsub, the training and test sets were each provided by 53 subjects. For xset, the training and test sets were provided by numbered even and odd cameras, separately. In addition, there are 532 erroneous samples to be ignored during training and evaluation.

The PKU-MMD dataset [48] is a large-scale benchmark for continuous multimodal 3D human movement recognition that covers 51 classes of complex human activities. The dataset contains two stages of increasing difficulty. Part I is an easier version for action recognition, while part II is more challenging because view changes generate more noise. Experiments were conducted under the Cross-Subject protocol.

4.2 Implementation details

The hardware platform used for the experiments in this study consists of 128 GB memory and 4 TITAN XP graphics cards. The software platform consists of Python 3.7 and Pytorch 1.6.0 framework. The parameter configurations were consistent with those in [29]. For data preprocessing, this

study followed the methodology of SkeletonCLR [34] for fair comparisons. Encoders Φ_{Query} and Φ_{Key} same as the ST-GCN [10] network. For the optimizer SGD with momentum 0.999 and weight decay 0.0001. The model was trained for a total of 300 epochs, with the first 250 epochs at a learning rate of 0.1, which was reduced to 0.01 for the last 50 epochs. In addition, experiments were conducted on all three skeleton views, namely, joint, bone, and motion. For all reported results for the three skeleton views, weighted fusion was performed using weights [0.6, 0.6, 0.4], as in other multi-stream GCN methods [29].The size of the memory bank is the same as in the literature [34], set to 32768. The linear evaluation was run for 100 epochs with the initial value of the learning rate set to 3. After evaluating 80 epochs, the learning rate decreased to 0.3.

4.3 Analysis of experimental results

4.3.1 Quantitative results

Tables 1 and 2 compare the proposed method with a wide range of existing unsupervised methods on two large datasets, NTU-60 and NTU-120, including results from fusing multiple data views (3s).

The results in Table 1 show that the proposed method significantly outperformed previous methods, both on a single data view (joint) and in the case of fusing three views. For a single view, the accuracy of the proposed method was improved by 2.5% and 2.3% on the xsub/xview benchmarks relative to the AimCLR method, separately. When three views are fused, the accuracy of the proposed method was raised by 1.4% and 1.6% on the xsub and xset benchmarks relative to the 3s-AimCLR method, separately. These results demonstrate the effectiveness of the proposed method.

Table 1 Linear evaluation results on NTU-60

Method	NTU-60(%)	
	xsub	xview
Single-stream	–	–
LongT GAN(AAAI 18) [26]	39.1	48.1
MS^{2}L(ACM MM 20) [49]	52.6	–
AS-CAL(Information Sciences 21) [28]	58.5	64.8
P&C (CVPR 20) [24]	50.7	76.3
SeBiReNet(ECCV 20) [50]	–	79.7
SkeletonCLR(CVPR 21) [34]	68.3	76.4
AimCLR(AAAI 22) [29]	74.3	79.7
PSTL(AAAI 23) [51]	**77.3**	81.8
MsMCLR	76.8	**82.0**
Three-stream	-	-
3s-SkeletonCLR(CVPR 21) [34]	75.0	79.8
3s-Colorization(ICCV 21) [25]	75.2	83.1
3s-CrosSCLR(CVPR 21) [34]	77.8	83.4
3s-AimCLR(AAAI 22) [29]	78.9	83.8
3s-PSTL(AAAI 23) [51]	79.1	82.6
3s-MsMCLR	**80.3**	**85.4**

Bold text in the tables indicates our method or optimal results

Table 2 Linear evaluation results on NTU-120

Method	NTU-120(%)	
	xsub	xset
LongT GAN(AAAI 18) [26]	35.6	39.7
P &C(CVPR 20) [24]	42.7	41.7
AS-CAL(Information Sciences 21) [28]	48.6	49.2
3s-SkeletonCLR (CVPR 21) [34]	60.7	62.6
3s-CrosSCLR(CVPR 21) [34]	67.9	66.7
ISC(ACM MM 21) [52]	67.9	67.1
3s-AimCLR(AAAI 22) [29]	68.2	68.8
3s-PSTL(AAAI 23) [51]	69.2	70.3
3s -MsMCLR	**70.3**	**71.8**

Bold text in the tables indicates our method or optimal results

The results in Table 2 display that the proposed MsMCLR outperforms other unsupervised methods, such as LongT GAN [26], P&C [24], AS-CAL [28], SkeletonCLR [34], CrosSCLR [34], ISC [52], AimCLR [29] and PSTL [51], on NTU-120 dataset. The results of the three-view fusion achieve an accuracy of 70.3% and 71.8% on the xsub and xset benchmarks, respectively. This indicates that the proposed MsMCLR is competitive on multiple classes of large-scale datasets.

To provide a more intuitive comparison, the accuracy rates for four benchmark datasets are plotted in Fig. 6, clearly demonstrating that MsMCLR outperforms several methods.

Table 3 presents the experimental results of our method on the PKU-MMD dataset. The proposed approach outperforms existing unsupervised methods in Part I of the dataset. It is worth noting that our approach is not as good as ISC and 3s-AimCLR on Part II of the PKU-MMD dataset, where the data viewpoints are highly variable, the action intervals are short, and there are also concurrent actions that make the boundaries of the actions less clear. In fact, Part II better reflects the situation of real application scenarios. This also means that the robustness of this paper's method to cope with skeleton noise needs to be improved. This is also the future research direction of this paper.

4.3.2 Qualitative results

To demonstrate the effect of the proposed MsMCLR model more intuitively. The visual embedding distributions of the MsMCLR, SkeletonCLR, and AimCLR models after pre-training for 300 epochs were compared, using the t-SNE [54]

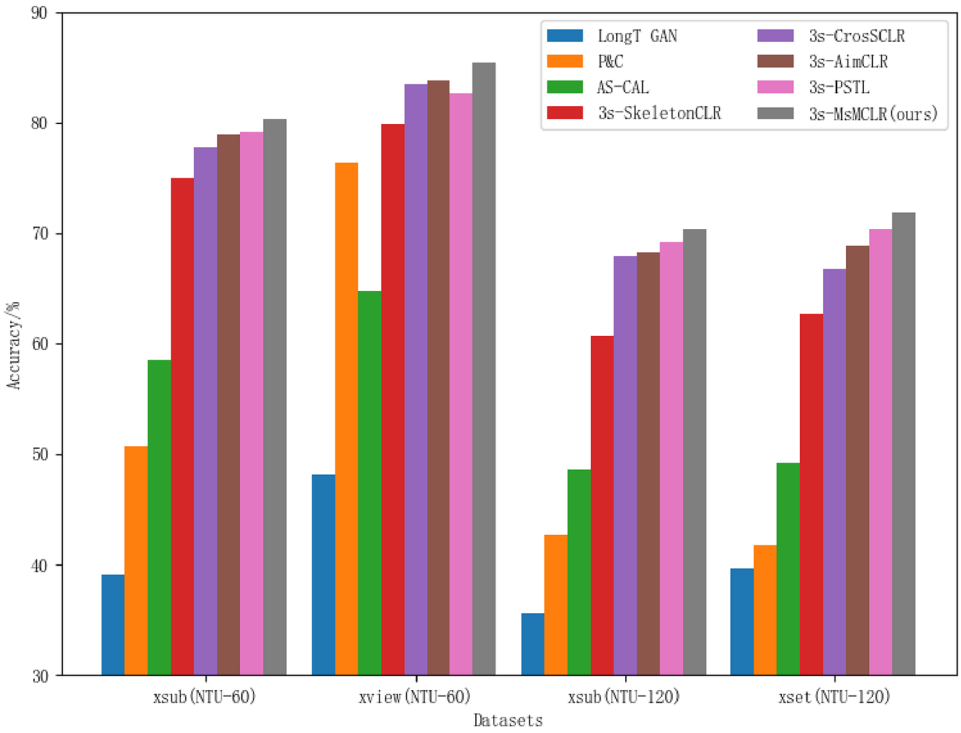

Fig. 6 Comparison of the accuracy of MsMCLR with other models under different benchmarks for NTU-60 and NTU-120 datasets

Table 3 Linear evaluation results on PKU-MMD

Method	PKU-MMD(%)	
	Part I	Part II
Supervised	-	–
ST-GCN (AAAI 18) [10]	84.1	48.2
VA-LSTM (TPAMI 19) [53]	84.1	50.0
Self-supervised	-	–
LongT GAN(AAAI 18)[26]	67.7	26.0
MS^{2}L (ACM MM 20)[49]	64.9	27.6
3s-CrosSCLR(CVPR 21)[34]	84.9	21.2
ISC(ACM MM 21)[52]	80.9	36.0
3s-AimCLR(AAAI 22)[29]	87.8	**38.5**
3s-MsMCLR	**88.6**	32.5

Bold text in the tables indicates our method or optimal results

dimensionality reduction algorithm. The feature embeddings of 10 among 60 classes on the xview benchmark of NTU-60 dataset were selected for comparison, and the results are presented in Fig. 7. It can be seen that compared to the embedding distributions of SkeletonCLR and AimCLR, the MsMCLR model with multiscale motion attention exhibits tighter grouping among embeddings of the same class, and greater separation between embeddings of different classes.

4.4 Analysis of ablation study results

The multi-scale motion attention mechanism and data augmentation are significant modules in the MsMCLR model. Therefore, an ablation study was conducted to validate the effectiveness of multi-scale, MsM Attention, and data augmentation in a self-supervised action recognition task.

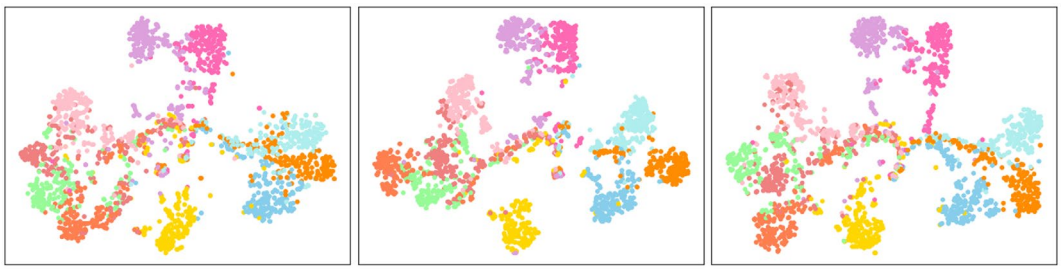

Fig. 7 Embeddings from 10 classes are sampled on the NTU-60 (xview) dataset and visualised with different colours for t-SNE. From the visual results, it can be concluded that MsMCLR is better at clustering embeddings of the same class than AImCLR and SkeletonCLR

Table 4 Ablation studies of multi-scale motion attention

Method	Att	Ms	NTU-60(%)	
			xsub	xview
MsMCLR w/o MsM Attention	✗	✗	79.2	85.6
MsMCLR w/o Ms	✓	✗	79.5	85.2
MsMCLR	✓	✓	**80.3**	**85.4**

4.4.1 Effectiveness of multi-scale motion attention mechanisms

As shown in Table 4, the experimental effects of three different baselines, the MsMCLR model without attention (MsMCLR w/o MsM Attention), MsMCLR model with attention but without a multi-scale skeleton (MsMCLR w/o Ms), and full MsMCLR model, are compared on the two benchmarks, xsub/xview, of the NTU-60 dataset. From Table 4, it can be observed that the MsMCLR model, without the inclusion of the multi-scale skeleton, increases by 0.3% on the xsub benchmark over the no-attention model. However, there is a slight decrease in accuracy on the xview benchmark. This is due to the strong augmentation strategy employed by the model, which generates a large number of new motion models, while the attention module without multi-scale makes it difficult to capture the inherent motion features of the actions in the new motion patterns. Meanwhile, the MsMCLR model containing MsM Attention achieved 80.3% and 85.4% accuracy on the xsub/xview benchmarks respectively, indicating the effectiveness of MsM Attention.

In the study of MsM Attention, we designed three modules to obtain the sample motion information. As displayed in Fig. 5, the methods represent inter-frame difference (only inter-frame), inter-node difference (only inter-node), and simultaneous inter-frame and inter-node difference (all) from top to bottom. Table 5 compares the effects of the three methods. The experimental results indicated that inter-frame and inter-node differences together can bring better results. Thus, it serves as the baseline for the study.

4.4.2 Data augmentation combination effects

Strong augmentation strategy for proposed MsMCLR model consists of eight data augmentations. Table 6 compares the effects of MsMCLR with and without (MsMCLR w/o strong aug) the strong augmentation strategy. The results revealed that the strong augmentation strategy brings 3.5% (xsub) and 2.6% (xview) accuracy boosts. Meanwhile, the accuracy of MsMCLR† versus AimCLR is also given in Table 6. † indicates that the same data enhancement as Aimclr was used. AimCLR has a different strong augmentation strategy from that proposed in this study. It can be seen that the accuracy of MsMCLR† exceeded that of AimCLR, which

Table 5 Comparison of three semantic capture approaches in MSM attention

Method	NTU-60(%)	
	xsub	xview
MsMCLR (only inter-frame)	80.0	84.6
MsMCLR(only inter-node)	79.7	85.1
MsMCLR(all)	**80.3**	**85.4**

Table 6 Comparison of the effects of strong augmentation strategies in MsMCLR models

Method	Strong aug method		NTU-60(%)	
	AimCLR strong aug	Our strong aug	xsub	xview
MsMCLR w/o strong aug	✗	✗	76.8	82.8
AimCLR [29]	✓	✗	78.9	83.8
MsMCLR†	✓	✗	79.8	84.6
MsMCLR	✗	✓	**80.3**	**85.4**

Table 7 Importance of L_{Ms}. Comparison between MsMCLR with L_{NCE} only and MsMCLR incorporating L_{Ms}

Method	NTU-60(%)	
	xsub	xview
AimCLR [29]	78.9	83.8
MsMCLR(only L_{NCE})	78.8	83.8
MsMCLR (L_{NCE}+L_{Ms})	**80.3**	**85.4**

demonstrates the effectiveness of our multi-scale motion attention. MsMCLR† is not as accurate as MsMCLR, indicating that our strong augmentation combination leads to more efficient motion patterns.

4.4.3 Importance of loss function

Table 7 compares the effect of using Eq. (10) and using Eq. (3) for our model. That is, the comparison between L_{Ms} loss and L_{NCE} loss. It can be seen that it is difficult to represent improvements in our model if only L_{NCE} is used. The reason is that a large number of new motion patterns brought by strong enhancement are not handled, which is detrimental to our model. Observing the third row, the performance of our model is shown after using L_{Ms} loss.

In addition, in order to observe the convergence process of the model, we plotted the loss change curves of the training process and the model accuracy of the linear evaluation, as shown in Fig. 8. From Fig. 8, we can see that the training loss decreases gradually with the training, and finally tends to stabilise, which indicates that the model is converging.

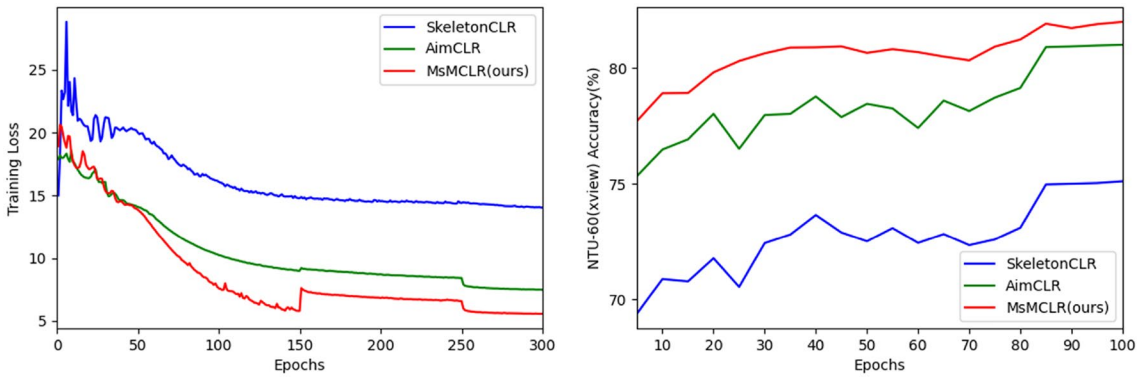

Fig. 8 Training loss curves and linear evaluation accuracy curves for SkeletonCLR, AimCLR and MsMCLR, baselined on NTU-60(xview) dataset

Table 8 Accuracy comparison of the results of different views (J, B, M) and the fusion of the three data views (all) on the NTU-60, NTU-120, and PKU-MMD datasets

Method	Stream	NTU-60(%)		NTU-120(%)		PKU(%)
		xsub	xview	xsub	xset	part I
SkeletonCLR [34]	Joint	68.3	76.4	56.8	55.9	80.9
AimCLR [29]	Joint	74.3	79.7	63.4	63.4	83.4
MsMCLR	Joint	**76.8**	**82.0**	**66.7**	**67.6**	**85.8**
SkeletonCLR [34]	Bone	69.4	67.4	48.4	52.0	72.6
AimCLR [29]	Bone	73.2	77.0	62.9	63.4	82.0
MsMCLR	Bone	**74.0**	**80.1**	**67.2**	**68.3**	**86.7**
SkeletonCLR [34]	Motion	53.3	50.8	39.6	40.2	63.4
AimCLR [29]	Motion	66.8	70.6	**57.3**	54.4	72.0
MsMCLR	Motion	**68.3**	**72.7**	56.1	**57.6**	**76.2**
3s-SkeletonCLR [34]	all	75.0	79.8	60.7	62.6	85.3
3s-AimCLR [29]	all	78.9	83.8	68.2	68.8	87.8
3s-MsMCLR	all	**80.3**	**85.4**	**70.3**	**71.8**	**88.6**

MsMCLR converges faster than AimCLR and SkeletonCLR, and the accuracy of MsMCLR is consistently ahead of the other two models when evaluated linearly, side by side confirming the superiority of our model.

4.4.4 Validity of MsMCLR model

As revealed in Table 8, the MsMCLR model was pre-trained for the joint, bone, and motion views, and linear evaluation was applied to each view of the model. The SkeletonCLR [34] and AimCLR [29] methods were used as baselines for comparison. It can be seen that MsMCLR performed much better than SkeletonCLR and AimCLR, with significant gains in the joint and bone views. For example, on the NTU-120 dataset, xset benchmark, and bone data view(B), the MsMCLR model accuracy exceeded that of the AimCLR baseline by 4.9% and of the SkeletonCLR baseline by 16.3%. This indicates the validity of the proposed model design.

Table 9 compares the number of parameters and the GFlops of the proposed method with those of AimCLR and SkeletonCLR. Compared with AimCLR, the addition of the MsM Attention module only brings a small number of parameters and calculations to the model, but further improves the accuracy of the model on this basis. The GFlops of the whole MsMCLR model were 2.2×10^2, which is lower than CrosSCLR(J+M), but the accuracy of MsMCLR was significantly higher, reflecting the effectiveness of the proposed method.

Table 9 Comparison of model parameters (Params) and floating-point operations (FLOPs). J+M represents the simultaneous input of joint and motion views

Method	NTU-60(%)		Params(M)	FLOPs(G)
	xsub	xview		
SkeletonCLR [34]	75.0	79.8	**1.8**	1.5×10^2
CrosSCLR(J+M) [34]	74.5	82.1	3.7	$\mathbf{2.9 \times 10^2}$
AimCLR [29]	78.9	83.8	1.8	2.2×10^2
MsMCLR	**80.3**	**85.4**	2.1	2.2×10^2

5 Conclusions

To focus more on the motion structure at each scale level, the MsMCLR model was proposed in this study. Firstly, several different data augmentations were considered to generate a large number of new motion patterns, which prompted the model to excavate more motion context and enhance its performance. The MsM attention mechanism classifies skeleton features into joint level, part level, and body level. Then, the attention mechanism incorporates multi-scale skeleton information, inter-frame motion features, and inter-node motion features, enabling concentrated attention on motion features at different scale levels to recognize action categories more accurately in videos. Finally, the embedding distribution of the ordinary augmentation branch is utilized to supervise the loss computation of the strong augmentation branch. The accuracy of the MsMCLR model in each linear evaluation protocol exceeds that of the other state-of-the-art models, illustrating its effectiveness. On the datasets NTU-60's xsub/xview, NTU-120's xsub/xview, and PKU-MMD's Part I, the accuracy of the MsMCLR model exceeds that of the other frontier models covered in the paper, which illustrates the effectiveness of the method. However, when facing the more challenging PKU-MMD Part II dataset, our method is slightly inferior, indicating that our method still needs to improve the robustness of the model in complex scenarios. The optimization of data processing and network structure will be our future improvement direction. The design of different attention mechanism modules will also be considered to improve the model performance.

Acknowledgements The authors would like to thank the reviewers and editor for their valuable comments and suggestions.

Author contributions Y. W. conceived the original idea and wrote the paper; M. Y. collected and tested skeleton-based motion recognition; Z. X., T. T., R. M., and Z. W. analyzed the data and revised the paper. All authors have read and agreed to the published version of the manuscript.

Funding This research was supported in part by the Guangxi Natural Science Foundation (2024GXNSFAA010493), National Natural Science Foundation of China (61862015, 62371350), Science and Technology Project of Guangxi (AD23023002, AD21220114), Guangxi Key Research and Development Program (AB17195025).

Data availability Data is openly available in a public repository. The NTU RGB+D 60 and NTU RGB+D 120 datasets that support the findings of this study are openly available in Rose Lab at https://doi.org/10.48550/arXiv.1604.02808; and https://doi.org/10.1109/TPAMI.2019.2916873. The PKU-MMD dataset that supports the findings of this study is openly available in PKU-NIP-Lab at https://doi.org/10.48550/arXiv.1703.07475.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. Qin, Z., Liu, Y., Perera, M., Gedeon, T., Ji, P., Kim, D., Anwar, S.: Anubis: Skeleton action recognition dataset, review, and benchmark. arXiv preprint arXiv:2211.09590. (2022)
2. Khan, M.A., Mittal, M., Goyal, L.M., Roy, S.: A deep survey on supervised learning based human detection and activity classification methods. Multimed. Tools and Appl. **80**(18), 27867–27923 (2021)
3. Varshney, N., Bakariya, B., Kushwaha, A.K.S.: Human activity recognition using deep transfer learning of cross position sensor based on vertical distribution of data. Multimed. Tools Appl. **81**(16), 22307–22322 (2022)
4. Guo, Z., Hou, Y., Xiao, R., Li, C., Li, W.: Motion saliency based hierarchical attention network for action recognition. Multimed. Tools Appl. **82**(3), 4533–4550 (2023)
5. Gao, J., Zhang, T., Xu, C.: I know the relationships: zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. Proc. AAAI Conf. Artif. Intell. **33**(1), 8303–8311 (2019)
6. Gao, J., Chen, M., Xu, C.: Vectorized evidential learning for weakly-supervised temporal action localization. IEEE transactions on pattern analysis and machine intelligence (2023)
7. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
8. Jalal, A., Kim, Y.H., Kim, Y.J., Kamal, S., Kim, D.: Robust human activity recognition from depth video using spatiotemporal multi-fused features. Pattern Recognit. **61**, 295–308 (2017)
9. Akula, A., Shah, A.K., Ghosh, R.: Deep learning approach for human action recognition in infrared images. Cogn. Syst. Res. **50**, 146–154 (2018)
10. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
11. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 667–676 (2019)
12. Karianakis, N., Liu, Z., Chen, Y., Soatto, S.: Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 715–733 (2018)
13. Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In: Proceedings of the Annual Conference on Neural Information Processing Systems (2020)
14. Wang, Y., Li, M., Cai, H., Chen, W., Han, S.: Lite pose: Efficient architecture design for 2d human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 13126–13136 (2022)
15. Zhou, Y., Li, C., Cheng, Z.Q., Geng, Y., Xie, X., Keuper, M.: Hypergraph transformer for skeleton-based action recognition. arXiv preprint arXiv:2211.09590 (2022)
16. Zhang, J., Lin, L., Liu, J.: Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. Proc. AAAI Conf. Artif. Intell. **37**(3), 3427–3435 (2023)
17. Peng, K., Yin, C., Zheng, J., Liu, R., Schneider, D., Zhang, J., Yang, K., Saquib Sarfraz, M., Stiefelhagen, R., Roitberg, A.: Navigating open set scenarios for skeleton-based action recognition. Proc. AAAI Conf. Artif. Intell. **38**(5), 4487–4496 (2024)
18. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for

skeleton-based action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13359–13368 (2021)
19. Chi, H., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 20186–20196 (2022)
20. Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., Tang, H.: Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In: Proceedings of the 28th ACM international conference on multimedia, pp. 55–63 (2020)
21. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1112–1121 (2020)
22. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1113–1122 (2021)
23. Kim, B., Chang, H.J., Kim, J., Choi, J.Y.: Global-local motion transformer for unsupervised skeleton-based action learning. In: Proceedings of the European Conference on Computer Vision, pp. 209–225 (2022)
24. Su, K., Liu, X., Shlizerman, E.: Predict cluster: Unsupervised skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9631–9640 (2020)
25. Yang, S., Liu, J., Lu, S., Er, M.H., Kot, A.C.: Skeleton cloud colorization for unsupervised 3d action representation learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 13423–13433 (2021)
26. Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
27. Wu, W., Hua, Y., Zheng, C., Wu, S., Chen, C., Lu, A.: Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. In: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops, pp. 224–229 (2023)
28. Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B.: Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Inf. Sci. **569**, 90–109 (2021)
29. Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R.: Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. Inf. Sci. **36**(1), 762–770 (2022)
30. Dang, L., Nie, Y., Long, C., Zhang, Q., Li, G.: Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 11467–11476 (2021)
31. Rao, H., Miao, C.: Skeleton prototype contrastive learning with multi-level graph relation modeling for unsupervised person re-identification. arXiv preprint arXiv:2208.11814. (2022)
32. Xu, B., Shu, X.: Pyramid self-attention polymerization learning for semi-supervised skeleton-based action recognition. arXiv preprint arXiv:2302.02327 (2023)
33. Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., Fu, Y.: Skeleton aware multi-modal sign language recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3413–3423 (2021)
34. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4741–4750 (2021)
35. Chen, Z., Liu, H., Guo, T., Chen, Z., Song, P., Tang, H.: Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition. arXiv preprint arXiv:2207.03065 (2022)
36. R., V., Chellapa, R.: Rolling rotations for recognizing human actions from 3d skeletal data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4471–4479 (2016)
37. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
38. Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser., Polosukhin, I.: Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems, 30 (2017)
40. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. **54**(10s), 1–41 (2021)
41. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multi-scale graph neural networks for 3d skeleton based human motion prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 214–223 (2020)
42. Gebotys, B., Wong, A., Clausi, D.A.: M2a: Motion aware attention for accurate video action recognition. In: Proceedings of the 19th IEEE Conference on Robots and Vision, pp. 83–89 (2022)
43. Wang, X., Qi, G.J.: Contrastive learning with stronger augmentations. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5549–5560 (2022)
44. Yoon, Y., Yu, J., Jeon, M.: Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition. Appl. Intell. **52**, 2317–2331 (2022)
45. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 60 (2019)
46. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
47. Liu, J., Shahroudy, A., Perez, M.L., Wang, G., Duan, L.Y., Chichung, A.K.: Ntu rgb+d 120: A large-scale benchmark for 3D human activity understanding. In: IEEE transactions on pattern analysis and machine intelligence (2019)
48. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: Pku-mmd: A large scale benchmark for skeleton-based human action understanding. In: Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, pp. 1–8 (2017)
49. Lin, L., Song, S., Yang, W., Liu, J.: Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2490–2498 (2020)
50. Nie, Q., Liu, Z.W., Liu, Y.H.: Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In: Proceedings of the European Conference on Computer Vision, pp. 102–118 (2020)
51. Zhou, Y., Duan, H., Rao, A., Su, B., Wang, J.: Self-supervised action representation learning from partial spatio-temporal skeleton sequences. Proc. AAAI Conf. Artif. Intell. **37**(3), 3825–3833 (2023)
52. Thoker, F.M., Doughty, H., Snoek, C.G.M.: Skeleton-contrastive 3D action representation learning. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1655–1663 (2021)
53. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1963–1978 (2019)

54. Van Der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.