AI

MDPI

Article

OTM-HC: Enhanced Skeleton-Based Action Representation via One-to-Many Hierarchical Contrastive Learning

Muhammad Usman [1,2], Wenming Cao [1,2], Zhao Huang [3], Jianqi Zhong [1,2] and Ruiya Ji [4,*]

[1] College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China; usmanmuhammad2022@email.szu.edu.cn (M.U.); wmcao@szu.edu.cn (W.C.); zhongjianqi2017@email.szu.edu.cn (J.Z.)
[2] Guangdong Key Laboratory of Intelligent Information Processing & Shenzhen University, Shenzhen 518060, China
[3] Department of Computer and Information Science, Northumbria University, Newcastle NE1 8ST, UK; zhao.huang@northumbria.ac.uk
[4] Department of Computer Science, Queen Mary University of London, London E1 4NS, UK
* Correspondence: r.ji@se24.qmul.ac.uk

Abstract: Human action recognition has become crucial in computer vision, with growing applications in surveillance, human–computer interaction, and healthcare. Traditional approaches often use broad feature representations, which may miss subtle variations in timing and movement within action sequences. Our proposed One-to-Many Hierarchical Contrastive Learning (OTM-HC) framework maps the input into multi-layered feature vectors, creating a hierarchical contrast representation that captures various granularities within a human skeleton sequence temporal and spatial domains. Using sequence-to-sequence (Seq2Seq) transformer encoders and downsampling modules, OTM-HC can distinguish between multiple levels of action representations, such as instance, domain, clip, and part levels. Each level contributes significantly to a comprehensive understanding of action representations. The OTM-HC model design is adaptable, ensuring smooth integration with advanced Seq2Seq encoders. We tested the OTM-HC framework across four datasets, demonstrating improved performance over state-of-the-art models. Specifically, OTM-HC achieved improvements of 0.9% and 0.6% on NTU60, 0.4% and 0.7% on NTU120, and 0.7% and 0.3% on PKU-MMD I and II, respectively, surpassing previous leading approaches across these datasets. These results showcase the robustness and adaptability of our model for various skeleton-based action recognition tasks.

Keywords: skeleton-based action representation learning; unsupervised learning; hierarchical contrastive learning; one-to-many

Citation: Usman, M.; Cao, W.; Huang, Z.; Zhong, J.; Ji, R. OTM-HC: Enhanced Skeleton-Based Action Representation via One-to-Many Hierarchical Contrastive Learning. *AI* **2024**, *5*, 2170–2186. https://doi.org/10.3390/ai5040106

Academic Editor: Demos T. Tsahalis

Received: 11 September 2024
Revised: 22 October 2024
Accepted: 29 October 2024
Published: 1 November 2024

1. Introduction

Recognizing human actions is extremely important in various domains such as human–computer interaction, intelligent surveillance, video content analysis, game control, and others [1–3]. In recent years, significant advancements have been observed in 3D skeleton-based action recognition using deep learning networks [4–6]. However, acquiring a more critical representation of skeletons remains unresolved in skeleton-based action recognition. To address this, a significant number of studies [5,7–11] have been carried out that employ a fully supervised approach to train a network. However, such approaches required substantial annotated 3D skeleton sequential data, demanding high costs and computational time. Unsupervised learning, on the other hand, has emerged as a promising alternative, allowing models to learn from data by leveraging intrinsic correlations without requiring labeled samples. Recently, there has been an emerging trend in the field of action recognition learning known as unsupervised skeleton-based action representation learning (SKARL) [12–14]. This unsupervised SKARL scheme has gained remarkable attention due to its potential to alleviate the burden of manual annotation.

Existing efforts in action representation learning using SKARL fall into three distinct categories: encoder–decoder-based methods [11,14], contrasting learning methods [15–17], and hybrid methods [18,19]. The encoder–decoder frameworks initially transform the input skeleton sequence into hidden features, which encompass skeleton reconstruction [11], skeleton colorization prediction [20], and forecasting skeleton displacement prediction [21]. However, the contrasting learning techniques typically include the augmentation of an input skeleton sequence, resulting in two augmented instances. The objective is to train an encoder that produces more comparable representations for instances with the same skeleton while ensuring those instances with different skeletons have dissimilar representations. On the other hand, the hybrid approaches incorporate the features of encoder–decoder and contrasting learning. The initial approach of unsupervised SKARL [15] involved the adaptation of momentum contrast learning (MoCo) [22], originally designed for recognizing images without supervision, which were adapted for identifying skeletal movements without supervision. Subsequently, several enhanced contrastive learning methodologies were introduced, which included the incorporation of uncertainty modeling for skeletons [23], the investigation of additional positive pairings [22], and the use of advanced skeleton-specific enhancements [16]. Contrastive learning methods typically begin by transforming skeleton sequences denoted as X into features at the instance level that conduct holistic instance-level contrast. Although the effectiveness of contrastive learning approaches has been demonstrated, it is worth considering that holistic instance-level contrast may not fully optimize the hierarchical structures inherent in human skeletons. A skeleton sequence is commonly interpreted as either a temporal sequence of complete skeleton frames or a spatial arrangement of skeleton joints. The frames or joints are fundamental elements in the temporal or spatial domains. These can be further organized into larger-scale structures, such as frame clips in the temporal domain or body parts in the spatial domain.

In this paper, we introduce a novel contrastive learning model termed OTM-HC (One-to-Many Hierarchical Contrast) to enhance unsupervised SKARL, drawing inspiration from the hierarchical organization of the human skeletal structure. The OTM-HC model (as presented in Figure 1) strategically combines three input modalities, offering respective advantages. The first input is the *skeleton sequence* X , where skeleton sequences are collected that represent the subject's movements. The collected data are derived from both the temporal and spatial domains. Capturing temporal and spatial data allows for a more detailed and nuanced analysis of dynamic patterns and movement characteristics. The second input is the *velocity calculation (vel)*. We calculate velocity data from the original skeleton sequence to enhance the model's understanding of dynamic components, visually representing how actions evolve and transform over time. This emphasizes the model's proficiency in decoding and identifying dynamic elements, a critical aspect of tasks related to human motion analysis. In a harmonized integration, the two inputs above are meticulously merged through concatenation. This fusion endows the model with an enhanced and adaptable feature representation. Input three is the *reference skeleton sequence* X, an identical duplicate of the initial data retained for reference, debugging, or future use. Third, input incorporation is also useful for method control and preventing mislearning. By merging these inputs through hierarchical contrastive learning, OTM-HC improves upon traditional methods by capturing multi-level action representations. Specifically, the hierarchical approach allows the model to process data at instance, domain, clip, and part levels, which enables it to capture both fine-grained and high-level details of human movements. Unlike previous flat contrastive learning models, OTM-HC's hierarchical structure allows it to separate and recognize actions with higher granularity, making it more robust and adaptable to complex action recognition tasks. In summary, OTM-HC offers the following contributions:

1. The OTM-HC enhances hierarchical encoder networks for skeleton sequences by incorporating advanced Seq2Seq encoders with unified downsampling modules. This model effectively captures complex temporal and spatial details, ensuring full congruence with advanced Seq2Seq transformer encoders.

2. The OTM-HC is a contrastive-learning-based model designed to enhance unsupervised SKARL by incorporating three inputs: skeleton sequences (X), velocity data (Vel), and reference skeleton sequences (X). These inputs capture actions' static and dynamic features, enriching the OTM-HC ability to identify comprehensive patterns and further strengthening unsupervised SKARL.
3. Comprehensive experiments on the four datasets show the uniqueness of OTM-HC and set a new benchmark for unsupervised SKARL, showing the strong transferability of the learned representations.

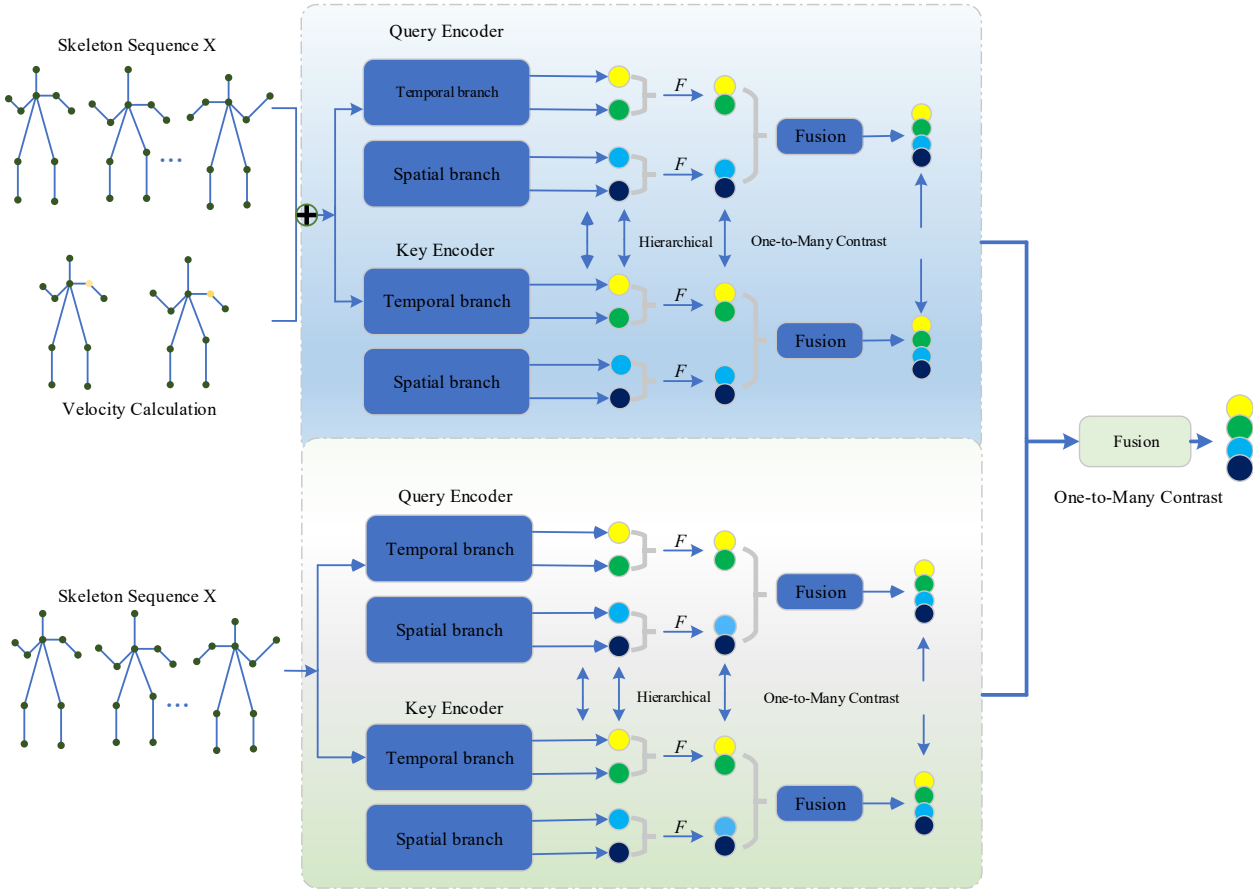


Figure 1. The proposed OTM-HC methodology employs a one-to-many hierarchical contrastive learning model, utilizing three key inputs: original skeleton sequences, velocity data, and reference skeleton sequences. This approach efficiently develops unsupervised action representations, focusing on both static and dynamic aspects of skeletal actions.

The rest of the paper is organized as follows: Section 2 reviews the existing studies in related work. Section 3 introduces our proposed model for unsupervised action representation learning. Section 4 details the experimental environment and the employed methodologies. Section 5 discusses the insights from the ablation study. Section 6 encapsulates our findings in the conclusion.

2. Related Work

2.1. Skeleton-Based Action Recognition

In computer vision research, recognizing actions based on skeletal information is a fundamental and complex study area. Motion recognition algorithms depending on skeletal structure are typically implemented by leveraging geometric interconnections among skeleton joints [24–27]. Contemporary approaches greatly emphasize deep neural networks; for instance, in [28], a hierarchical recurrent neural network (RNN) is used for processing body key points. In addition, attention-based techniques suggest different pathways of

automatically identifying significant skeleton joints [29–32] and video frames [30,31] to improve adaptive learning of the concurrent manifestation of skeleton joints. Recurrent neural networks frequently encounter the issue of gradient vanishing [33], which can lead to challenges in optimization. Recently, the interest has been in graph convolution networks (GCNs) [34,35] for skeleton-based action recognition. This approach introduces spatial–temporal graph convolution networks to extract spatial and temporal structural characteristics from skeleton data. Building upon the success of the transformer model, as reported in [17,36], recent studies [37,38] have started employing its proficient capabilities in sequence processing for tasks involving skeleton data.

2.2. Contrastive Learning

The study in [15] initiated the application of contrastive learning techniques within the domain of unsupervised SKARL. Various techniques [39–42] utilize a representation learning method that compares positive and negative pairs. The technique aims to enhance the similarity between representations of positive pairs while increasing the dissimilarity between representations of negative pairs. The primary area of attention is to explore the construction of pairs to get resilient representations. The simple framework for contrastive learning (SimCLR), as suggested by [43], employs various data augmentation techniques, including random cropping, Gaussian blur, and color distortion, to generate positive samples. In the study [43], a memory module with a queue structure is implemented to store negative samples. This queue undergoes regular updates throughout the training procedure. Notably, contrastive learning has attracted considerable interest among researchers in unsupervised SKARL. The momentum contrast (MoCo) method enables contrastive learning through a singular stream approach [22]. To leverage the knowledge from the other streams, a multiview contrastive learning approach is introduced in [44], while multiple models to acquire knowledge from various representations of skeletons are studied in [45]. The study in [16] introduced targeted spatial and temporal augmentations for skeleton data, aiming to enhance the spatio-temporal consistency in the learned representations, wherein three different types of encoders are used to make contrasts across various architectures. In another study by [23], skeleton sequences were converted into a probabilistic embedding space. They measured similarity based on how close these probabilistic distributions were to each other.

2.3. Hierarchical Human-Skeleton-Based Modeling

Many successful attempts have been made to explore the hierarchical representation of human skeletons; these include several studies like [6,26,46], where hand-crafted designs are employed to represent the spatial hierarchical structure. However, they did not examine the temporal hierarchical structure, which is crucial for capturing the dynamic aspects of human actions over time. In [19], temporal hierarchical cues are highlighted at the frame, clip, and video levels to enhance their methodology. The Hierarchical Transformer framework helps us model spatial and temporal features [46], and they are further inspired by hierarchical transformer's (Hi-TRS's) self-supervised hierarchical pre-training scheme, which captures these connections at multiple levels [19]. Previous works have modeled human skeletons with hand-crafted designs, focusing primarily on the spatial hierarchical structure while neglecting the temporal dimension, highlighting key limitations in these approaches. Conversely, our modified approach represents both temporal and spatial hierarchical structures, facilitating easier implementation and leading to a more comprehensive understanding of action dynamics. Additionally, unlike prior models that depicted skeletons solely at the instance level, our proposed scheme captures the representation of skeletons across multiple levels: instance, domain, clip, and part.

3. The Proposed Framework

The flowchart of the proposed OTM-HC model, illustrated in Figure 2, contains two primary components: (A) a hierarchy-based encoder network that effectively captures

comprehensive information from temporal and spatial domains, resulting in hierarchical multi-level representations and (B) One-to-Many hierarchical contrastive learning, a method that improves the contrastive learning process by facilitating one-to-many comparisons across four levels. This approach aligns with the hierarchical structure of human skeleton systems, providing an efficient and unsupervised learning environment. In contrast, the flexibility of graphic representation is enhanced by employing the attention mechanisms [29,47,48]. These mechanisms are designed to dynamically identify unique attributes by analyzing spatial arrangements and temporal variations within the data. Next, the two main components of OTM-HC are discussed in detail.

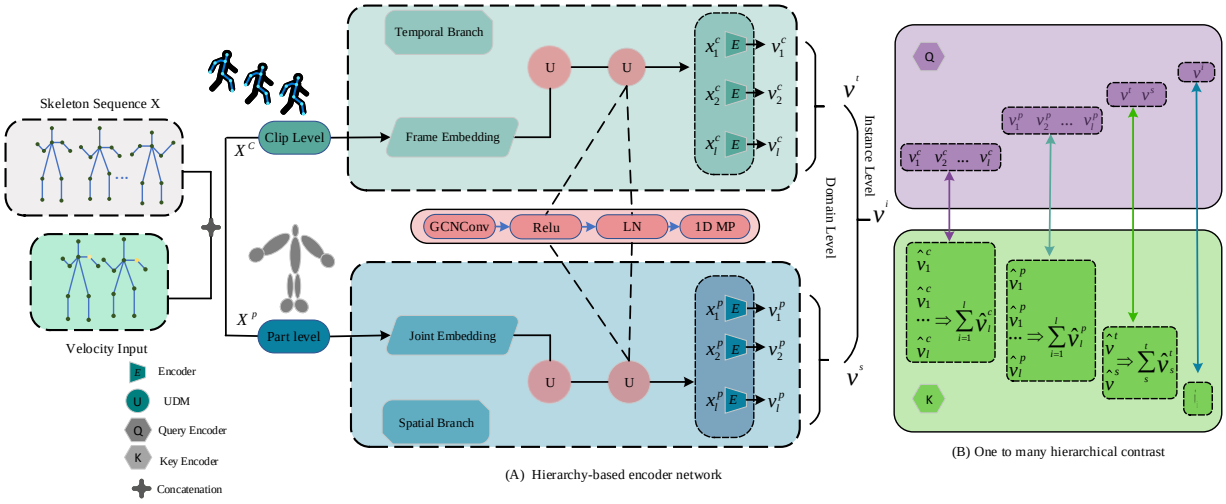


Figure 2. (**A**) The OTM-HC model is built around a hierarchical encoder network, which seamlessly integrates a Seq2Seq encoder with innovatively designed Unified Downsampling Modules (UDM). (**B**) OTM-HC employs a hierarchical one-to-many contrast approach, conducting contrastive learning across four levels, utilizing a shared encoder within the same branch.

3.1. A Hierarchy-Based Encoder Network

The hierarchy-based encoder network is composed of temporal and spatial branches, each tailored to fulfil a distinct role in processing the concatenated input data, which consists of the *skeleton sequence X* and the calculated *velocity (vel)*. The temporal branch is tasked with encoding data from the temporal domain, with a specific focus on capturing the dynamics of time across various levels of granularity, leading to the creation of clip-level representations. On the other hand, the spatial branch is dedicated to processing spatial domain information, generating part-level representations by encoding at multiple spatial granularities. These clip-level and part-level representations are then utilized to construct domain-level and instance-level representations. The architectural design of this network effectively combines the temporal and spatial characteristics of the input data, thereby enabling robust feature extraction for downstream tasks such as action recognition and retrieval analysis.

3.1.1. Clip-Level Representation

In the context of a skeleton sequence denoted by $X \in \mathbb{R}^{T\times J\times 3}$, where T denotes the number of frames and J represents the number of joints, with each joint defined by three spatial coordinates, the initial step involves transforming the sequence into a time-major format. This results in a sequential list of frames $X^c = \{x_i^c\}_{i=1}^{T}, \in \mathbb{R}^{T\cdot 3J}$, where x_i^c denotes the i-th frame, representing a complete human skeleton. A frame embedding transforms the input into a C-dimensional dense feature space using two fully connected layers. To be more specific, a transformed feature is generated from this process x_i^c and is obtained as:

$$x_i^c = W_2(\sigma(W_1 x_i^c + b_1)) + b_2. \quad (1)$$

where the transformation matrices are $W_1 \in \mathbb{R}^{C \cdot 3J}$ and $W_2 \in \mathbb{R}^{C \cdot C}$, the bias vectors are $b_1 \in \mathbb{R}^{\mathbb{C}}$ and $b_2 \in \mathbb{R}^{\mathbb{C}}$, and the ReLU activation function is denoted by σ. This embedding encapsulates the distinct frame-level attributes of each alignment, maintaining information on joint positions within an organized feature space. Moreover, the aggregate representation of all predicted attributes is expressed as $X_1^c = \{x_i^c\}_{i=1}^{T} \in \mathbb{R}^{T \cdot C}$. These characteristics are designated as the basic frame representation for ease of reference. However, frame-level features individually are insufficient for capturing the temporal relationships necessary for representing dynamic events. We propose dividing the sequence into many temporal segments, enabling the model to analyze movement variations across different timescales. The first step involves creating clips with varying temporal granularities, after which Seq2Seq encoders are employed to capture and represent temporal dependencies within each clip.

3.1.2. Segment Creation

A unified downsampling module (UDM) is introduced to generate clips with varying temporal granularities, which effectively combines consecutive frame or clip data sequentially, creating increasingly coarse-grained clips. The organizational structure of UDM is illustrated in Figure 2 and can precisely be described by the following equation:

$$U(\cdot) = MaxPool1D[\ln(\sigma(GCNConv(\cdot)))] \tag{2}$$

where $GCNConv$ denotes a one-dimensional graph convolutional operation with a kernel size of 5 and a stride size of 1. The symbol σ represents the ReLU activation function, which applies non-linearity by filtering out negative values from the $GCNConv$ output, thereby retaining only positive activations. LN stands for Layer Normalization, and $MaxPool1D$ refers to a one-dimensional max pooling with a kernel size of 2. A one-dimensional convolutional layer allows for collecting contextual information nearby, while max pooling helps aggregate this information. UDM can be configured in series to produce clips with a more coarse-grained structure. Additionally, frames are treated as individual clips with a granularity of 1 for descriptive purposes. Although the resulting feature vectors differ in detail, they are expected to exhibit similarities in their predominant sense because they represent related actions or features. These shared characteristics reflect underlying patterns or behaviours common to the overall structure of the activity, which the UDM captures. Using UDM, it is straightforward to obtain clips with higher granularity, as shown in the following equation:

$$X_{n+1}^c = U(X_n^c). \tag{3}$$

We can obtain clips with various temporal granularities by applying UDM $L-1$ times in series, such as $X_1^c, X_2^c, \ldots, X_L^c$.

3.1.3. Temporal Maximum Pooling

The temporal relationship for each granularity level of clips is further simulated using a Seq2Seq encoder. To elaborate, the clips X_n^c are input into a Seq2Seq encoder transformer, followed by a temporal max pooling layer ($TMaxP$) for feature aggregation. As a result, the clip-level feature v_n^c at the granularity of n is produced as:

$$v_n^c = TMaxP(Seq2Seq(X_n^c)). \tag{4}$$

This method leverages the outputs from all time steps of the Seq2Seq encoder. Thus, when provided with a set $X_1^c, X_2^c, \ldots, X_L^c$ of L granularities, we can derive the ultimate clip-level representation, denoted as V^c, which encompasses L feature vectors:

$$V^c = \{v_1^c, v_2^c, \ldots, v_L^c\}. \tag{5}$$

3.1.4. Part-Level Representation

The process of acquiring the part-level representation is similar to obtaining the clip-level representation. Therefore, we primarily focus on specifying the distinct choices at the part level. Here, we take the skeleton sequence $X \in \mathbb{R}^{T \cdot J \cdot 3}$ and transform it into a space-major domain list of joints $X^p = \{x_i^p\}_{i \in \mathbb{J}}, \in \mathbb{R}^{J \cdot 3T}$, where x_i^p represents the i-th joint, essentially a progression of a joint over time. Similarly, a joint embedding is employed to derive the first joint representation, $X_1^P \in \mathbb{R}^{J \cdot C}$. We utilize the Unified Downsampling Module (UDM) to capture diverse spatial granularities by aggregating nearby joints into increasingly larger body segments. In this context, UDM combines adjacent joints or components to form larger-sized parts. Sequentially using UDM $L-1$ times enables us to obtain components with varying spatial resolutions $X_1^p, X_2^p, \ldots, X_L^p$. Each level's spatial dependencies are then encoded using a Seq2Seq encoder combined with max pooling, resulting in the final part-level representation $V^p = \{v_1^p, v_2^p, \ldots, v_L^p\}$. Pooling operations at each granularity reduce noise by focusing on the most significant movements within each part, yielding a robust feature representation.

3.1.5. Domain-Level and Instance-Level Representation

We currently have a clip-level representation encompassing diverse temporal resolutions, symbolized as V^t, along with a part-level representation that captures a range of spatial resolutions, indicated as V^s. These multi-granular representations are systematically integrated to construct the domain-level and instance-level representations. The temporal-domain representation, v^t, is derived from synthesizing all L feature vectors at the clip level. Similarly, the spatial-domain representation, v^s, is obtained by combining all corresponding feature vectors from the part level:

$$\begin{aligned} v^t &= F(V^c) = F(v_1^c, v_2^c, \ldots, v_L^c), \\ v^s &= F(V^p) = F(v_1^p, v_2^p, \ldots, v_L^p), \end{aligned} \tag{6}$$

where $F(\cdot)$ denotes the fusion operator applied across several feature vectors, with concatenation being on the method used in our implementation. Furthermore, the integration of temporal-domain and spatial-domain representations into the instance-level representation v^i is accomplished by concurrently combining them:

$$v^i = F(v^t, v^s). \tag{7}$$

It is important to note that while several hierarchical representations have been developed [5,19], the ultimate skeletal sequence representation relies solely on the instance-level representation v^i. As explained in a subsequent section, alternative representations convey the one-to-many hierarchical contrasts.

3.2. One-to-Many Hierarchical Contrasts

We introduce an innovative method beyond traditional instance-level contrast in a one-to-many hierarchical contrastive approach. The method extends the contrastive analysis to multiple feature levels within the data, encompassing instance-level, domain-level, clip-level, and part-level features. To achieve this, our approach draws inspiration from the evolving MoCo framework, as detailed in [43]. Our approach utilizes a query and a key encoder, operating alongside a dynamic dictionary queue and a moving averaged update mechanism. Moreover, following the scheme in [49], we employ a two-layer multilayer perceptron for feature projection before contrastive analysis. We maintain the same notation for the projected features as the original work for clarity and brevity in our descriptions. We use a hat symbol ($\hat{}$) to signify features derived from the key encoder and m to denote features from the queue. This one-to-many hierarchical contrastive methodology represents a significant advancement in contrastive learning, enabling a more

comprehensive understanding of data patterns and relationships across multiple levels of abstraction.

3.2.1. One-to-Many Instance-Level Contrast

Following the study conducted by [50], we employ the noise contrastive estimating loss InfoNCE [51] as a means of contrast. The contrastive loss is calculated at the instance level as follows:

$$L_{instance} = -log\frac{exp(v^i.\hat{v}^i/\tau)}{exp(v^i.\hat{v}^i/\tau) + \sum_{m_i^j \in M^i} exp(v^i.m_j^i/\tau)}, \tag{8}$$

where τ represents the temperature hyper-parameter and the term m_{i_j} refers to the jth negative sample from the first-in-first-out queue M_i, which contains previously projected features at the instance level.

3.2.2. One-to-Many Domain-Level Contrast

At the domain level, it is postulated that a skeleton sequence demonstrates parallelisms in representations across both temporal and spatial domains. This assumption is based on the idea that temporal and spatial representations provide distinct yet complementary insights into the same skeleton sequence and should, thus, align with high-level semantics. To realize this, we introduce a hierarchical contrastive loss function that targets two cross-domain feature sets, representing temporal-spatial (v_t-v_s) domain features, respectively. This loss function is designed to encourage the network to align the temporal and spatial features of the same skeleton sequence while facilitating the differentiation of features from different skeleton sequences:

$$\begin{aligned} L_{domain} = &-log\frac{exp(v^t.\hat{v}^s/\tau)}{exp(v^t.\hat{v}^s/\tau) + \sum_{m_j^s \in M^s} exp(v^t.m_j^s/\tau)} \\ &-log\frac{exp(v^s.\hat{v}^t/\tau)}{exp(v^s.\hat{v}^t/\tau) + \sum_{m_j^t \in M^t} exp(v^s.m_j^t/\tau)}. \end{aligned} \tag{9}$$

3.2.3. One-to-Many Clip-Level Contrast

It is crucial to note that the clip-level representation comprises multiple feature vectors, each varying in temporal granularity. Although these vectors differ in detail, they should exhibit similarities in their predominant sense. Consequently, we consider clip-level attributes affirmatively across various granularity levels within the same input instance. For implementation purposes, the anchor sample is considered the clip-level feature at a granularity of 1 and is deemed positive with other granularities of the same input. The contrastive loss at the clip level is calculated using the InfoNCE method, which includes a greater number of positive pairs:

$$L_{clip} = -log\frac{\sum_{l=1}^{L} exp(v_1^c.\hat{v_l^c}/\tau)}{\sum_{l=1}^{L} exp(v_1^c.\hat{v_l^c}/\tau) + \sum_{m_j^c \in M^c} exp(v_1^c.m_j^c/\tau)}, \tag{10}$$

3.2.4. One-to-Many Part-Level Contrast

The contrastive loss at the part level is computed based on the representation of individual parts, similar to the approach used for clip-level contrast. It is expressed as:

$$L_{part} = -log\frac{\sum_{l=1}^{L} exp(v_1^p.\hat{v_l^p}/\tau)}{\sum_{l=1}^{L} exp(v_1^p.\hat{v_l^p}/\tau) + \sum_{m_j^p \in M^p} exp(v_1^p.m_j^p/\tau)}. \tag{11}$$

Ultimately, the network is trained by minimizing the cumulative losses described above. The calculation for deriving the total loss is specified as follows:

$$L_{\text{total}} = L_{\text{instance}} + L_{\text{domain}} + L_{\text{clip}} + L_{\text{part}}.$$

4. Experimental Environment

4.1. Datasets

NTU-60 [42]. NTU-RGB+D is a widely used human interaction dataset featuring multiple views and subjects. It is one of the most popular datasets for skeleton-based tasks, including more than 56,880 action samples from NTU-60, which covers 60 different action categories. We follow two suggested protocols: (a) Cross-Subject (xsub), where the data for training and testing are collected from different subjects to ensure a cross-subject approach, and (b) Cross-View (xview), where the data used for training and testing are gathered from various camera perspectives.

NTU-120 [5]. NTU-120 covers 120 action categories and includes more than 114,480 action examples. Two recommended protocols are adopted: (a) Cross-Subject (xsub), where the training and testing data are obtained from 106 different subjects, and (b) Cross-Setup (xset), where the data for training and testing are gathered from 32 different setups.

PKU Multi-Modality [52]. PKU-MMD is an extensive dataset that provides a comprehensive 3D analysis of human actions, consisting of nearly 20,000 scenarios and 51 action labels. The dataset is divided into two subsets: Part I, a simplified version, and Part II, which presents more complex data due to significant variations in perspective and utilizes a cross-subject protocol.

4.2. Implementation Details

To ensure an equal comparison, we utilize the data augmentation techniques outlined by [16], which include shearing, joint jittering, and temporal cropping. Shearing and joint jittering are spatial transformations: shearing randomly rotates the poses of the skeleton, while joint jittering randomly shifts the joints around their original positions. Temporal cropping improves sequential data by randomly selecting a starting frame and resampling it to a predefined length at different intervals. We employ a Seq2Seq Transformer encoder, using a one-layer Transformer for S2S encoding. Note that the encoders for all granularities within a branch are shared. We set the model dimension C and encoder output size to 512 to balance computational efficiency with representational capacity. The output dimension for MLP projection is set to 128, allowing for compact feature embeddings during contrastive learning. The dynamic dictionary queue size is 2048, providing diverse negative samples, while the temperature value τ is set to 0.2 to sharpen distinctions between positive and negative pairs. For optimization, we use SGD with a momentum of 0.9 and a weight decay of 0.0001 to stabilize convergence and reduce overfitting. The model is trained for 450 epochs with an initial learning rate of 0.01, decayed by a factor of 0.1 at epoch 350. We employ a mini-batch size of 64, balancing memory efficiency with convergence stability. Performance is measured using Top-1 Accuracy for classification tasks.

4.3. Quantitative Evaluation

Two common evaluation techniques, cross-subject (xsub) and cross-view (x-view), are implemented on NTU-60 and NTU-120 datasets. In reference [13], the x-sub analysis findings are presented on the PKU-MMD I and II datasets. The metric of highest accuracy, denoted as Top-1 accuracy, is commonly employed to evaluate performance.

4.4. Comparison with the Previous State-of-the-Art Methods

In this study, our objective is to precisely compare the robustness of the OTM-HC approach against a range of contemporary unsupervised methods that currently define the state-of-the-art. The analyzed methods include encoder–decoder models, contrastive learning techniques, and innovative hybrid models. Specifically, our analysis encompasses meth-

ods such as ISC [16], MS^{2}L [13], Colorization [20], H-Transformer [46], GL-Transformer [21], SkeAttnCLR [45], ActCLR [44], Predict & Cluster [12], AimCLR [17], HYSP [53], HaLP [54], Masked Colorization [55], and Hico-T [56]. Our evaluation framework focuses on two tasks central to the current research: skeleton-based action recognition and retrieval. These tasks have been the focal points of several notable studies, including MS^{2}L [13] and Skeleton-Contrastive [16], among others. It is crucial to emphasize that to tackle these downstream tasks effectively, each model undergoes an initial phase of unsupervised training conducted without reliance on labeled data, highlighting the genuinely unsupervised nature of the models. Once this foundational training phase is complete, the models are well-prepared for subsequent evaluative tasks.

4.4.1. Joint-Based Action Recognition

In this task, an additional linear classifier, specifically a fully connected layer, is integrated into the skeleton sequence representation derived from the pre-training model corresponding to the target dataset. The classifier is initially trained on the preceding dataset. Following the methodologies of previous studies [13,17,57], the pre-training model is kept in a frozen state, with the primary focus of training being on the linear classifier. Table 1 summarizes the performance achieved on the NTU-60, NTU-120, PKU-MMD I, and PKU-MMD II datasets. The OTM-HC models we propose consistently show superior performance compared to earlier techniques across various categories, often by a significant margin. Among the four datasets, the OTM-HC models exhibit the lowest performance on PKU-MMD II. This is attributed to the greater challenges presented by this dataset, primarily due to increased noise from variations in viewpoint [17]. Even on this challenging dataset, our state-of-the-art model outperforms the existing methodologies. These results underscore the usefulness of the proposed OTM-HC framework for unsupervised skeleton-based action recognition.

4.4.2. Joint-Based Action Retrieval

Here, we simulated the conditions described in [16,17]. When dealing with an action query, the training samples are examined to identify the most similar action, using cosine similarity as the metric. The results for the NTU-60 and NTU-120 datasets are presented in Table 2. Our OTM-HC model, which employs a single transformer encoder, consistently prevails over the existing methods. These findings prove that the action representation derived from the OTM-HC model exhibits a higher level of discrimination.

Table 1. Comparisons to the best state-of-the-art methods for the downstream task of skeleton-based action recognition across four different datasets. A plus sign ($^+$) indicates that the results incorporate multiple modalities, such as joint, bone, and motion views. In contrast, other results use only a single joint view as input. The best result is highlighted in bold, while the second-best result is underlined.

Method	Type	Encoder	NTU-60		NTU-120		PKU-MMD I	PKU-MMD II
			x-Sub	x-View	x-Sub	x-Setup	x-Sub	x-Sub
ISC (2021)[16]	Contrastive Learning	GRU&CNN&GCN	76.3	85.2	67.1	67.9	80.9	36.0
MS2L (2020) [13]	Hybrid Learning	GRU	52.6	-	-	-	64.9	27.6
Colorization (2021) [20]	Encoder–Decoder	GCN	75.2	83.1	-	-	-	-
H-Transformer (2021) [46]	Encoder–Decoder	Transformer	69.3	72.8	-	-	-	-
GL-Transformer (2022) [21]	Encoder–Decoder	Transformer	76.3	83.8	66.0	68.7	-	-
AimCLR (2022) [17]	Contrastive Learning	GCN	74.3	79.7	63.4	63.4	87.8$^+$	38.5$^+$
SRCL (2022) [58]	Contrastive Learning	GCN	76.7	82.5	67.1	67.5	-	-
HYSP (2023) [53]	Contrastive Learning	GCN	78.2	82.6	61.8	64.6	83.8	-
HaLP (2023) [54]	Contrastive Learning	GRU	79.7	86.8	71.1	72.2	-	43.5
SkeAttnCLR (2023) [45]	Contrastive Learning	GCN	80.3	86.1	66.3	<u>74.5</u>	87.3	**52.9**

Table 1. *Cont.*

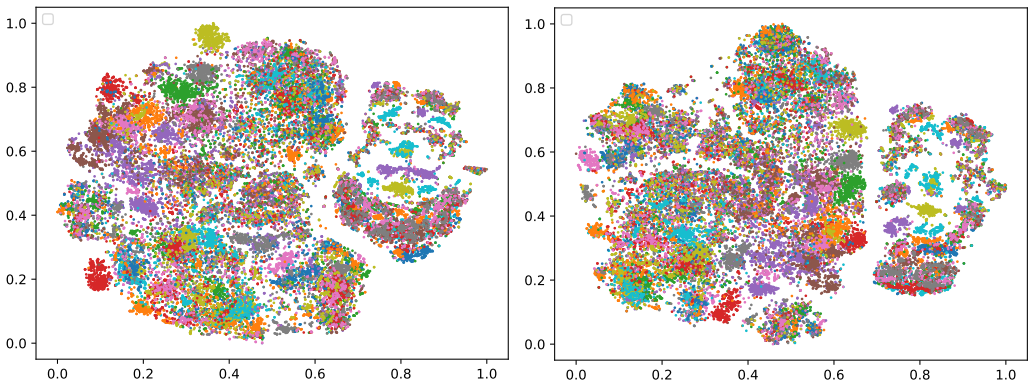
Method	Type	Encoder	NTU-60		NTU-120		PKU-MMD I	PKU-MMD II
			x-Sub	x-View	x-Sub	x-Setup	x-Sub	x-Sub
ActCLR (2023) [44]	Contrastive Learning	GCN	80.9	86.7	69.0	70.5	-	-
Hico-T (2023) [56]	Contrastive Learning	Transformer	81.1	88.6	72.8	74.1	89.3	49.4
Masked Colorization (2024) [55]	Contrastive Learning	DGCNN	79.1^{+}	87.2^{+}	69.2^{+}	70.8^{+}	89.2^{+}	49.8^{+}
Our work OTM-HC	Contrastive Learning	Transformer	**82**	**89.2**	**73.2**	**74.8**	**89.9**	50.1

Table 2. This study aims to compare state-of-the-art algorithms for skeleton-based action retrieval using two datasets: NTU-60 and NTU-120. The best result is highlighted in bold, and the next best result is underlined.

Methods	Encoder	NTU-60		NTU-120	
		x-Sub	x-View	x-Sub	x-Set
AimCLR (2022) [17]	GCN	62.0	-	-	-
Predict&Cluster (2020) [12]	GRU	50.7	76.3	39.5	41.8
ISC (2021) [16]	CNN	62.5	82.6	50.6	52.3
HaLP (2023) [54]	GCN	65.8	83.6	55.8	59.0
SkeAttnCLR (2023) [45]	Transformer	69.4	76.8	46.7	58.0
Hico-T (2023) [56]	Transformer	68.3	84.8	56.6	59.1
Our work OTM-HC	Transformer	**70.74**	**85.44**	**57.1**	**61.07**

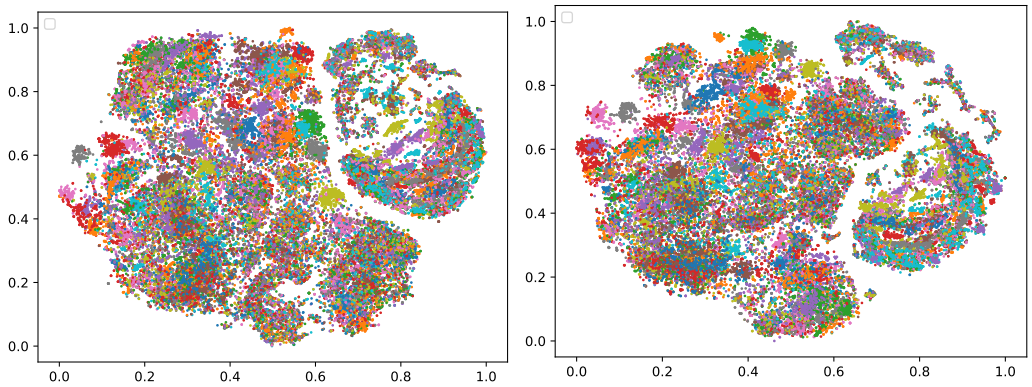
4.5. Visualization Results

The t-SNE (t-distributed Stochastic Neighbor Embedding) visualizations were utilized to project high-dimensional joint stream data into a 2D space, revealing the clustering patterns within the NTU60 and NTU120 datasets. This technique highlights how well the model separates different action classes. We optimized the t-SNE perplexity by minimizing the Davies-Bouldin Index (DBI), ensuring clearer cluster separation. This approach allowed us to observe distinctive patterns and relationships, emphasizing the model's robustness in capturing complex human activities, as depicted in Figure 3.



(**a**) (**b**)

Figure 3. *Cont.*



(**c**) (**d**)

Figure 3. t-SNE visualizations of feature embeddings derived from the one-to-many hierarchical contrast model on the NTU-60 and NTU-120 datasets, focusing on the joint stream. The distinct clusters reveal the model's ability to capture global action patterns and fine-joint-specific nuances. These visualizations illustrate how the model effectively balances the representation of local movement details with global structural patterns, enhancing its capability to differentiate complex human actions. (**a**) NTU60 x-sub; (**b**) NTU60 x-view; (**c**) NTU120 x-sub; (**d**) NTU120 x-set.

5. Ablations Studies

We thoroughly examine all the improvements made to the proposed OTM-HC model. We perform ablation studies on the NTU-60 dataset, specifically within the framework of the skeleton-based action recognition downstream task. In these experiments, we use a transformer as the standard encoder for our model.

5.1. Efficiency of One-to-Many Hierarchical Contrast

The results of the one-to-many hierarchical comparison across various features at different levels, namely, instance, domain, and clip & part levels—are presented in Table 3. All models utilize the hierarchical encoder network. The instance level represents individual sample features, while the domain level aggregates features across broader categories, and the clip & part level captures finer-grained segmentation within each action sequence. Notably, the model employs instance-level contrast using the same contrastive learning methods as CrosSCLR [57] and AS-CAL [15]. Including domain-level and clip & part-level one-to-many contrasts leads to improved outcomes, demonstrating the effectiveness of the one-to-many hierarchical contrast approach. Performance is evaluated using Top-1 Accuracy, which measures the model's classification accuracy by comparing the highest-probability predictions with the ground truth.

Table 3. Performance of one-to-many hierarchical contrast on multi-level characteristics. All variants apply different loss functions to a shared representation that captures multiple levels of detail. The best result is highlighted in bold.

Instance Level	Domain Level	Clip & Part Level	W/OTM-HC	x-Sub	x-View
yes	yes	yes	NO	80.5	88.3
yes	yes	yes	yes	**82**	**89.2**

5.2. Efficiency of Different Fusion Methods

We also examine various fusion methods for combining two branches: element-wise sum, Hadamard product, and weighted sum. The findings are presented in Table 4, indicating that the concatenation operation yields the most favourable outcomes.

Table 4. The evaluation of our model performance utilizing various fusion techniques on the NUT-60 dataset. The best result is highlighted in bold.

Fusion Methods	x-Sub	x-View
Hadamard product	79.81	86.05
weighted sum	80.72	87.94
Element-wise sum	80.62	87.83
Concatenation	**82**	**89.2**

5.3. Success of Different UDM Structures

It is important to note that the Unified Downsampling Module (UDM) operates through a Graph Convolution Layer (GCNConv), followed by a max-pooling process. As shown in Table 5, an experiment was conducted where max-pooling was replaced with mean-pooling. The results revealed that the mean-pooling approach produced performance comparable to the max-pooling method. Additionally, it is observed that omitting GCNConv in these two models leads to a consistent decline in performance. These outcomes underline the importance of incorporating the GCNConv technique within the UDM framework. According to reference [16], our models consistently surpass the current best model in terms of UDM structures. This further validates the effectiveness of our proposed OTM-HC architecture.

Table 5. The performance of our model with varying UDM arrangements on the NTU-60 dataset is evaluated.The best result is highlighted in bold.

Arrangments	x-Sub	x-View
GCNConv + Adaptive-MAxpooling	80.2	88.1
GCNConv + max-Pooling	**82**	**89.2**
Conv1d + mean-pooling	80.8	88.4
Max-pooling	78.8	87.3

5.4. Effectiveness of Skeleton Sequence Views

We report the performance solely based on the collective perspective of skeletons as input. For a comprehensive comparison, we also explored different perspectives on skeleton sequences and a three-stream fusion approach involving joints, motions, and bones, as previously investigated in [17,57]. Regarding fusion, our methodology aligns with the parameters set forth by [57]. The results are briefly displayed in Figure 4. The models we have developed constantly demonstrate superior performance compared to their rivals, with noticeable margins, further confirming our suggested methodology's efficiency.

5.5. Different Level Representation

We conducted the ablation study for each hierarchical level in our model by focusing on the instance, domain, clip, and part levels. By systematically disabling each level and observing the impact on performance, we demonstrated the significant contributions of each component. As shown in Table 6, although each level independently enhances performance, the combination of all three levels produces the optimal results. Following this scheme, the accuracy of the NTU-60 x-sub and x-view protocols is improved to 82 and 89.2, respectively, signifying the complementary nature of these hierarchical representations.

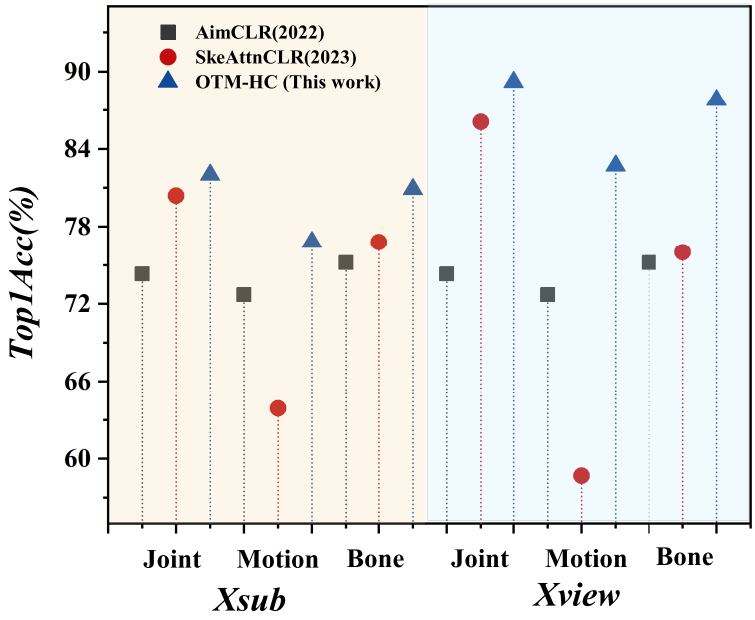


Figure 4. This study compares performance by utilizing various views of skeletal sequences, specifically focusing on x-sub and x-view, within the context of the NTU-60 dataset.

Table 6. Comparative analysis of performance metrics across different levels of abstraction.

Instance Level	Domain Level	Clip & Part Level	NTU-60	
			x-Sub	x-View
✓	✓	✓	**82**	**89.2**
✓	-	-	80.3	88.1
✓	✓	-	80.9	88.6

6. Conslusions

In summary, we proposed a new contrastive learning framework, OTM-HC, for unsupervised SKARL that can learn more discriminative representation in unsupervised circumstances and possesses strong transferability by combining multiple-level representation and hierarchical contrast. In addition, through an extensive series of experiments, we have validated the effectiveness of OTM-HC, signifying its success in state-of-the-art performance when dealing with unsupervised SKARL tasks. Given the simplicity and effectiveness of OTM-HC, we believe OTM-HC can be utilized as a new robust baseline for unsupervised SkARL. Future research may investigate adaptive domain weighting and domain-specific tuning to enhance the framework for diverse real-world applications.

Author Contributions: M.U.: Methodology, Software, Data Curation, Writing—Original Draft, Writing—Review & Editing; W.C.: Formal Analysis, Supervision, Project Administration; Z.H.: Writing—Review & Editing; J.Z.: Writing – Review & Editing; R.J.: Funding Acquisition, Validation. All authors have read and agreed to the published version of the manuscript.

Funding: The Fundamental Research Foundation of Shenzhen under Grant JCYJ20230808105705012 and the National Natural Science Foundation of China, Grant No. 61971290.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cob-Parro, A.C.; Losada-Gutiérrez, C.; Marrón-Romera, M.; Gardel-Vicente, A.; Bravo-Muñoz, I. A new framework for deep learning video based Human Action Recognition on the edge. *Expert Syst. Appl.* **2024**, *238*, 122220. [CrossRef]
2. Kulbacki, M.; Segen, J.; Chaczko, Z.; Rozenblit, J.W.; Kulbacki, M.; Klempous, R.; Wojciechowski, K. Intelligent video analytics for human action recognition: The state of knowledge. *Sensors* **2023**, *23*, 4258. [CrossRef] [PubMed]
3. Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; Wang, M. Dual encoding for video retrieval by text. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4065–4080. [CrossRef]
4. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
5. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef]
6. Wang, M.; Ni, B.; Yang, X. Learning multi-view interactional skeleton graph for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *45*, 6940–6954. [CrossRef] [PubMed]
7. Moliner, O.; Huang, S.; Åström, K. Bootstrapped representation learning for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4154–4164.
8. Zhou, Y.; Yan, X.; Cheng, Z.Q.; Yan, Y.; Dai, Q.; Hua, X.S. BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2049–2058.
9. Qu, H.; Cai, Y.; Liu, J. Llms are good action recognizers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 18395–18406.
10. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2117–2126.
11. 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, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
12. Su, K.; Liu, X.; Shlizerman, E. Predict & cluster: Unsupervised skeleton based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9631–9640.
13. 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, Seattle, WA, USA, 12–16 October 2020; pp. 2490–2498.
14. Nie, Q.; Liu, Z.; Liu, Y. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 102–118.
15. 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.* **2021**, *569*, 90–109. [CrossRef]
16. Thoker, F.M.; Doughty, H.; Snoek, C.G. Skeleton-contrastive 3D action representation learning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 1655–1663.
17. Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; Ding, R. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 762–770.
18. Su, Y.; Lin, G.; Wu, Q. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13328–13338.
19. Chen, Y.; Zhao, L.; Yuan, J.; Tian, Y.; Xia, Z.; Geng, S.; Han, L.; Metaxas, D.N. Hierarchically self-supervised transformer for human skeleton representation learning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 185–202.
20. 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/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13423–13433.
21. 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, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 209–225.
22. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
23. Su, Y.; Lin, G.; Sun, R.; Hao, Y.; Wu, Q. Modeling the uncertainty for self-supervised 3d skeleton action representation learning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 769–778.
24. Vemulapalli, R.; Chellapa, R. Rolling rotations for recognizing human actions from 3d skeletal data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4471–4479.

25. Usman, M.; Zhong, J. Skeleton-based motion prediction: A survey. *Front. Comput. Neurosci.* **2022**, *16*, 1051222. [CrossRef] [PubMed]
26. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
27. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1227–1236.
28. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
29. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE Trans. Image Process.* **2018**, *27*, 3459–3471. [CrossRef] [PubMed]
30. Zhang, P.; Xue, J.; Lan, C.; Zeng, W.; Gao, Z.; Zheng, N. Adding attentiveness to the neurons in recurrent neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 135–151.
31. Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*; Kremer, S.C., Kolen, J.F., Eds.; IEEE Press: New York, NY, USA, 2001.
32. 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, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
33. 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/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13359–13368.
34. Chen, H.; Jiang, Y.; Ko, H. Pose-guided graph convolutional networks for skeleton-based action recognition. *IEEE Access* **2022**, *10*, 111725–111731. [CrossRef]
35. Lee, J.; Lee, M.; Lee, D.; Lee, S. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 10444–10453.
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. Zhang, H.; Hou, Y.; Zhang, W.; Li, W. Contrastive positive mining for unsupervised 3d action representation learning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 36–51.
38. Zhang, J.; Jia, Y.; Xie, W.; Tu, Z. Zoom transformer for skeleton-based group activity recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8646–8659. [CrossRef]
39. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 776–794.
40. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3733–3742.
41. Ye, M.; Zhang, X.; Yuen, P.C.; Chang, S.F. Unsupervised embedding learning via invariant and spreading instance feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6210–6219.
42. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
43. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
44. Lin, L.; Zhang, J.; Liu, J. Actionlet-Dependent Contrastive Learning for Unsupervised Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2023; pp. 2363–2372.
45. Hua, Y.; Wu, W.; Zheng, C.; Lu, A.; Liu, M.; Chen, C.; Wu, S. Part Aware Contrastive Learning for Self-Supervised Action Recognition. *arXiv* **2023**, arXiv:2305.00666.
46. Cheng, Y.; Chen, X.; Chen, J.; Wei, P.; Zhang, D.; Lin, L. Hierarchical Transformer: Unsupervised Representation Learning for Skeleton-Based Human Action Recognition. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
47. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 2, pp. 1735–1742.
48. Bachman, P.; Hjelm, R.D.; Buchwalter, W. Learning representations by maximizing mutual information across views. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 15535–15545.
49. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.

50. Goutsu, Y.; Takano, W.; Nakamura, Y. Motion recognition employing multiple kernel learning of fisher vectors using local skeleton features. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 79–86.
51. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 588–595.
52. Liu, C.; Hu, Y.; Li, Y.; Song, S.; Liu, J. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv* **2017**, arXiv:1703.07475.
53. Franco, L.; Mandica, P.; Munjal, B.; Galasso, F. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. *arXiv* **2023**, arXiv:2303.06242.
54. Shah, A.; Roy, A.; Shah, K.; Mishra, S.; Jacobs, D.; Cherian, A.; Chellappa, R. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18846–18856.
55. Yang, S.; Liu, J.; Lu, S.; Hwa, E.M.; Hu, Y.; Kot, A.C. Self-Supervised 3D Action Representation Learning With Skeleton Cloud Colorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 509–524. [CrossRef]
56. Dong, J.; Sun, S.; Liu, Z.; Chen, S.; Liu, B.; Wang, X. Hierarchical contrast for unsupervised skeleton-based action representation learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 525–533.
57. 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/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4741–4750.
58. Zhang, W.; Hou, Y.; Zhang, H. Unsupervised skeleton-based action representation learning via relation consistency pursuit. *Neural Comput. Appl.* **2022**, *34*, 20327–20339. [CrossRef]