electronics

MDPI

Article

Human Motion Prediction Based on a Multi-Scale Hypergraph for Intangible Cultural Heritage Dance Videos

Xingquan Cai, Pengyan Cheng, Shike Liu, Haoyu Zhang * and Haiyan Sun

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; caixingquan@ncut.edu.cn (X.C.); chengpengyan@mail.ncut.edu.cn (P.C.); shikeliu@mail.ncut.edu.cn (S.L.); sunhaiyan@ncut.edu.cn (H.S.)

* Correspondence: zhanghaoyu@mail.ncut.edu.cn; Tel.: +86-15216262561

Abstract: Compared to traditional dance, intangible cultural heritage dance often involves the isotropic extension of choreographic actions, utilizing both upper and lower limbs. This characteristic choreography style makes the remote joints lack interaction, consequently reducing accuracy in existing human motion prediction methods. Therefore, we propose a human motion prediction method based on the multi-scale hypergraph convolutional network of the intangible cultural heritage dance video. Firstly, this method inputs the 3D human posture sequence from intangible cultural heritage dance videos. The hypergraph is designed according to the synergistic relationship of the human joints in the intangible cultural heritage dance video, which is used to represent the spatial correlation of the 3D human posture. Then, a multi-scale hypergraph convolutional network is constructed, utilizing multi-scale transformation operators to segment the human skeleton into different scales. This network adopts a graph structure to represent the 3D human posture at different scales, which is then used by the single-scalar fusion operator to spatial features in the 3D human posture sequence are extracted by fusing the feature information of the hypergraph and the multi-scale graph. Finally, the Temporal Graph Transformer network is introduced to capture the temporal dependence among adjacent frames within the time domain. This facilitates the extraction of temporal features from the 3D human posture sequence, ultimately enabling the prediction of future 3D human posture sequences. Experiments show that we achieve the best performance in both short-term and long-term human motion prediction when compared to Motion-Mixer and Motion-Attention algorithms on Human3.6M and 3DPW datasets. In addition, ablation experiments show that our method can predict more precise 3D human pose sequences, even in the presence of isotropic extensions of upper and lower limbs in intangible cultural heritage dance videos. This approach effectively addresses the issue of missing segments in intangible cultural heritage dance videos.

Keywords: human motion prediction; hypergraph; multi-scale hypergraph convolutional network; transformer

check for updates

Citation: Cai, X.; Cheng, P.; Liu, S.; Zhang, H.; Sun, H. Human Motion Prediction Based on a Multi-Scale Hypergraph for Intangible Cultural Heritage Dance Videos. *Electronics* **2023**, *12*, 4830. https://doi.org/10.3390/electronics12234830

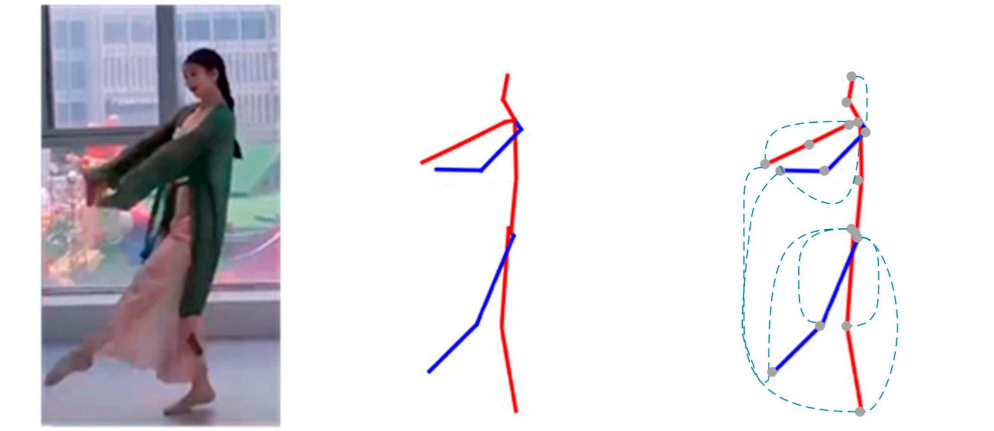
Academic Editor: Hyunjin Park

Received: 25 October 2023
Revised: 24 November 2023
Accepted: 27 November 2023
Published: 29 November 2023

1. Introduction

Intangible cultural heritage dances are produced in the process of labor and daily life entertainment within communities, inheriting and developing the original dances of various nationalities. Their artistic expressions and cultural significance are deeply influenced by the local regional environment, national culture, historical context, and customs, which are reflected in various dance movements, costumes, and props [1]. However, due to the deficiency of video documentation, primarily attributable to the limitations of early video recording equipment, these intangible cultural heritage dances are facing the risk of being lost [2]. Therefore, there is an urgent need to safeguard and preserve the cultural heritage of intangible cultural heritage dances.

In recent years, with the development of deep learning, 3D skeleton-based human motion prediction methods have been used to predict future pose sequences from observed motion sequences, which are applied in various domains, including autonomous driving [3], health care [4], and pedestrian tracking [5]. Motion prediction can be broadly classified into two categories: short-term prediction and long-term prediction. Short-term prediction involves estimating motion shortly based on the current motion, while long-term prediction predicts motion in subsequent moments based on repeated iterations of the predicted motion segments [6]. Modern dance styles such as jazz and urban are characterized by repetition and symmetry in dance movement choreography, such as repeating the same movement within an eight-beat period, with the same hand and foot movements executed on both sides but in opposite directions. However, in intangible cultural heritage dances such as Miao dance and Wa dance [1], the choreography of movements often extends in the same direction for both upper and lower limbs, as shown in Figure 1a,b. This choreographic feature causes the lack of interaction of distant joints (e.g., the connection from the right hand to the right foot), leading to the problem of low accuracy of the prediction results.



(**a**) Dance Movement Schematic

(**b**) 3D Skeleton Schematic

(**c**) Hypergraph Skeleton Schematic

Figure 1. Schematic diagram of upper and lower limb isotropic extension of intangible cultural heritage dance.

To address the above problems, existing methods [7–9] focus on constructing graph convolution networks to represent the spatial correlations between human joint points. However, these graph convolution-based approaches only consider physical constraints of the body (e.g., the display of joint angles) [10], and lack the modeling of interactions across jointed limb segments. Therefore, we propose to use hypergraphs to represent the interactions among human joint points.

Therefore, we propose a human motion prediction method based on a multi-scale hypergraph convolutional network for intangible cultural heritage dance videos. Additionally, we design an array of joint point hypergraphs for the interactions among different joint points of performer's intangible cultural heritage dances (as illustrated in Figure 1c). A multi-scale hypergraph convolutional network is constructed to extract spatial features of the 3D gesture sequences. Subsequently, the Temporal Graph Transformer module is introduced to extract the temporal information within the action sequences. Ultimately, this method outputs the predicted 3D human joint point coordinates.

We evaluate our model on the publicly available large-scale 3D human pose estimation datasets Human 3.6M [11] and 3DPW [12] and on a small homemade non-legacy dance movement dataset. Our approach achieves superior performance compared to several representative 3D human pose estimation methods and is effective in overcoming the problem of low accuracy in motion prediction due to the lack of long-range joint point interactions in dance movements. In addition, we conducted ablation studies to demonstrate that multi-scale hypergraphs can better focus on the long-distance interactions between multiple joint points.

In summary, our main contributions are listed as follows:

(1) We design joint point hypergraphs for representing 3D human gesture sequence spatial information for intangible cultural heritage dance videos' joint point interaction information.

(2) A multi-scale hypergraph convolutional network is constructed for the joint hypergraph, which extracts the spatial features of the 3D human posture sequence represented by the multi-scale hypergraph.

(3) A Temporal Graph Transformer is introduced for the multi-scale hypergraph convolutional network, to extract the temporal features among 3D human posture sequences.

2. Related Work

2.1. Motion Prediction

Motion prediction is the inference of human motion from temporally incomplete video data [6]. 3D human motion prediction [13] using RNNs has been extensively studied in the past few years. In 2015, Fragkiadaki et al. [14] proposed a recursive encoder–decoder model that introduces a non-linear network of encoders and decoders to enable the integration of representation learning and dynamic learning in the space and time domain. In 2017, Martinez et al. [7] introduced a recursive encoder–decoder model with the RNN units by adding residual connections. With the development of convolutional networks, they gained significant achievements in hierarchical structure and capturing spatio-temporal correlation. In 2018, Li et al. [15] proposed a hierarchical structure containing convolutional long-term encoders and decoders, efficiently capturing spatial and temporal correlation. In 2021, Sodianos et al. [8] proposed a spatio-temporally separable convolutional network to solve the temporal and spatial interaction complexity prediction. Dang et al. [10] proposed a multi-scale residual graph network with descending and ascending GCNs to extract features in a fine-to-coarse manner. In 2022, A. Bouazizi et al. [9] used a multilayer perceptron (MLP) architecture alone to perform short-term and long-term human motion prediction with good performance.

2.2. Hypergraph

Meanwhile, hypergraph learning [16] has also achieved good performance in many applications. In 2005, S. Agarwal et al. [17] used group averaging to transform hypergraphs into simple graphs, applying hypergraphs for clustering. In 2009, Tian et al. [18] proposed a semi-supervised learning method, HyperPrior, to classify gene expression data by using biological knowledge as a constraint for classification. In 2010, Bu et al. [19] developed music recommendations by modeling the relationship of different entities including music, tags, and users through the hypergraph.

Since graph-structured neural networks can only focus on the connection structure between neighboring nodes, they fail to capture the interaction relationship of distant joints. For skeleton-based motion prediction tasks, where actions usually need to be coordinated across multiple joints, previous studies have used graph convolution methods that only consider the physical connections between joints, but ignore the unique characteristics of the three-dimensional skeleton, where each type of body joint has its own unique physical function. Therefore, in order to better represent the interactions between different joint groups (as illustrated in Figure 1c), we designed the hypergraph structure to divide the human joints into different groups, which can better extract the features of the joint groups. In fact, for the movements in the intangible cultural heritage dance video, it is necessary to consider not only the connection relationship between adjacency joints but also the interaction relationship between distant joints (e.g., in the intangible cultural heritage dance movement, the left arm and the left leg extended in the same direction). Therefore, this paper proposes a human motion prediction method for intangible cultural heritage dance videos based on convolutional networks with multi-scale hypergraphs. Compared to traditional graph-structured neural networks, the convolutional network based on the multi-scale hypergraph can extract rich spatial features of 3D joint points and can capture rich multi-scale relationships to extract action dependencies between neighboring frames for motion prediction.

2.3. Multi-Scale Convolutional Networks

Graph Convolutional Networks can convolve irregularly structured data like 3D human skeleton compared to traditional CNNs. Meanwhile, in order to synthesize the 3D human skeleton information at different scales, multi-scale graph convolutional neural network is proposed to solve this problem. For example, Li et al. [20] used multi-scale multistreaming GCN to obtain more discriminating temporal features. Fan et al. [21] selectively fused different scale features. Li et al. [22] generated the next scale by removing some joints in the middle position. Dang et al. [10] performed scale generation by selecting the middle of these sites. However, the above methods are mainly suitable for classification tasks due to their simplicity and focus on information extraction only.

2.4. Transformer Network

Human motion prediction needs to consider not only the spatial correlation of the human skeleton in each frame but also the temporal continuity of the action in the time domain. Early prediction methods LSTM [23] extract temporal information by extracting sequential time cues between frames, and Seq2Seq [15] obtains motion prediction results by constructing encoder and decoder architectures that are jointly trained based on the loss of previous sampling. However, the two approaches mentioned above lack the understanding of extended temporal information. In 2017, Vaswani et al. [24] applied the self-attention mechanism to a wide range of applications in the field of NLP. Inspired by the other of this application, many explored its application in the field of human motion prediction. In 2021, Cheng et al. [25] designed Motion-Transformer to capture temporal dependence by pre-training on self-detection of human actions. Lin et al. [26] proposed a novel Transformer for cross-attention, which can capture temporal dependence within an image block by alternately applying attention within and between image blocks to build an efficient hierarchical network. In 2020, Wu et al. [27] learned complex patterns and dynamics of time series data through a self-attention mechanism, which can be applied to unused types of time series data. Therefore, we consider using the Transformer model [28] to accomplish the extraction of time-series features in more complex human movement prediction problems.

In the current landscape, the preservation and inheritance of intangible cultural heritage dances face a significant challenge due to the acute shortage of skilled individuals. As a result, some of these dance forms have remained unpassed, with only a few early recorded performance videos exist. Unfortunately, the limitations of the early video equipment and the suboptimal methods for image preservation have led to the situation that there are missing dance segments in the video transitions and narration, among other segments [1]. Therefore, utilizing the method of 3D human posture estimation to obtain the 3D human key points from the video [29], and predicting the missing part of the action frames in the video based on the obtained coordinates of the joints, can assist with the inheritance and preservation of the intangible cultural heritage dance videos.

Therefore, based on the above analysis, we propose a human motion prediction method based on a multi-scale hypergraph convolutional network for intangible cultural heritage dance videos. We first design the hypergraph based on the body joints, then construct a multi-scale hypergraph convolutional network to extract the spatial feature information of the 3D human joint sequences, and then extract the temporal feature information in the 3D human joint sequences by constructing a Temporal Graph Transformer network, and then ultimately output the 3D human joint sequences obtained by prediction.

3. The Proposed Method

Aiming at the current 3D human motion prediction methods for intangible cultural heritage dance videos, particularly the challenges posed by the extended choreography style involving isotropic extensions of the upper and lower limbs, which often results in significant motion prediction errors, our paper proposes a novel human motion prediction method based on the hypergraph multi-scale fusion convolutional network for intangible cultural heritage dance videos.

At the stage where video data are obtained from the 3D skeleton information, there exist two types of 3D human pose estimation methods: single-stage methods [30] and two-stage methods [31]. Since two-stage methods can further benefit from 2D pose information and large-scale 2DHPE (human pose estimation) datasets [32], generally exhibit better performance [29]. For example, the multi-hypothesis transformer (MHFomer) model [33] in the two-stage approach can extract accurate 3D human pose sequences, even in complex environmental settings. As shown in Figure 2, the specific operational steps of our proposed human motion prediction method, grounded in the hypergraph multi-scale fusion convolutional network for non-legacy videos are as follows: First and foremost, we initiate the process by inputting the folk-dance video. Subsequently, we use the multiple hypothesis transformer (MHFomer) algorithm to obtain the 3D human posture sequence. Next, to harness the spatial correlations within the 3D human joints more effectively, we design a set of multi-scale joint hypergraphs. These hypergraphs are instrumental in facilitating a more comprehensive understanding of the connections among joints. Then, a multi-scale convolutional network is constructed to fuse the spatial features at different scales to better deal with the problem under the upper and lower limbs in the same direction extension choreography mode. Finally, the Temporal Graph Transformer model is introduced to extract temporal features from the resultant feature graph sequence. This extracted information is then utilized to predict the future coordinate sequence of the 3D joint points.

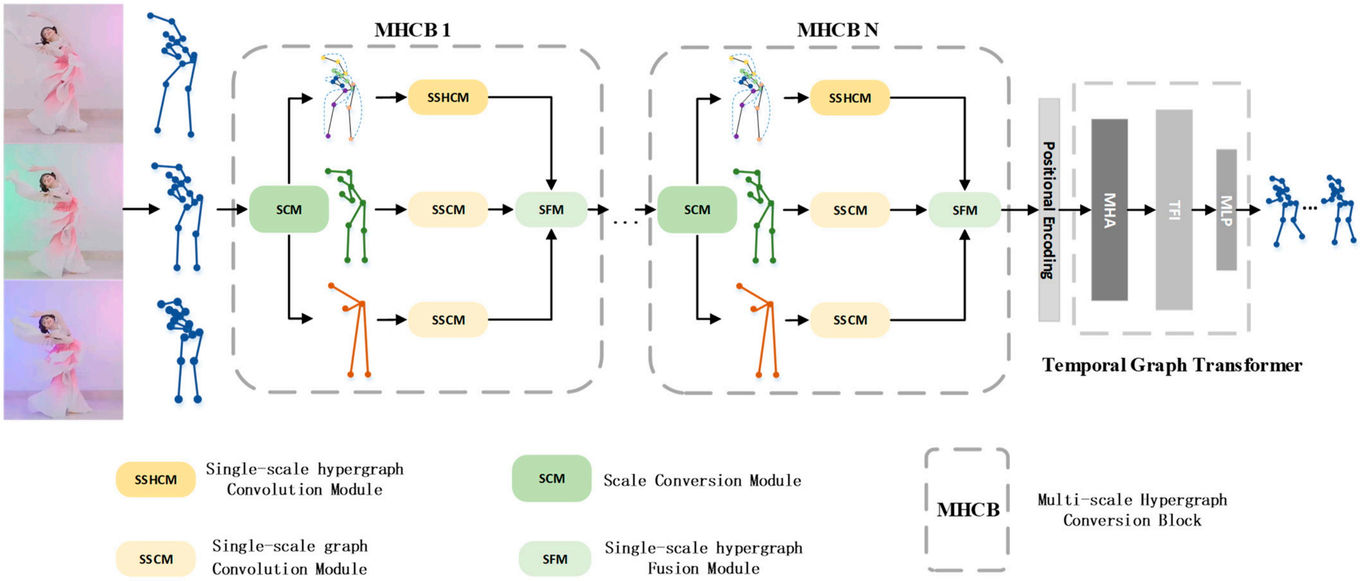
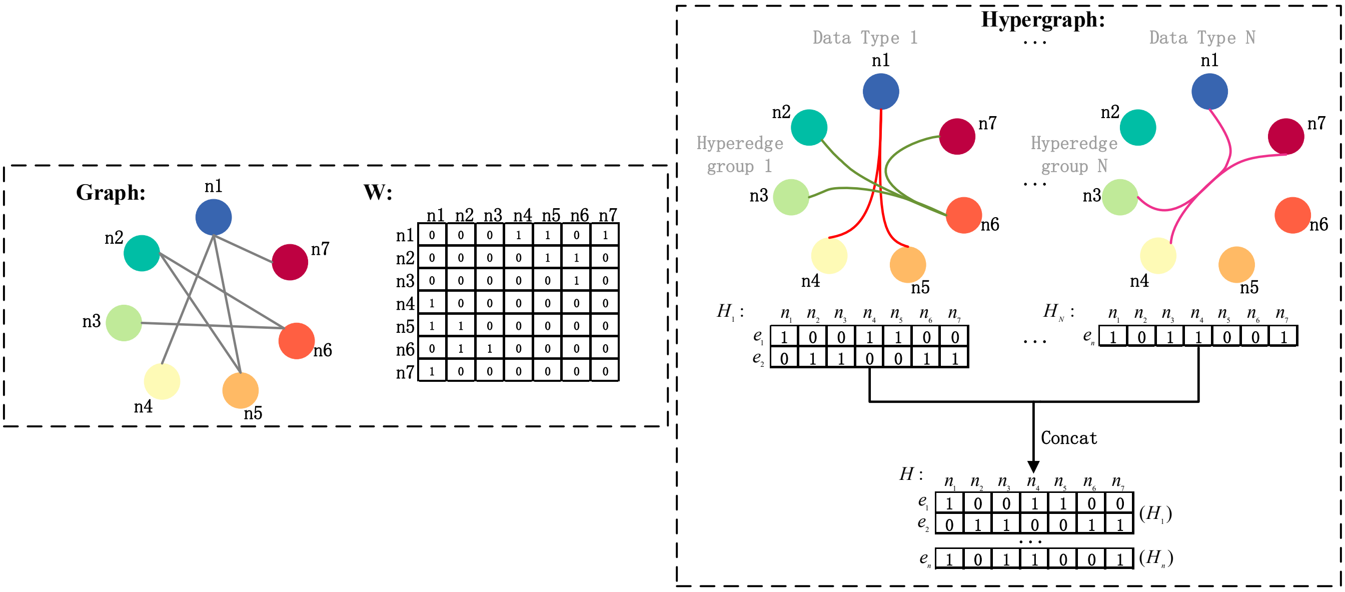

MHCB 1
MHCB N
SSHCM
SCM
SSCM
SFM
SSHCM
SCM
SSCM
SFM
Positional Encoding
MHA
TFI
MLP
Temporal Graph Transformer
SSHCM Single-scale hypergraph Convolution Module
SSCM Single-scale graph Convolution Module
SCM Scale Conversion Module
SFM Single-scale hypergraph Fusion Module
MHCB Multi-scale Hypergraph Conversion Block

Figure 2. Hypergraph multi-scale fusion convolutional network architecture diagram.

3.1. Joint Hypergraphs Generation

The traditional graph structure builds a spatial graph based on body joints in each frame [10,34]. However, this modeling approach only considers the physical constraints between body joints. This makes it difficult to effectively capture the connections across joints, thereby failing to provide an accurate depiction of the overarching relationships within the human skeleton and interactions among distant joints. Therefore, we propose a method to design a hypergraph based on human joints to model the connectivity between nodes in the hypergraph. This method can better capture the spatial correlation between multiple joints. Moreover, the use of hypergraphs can fully consider the connectivity relationships between different joints and the interaction relationships at a distance.

The multi-scale node hypergraph expresses the spatial information among joint points. Since the intangible cultural heritage dance movements need to be accomplished by interaction and collaboration among multiple joint points, our designs the hypergraph structure with 3D human joint points as graph nodes, and the hypergraph structure takes the human joint points as vertices, and connects them with hyperedges according to the interaction relationship among different joint points, to construct the hypergraph $G_{spa} = (V_{spa}, \varepsilon_{spa}, W_{spa})$ as shown in Figure 3.



W:

	n1	n2	n3	n4	n5	n6	n7
n1	0	0	0	1	1	0	1
n2	0	0	0	0	1	1	0
n3	0	0	0	0	0	1	0
n4	1	0	0	0	0	0	0
n5	1	1	0	0	0	0	0
n6	0	1	1	0	0	0	0
n7	1	0	0	0	0	0	0

H_1 :

	n_1	n_2	n_3	n_4	n_5	n_6	n_7
e_1	1	0	0	1	1	0	0
e_2	0	1	1	0	0	1	1

H_N :

	n_1	n_2	n_3	n_4	n_5	n_6	n_7
e_n	1	0	1	1	0	0	1

H :

	n_1	n_2	n_3	n_4	n_5	n_6	n_7	
e_1	1	0	0	1	1	0	0	(H_1)
e_2	0	1	1	0	0	1	1	
...								
e_n	1	0	1	1	0	0	1	(H_n)

Figure 3. Schematic diagram of the structure of graph and hypergraph.

Where V_{spa} denotes the vertex set $V_{spa} = \{v_n | n = 1, \ldots N\}$ of the hypergraph, it represents the number of human joints, and 17 joints are used in this paper for 3D joint coordinate prediction. ε_{spa} denotes the set of joint-space hyperedges, and W_{spa} denotes the weight of each hyperedge in the set of joint-space hyperedges.

The construction of the hypergraph is divided into four steps. Initially, the process entails defining the initial spatial correlation matrix. Subsequently, we calculate the number of hyper-edges of the joints. Furthermore, we calculate the count of joints contained within these hyper-edges. Finally, we generate the spatially regularized hypergraph Laplace matrix. The specific steps are as follows:

Step 1. Define the initial spatial association matrix. Define the association matrix H_{spa} with initial size $|v| \times |\varepsilon|$ as shown in Equation (1).

$$H_{spa}(v, e) = \begin{cases} 1, & if v \in e \\ 0, & if v \notin e \end{cases} \tag{1}$$

where $v \in V_{spa}$ and $e \in \varepsilon_{spa}$.

Step 2. Calculate the number of hyperedges at the joints. Based on the association matrix H_{spa}, the degree of the node $v \in V_{spa}$ is computed to represent the number of hyperedges containing that joint, as shown in Equation (2).

$$d(v) = \sum_{e \in \varepsilon_{spa}} W_{spa}(e) H_{spa}(v, e) \tag{2}$$

Step 3. Calculate the number of joints contained in the hyperedge. The degree of the computed hyperedge $e \in \varepsilon_{spa}$ indicates the number of joints contained in that hyperedge, and the computation process is shown in Equation (3).

$$\delta(e) = \sum_{v \in V_{spa}} H_{spa}(v, e) \tag{3}$$

Step 4. Generate a spatially regularized hypergraph Laplace matrix. To utilize the high-order joints interaction information in the hypergraph for feature extraction, the hypergraph Laplacian matrix G_{spa} is generated based on the association matrix H_{spa}, which is calculated as shown in Equation (4).

$$G_{spa} = D_v^{-1/2} H_{spa} W_{spa} D_e^{-1} (H_{spa})^T D_{v_t}^{-1/2} \tag{4}$$

where D_v and D_e denote the diagonal matrices of vertex degree $d(v)$ and hyperedge degree $\delta(e)$ in the hypergraph. Respectively, based on the above steps, a hypergraph $G_{spa} = \left(V_{spa}, \varepsilon_{spa}, W_{spa}\right)$ with 3D joints as graph nodes is constructed to represent the non-physical dependencies of multiple 3D joints in the spatial domain.

3.2. Multi-Scale Hypergraph Convolution Module Construction

The human skeleton can be scaled to different scales according to limb segments (e.g., legs, torso), we establish three scales for representing the human skeleton, namely, torso scale, limb scale, and joint scale. Therefore, we propose a multi-scale architecture that effectively utilizes the skeleton information at different scales for feature extraction. Using a multi-scale hypergraph convolution module, we construct a multi-scale hypergraph module to extract spatial information from different scales based on the designed joint hypergraph.

The steps of constructing the multi-scale hypergraph convolution network are divided into three steps: Firstly, the multi-scale segmentation operator is constructed. Then, the single-scale graph convolution and hypergraph convolution modules are constructed for the spatial information extraction at each scale. Finally, the single-scale hypergraph fusion operator is designed to fuse the information from different scales.

3.2.1. Constructing Multi-Scale Segmentation Operator Construction

In contrast to previous approaches [35], which categorize joints based on joints alone to obtain multi-scale maps, our method categorizes human joints into three scales based on human skeleton relationships. These scales include the torso, limb, and joint scales. The torso scale focuses on the global information of the human skeleton, the limb scale focuses on the overall connections among limb segments, and the joint scale focuses on the connectivity among joint points. In our paper, we construct a multi-scale spatial map, as shown in Figure 4. Specifically, the joint scale contains 17 joints, while the limb scale and torso scale each consist of 11 and 5 joint points, respectively.

The human skeleton map contains rich connectivity among joints, and by aggregating multiple connected joints in proximity, different-scale skeleton maps representing global information can be obtained.

Maximum pooling is used in a general approach to classify different scales of skeleton maps [35], and maximum pooling selects joints that contain the most information among the neighboring connected joints as the representative to obtain the torso scale map that contains global information. However, maximum pooling tends to ignore the joints with less information. Therefore, we use average pooling to aggregate the information between adjacent connected joints. Compared to maximum pooling, average pooling can focus on the information in each joint point, making the limb-scale and torso-scale skeleton maps more complete global information.

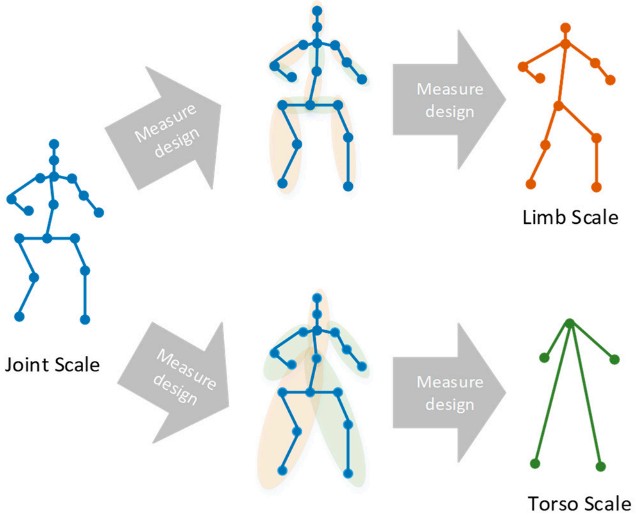


Figure 4. Schematic diagram of joint scale segmentation. Three spatial body scale graphs: joint, limb and torso scale.

Constructing the multi-scale segmentation operator is specifically divided into two steps. Initially, we define the spatial map convolution. Then, design the scale transformation operator for scaling purposes. The specific steps are as follows:

Step 1. Calculate spatial graph convolution.

We represent the human skeleton joints as a spatial graph, where the joints are used as graph nodes and the neighboring connections among the nodes are used as the edges of the graph, defining $g = (z, e)$, where $z = v \times t$, $g \in R^{V \times V \times t}$ is the human skeleton graph containing v joints in t frames, and we define the adjacency matrix $g_k(i, j)$ as shown in Equation (5).

$$g_{k(i,j)} = \begin{cases} 1 & if \quad d(i,j) = k \\ 0 & otherwise \end{cases} \tag{5}$$

where k is the path between node i and node j. To solve the problem of too little information, we superimpose the neighbor matrix $g_{k(i,j)}$ obtained from different values of k. The calculation process is shown in Equation (6).

$$g = \sum_{k=0}^{K} g_k \tag{6}$$

Meanwhile, considering the spatio-temporal graph $g = S \otimes T \in R^{(TV)\times(TV)}$ as a single-scale down graph information, we set $X \in R^{T \times V \times D}$ as the motion tensor, and based on the decomposability assumption, a spatial graph convolution is defined, as shown in Equation (7).

$$X' = V_{*T}(U_{*S}X) \tag{7}$$

where $*S$ denotes the spatial graph convolution for decomposition and U and V denote the graph filters. Equation (7) indicates that the spatio-temporal convolution map can be decomposed into a spatial and temporal graph convolution. Based on Equation (7), the spatial convolution processes each data frame individually, and works as shown in Equation (8) for the tth timestamped segment in X.

$$(U_{*S}X)^{[t,:,:]} = \sum_{\ell=0}^{L} S^{\ell} X^{[t,:,:]} U_{\ell} \in R^{V \times D'} \tag{8}$$

where $U \in R^{L\times D\times D'}$, the ℓth fragment $U_\ell \in R^{V\times D'}$ is the trainable weight matrix corresponding to the ℓth order. Obtained after the above process is the spatial feature X obtained by the spatial convolution operator.

Step 2. Scale conversion operator.

In order to convert the obtained joint-scale spatial graph into any set scale, we proposes a trainable average pooling operator, let $X \in R^{T\times V\times d}$ be the spatial data at the joint scale, S_0 be the spatial graph adjacency matrix, and at the rth spatial scale, the spatial pooling operator $\psi_{0\to r} \in [0,1]^{\overline{V}\times V_r}$ is expressed as shown in Equation (9).

$$\psi_{0\to r} = \sigma\left(S_0\left[ReLU\left(U_{*S_0}X\right)_{13}W_{0\to r}\right]\right) \tag{9}$$

where U_{*S_0} can be obtained from the above equation, $[\cdot]_{13} : R^{T\times V\times d} \to R^{V\times (dT)}$ denotes the conversion of features from temporal dimension to spatial dimension. $W_{0\to r} \in R^{(dT)\times M_r}$ is the trainable weights, $\sigma(\cdot)$ is the softmax operation performed on each dimension. $(\psi_{0\to r})_{i,j}$ denotes the assignment of the ith joint of the joint scale to the jth group of the rth spatial scale. The original image features and spatial map adjacency matrix can be converted to any r spatial scale by the scale conversion operator obtained above, as shown in Equations (10) and (11).

$$\mathrm{X}_r^{[t,:,:]} = \psi_{0\to r}^T X^{[t,:,:]} \tag{10}$$

$$S_r = \psi_{0\to r}^T S_0 \psi_{0\to r} \tag{11}$$

After the above steps, the spatial features of the body parts in the rth scale can be obtained by fusing the features of the plurality of body joints by Equation (14). A new connectivity spatial matrix diagram for the rth scale of the coarsened scale can be obtained by Equation (15). for representing the physical connections of the multi-joint set at the rth scale.

3.2.2. Single-Scale Graph Convolution and the Hypergraph Convolution Module Construction

To fully extract the spatial features of the 3D human skeleton at each scale, we propose a single-scale graph convolution module. Its structure is shown in Figure 5.

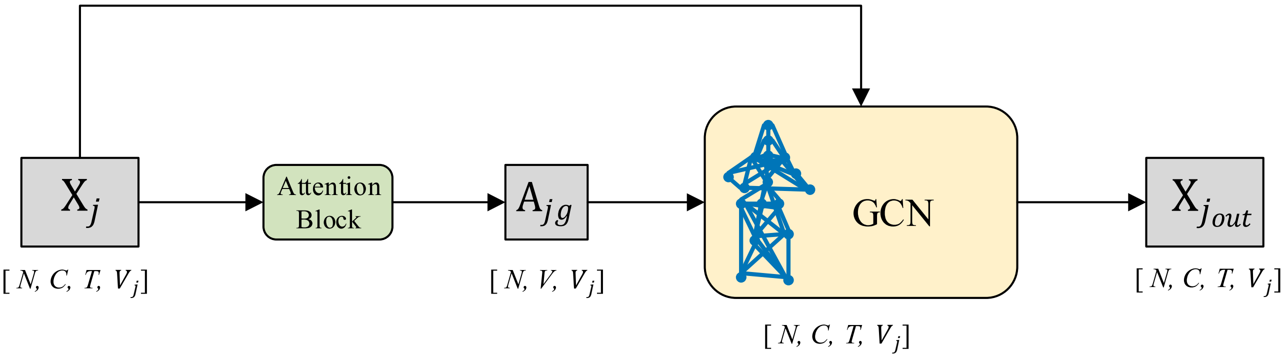


Figure 5. Schematic diagram of the single-scale graph convolution module.

We use the limb scale in it as an example, where the trainable neighbor matrix in the single-scale graph convolution is $A_{Sr} \in R^{Vr\times Vr}$. V_r is the number of limb scales, and all the joints of this matrix are connected to each other, which need to be obtained by training. During the training process, the weights between each neighboring joint point in A_{Sr} are adaptively adjusted, which is calculated as shown in Equation (12).

$$X_{S,SP} = ReLU\left(conv_{ja}\left(A_{Sr}S_rW_S + S_rU_S\right)\right) \in R^{V_r\times V\prime_r} \tag{12}$$

where $S_r \in R^{V_r \times D_X}$ is the spatial matrix represented by the input graph and W_S, $U_S \in R^{V_j \times V_j}$ is the trainable parameter matrix, after which spatial features are extracted from the limb scale. Then, the obtained spatial feature matrix $X_{S,SP}$ is input into two parallel convolutional layers with convolutional kernel size 1 to obtain the intermediate features, and then the two sets of intermediate features are multiplied together to output the adjacency matrix, which is computed as shown in Equation (13).

$$A_{ja} = conv_1(X_j)^T \cdot conv_2(X_j) \in R^{N \times V_j \times V_j} \tag{13}$$

where N is the batch size. Similarly, we designed the Single-Scale hypergraph convolution module. During the training process, its computation is shown in Equation (14).

$$X_{S,SP} = ReLU\big(conv_{ja}\big(H_{spa} X_S W_S\big)\big) \in R^{V \times D'_X} \tag{14}$$

where H_{spa} is the association matrix of the hypergraph, $X_S \in R^{V \times D_X}$ is the matrix represented by the input hypergraph, and $W_S \in R^{D_X \times D'_X}$ is the trainable parameter matrix after which spatial features are extracted from the limb scale. After the above steps, the spatial feature matrix X_{SP} in r scale can be obtained.

We compare the difference between Equations (12) and (14), where the trainable matrix W_S in Equation (14) yields richer information about crotch-joint interactions, due to the fact that when we designed the hypergraph structure, the Laplace matrix of the hypergraph (shown in Equation (4)) is more biased towards focusing on interactions between remote joint points.

3.2.3. Single-Scale Hypergraph Fusion Operator Construction

The torso-scale features, limb-scale features obtained by graph convolution processing, and joint-scale features obtained by single-scale hypergraph convolution processing are fused, and we design a trainable fusion parameter for automatically adjusting the fusion operator, which is shown in Equations (15) and (16).

$$X_2^+ = \alpha W_{32} X_3 + (1-\alpha) X_2 \tag{15}$$

$$X^+ = \alpha W_{21} X_2^+ + (1-\alpha) X_1 \tag{16}$$

where X_i denotes a feature of order i, W_{ij} denotes a weight matrix of the up-adopted features from order i to order j, α is a fusion coefficient, and + represents a fusion feature. After the above steps, by constructing a multi-scale hypergraph convolution module, the 3D human body pose represented by the multi-scale hypergraph is subjected to feature extraction, and 3D joint point spatial features of dimension $X \in R^{V \times D_X}$ are obtained.

3.3. Temporal Graph Transformer to Extract Spatio-Temporal Features Introduction

Through the utilization of a multi-scale hypergraph convolutional network, the method can extract spatial features from the 3D human skeleton, employing a multi-scale hypergraph representation. However, human motion prediction needs to consider not only the spatial correlation of the human skeleton for each frame but also the temporal coherence across actions in the time domain.

Compared with general time series methods (e.g., LSTM [23] and seq2seq [15]), the Transformer [28] model can learn complex dependencies with different temporal sequence lengths. In the human motion prediction task, there is a strong continuity of actions between the frames of human gesture sequences, so we propose the Temporal Graph Transformer method for extracting temporal features of human motion. The method consists of four core components, which are position coding, Multihead Self Attention (MHA) module, Temporal Feature Interaction (TFI) module, and Multilayer Perceptron (MLP) module. The structure of the Temporal Graph Transformer is shown in Figure 6.

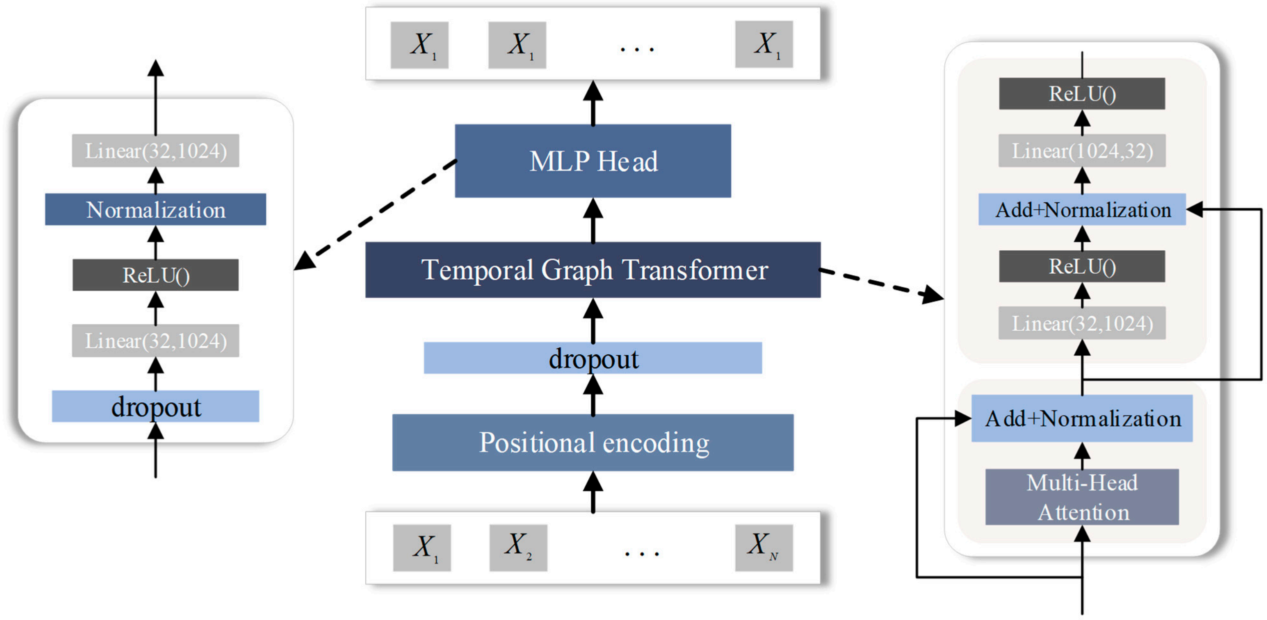


Figure 6. Structure of Temporal Graph Transformer.

(1) Position coding

In order to preserve the positional information of human skeletal joints, we introduces positional encoding $E_{SPos1} \in R^{(J \cdot 2) \times V}$, which is used to distinguish different history frames. Since the position of the key points of each frame in the input sequence is fixed, we choose the local timestamp method to encode the position of the data. We let t be the desired position in the input frame. The calculation process is shown in Equation (17).

$$E_{SPosi} = \begin{cases} \sin(pos/(1000)^{2k/d}) & if \;\; i = 2k, \\ \cos(pos/(1000)^{2k/d}) & if \;\; i = 2k+1, \end{cases} \tag{17}$$

where k takes the value of $\{0, 1, \ldots, d/2-1\}$, d is the number of feature channels of the input fusion feature X, and t is the desired position in the input frame.

(2) The MHA module

In the MHA module, the input multi-scale hypergraph features, $X^T \in R^{n \times d}$ are mapped to the vectors Queries $Q \in R^{n \times d}$, Keys $K \in R^{n \times d}$ and Values $V \in R^{n \times d}$ through three different weight matrices as shown in Equation (18).

$$Q^T,\; K^T,\; V^T = \hat{X}^T W_q^T,\; \hat{X}^T W_k^T,\; \hat{X}^T W_v^T \tag{18}$$

where W_q, W_k, and W_v are learnable weights. A dot product is performed for each query Q and key value K to measure the degree of association between the query and the key. Then, in order to control the computational scale of the dot product, a scaling operation is performed, dividing by the dimension $\sqrt{d}$ of the vector k. Finally, the temporal feature information of the 3D pose is obtained by transforming it into attention weights and multiplying it with V^T through softmax. The calculation process is shown in Equation (19).

$$M^T = \text{softmax}(Q^T (K^T)^T / \sqrt{d_A^T}) V^T \tag{19}$$

where $\sqrt{d}$ is the scaling factor.

(3) The TFI module

In order to explore the connection among potential features, a shared three-layer feed-forward neural network model is designed. Its computational procedure is shown in Equation (20).

$$U^T = ReLU(BN(ReLU(\hat{M'}^T W_0^T))W_1^T)W_2^T \tag{20}$$

Meanwhile, in order to solve the problems of gradient vanishing and gradient explosion in neural network training, we use residual connection $M'^T = M^T + X^T$ for training.

(4) The MLP module

After the MHA module and the TFI module, the feature information is transformed using the MLP module. The MLP module contains two linear layers, the ReLU activation function, and the batch normalization layer. The calculation process is shown in Equations (21)–(23).

$$Z = W_1 \cdot X + b_1 \tag{21}$$

$$\hat{A} = BN(ReLU(Z)) \tag{22}$$

$$Z_2 = W_2 \cdot \hat{A} + b_2 \tag{23}$$

where W_1 and W_2 denote the weights of the two linear layers, respectively, and b_1 and b_2 are the two bias terms. Following the steps above, the outcome is the extraction of spatiotemporal features from the input 3D skeleton sequence, enabling the prediction of the 3D human skeleton sequence for a future period.

4. Experimental Verification and Analysis

To verify the feasibility and effectiveness of the method in this paper, a human motion prediction method based on a multi-scale hypergraphic convolutional network for non-legacy dance videos is designed and implemented. For the experimental validation, the computer hardware environment used is Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz, 64GB RAM, and NVIDIA Quadro RTX 6000 graphics card; the software environment is Windows 10 operating system; and the runtime environments are Python3.8, PyTorch1.7.1 and Pycharm2022.2.1.

4.1. Datasets and Evaluation Indicators

In this paper, the method is evaluated on the Human3.6M [11] dataset, which is more popular in 3D human motion prediction. The Human3.6M dataset is widely used for human motion prediction. He consists of 3.6 million 3D human pose images with 32 joints per 3D pose. 7 professional subjects are performing 15 different daily actions (e.g., walking, eating, talking on the phone). We follow previous paradigms [7,8,15] and construct five of the subjects (S1, S5, S6, S7, S8) as the training dataset and two subjects (S9, S11) as the test dataset.

Meanwhile, the method in this paper is evaluated for generalization performance on the 3DPW dataset, which is more prevalent in the field of motion prediction. The 3D Pose in the Wild dataset (3DPW) [12] is a large-scale dataset consisting of video sequences acquired by a moving cell phone camera, containing more than 51k frames of 3D poses for challenging indoor and outdoor activities. challenging indoor and outdoor activities. We use the training, testing, and validation separation suggested by the official setup. The frame rate of the 3D poses is 30 Hz.

In addition, this paper has curated a small-scale dataset of intangible cultural heritage dance movements to facilitate a more effective exploration of the motion characteristics within intangible cultural heritage dance videos. We engaged the expertise of a master in intangible cultural heritage dance as our subject. We employed the NOKOV optical motion capture system to collect a dataset of their intangible cultural heritage dance movements. The capture duration for each dance segment was set at 20 s, with a frame rate of 120 frames per second, covering an approximate area of 1 m × 1 m. This intangible cultural heritage dance dataset, developed within this paper's context, comprises 12,000 images, representing five distinct intangible cultural heritage dance forms.

This paper uses MPJPE (Mean Per Joint Position Error) as the evaluation index. By calculating the average error between the predicted joint coordinates and the real joint coordinates to the ground distance after heel joint alignment, the MPJPE calculation process is shown in Equation (24).

$$L_{MPJPE} = \frac{1}{V(T+K)} \sum_{k=1}^{T+K} \sum_{v=1}^{V} ||\hat{X}_{vk} - X_{vk}||_2 \tag{24}$$

where X_{vk} denotes the predicted 3D joint point coordinates of joint v in frame k, and $X_{vk} \in R^3$ is the corresponding true 3D joint point coordinates. $|\cdot||_2$ denotes the ℓ_2 paradigm. For the angle-based representation, the loss function between the predicted joint angles and the real situation is calculated as shown in the formula (25).

$$L_{pred} = \frac{1}{J \times T_f} \sum_{j=1}^{J} \sum_{t=T_h+1}^{T_h+T_f} ||\hat{X}_{t,j} - X_{t,j}||_2 \tag{25}$$

where $\hat{X}_{t,j}$ denotes the predicted angle of the joint j at frame t and $X_{t,j}$ is the corresponding true joint angle.

4.2. Comparative Experiments on 3D Motion Prediction

To verify the feasibility and effectiveness of the human motion prediction method based on a multi-scale hypergraph convolutional network for intangible cultural heritage dance videos proposed in this paper, a 3D motion prediction comparison experiment is designed. The deep learning framework Pytorch is used in the experiment to construct the multi-scale hypergraph convolutional network model. The initial learning rate of the model is set to 0.001, and after the 20th epoch, it is reduced by a factor of 0.1 for every 5 epochs. batch size is 256. this paper compares it with the existing algorithms on the Human3.6M dataset. The experimental results are shown in Tables 1 and 2.

We quantitatively evaluate our proposed model for current state-of-the-art short-term (<500 ms) and long-term (>500 ms) predictions.

Table 1. Performance comparison between different methods for short-term prediction (400 ms) of MPJPE (mm) for each activity on the Human3.6M dataset.

Motion	Res.Sup [7]	convSeq2Seq [15]	LTD-10-25 [13]	MotionMixer [9]	STSGCN [8]	SPGSN [36]	Ours
Walking	66.1	63.6	44.4	42.4	45.9	41.5	**38.4**
Eating	61.7	48.4	38.6	36.1	45.0	38.0	**35.8**
Smoking	65.4	48.9	39.5	36.8	44.7	**34.6**	36.0
Discussion	91.3	77.6	68.1	64.1	68.5	67.1	**63.9**
Direction	84.1	69.7	58.0	53.4	53.2	**50.3**	53.5
Greeting	108.8	96.0	82.6	82.2	87.6	86.4	**76.3**
Phoning	76.4	59.9	50.8	51.1	52.0	48.5	**48.2**
Waiting	87.7	69.7	44.4	56.4	59.2	54.1	**53.5**
WalkingDog	110.6	103.3	38.6	87.8	93.3	**84.9**	87.0
WalkingToge	67.3	61.2	39.5	43.5	43.9	40.9	**38.5**
Posing	114.3	92.9	79.9	79.5	73.1	76.5	**68.8**
Purchases	100.7	89.9	78.1	76.1	79.6	74.4	**73.7**
Sitting	91.2	63.1	58.3	54.5	57.8	**53.4**	54.8
Sitting down	112.0	82.7	76.4	74.5	76.8	**70.7**	73.7
Taking photo	87.6	63.6	54.3	51.6	56.3	52.7	**50.1**
Average	88.3	72.7	68.1	59.3	62.9	58.3	**56.8**

Table 2. Performance comparison between different methods for long-term prediction (1000 ms) of MPJPE (mm) for each activity on the Human3.6M dataset.

Motion	Res.Sup [7]	convSeq2Seq [15]	LTD-10-25 [13]	MotionMixer [9]	STSGCN [8]	SPGSN [36]	Ours
Walking	79.1	82.3	60.9	59.9	66.7	**53.6**	55.2
Eating	98.0	87.1	75.8	76.6	75.1	73.4	**73.1**
Smoking	102.1	81.7	72.1	**68.5**	74.1	68.6	70.2
Discussion	131.8	129.3	118.5	117.4	**107.7**	118.6	117.1
Direction	129.1	115.8	105.5	105.4	109.9	**100.5**	105.2
Greeting	153.9	147.3	136.8	136.5	**103.8**	143.2	136.7
Phoning	126.4	114.0	105.1	104.4	109.9	**102.5**	103.2
Waiting	135.4	117.7	108.3	107.7	118.3	**103.6**	103.8
WalkingDog	164.5	162.4	146.4	142.2	**118.3**	138.0	145.5
WalkingToge	98.2	87.4	65.7	65.4	95.8	**60.9**	61.8
Posing	183.2	187.4	174.8	174.9	107.6	**165.4**	168.4
Purchases	154.0	151.5	134.9	135.1	**119.3**	133.9	132.6
Sitting	152.6	120.7	118.7	115.7	119.8	116.2	**114.7**
SittingDown	187.4	150.3	143.8	141.1	**129.7**	149.9	141.5
TakingPhoto	153.9	128.1	115.9	114.6	119.8	118.2	**111.9**
Average	136.6	124.2	112.4	111.0	113.3	109.6	**109.4**

As can be seen from Tables 1 and 2, the short-term and long-term prediction performance of this paper's method in estimating average frames for each activity on the Human3.6M dataset outperforms previous methods. In particular, the error of this method in short-term prediction of MPJPE for various types of action behaviors on the Human3.6M dataset is reduced by 11.3 mm and 2.5 mm compared to the existing LTD-10-25 [13] and MotionMixer [9] methods. The error of long-term prediction of MPJPE for various types of action behaviors on the Human3.6M dataset is reduced by 11.3 mm and 2.5 mm compared to the existing LTD-10-25 and MotionMixer methods were reduced by 3 mm and 0.6 mm.

As can be seen from Table 3, the performance of this paper's method in short-term prediction and long-term prediction in the estimation of the average frames of each activity on the 3DPW dataset is better than the previous methods. Comparing with the existing MLP-based motion prediction method Motion-Attention [9] and Transformer-based method AuxFomer [36], the reductions are 4.2 mm and 18.3 mm in short-term prediction and 3.5 mm and 39.2 mm in long-term prediction, respectively. In order to show the detailed information in the IHC dance videos more clearly, we pre-processed the input dance videos, retained the first 15s of the videos for input into the model, and compared the output with the intercepted videos (Truth). We visualized the sequence of individual joint coordinates output from the network model, where the visualization results for the Hmong dance, the Korean dance and the Mongolian dance are shown in Figures 7 and 8.

Table 3. Performance comparison between different methods for short-term forecasting and long-term forecasting of MPJPE (mm) for each activity on the 3DPW dataset.

Millisecond	400	1000
convSeq2Seq [15]	58.8	87.8
LTD-10-25 [13]	46.6	75.5
Motion-Attention [9]	44.4	71.8
AuxFormer [36]	58.5	107.5
Ours	**40.2**	**68.3**

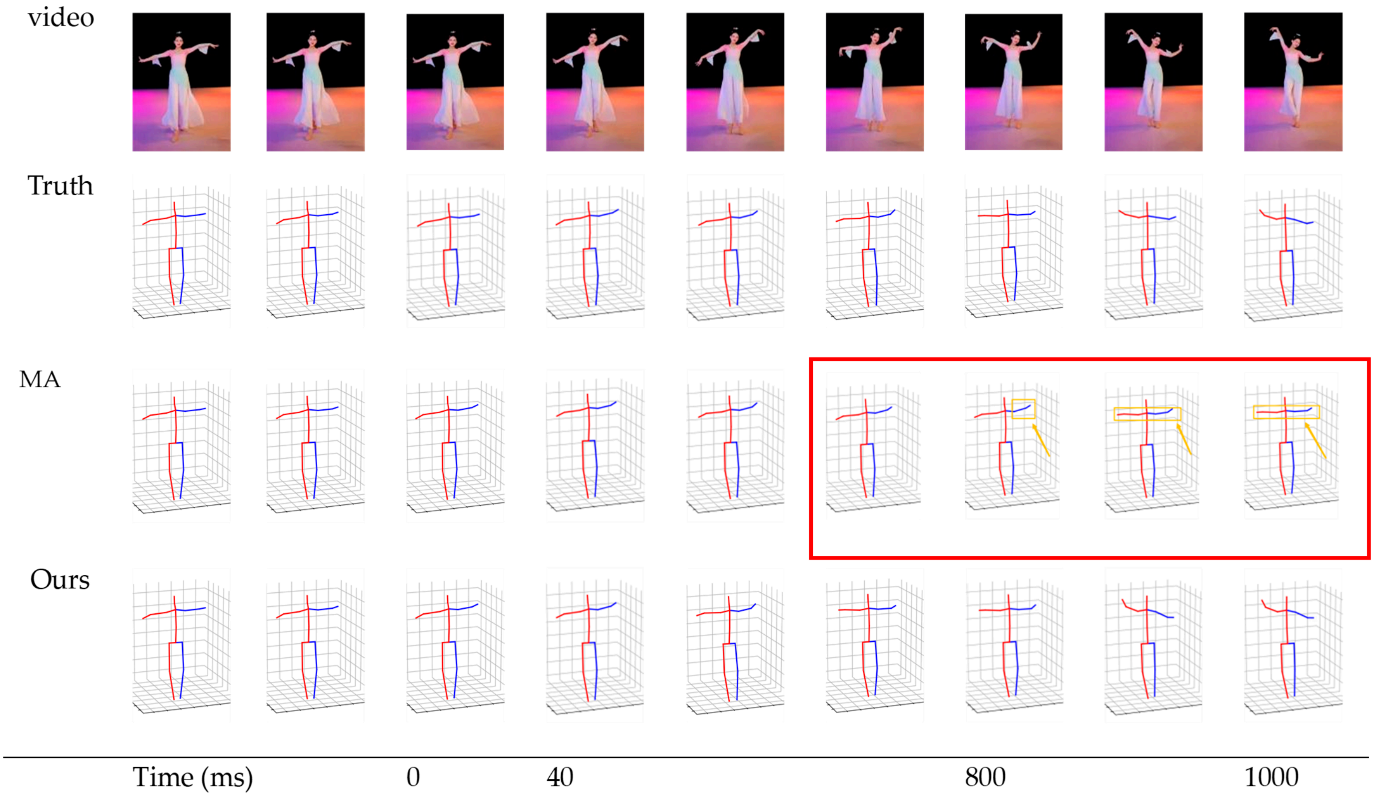


Figure 7. Visual results of short-term and long-term prediction sequences for Dai ethnic dance [13].

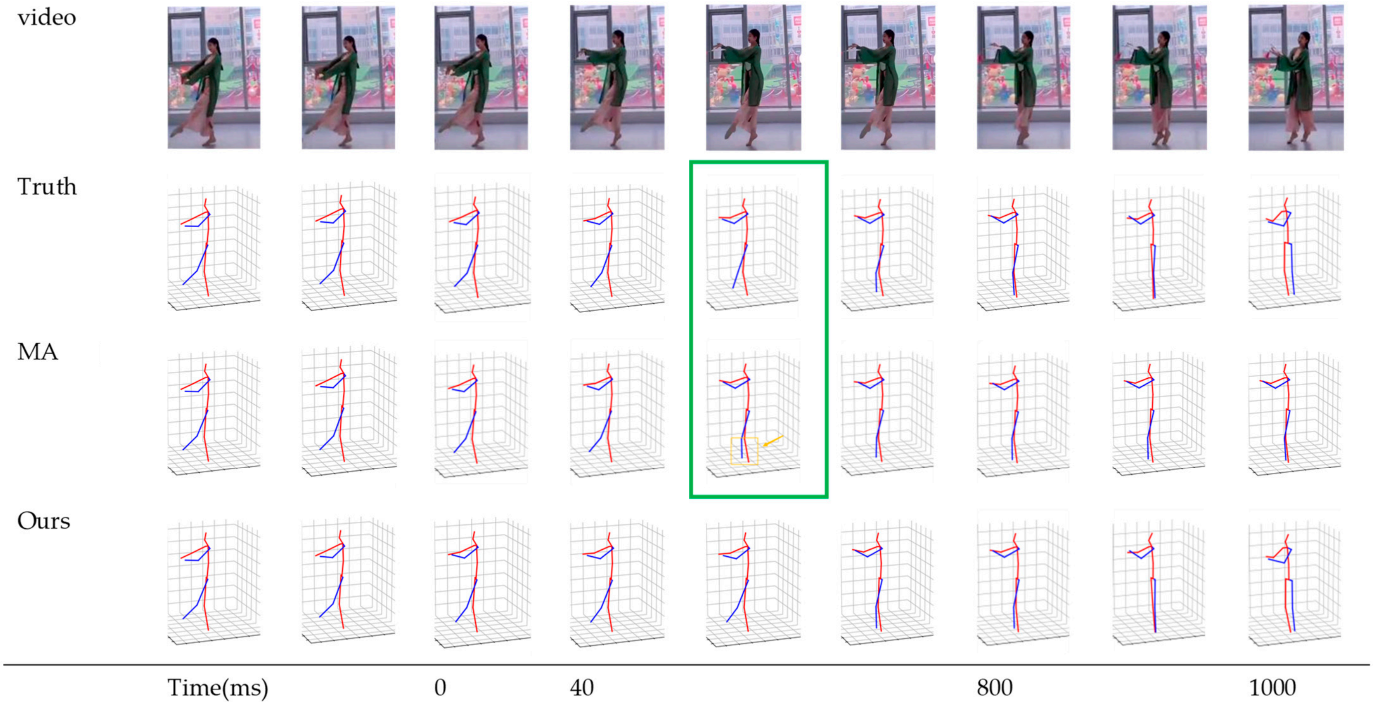


Figure 8. Visual results of short-term and long-term prediction sequences under the condition of simultaneous upper and lower limb extension [13].

In Figures 7 and 8, the first row represents the frames of the intangible cultural heritage dance videos after splitting, the second row showcases the visual sequence of three-dimensional human pose coordinates from our self-made dataset, the third row displays the prediction video frame sequence generated by the Motion-Attention method, and the fourth row exhibits the prediction video sequence produced by our proposed method. In Figures 7 and 8, it can be observed that significant errors in the predictions from Motion-Attention become noticeable at the 500 ms mark (highlighted in red) and the 100 ms mark (highlighted in green). Therefore, it is evident that our method outperforms Motion-Attention by accurately and reasonably completing the actions, providing a superior prediction of the motion sequences.

4.3. Ablation Experiments

To verify the impact brought by each module and design in the method, we conduct ablation experiments based on the MPJPE evaluation metrics on the Human3.6M dataset. The first is the impact brought by different graph structures on the accuracy of motion prediction. Table 4 demonstrates that the method in this paper employs the traditional graph structure [10], multi-scale graph [35], hypergraph [17], and multi-scale hypergraph for short-term and long-term prediction. It can be seen that the multi-scale hypergraph designed in this paper reduces the MPJPE errors in short-term prediction and long-term prediction by 2.1 mm and 4.1 mm, respectively, compared to the traditional graph structure. It indicates that the multi-scale hypergraph designed in this paper for joints is effective.

Table 4. Comparative depletion experiments of different graph structures.

Graph Structures	**MPJPE**	
	400	**1000**
Traditional Graphs [10]	58.9	113.5
Multi-Scale Graphs [35]	58.6	110.1
Spatial Hypergraphs [17]	57.2	109.8
Multi-Scale Hypergraphs	**56.8**	**109.4**

Then, ablation experiments were performed for the scale segmentation operator. We compared the method of this paper with the method of replacing scale generation with joint removal [37], and Table 5 demonstrated that the method of this paper and the MSGC method for short-term prediction and long-term prediction, the multi-scale segmentation operator designed in this paper reduces the MPJPE error by 0.8 mm and 0.5 mm in short-term prediction and long-term prediction, respectively, in comparison with the MSGC method, to prove that our segmentation method is effective.

Table 5. Comparative depletion experiments of different scale segmentation operators.

Segmentation Method/Scale	**MPJPE**				
	Joint Scale	**Skeleton Scale**	**Component Scale**	**400**	**1000**
MSGC [38]	√	√	√	69.2	119.4
Ours-1L	√			69.6	119.8
Ours-2L	√	√		69.1	119.3
Ours	√	√	√	**68.4**	**118.9**

Next, ablation experiments of segmentation were performed on different scales of segmentation. To verify the effectiveness of the multi-scale mechanism, we set the segmentation scales as joint scale, bone scale, and part scale, respectively, and conducted experiments on different numbers of scales. From Table 5, compared with the short-term prediction and long-term prediction results of Ours-1L, Ours-2L, and Ours, it can be seen that the three scales segmentation of this paper's method gives the optimal results, which verifies the effectiveness of the multi-scale architecture.

At the same time, we conducted ablation experiments on the number of multi-scale hypergraph convolution modules. To verify the effect of the stacking of multi-scale hypergraph convolution modules, we conduct experiments on different numbers of multi-scale hypergraph convolution, adjust the number from 1 to 4, and show the MPJPE on the H3.6M dataset. As can be seen in Table 6, the MPJPE gradually decreases when we use 1 to 3 modules, and when the number of multi-scale hypergraph convolution modules is increased to 4, the MPJPE value tends to increase, so we finally choose to stack 3 multi-scale hypergraph convolution modules.

Table 6. Comparative depletion experiments of different numbers of multi-scale convolution modules.

Number of (MCM)	**MPJPE**				
	1	**2**	**3**	**4**	**0**
400 ms	57.4	56.9	**56.8**	57.2	57.6
1000 ms	109.9	109.8	**109.4**	110.7	110.9

Finally, the effectiveness of the TGT module is tested. To verify the effectiveness of our introduced Temporal Graph Transformer, we have experimented with this paper's method with other time series models [37,39]. From Table 7, we can see that the introduction of the TGT module is effective.

Table 7. Depletion experiments using different time series models.

Decoder	**400**	**1000**
TCN [39]	58.7	112.9
GRU [37]	58.2	112.7
Ours	**56.8**	**109.4**

In order to verify the validity of our movement prediction method, we invited 100 dance majors and 100 non-dance majors as well as 6 professional modern dance teachers and 3 professional intangible cultural heritage dance teachers to score our method by means of a questionnaire. The composition of the experimental subjects is shown in Figure 9.

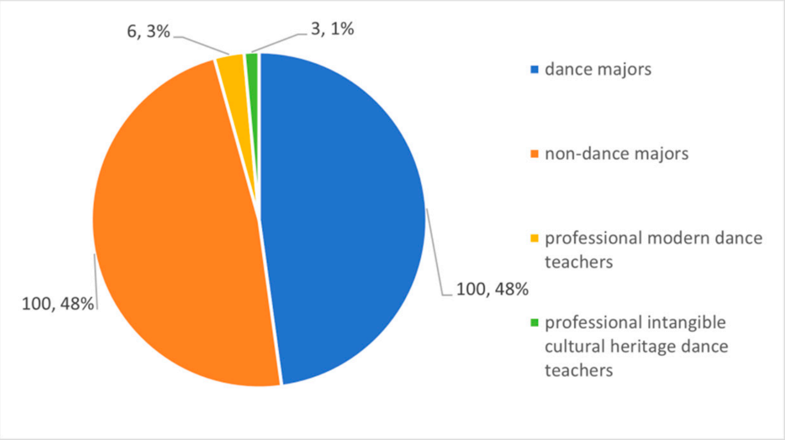


Figure 9. Number of subjects and their composition.

Subjects were allowed to score the predicted segments after learning them, marking them out of 5. Subjects could choose 0–5 to rate the results. A 15 s video of each of the five NRL performance forms, namely Dai, Miao, Wei, Tibetan dance and Jiaozhou Yangge, is selected as the experimental data, and the prediction results are uniformly outputted for one kind of video in each experiment. The experiment recorded the average of all subjects' scores on the predicted dance segment. The statistical results of the experiment are shown in Figure 10.

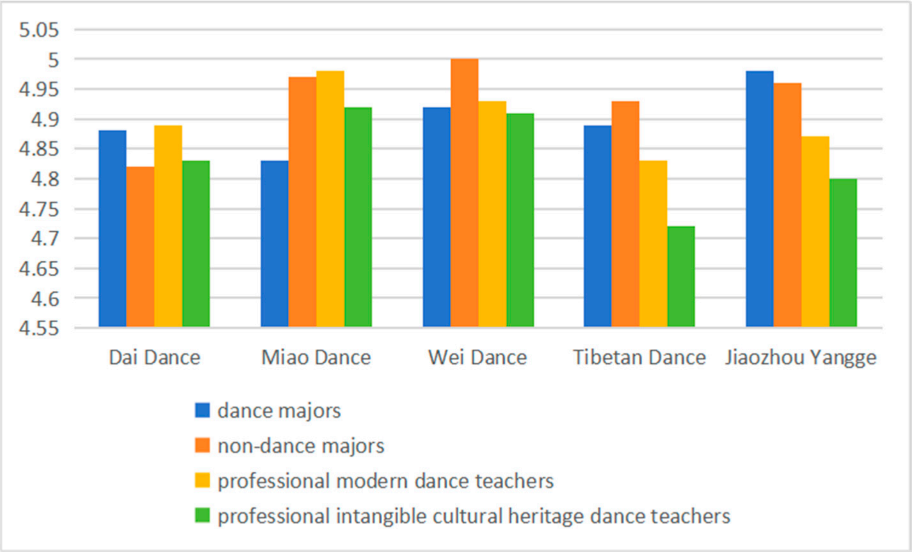


Figure 10. Statistical table of subjects' scores on the prediction results.

From Figures 9 and 10, it can be seen that the scores of the Viennese dance are higher than other dances, which may be analyzed due to the fact that the Viennese dance has more repetitive movements within one octave, which makes it easier for learners to learn. The overall scores are higher than 4.7, which shows that our method can meet the needs of the experimental use scenarios for dance movement prediction, and the present approach is effective.

5. Discussion

In this section, we summarize the results of our study on HPP (Human Pose Prediction) and compare them with the results of previous studies. We will then analyze the limitations of the proposed multi-scale hypergraph convolutional network and propose future research directions regarding HPP.

We focus on using hypergraphs to represent interactions across joints and have achieved satisfactory results on Human3.6m and 3DPW. The main underlying network of our approach is the multi-scale graph convolutional network. The multi-scale network can extract richer spatial information of the 3D human skeleton, combined with the improved Transformer to extract rich temporal information.

Our method has satisfactory results in short-term prediction, but only the transformer module handles temporal information in our method, and our multi-scale hypergraph structure is only concerned with the processing of spatial features, which leads to poor results in long-term prediction.

6. Conclusions

We tackled the problem that intangible cultural heritage dance platoon characteristics have a lack of interaction of remote joints, which leads to the low accuracy of current human motion prediction methods. Compared to previous methods that only consider extracting rich-scale information while ignoring cross-joint interaction information, this paper proposes a human motion prediction method based on multi-scale hypergraphical convolutional networks for intangible cultural heritage dance videos. We input the intangible cultural heritage dance video, and the 3D pose sequence is extracted by the 3D pose estimation algorithm MHFormer. Motion prediction is performed on the extracted 3D pose sequence of the dance action video. Firstly, we input the 3D human posture sequence of the intangible cultural heritage dance video, and designed the spatial hypergraph structure according to the interaction relationship of multiple human joints in the dance video, to enrich the connection between the 3D skeletal joint points. Then, we constructed a multi-scale hypergraph convolutional network by means of a spatial hypergraph, using the scale transformation operator to replace the simple way of taking a certain node to replace the part information, the extract spatial features in 3D human posture sequences. Finally, Temporal Graph Transformer network was introduced to extract the temporal features in the 3D human posture sequence and output the predicted 3D human posture coordinate sequence. It was experimentally verified that this method can obtain more accurate predicted 3D human posture sequence results in the case of intangible cultural heritage dance videos where there are isotropic extensions of the upper and lower limbs.

The next phase of research will be to reduce the computational complexity by adopting a lighter weight network structure. For example, alleviating some redundant layers of Temporal Graph Transformer, streamlining the number of multi-scale hypergraph modules, etc. In addition, the algorithm will be extended to multi-person dance videos, and a 3D human movement prediction system for dance videos will be developed and promoted for more convenient application in the field of non-heritage dance teaching.

Author Contributions: Conceptualization, X.C. and P.C.; methodology, X.C. and P.C.; software, X.C.; validation, X.C., P.C. and S.L.; formal analysis, X.C.; investigation, P.C.; resources, S.L.; data curation, P.C.; writing—original draft preparation, P.C.; writing—review and editing, X.C. and H.S.; visualization, P.C.; supervision, S.L.; project administration, H.Z.; funding acquisition, X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Funding Project of National Natural Science Foundation of China (No. 61503005) and the Funding Project of Humanities and Social Sciences Foundation of the Ministry of Education in China (No. 22YJAZH002).

Data Availability Statement: Data are contained within the article.

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

References

1. Yang, X.N. The physical anthropological value of "intangible cultural heritage" dances. *House Dra.* **2018**, *32*, 110.
2. Li, K.N. Protection and inheritance of ethnic folk dances from the perspective of intangible cultural heritage. *Dancefahion* **2022**, *12*, 98–100.
3. Chen, S.; Liu, B.; Feng, C.; Vallespi-Gonzalez, C.; Wellington, C. 3D point cloud processing and learning for autonomous driving. *IEEE Signal Process* **2020**, *38*, 68–86. [CrossRef]
4. Troje, N.F. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *J. Vis.* **2002**, *2*, 371–387. [CrossRef] [PubMed]
5. Ankur, G.; Julieta, M.; James, L.; Robert, W. 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2061–2068.
6. Kong, Y.; Fu, Y. Human action recognition and prediction: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 1366–1401. [CrossRef]
7. Martinez, J.; Black, M.J.; Romer, J. On human motion prediction using recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2891–2900.
8. Sofianos, T.; Sampieri, A.; Franco, L.; Galasso, F. Spacetime-separable graph convolutional network for pose forecasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11209–11218.
9. Bouazizi, A.; Holzbock, A.; Kressel, U.; Dietmayer, K.; Belagiannis, V. Motionmixer: Mlp-based 3d human body pose forecasting. *arXiv* **2022**, arXiv:2207.00499.
10. 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/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11467–11476.
11. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]

12. Von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering accurate 3d human pose in the wild using imus and a moving camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 601–617.
13. Mao, W.; Liu, M.; Salzmann, M. History repeats itself: Human motion prediction via motion attention. In Proceedings of the 2020 16th Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 474–489.
14. Fragkiadaki, K.; Levine, S.; Felsen, P.; Malik, J. Recurrent network models for human dynamics. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4346–4354.
15. Li, C.; Zhang, Z.; Lee, W.S.; Lee, G.H. Convolutional sequence to sequence model for human dynamics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5226–5234.
16. Zhou, D.; Huang, J.; Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Adv. Neural Inf. Process. Syst.* **2006**, *19*.
17. Agarwal, S.; Lim, J.; Zelnik, M.L.; Perona, P.; Kriegman, D.; Belongie, S. Beyond pairwise clustering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 838–845.
18. Tian, Z.; Hwang, T.H.; Kuang, R. A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge. *Bioinformatics* **2009**, *25*, 2831–2838. [CrossRef]
19. Bu, Y.; Howe, B.; Balazinska, M.; Ernst, M.D. HaLoop: Efficient iterative data processing on large clusters. In Proceedings of the VLDB Endowment, Seattle, WA, USA, 29 August–3 September 2010; Volume 3, pp. 285–296.
20. Li, W.; Liu, X.; Liu, Z.; Du, F.; Zou, Q. Skeleton-based action recognition using multi-scale and multi-stream improved graph convolutional network. *IEEE Access* **2020**, *8*, 144529–144542. [CrossRef]
21. Fan, Y.; Wang, X.; Lv, T.; Wu, L. Multi-scale adaptive graph convolutional network for skeleton-based action recognition. In Proceedings of the 15th International Conference on Computer Science & Education (ICCSE), Delft, The Netherlands, 18 August 2020; pp. 517–522.
22. Li, T.; Zhang, R.; Li, Q. Multi scale temporal graph networks for skeleton-based action recognition. *arXiv* **2020**, arXiv:2012.02970.
23. Yuan, Y.; Kitani, K. Ego-pose estimation and forecasting as real-time pd control. In Proceedings of the the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10082–10092.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; p. 30.
25. Cheng, Y.B.; Chen, X.; Zhang, D.; Lin, L. Motion-transformer: Self-supervised pre-training for skeleton-based action recognition. In Proceedings of the 2nd ACM International Conference on Multimedia in Asia, Beijing, China, 16–18 December 2021.
26. Lin, H.; Cheng, X.; Wu, X.; Shen, D. Cat: Cross attention in vision transformer. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
27. Wu, N.; Green, B.; Ben, X.; O'Banion, S. Deep transformer models for time series forecasting: The influenza prevalence case. *arXiv* **2020**, arXiv:2001.08317.
28. Kanchana, R.; Muzammal, N.; Salman, K.; Fahad, S.K.; Michael, R. Self-supervised video transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022.
29. Lan, G.; Wu, Y.; Hu, F.; Hao, Q. Vision-based human pose estimation via deep learning: A survey. *IEEE Trans. Hum. Mach. Syst.* **2022**, *53*, 253–268. [CrossRef]
30. Li, S.; Chan, A.B. 3D human pose estimation from monocular images with deep convolutional neural network. In Proceedings of the 2015 Computer Vision (ACCV), Singapore, 1–5 November 2015; Springer: Berlin/Heidelberg, Germany; pp. 332–347.
31. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3D human pose estimation with spatial and temporal transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11656–11665.
32. Sapp, B.; Taskar, B. Modec: Multimodal decomposable models for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3674–3681.
33. Li, W.; Liu, H.; Tang, H.; Wang, P.; Van Gool, L. Mhformer: Multi-hypothesis transformer for 3D human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 13147–13156.
34. Liao, J.; Xu, J.; Shen, Y.; Lin, S. THANet: Transferring Human Pose Estimation to Animal Pose Estimation. *Electronics* **2023**, *12*, 4210. [CrossRef]
35. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3316–3333. [CrossRef] [PubMed]
36. Li, M.; Chen, S.; Zhang, Z.; Xie, L.; Tian, Q.; Zhang, Y. Skeleton-parted graph scattering networks for 3d human motion prediction. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 18–36.
37. Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; Tian, Q. Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction. *IEEE Trans. Image Process.* **2021**, *30*, 7760–7775. [CrossRef] [PubMed]

38. Gui, Z.; Peng, D.; Wu, H.; Long, X. MSGC: Multi-scale grid clustering by fusing analytical granularity and visual cognition for detecting hierarchical spatial patterns. *Future Gener. Comput. Syst.* **2020**, *112*, 1038–1056. [CrossRef]
39. Zhai, D.H.; Yan, Z.; Xia, Y. Lightweight Multiscale Spatiotemporal Locally Connected Graph Convolutional Networks for Single Human Motion Forecasting. *IEEE Trans. Autom. Sci. Eng.* **2023**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.