IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 September 2023, accepted 8 October 2023, date of publication 17 October 2023, date of current version 23 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325291

RESEARCH ARTICLE

Human Motion Prediction Based on Space-Time-Separable Graph Convolutional Network

RUI LI[1,2], DUO HE[1], SHIQIANG YANG[1], AN YAN[1], XIN ZENG[1], AND DEXIN LI[1]
[1]School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, Xi'an 710048, China
[2]Xi'an People's Hospital, Xi'an 710100, China

Corresponding author: Shiqiang Yang (yangsq@126.com)

This work was supported in part by the Key Laboratory of Unmanned Aerial Vehicle (UAV), Northwestern Polytechnical University (NWPU), under Grant 2022-JCJQ-LB-719; and in part by the Foundation Strengthening Programme Technical Area Fund under Grant 2020-JCJQ-JJ-367.

ABSTRACT Human motion prediction is a popular method to predict future motion sequences based on past sequences, which is widely used in human-computer interaction. Space-time-separable graph Convolutional Network (STS-GCN) is a conventional mathematical model for human motion prediction. However, the uncertainty of human movements often leads to the problem of significant prediction error in the prediction results. This paper first proposed a Multi-scale STS-GCN (MSTS-GCN) model based on the conventional STS-GCN method to find the relevant factors that affect the prediction results. In our study, the constructed Multi-scale Temporal Convolutional Network (MTCN) decoder effectively reduced the human motion prediction error at specific time nodes. To expand the transmission and utilization performance in a larger receptive field, a Gated Recurrent Unit-TCN decoder was also designed. Finally, a new STS-GCN (NSTS-GCN) human motion prediction model was proposed, which realized the transmission and utilization of motion sequence features under a larger temporal perceptual field. To verify the effectiveness of NSTS-GCN, the Human3.6M dataset, AMASS, and 3DPW dataset were tested. The experimental results show that the MPJPE error of the proposed model for human joint prediction at each time node is reduced compared with the conventional STS-GCN model, and the mean reduction was achieved by 3.0mm. All the experimental results validated the effectiveness of the proposed NSTS-GCN model, which further improved the performance of human motion prediction.

INDEX TERMS Human motion prediction, decoder, STS-GCN, GRU.

I. INTRODUCTION

In recent years, motion capture data has attracted considerable attention for the prediction of human motion, which has been applied in a wide range of applications. With China's aging population, using human motion prediction in-home monitoring can help reduce accidents among the elderly. Meanwhile, using human motion prediction in automatic driving is helpful in preventing traffic accidents and improving the safety performance of vehicles. In the field of Virtual Reality and Augmented Reality, this technology is used to track the user's body movement to realize an immersive experience. The core technology of human motion prediction aims to predict subsequent motion sequences based on historical human motion sequences, which can enhance the real-time and efficient human-robot interaction system by judging human motion changes and motion trajectories in advance [1]. By predicting human motion trends, the robot can perform path planning to avoid collision or drift with people and improve collaboration efficiency. Therefore, human motion prediction is of great significance in human-computer interaction [2], [3], healthcare [4], [5], intelligent driving [6], and other fields.

Early human motion prediction methods based on deep learning mostly use Recurrent Neural Networks (RNN) [7], [8] due to its advantages for time series tasks [9]. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

its performance was limited by human body dynamics, which caused its performance hard to improve further. Another famous study was performed by Abdullahi, who designed a bidirectional long-short term memory-fast fisher vector algorithm to train 3D hand skeletal information of motion and orientation angle features and further used it to classify dynamic sign words [10], [11]. One of the representative works [12], [13] from Sejong University proposed a cloud-assisted IoT computing framework for human activity recognition in uncertain low-lighting environments and applied a lightweight three-dimensional convolutional neural network architecture to extract spatiotemporal features from significant frames to easily identify violent behaviors in video. This group also pre-trained a vision transformer to extract frame features and did research on identifying abnormal behaviors in the video [14], [15].

Since (GCN) [16], [17], [18] can compensate for the inherent deficiency of weak spatial modeling ability that exists in RNNs, they have achieved good results in human motion recognition. In recent years, researchers have started to apply GCN to predict human motion [19], [20], especially to encode the spatial-temporal features of human skeleton sequences. Li et al. [21] proposed a Dynamic Multiscale Graph Neural Networks model to simulate the internal relationships of the human body extract single-scale features, and perform cross-scale feature fusion through multiscale graph computation units, It achieved good results in human motion prediction. Zhou and his colleagues [22] proposed a new Multiscale Graph Convolution Network to capture the correlation among human body components and deeply explored the correlation between human joints and components in the multiscale graph. However, most of the existing studies consider the dependencies between joints and ignore the interrelationships between bones, which affects the prediction accuracy. To solve this problem [23], a directed acyclic graph was used to represent the human skeleton, with joints as vertices and bones as directed edges and updated joint and bone features based on the observed human motion states. This approach successfully predicted human motion in realistic prediction scenarios. Zhang et al. [24] proposed a structured method to predict bone points. In their approach, they used motion features to predict the basic joints. Then they combined the upper predicted joint points, extracted motion features to predict the next joint, and then iteration to the whole bone in the short-term prediction of human movement.

Another basic model commonly used for human motion prediction is Transform, which uses a self-attention mechanism to process input information and generate the output. Based on transformer global attention architecture and progressive decoding strategy, Cai et al. [25] predicted the human motion according to the kinematic tree gradual prediction target joint DCT coefficient, the centre of the eight joints as seed joints, and then estimated the structural connectivity of the body skeleton, the joint prediction from the centre to the periphery. Guo et al. [26] proposed a lightweight network based on multi-layer perceptron (MLP), which used DCT transform to encode time information. This method also combined the prediction of joint residual displacement and optimization speed as auxiliary loss. It achieved an excellent prediction effect with only three components: a fully connected layer, a normalized layer, and a transpose operation.

Combine ''spatial attention'' and ''temporal attention'' mechanisms, Aksan et al. [27] use a transformer model to decouple temporal and spatial self-attention mechanisms composed of temporally coherent postures, then generating more reasonable future skeletons in the short and long term. Dang and his colleagues [28] designed a multiple GCN with multiscale architecture to compensate for the ability of GCN modelling stratification and context information of human posture. In detail, a set of GCN forms to extract features and another set of GCN to add residual connections between input and output pose, which allows the whole framework to learn more representative features. Li et al. [29] proposed a Multiscale Spatio-Temporal Graph Neural Network (MST-GNN), whose core is a multiscale spatiotemporal graph that simulates motion relationships on different spatial and temporal scales and successfully implemented human motion sequences based on skeletal features under motion category uncertainty. Considering the role of time and space dimensions separately will limit complex motion and the understanding of the spatio-temporal dynamics of the human body. Sofianos et al. [30] proposed the Space-Time-Separable Graph Convolutional Network (STS-GCN), the first spatio-temporal separable GCN that decomposes the spatio-temporal graph connectivity matrix into temporal and spatial affinity matrices while achieving full exploitation of joint-joint and time-time correlations. Although STS-GCN achieves a better human motion prediction performance, due to the uncertainty and randomness of human motion, there is still the problem of long-term prediction inaccuracy inherent in prediction due to the lack of effective features when performing long-time motion sequence prediction.

This paper constructs the Multi-scale Space-Time-Separable Graph Convolutional Network (MSTS-GCN) model based on the STS-GCN model by designing the multiscale decoder MTCN to obtain the motion sequence characteristics at different time scales, and the human motion prediction performance at some time nodes are improved, thus precisely the decoder is an important factor to improve the model accuracy effectively; based on this analysis result, the TCN decoder of STS-GCN model is fused with the GRU [31] to establish the GRU-TCN decoder, which combines the advantages of parallel data processing and higher efficiency of TCN with the ability of GRU to preserve the effective information in long-term sequences, to achieve efficient transmission and utilization of motion sequence features under a larger time perception field and construct a complete NSTS-GCN human motion prediction model.

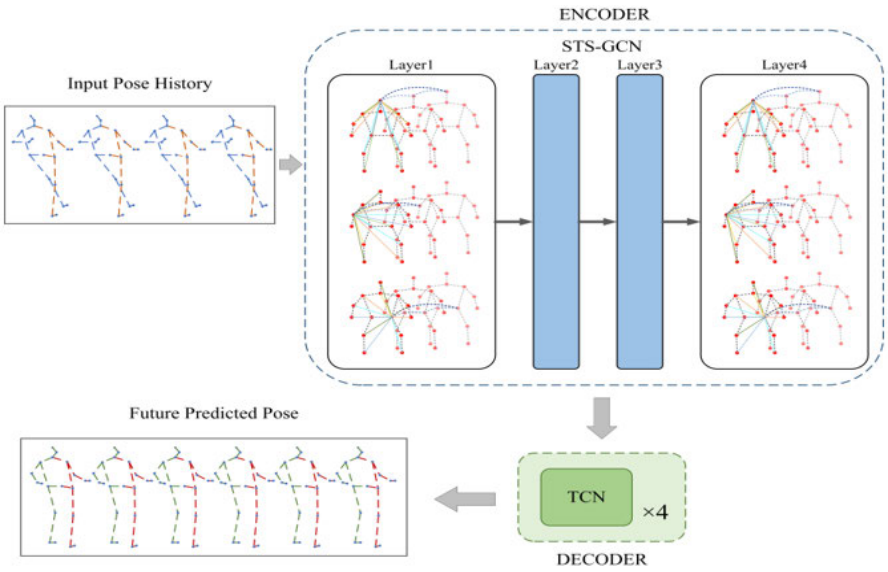


FIGURE 1. Framework of the STS-GCN model.

The main contributions of this paper include the following:

(1) In order to verify the decoding ability of the STS-GCN decoder, the MTCN decoder was designed, and the MSTS-GCN human motion prediction model was constructed, which improved some prediction results, and proved that the decoder is an important factor affecting the accuracy of human motion prediction;

(2) In order to reduce the error of human motion prediction, the GRU-TCN decoder is designed, and the NSTS-GCN human motion prediction model is constructed, realizing the transmission and utilization of motion sequence features in a larger time perception field, and obtaining more abundant human motion features;

(3) To verify the effectiveness of NSTS-GCN, we tested it in the Human3.6M dataset, AMASS, and 3 DPW datasets and demonstrated that the model can effectively reduce the error of human motion prediction, reducing the error at each time node.

II. METHOD

Human motion prediction methods output predicted future human motion sequences based on the observed historical motion sequences. Two famous frameworks of encoder-decoder and autoencoder were usually used in this field. The encoder-decoder framework is typically used for Sequence-to-Sequence (Seq2Seq) tasks, where the input sequence is encoded into a fixed-length vector. Then, the decoder converts that vector into the target sequence. In contrast, autoencoders are usually used for dimensionality reduction and feature learning, which usually do not involve sequence data processing, but use structures such as fully connected layers to process the input data. Hence, most graph convolution-based human motion prediction methods use encoder-decoder architecture [32], [33], which encodes the spatio-temporal information of motion sequences by an encoder, and then decodes the resulting feature vectors to predict future motion sequences.

The traditional decoder cannot utilize more effective information in the decoding process due to the defect of a single structure, so it is still necessary to construct a novel decoder to obtain the deep features of historical motion sequences and reduce the error of human motion prediction. This paper takes the STS-GCN model as the basis, by designing a decoder architecture based on multi-scale temporal convolution to obtain features at different scales, which can reduce the prediction error at some time nodes. Furthermore, a GRU-TCN decoder was designed, which combines the advantages of GRU and TCN in processing motion sequence features to obtain a lower prediction error.

A. STS-GCN MODEL

STS-GCN [30] is a human motion prediction model based on the Encoder-Decoder framework. The input human motion history sequence is encoded by the Encoder and converted into a feature vector of a specific length. Then, the extracted feature vector is decoded by the Decoder to obtain the predicted motion sequence. The flow framework of the STS-GCN model is shown in Figure 1, consisting of an encoder and 4-layers of temporal convolutional modules. The detailed procedure is as follows.

The input data of the STS-GCN model is a T-frame historical motion sequence represented by $\Xi_{in} = [X_1, X_2, \ldots, X_T]$, where $X_i \in \mathbb{R}^{3\times V}$ is a vector consisting of 3D coordinates or angles of all human joints in frame i. The number of joints in this step depends on the selected data set. Suppose we selected the AMASS dataset, the joint number is decided as 18.

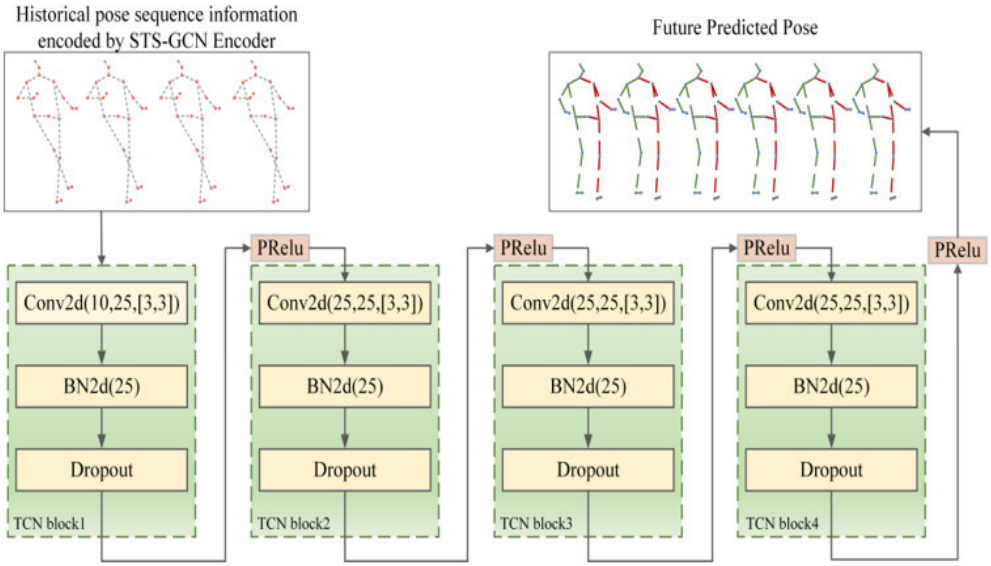


FIGURE 2. Structure of the TCN decoder.

The encoder of STS-GCN uses temporal separable graph convolution to model the input Ξ_{in} and encodes the motion sequence into a graph structure $G = (v, \varepsilon)$. When encoding all joint information is completed, it decoded by the TCN decoder to obtain the joint coordinate information of the predicted skeleton sequence and output the human motion prediction sequence for the next K-frames, which is represented by out $\Xi_{out} = [X_{T+1}, X_{T+2}, \ldots, X_{T+K}]$.

The encoder of the STS-GCN model uses a 4-layer spatio-temporal separable graph convolution with residual connected Parametric Rectified Linear Unit (PReLU) activation function, to obtain rich motion features in the spatial and temporal domains through the separable graph convolution, respectively. The decoder of the STS-GCN model uses a TCN module, which has a simple structure and can receive sequence inputs of any length and produce output features of the same length. The decoder's specific structure is shown in Figure 2. The historical human motion sequences are modeled by 4-layer temporal separable graph convolution, and after obtaining the encoded human motion sequence information, a temporal convolutional decoder consisting of 4 TCN modules with the same structure is used to map the encoded output information to the future time range and predict the future 3D coordinates or angles of human joints.

From Figure 2, it is clearly found that the decoder uses 4 TCN modules with the same architecture, each of which contains a two-dimensional convolution with a kernel size of [3, 3], a BN layer, and a Dropout layer, with BN used to speed up the convergence of the network and Dropout used to prevent model overfitting. Each module is followed by a PReLU activation function to further improve the model fitting ability. More accurate human motion features can be extracted through the iterative motion of the four temporal convolution modules, and the predicted future motion sequences will be output.

Since the four temporal convolutions are the same size, the decoder convolutions' receptive field is single, which cannot extract rich feature correlations between different motion sequences. It also lacks the feature interaction between different timing information. To solve the above problems, this study proposes the multi-scale decoder MSTS-GCN to realize feature fusion between different receptive field ranges.

B. MSTS-GCN MODEL WITH THE INTRODUCTION OF MTCN DECODER

To improve the performance of human motion prediction, most researchers have devoted themselves to making full use of relevant spatial and temporal information in the process of encoding historical motion sequences, ignoring the subsequent process of obtaining predicted motion sequences by decoding. How to effectively utilize the encoding information obtained from the encoder to obtain human motion sequences with smaller errors by an efficient decoder needs further study.

The STS-GCN model's decoder only uses four identical [3, 3] temporal convolutional layers for decoding motion sequences, which have a simple structure and a small number of parameters. The decoding process has a small convolutional field of perception, can only correlate the previous frame and the next frame, and lacks remote information interaction between motion sequences. This study resets the convolution kernel of each layer of temporal convolution in the TCN decoder and establishes the MTCN decoder to solve the problem that the TCN cannot obtain effective remote correlation information, and the MTCN structure is shown in Figure 3.

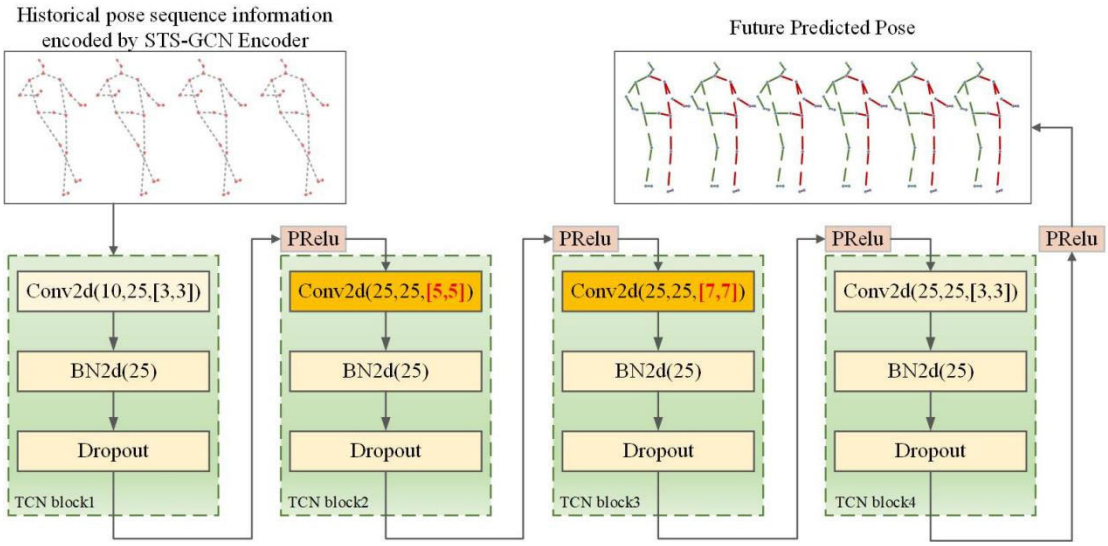


FIGURE 3. Structure of MTCN decoder.

The MTCN sets the convolution kernel sizes of the first to fourth TCN modules to [3, 3], [5, 5], [7, 7], and [3, 3], respectively, to equip the network with flexible temporal perceptual field ranges and thus obtain motion features at different scales. For the motion sequence information input to the decoder, the first TCN module keeps the original convolution kernel setting and obtains the neighboring valid information through smaller convolution kernels, to obtain highly correlated motion features. The second and third TCN modules use incremental convolution kernel sizes to obtain relevant motion information at farther distances. Finally, the [3, 3] convolution kernels are used to aggregate the key information of the before and after frames to obtain more effective decoding data. The MTCN decoder is applied to the STS-GCN model, and the MSTS-GCN model is constructed.

Compared to the performance of two decoders, the major drawback of the traditional TCN decoder is the single receptive field. It also caused the extraction of the feature associations between motion sequences poorly. Hence, it still has a motivation to improve its performance. The improved MTCN owns a multi-scale temporal receptive field range, which can effectively realize the feature interaction between different temporal information and obtain more effective decoding information.

To further investigate its advantages, we conducted the experiments. Through the experiments, it can be seen that the MSTS-GCN model based on MTCN achieves good results in some of the metrics of the human motion prediction task, which indicates that the performance of the decoder is an important factor leading to the error, and the prediction error can be effectively reduced by improving the results of the decoder, but the MSTS-GCN model still needs to be improved in some performance metrics.

In the MSTS-GCN model, only on multi-scale time convolution feature interaction between frames is limited. with the deepening of network layers, the feature information after multiple extractions will cause inevitable losses, especially under the premise of long sequence information, the information loss is more obvious. In order to make the network retain more features in the long sequence information, this study proposed an NSTS-GCN fusion GRU decoder model.

C. NSTS-GCN MODEL WITH THE INTRODUCTION OF GRU-TCN DECODER

TCN has achieved excellent performance in sequence modeling tasks because they allow parallel computation, but TCN can only utilize nearest-neighbor sequence information, and the receptive field size is limited to capture relevant information of arbitrary length. GRU as a variant of RNN, can effectively correlate long-time sequence information, which makes up for the deficiency of RNNs that are prone to gradient disappearance or explosion and is a common structure for today's sequence modeling tasks.

The improved MTCN decoder demonstrates that a multi-scale temporal convolutional decoder could improve motion sequence prediction. Based on this experience, we combine the advantages of TCN and GRU to construct a GRU-TCN decoder that fuses RNN and temporal convolutional neural networks. Then, the proposed GRU-TCN decoder was applied to STS-GCN networks to construct the NSTS-GCN human motion prediction model further.

Gated Recurrent Unit (GRU) [31] can obtain the semantic correlation between long-term time series efficiently and suppress gradient dispersion or explosion phenomenon, which has a simple structure and low training difficulty. GRU has two inputs (the input of the current time step and the implied output of the previous time step), two outputs (the output of the current time step and the implied output passed to

the next time step), and a core structure consisting of two gating mechanisms: update gate and reset gate. The internal operation flow of basic GRU is shown in Figure 4 [31].

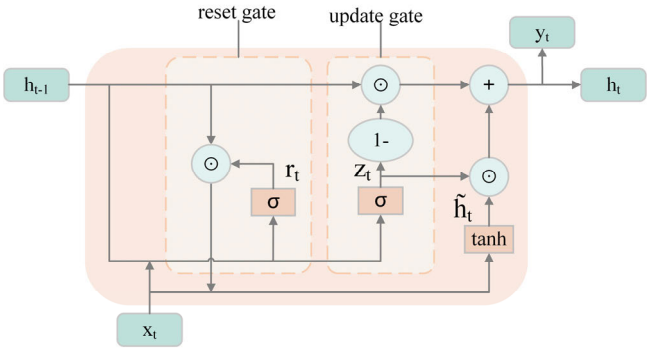


FIGURE 4. Flowchart of GRU [31].

The GRU can capture temporal correlations of varying scales in motion sequences, use update gates for forgetting and selecting memory information, and use reset gates to control the output information of the gated loop unit. The advantage of update and reset gates is the ability to always retain long-term sequence information without eliminating historical memory due to time change or forgetting useless information at the current stage.

Although LSTM can effectively capture the semantic association between long sequences and suppress the gradient disappearance or explosion phenomenon, its internal structure is relatively complex, and its training efficiency is low. To solve the above problem, the GRU structure is proposed. The GRU has the same effect as LSTM, but its structure and calculation are more straightforward and more accessible to train than the LSTM, which can significantly improve training efficiency.

When GRU is applied to human motion prediction, its input information contains the output motion sequence information h_{t-1} of the previous node and the input motion sequence information x_t of the current node. The GRU firstly splices x_t with h_{t-1} for different linear transformations and activates with the Sigmoid function to obtain the update gate value z_t and the reset gate value r_t. Then, the product calculation of r_t and h_{t-1} is carried out element by element, which controls the utilization of the implied motion sequence information h_{t-1} at the previous node. Next, the reset h_{t-1} is linearly transformed by splicing it with x_t and activated by the tanh function to scale the data to the range of -1∼1, to obtain new implicit motion sequence information $\tilde{h}_t$. Finally, the value z_t of the update gate is applied to $\tilde{h}_t$ and $1 - z_t$ is applied to the implied motion sequence information h_{t-1} at the previous time node, and the two results obtained by the update gate are summed to obtain the final output, which is the implied motion sequence information h_t transmitted to the next node. The whole process preserves the previous motion sequence information by the update gate z_t, and when the update gate value z_t tends to 1, the result without the motion sequence information of the previous node is output, and the implied motion sequence information h_t transmitted to the next node is only related to the input x_t, the implied motion sequence information h_{t-1} passed to the previous node is output when the update gate value z_t tends to 0. The calculation principle [31] of GRU can be expressed as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{1}$$

$$r_t = \sigma\,(W_r \cdot [h_{t-1}, x_t]) \tag{2}$$

$$\tilde{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{4}$$

where x_t is the input value of GRU; z_t is the update gate value, z_t has a value range of 0∼1, and its value near 1 means more data are retained, while near 0 means more data are discarded; W_z is the update gate weight matrix; r_t is the reset gate value; W_r is the reset gate weight matrix; σ is the Sigmoid activation function; h_{t-1} is the implied state of the previous time step; $\tilde{h}_t$ is the transition current node implied state; W is the weight matrix; h_t is the final output implied state of the current node; $\odot$ is the element-by-element product operation of two homotypic matrices.

The GRU enables the implied motion sequence information output by the human motion prediction model at the previous time step to be used as part of the input at the current time step, i.e., in addition to the normal input information, the input at the current time step also contains the implied motion sequence information of the previous time step, which positively influences the output at the current time step by using the forward motion sequence information.

The decoding process of human motion prediction can be regarded as a time series problem. When decoding motion sequence information, it is necessary to repeatedly utilize the motion sequence information obtained by encoding, and it is difficult to obtain sufficient motion sequence features by relying on temporal convolutional layers alone. TCN has the feature of parallel processing data with low memory occupation but is still inherently limited by the perceptual field size of convolutional networks, which only utilize the last module's output information, and cannot grasp and efficiently utilize the information related to longer distances. GRU as a method specifically made for processing time series data, has a powerful nonlinear fitting capability and can extract the data features of each output layer well through the cyclic mechanism, especially more effective for long-term time series. We introduce GRU based on the TCN decoder to build a GRU-TCN decoder, which can deeply utilize the information encoded by the STS-GCN encoder to obtain better human motion prediction results. The framework of the GRU-TCN decoder is shown in Figure 5.

After each TCN module extracts the corresponding time series information, the GRU selectively stores the motion features outputted by each TCN module. The third TCN module of the original decoder can only receive the output information of the second TCN module, but after introducing

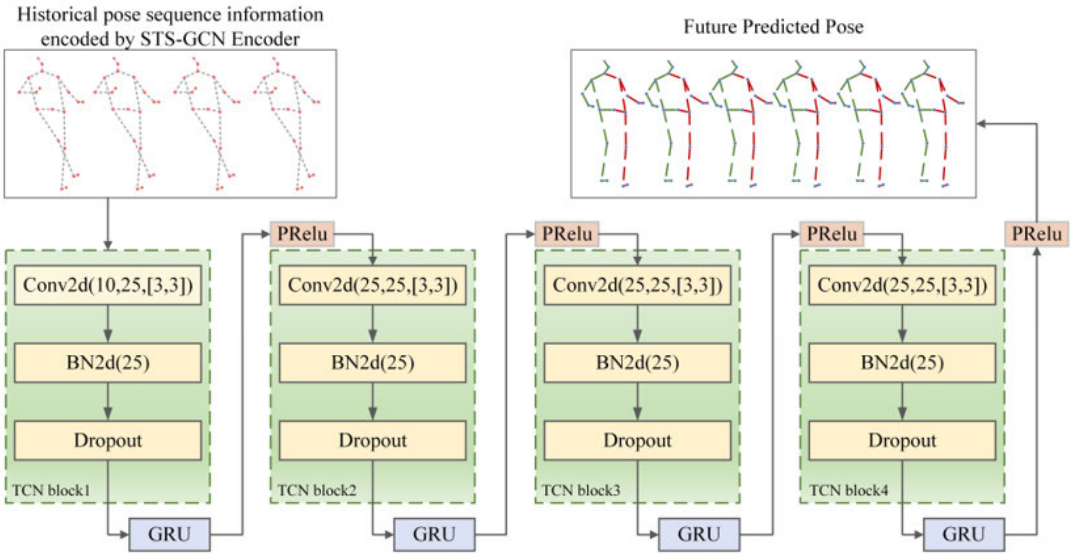


FIGURE 5. Framework of GRU-TCN decoder.

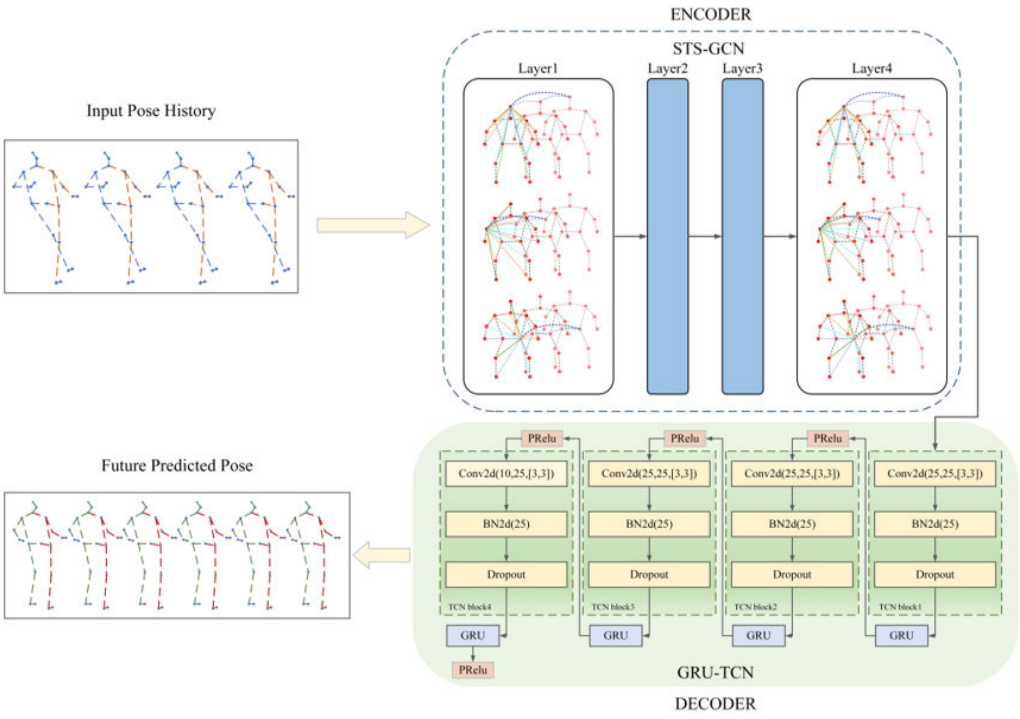


FIGURE 6. NSTS-GCN network flow diagram.

GRU, GRU can selectively store the output information of the first module as the implied state and use it as the input of the third module together with the output information of the second TCN module, to enrich the feature information received by the third module, and the fourth module and so on, thus enhance the decoder's ability to correlate long-range sequence information and better decode to obtain future motion sequences.

The GRU-TCN decoder was applied to the original STS-GCN network as a way to construct the NSTS-GCN human motion prediction model, and the flow diagram of NSTS-GCN is shown in Figure 6.

Figure 6 depicts the overall process of the proposed NSTS-GCN. The detailed steps are given as follows:

Step 1: A piece of skeletal sequence data from human movement is fed into the encoder NSTS-GCN.

Step 2: NSTS-GCN first skeleton encodes the input motion sequence through a 4-layer spatio-temporal separable encoder. The motion information of spatial dimension and temporal dimension is obtained, respectively. Furthermore, the obtained information is encoded into a feature vector.

Step 3: Based on the acquired feature vector, the new GRU-TCN decoder is used to infer the possible motion trend of the future skeleton, and the predicted motion sequence is obtained.

The traditional human motion prediction model STS-GCN has inherent training difficulties when dealing with human motion sequence data of a high dimensional and highly random nature. Therefore, the use of GRU allows further information propagation between different nodes for the encoded historical motion states of the encoder, retaining the current input motion sequence information at each step of the temporal convolutional neural network and adding new content to it, while forgetting all past states that no longer add additional information to the current state, thus reducing the quantization loss of short-term prediction and achieving reasonable NSTS-GCN facilitates both short-term and long-term human motion prediction by combining TCN and GRU to retain valid implicit motion information while processing motion sequence features in parallel.

III. EXPERIMENTS

To validate the performance of the proposed NSTS-GCN model, it was experimentally on three large-scale and challenging datasets, which were Human3.6M [34], AMASS [35], and 3DPW [36]. Considering the characteristics of the Human3.6M and AMASS datasets, each dataset has its own models. To further validate the generalization ability of the proposed model, the model trained by AMASS was used to predict human motions.

The graph encoder of the NSTS-GCN model retains the design of the STS-GCN model and consists of four layers of STS-GCN spatio-temporal separable graph convolution, each layer differing only in the number of channels C(l): the first layer from 3 to 64, the second from 64 to 32, the third from 32 to 64, and the last from 64 to 3. Each layer of the graph convolutional encoder adopts batch normalization and residual connection.

A. DATASET SETTING AND EVALUATION INDEX

1) HUMAN MOTION PREDICTION ON THE HUMAN3.6M DATASET

For the human motion prediction task, the Human3.6M dataset is divided into training, validation, and test sets.

Human3.6M is a huge human motion prediction dataset consisting of 3.6 million 3D human poses and corresponding images, in which eleven professional actors (six males and five females) perform 17 scenes of motions (e.g., walking, eating, etc.) from four different viewpoints in an indoor experimental environment, and only 15 of these motions were selected according to the STS-GCN model setup, with each actor's body skeleton being represented as 32 key points. Actors 1, 6, 7, 8, and 9 (S1, S6, S7, S8, S9) were used for training, actor 11 (S11) for validation, and actor 5 (S5) for testing. The optimal NSTS-GCN human motion prediction model trained on the training set is validated and tested on the validation and test sets to evaluate the model performance.

For prediction based on 3D joint coordinates, the 22 key points annotated with the dataset were selected according to the STS-GCN model, and the Mean Per Joint Position Error (MPJPE) proposed in Human3.6M was used as the loss function, and the Euclidean distance between each predicted 3D joint position and the real joint position is calculated, and the error is measured in millimeters (mm). The MPJPE is calculated as:

$$MPJPE = \frac{1}{V(T+K)} \sum_{k=1}^{T+K} \sum_{v=1}^{V} \| \hat{x}_{vk} - x_{vk} \|_2 \quad (5)$$

where V is the number of key points of the human skeleton; T is the number of observed frames; K is the number of predicted frames;$\hat{x}_{vk} \in \mathbb{R}^3$ is the predicted v-th joint coordinate of the k-th frame; $x_{vk} \in \mathbb{R}^3$ is the real v-th joint coordinate of the k-th frame. For prediction based on the angle representation, the 16 key points of the dataset labeled according to the STS-GCN model are selected, and the average L_1 distance between the predicted obtained joint angle and the real joint angle is used as the loss function, i.e., Mean Angle Error (MAE), and the measured angle error is measured in degrees (°). To facilitate the comparison of results with other studies, we also adopt this unit. The MAE error is calculated as follows:

$$MAE = \frac{1}{V(T+K)} \sum_{k=1}^{T+K} \sum_{v=1}^{V} |\hat{x}_{vk} - x_{vk}| \quad (6)$$

where $\hat{x}_{vk} \in \mathbb{R}^3$ is the angle of the v-th joint in the k-th frame predicted in the exponential mapping representation; $x_{vk} \in \mathbb{R}^3$ is the true value of the v-th joint angle in the k-th frame.

2) HUMAN MOTION PREDICTION ON THE AMASS DATASET

A total of eighteen existing motion capture datasets were collected in the AMASS dataset, and only thirteen of them were selected according to the original STS-GCN model. Eight of the thirteen datasets were used for training, four for validation, and one for testing (the dataset used for testing was named BMLrub). The AMASS dataset consisted of forty human subjects who performed walking movements, and the human body of each person's pose was represented by 52 joints, including 22 body joints and 30 hand joints, following the STS-GCN model focusing only on body joints, dropping four static joints, and predicting a human motion sequence containing 18 body joints.

3) HUMAN MOTION PREDICTION ON THE 3DPW DATASET

The 3DPW dataset consists of 60 video sequences captured by cell phone cameras with 51,000 frames, including indoor and outdoor activities. The generalizability of the NSTS-GCN model obtained by training from the AMASS dataset is tested using 3DPW.

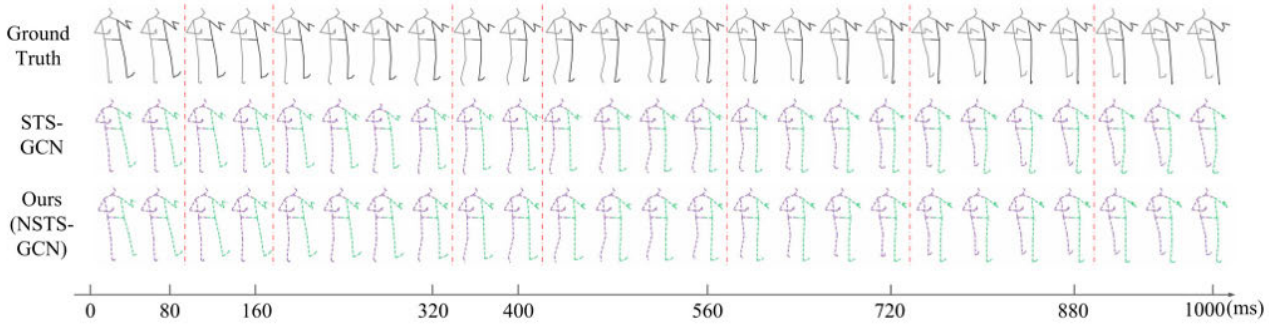


FIGURE 7. Comparison of visual results of eating motion prediction sequences.

TABLE 1. The number of frames in a motion sequence in relation to time.

human motion sequence predictive value correspondence							
short-term prediction				long-term prediction			
frame 2	frame 4	frame 8	frame 10	frame 14	frame 18	frame 22	frame 25
80ms	160ms	320ms	400ms	560ms	720ms	880ms	1000ms

TABLE 2. MPJPE index comparison between NSTS-GCN and STS-GCN models in eating motion.

msec	Eating MPJPE (mm)							
	80	*160*	*320*	*400*	*560*	*720*	*880*	*1000*
STS-GCN[30]	6.8	11.3	22.6	25.4	33.9	40.2	46.2	52.4
Ours (NSTS-GCN)	**6.0(-0.8)**	**10.3(-1.0)**	**18.5(-4.1)**	**22.7(-2.7)**	**30.9(-3.0)**	**38.7(-1.5)**	**45.2(-1.0)**	**48.5(-3.9)**

B. DATASET RESULTS

Quantitative and qualitative evaluation of the performance of the NSTS-GCN model and other advanced human motion prediction models on short-term predictions of less than 500 milliseconds (ms) and long-term predictions of greater than 500 milliseconds (ms). The comparison models include the ConvSeq2Seq [37] model that uses convolutional layers to encode long-term and short-term historical motion sequences separately; the LTD-X-Y [38] model using DCT to encode the frequency of video sequences (X denotes the number of observed frames and Y denotes the number of predicted frames); DCT-RNN-GCN [39] model that extends based on the LTD-X-Y with RNN and motion attention mechanism; BC-WGAIL-div [40] model with reinforcement learning; and STS-GCN, the base model adopted by NSTS-GCN, which uses temporally separable graph convolution for human motion prediction for the first time.

All algorithms use 10 frames of motion sequences as input for a total of 400ms, except for LTD, which uses multiple inputs, predicting 2∼10 frames (80∼400ms) of human motion sequences in the future is a short-term prediction, and predicting 14∼25 frames (560∼1000ms) of motion sequences in the future is a long-term prediction. For the sake of comparison, all algorithms use accepted standards for action prediction, The correspondence between the number of frames of motion sequences and time is shown in Table 1 [30].

The NSTS-GCN model is realized based on Pytorch1.7.1 deep learning framework. All the experiments use a single NVIDIA GeForce GTX 1050Ti graphics card, 4G video memory, CPU Intel (R) Core (TM) i5-2320 CPU @ 3.00GHz, and a Python version of 3.7.0. The model was optimized using the Adam optimizer for 50 epochs, the initial learning rate was set to 0.01, and the learning rate was reduced by 1/10 for every 5 epochs after the 20th epoch. The model uses 10 frames (400ms) of human motion sequences as observations to predict 25 frames (1000ms) of human motion sequences in the future.

1) HUMAN3.6M DATASET RESULTS

Table 2 and Figure 7 show the quantitative and qualitative prediction results of the eating motion in the Human3.6M test set, respectively. The quantitative evaluation gives the model's MPJPE joint coordinate prediction error at different time nodes (80ms, 160ms, 320ms, 400ms, 560ms, 720ms, 880ms, and 1000ms) for each motion comparison results, and qualitative evaluation gives the model prediction results comparison for motion sequences within 0∼1000ms.

As can be seen from Table 2, NSTS-GCN has reduced the joint coordinate error at each time node of motion prediction compared to STS-GCN, with the largest reduction of 4.1mm in MPJPE error at the 320ms and the smallest reduction

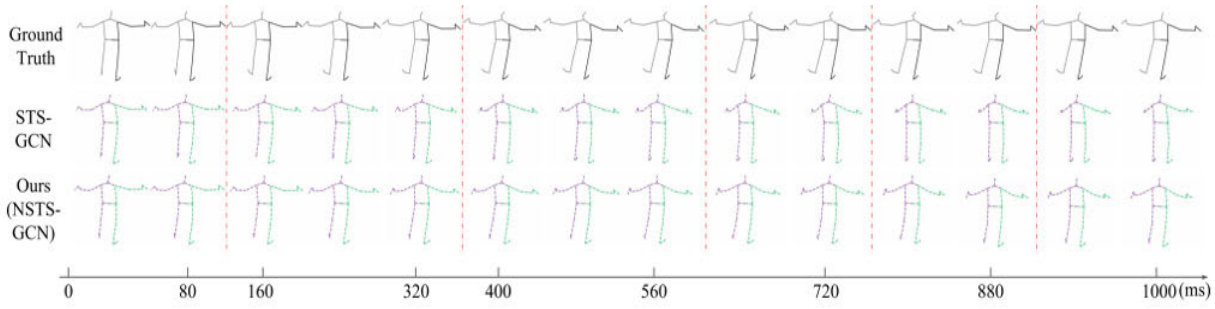


FIGURE 8. Comparison of visual results of posing motion prediction sequences.

TABLE 3. MPJPE index comparison between NSTS-GCN and STS-GCN models in posing motion.

msec	Posing MPJPE (mm)							
	80	*160*	*320*	*400*	*560*	*720*	*880*	*1000*
STS-GCN[30]	9.9	18.0	38.2	45.6	64.3	79.3	94.5	106.4
Ours (NSTS-GCN)	**8.2(-1.7)**	**15.4(-2.6)**	**31.2(-7.0)**	**39.3(-6.3)**	**55.4(-8.9)**	**69.7(-9.6)**	**82.3(-12.2)**	**89.4(-17.0)**

TABLE 4. Average MPJPE index comparison of human motion prediction model on Human3.6M test set.

msec	Average MPJPE (mm)							
	80	*160*	*320*	*400*	*560*	*720*	*880*	*1000*
ConvSeq2Seq[37]	16.6	33.3	61.4	72.7	90.7	104.7	116.7	124.2
LTD-50-25[38]	-	-	-	-	79.6	93.6	105.2	112.4
LTD-10-25[38]	-	-	-	-	79.5	94.0	105.6	112.7
LTD-10-10[38]	11.2	23.4	47.9	58.9	78.3	93.3	106.0	114.0
DCT-RNN-GCN[39]	10.4	22.6	47.1	58.3	77.3	91.8	104.1	112.1
STS-GCN[30]	10.1	17.1	33.1	38.3	50.8	60.1	68.9	75.6
Ours (MSTS-GCN)	11.9	18.4	31.6	37.8	47.8	57.6	66.2	72.5
Ours (NSTS-GCN)	**8.9**	**15.5**	**28.8**	**35.2**	**46.8**	**57.0**	**65.9**	**71.6**

of 0.8mm in MPJPE error at the 80ms, with an average error of 2.25mm at all moments.

The first row of motion sequences in Figure 7 shows the true human pose, the second row shows the predicted results of the STS-GCN model, and the third row shows the predicted results of the NSTS-GCN model. It can be seen that the coordinate positions predicted by NSTS-GCN and STS-GCN are almost the same before the 320ms, which are consistent with the real values of the joints, but between 320ms and 720ms, the leg position and arm position predicted by NSTS-GCN and STS-GCN gradually deviate from the real positions, and after the 720ms, compared with STS-GCN, the NSTS-GCN predicted the leg position and arm position closer to the real position than STS-GCN, which proved the validity of the model motion prediction.

Table 3 and Figure 8 show the quantitative and qualitative prediction results of the Human3.6M test focused on posing movements, respectively. The comparison of the results of the two networks illustrates the superiority of our proposed NSTS-GSN model.

From Table 3, it can be seen that the NSTS-GCN has a larger reduction in joint coordinate error compared to the STS-GCN model in both motion sequence predictions, especially in the long-term prediction phase, with the smallest reduction in MPJPE error of 1.7mm at the 80ms and the largest reduction in MPJPE error of 17.0mm at the 1000ms, with the average error at all moments of 8.2mm.

Due to the large lateral amplitude of the pose motion, to better present the prediction results, except for the 1st frame and the 25th frame, Figure 8 only lists the even frames of the motion sequence, i.e., between the 160ms and 320ms including the 5th, 6th, 7th and 8th frames, Figure 8 only gives the 6th and 8th frames, and so on for the remaining frames. From Figure 8, it can be seen that the prediction results of NSTS-GCN and STS-GCN models are approximately the same before the 160ms, and after the 160ms NSTS-GCN gradually shows better prediction results. For the arm position of the pose motion, NSTS-GCN can generate prediction results with smaller errors compared with STS-GCN.

Table 4 statistically shows the results of NSTS-GCN and MSTS-GCN human motion prediction models with other

TABLE 5. Average MAE index comparison of human motion prediction model on Human3.6M test set.

msec	Average MAE (°)							
	80	*160*	*320*	*400*	*560*	*720*	*880*	*1000*
LTD-10-25[38]	0.34	0.57	0.93	1.06	1.27	1.44	1.57	1.66
LTD-10-10[38]	0.32	0.55	0.91	1.04	1.26	1.44	1.59	1.68
BC-WGAIL-div[40]	0.31	0.57	0.90	1.02	1.23	-	-	1.65
DCT-RNN-GCN[39]	0.31	0.55	0.90	1.04	1.25	1.42	1.56	1.65
STS-GCN[30]	**0.24**	0.39	0.59	0.66	0.79	0.92	1.00	1.09
Ours (MSTS-GCN)	0.33	0.41	0.63	0.68	**0.78**	0.90	1.02	**1.05**
Ours (NSTS-GCN)	**0.24**	**0.38**	**0.57**	**0.65**	**0.78**	**0.87**	**0.98**	**1.05**

TABLE 6. Average MPJPE index comparison of human motion prediction model on BMLrub test set.

msec	Average MPJPE (mm)							
	80	*160*	*240*	*400*	*560*	*720*	*880*	*1000*
convSeq2Seq[37]	20.6	39.6	59.7	67.6	79.0	87.0	91.5	93.5
LTD-10-10[38]	10.3	19.3	36.6	44.6	61.5	75.9	86.2	91.2
LTD-10-25[38]	11.0	20.7	37.8	45.3	57.2	65.7	71.3	75.2
DCT-RNN-GCN[39]	11.3	20.7	35.7	42.0	51.7	58.6	63.4	67.2
STS-GCN[30]	10.0	12.5	21.8	**24.5**	**31.9**	38.1	42.7	45.5
Ours (MSTS-GCN)	10.1	12.5	21.2	24.7	31.4	37.1	41.5	44.6
Ours (NSTS-GCN)	**8.4**	**12.4**	**21.0**	24.9	**31.9**	**37.6**	**42.2**	**45.4**

advanced models on the Human3.6M test set, which shows the mean MPJPE comparison for 15 motions in the test set.

From Table 4, it can be learned that when tested on the Human3.6M test set, the MPJPE error of the MSTS-GCN model with the MTCN decoder decreased compared to both the STS-GCN and other human motion prediction models at prediction sequence lengths of 320ms (frame 8) and above and at the 80ms (frame 2) and 160ms (frame 4) of the short-term prediction, the prediction error increased compared to STS-GCN. The NSTS-GCN model with the GRU-TCN temporal convolution decoder achieves the lowest prediction error at all frames compared to the STS-GCN, MSTS-GCN, and other human motion prediction models. Compared with the STS-GCN model, the MPJPE errors at the 80ms (frame 2) to the 1000ms (frame 25) are reduced by 1.2mm, 1.6mm, 4.3mm, 3.1mm, 4.0mm, 3.1mm, 3.0mm, and 4.0mm, respectively, and the average errors at eight different moments are reduced by 3.0mm, which can be seen that the NSTS- GCN model performs better in long-term prediction.

A comparison of the MAE of the NSTS-GCN and MSTS-GCN human motion prediction models with other advanced models for 15 motions on the Human3.6M test set is shown in Table 5.

As can be seen from Table 5, the MAE metric on the Human3.6M test set, MSTS-GCN only has a slight advantage in the long-term prediction at greater than 500ms, and the short-term prediction error at less than 500ms has increased compared to the STS-GCN model.

The MAE of the NSTS-GCN model was reduced by 0.01°, 0.02°, 0.03°, 0.01°, 0.05°, 0.02°, and 0.04° for the rest of the prediction frames, except for the 80ms when the same error value was maintained, which proved the effectiveness of the NSTS-GCN model for human motion prediction.

2) AMASS DATASET RESULTS

The human motion prediction models were trained on the AMASS dataset and tested for performance on the BMLrub sub-dataset. Table 6 shows the comparison of short-term prediction and long-term prediction average MPJPE results based on 3D coordinates for different human motion prediction models on the BMLrub test set.

As can be learned from Table 6, when tested on the BML-rub, the MPJPE index of the MSTS-GCN model with the MTCN decoder is reduced compared to the STS-GCN model except for the 80ms, 160ms, and 400ms. The NSTS-GCN model with the GRU-TCN decoder increases the MPJPE by 0.4mm compared to the STS-GCN only at 400ms (frame 10), has the same error value at 560ms (frame 14), and decreases the MPJPE values at other frames by 1.6mm, 0.1mm, 0.8mm, 0.5mm, 0.5mm, 0.5mm, 0.1mm, and the average error at eight different moments was reduced by 0.4mm, proving that the NSTS-GCN model has good motion prediction performance. Although MSTS-GCN has slightly better long-term prediction than NSTS-GCN, it is known collectively that the NSTS-GCN model has more comprehensive human motion prediction performance.

3) 3DPW DATASET RESULTS

The best human motion prediction model obtained by training the AMASS dataset was tested on the 3DPW dataset to examine the generalization performance of the model Table 7

TABLE 7. Average MPJPE index comparison of human motion prediction model on 3DPW test set.

msec	Average MPJPE (mm)							
	80	*160*	*240*	*400*	*560*	*720*	*880*	*1000*
convSeq2Seq[37]	18.8	32.9	52.0	58.8	69.4	77.0	83.6	87.8
LTD-10-10[38]	12.0	22.0	38.9	46.2	59.1	69.1	76.5	81.1
LTD-10-25[38]	12.6	23.2	39.7	46.6	57.9	65.8	71.5	75.5
DCT-RNN-GCN[39]	12.6	23.1	39.0	45.4	56.0	63.6	69.7	73.7
STS-GCN[30]	**8.6**	**12.8**	**21.0**	**24.5**	**30.4**	**35.7**	**39.6**	**42.3**
Ours (MSTS-GCN)	9.6	13.4	21.6	24.6	30.6	35.7	39.6	42.5
Ours (NSTS-GCN)	8.9	13.3	21.4	24.8	30.9	35.8	40.0	42.6

shows the test results of the human motion prediction model on the 3DPW dataset.

As can be seen from Table 7, the MPJPE error metrics of the MSTS-GCN and NSTS-GCN models on the 3DPW test set are generally higher than those of the STS-GCN model, but compared with other advanced human motion prediction models, the NSTS-GCN model excels in the MPJPE error metrics corresponding to short-term prediction and long-term prediction, achieving a more desirable human motion prediction This indicates that the NSTS-GCN model is a good predictor of human movement. This indicates that the NSTS-GCN model has certain advantages in capturing motion coherence and long-term dependence, and can better model and predict the temporal evolution of human motion sequences. However, the overall prediction effect could be better than the original model. On the one hand, the data set includes indoor and outdoor environments. It is not only collected in a single laboratory environment, so the data set has higher requirements for the model's prediction performance. On the other hand, the model tested on this dataset is trained on the AMASS dataset, and the prediction has a specific difficulty. The model's generalization decreases after using GRU to optimize the original network.

In summary, the NSTS-GCN model with GRU-TCN decoder shows good performance in motion sequence prediction by testing on the Human3.6M test set, BMLrub test set, and 3DPW datasets, it is a good human motion prediction model.

IV. CONCLUSION

Due to the single convolution receptive field of the STS-GCN decoder, the feature association between different motion sequences cannot be well extracted, and the feature interaction during the extended timing information needs to be improved. To solve the problem, this study proposed the multi-scale time convolution decoder MTCN to obtain the motion sequence information of different time receptive fields and constructed the MSTS-GCN model. This model demonstrated that the decoder's performance has an essential impact on the human prediction results. Unfortunately, MSTS-GCN is also challenging to effectively obtain short-term motion sequence information. However, GRU can remember the long-term relevant information of the motion sequence. Hence combining the advantages of temporal convolution TCN and GRU, we design a GRU-TCN decoder, which captures richer motion sequence features, the NSTS-GCN human motion prediction model was constructed and tested on the Human3.6M, BMLrub, and 3DPW datasets. The experimental results on the Human3.6M test set showed that the NSTS-GCN model decreased on the MPJPE and MAE prediction index compared with the STS-GCN. The results of the BMLrub test set show that the NSTS-GCN only increases the MPJPE error at 400ms by 0.4mm, and the prediction errors at the rest of the time nodes are reduced. The results of the 3DPW test set show that the prediction error of NSTS-GCN increases slightly. By comparing the quantitative and qualitative results with other human motion prediction models, it is clear that NSTS-GCN is an effective human motion prediction model. NSTS-GCN adopts an encoder-decoder framework, which can utilize less effective action information and cannot achieve efficient human motion prediction. Subsequent consideration can be given to human motion prediction based on human intention, increasing the input motion prediction model to further improve human motion prediction performance. Despite the advantages of NSTS-GCN, several disadvantages exist when using this model to predict human motion. Due to the use of GRU to extract associations between longer motion sequences, the generalization of the model has decreased. Moreover, the generalization performance of the new model still has the motivation to further improve.

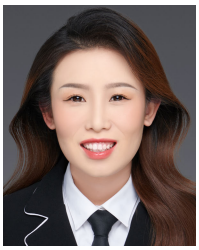
REFERENCES

[1] V. V. Unhelkar, P. A. Lasota, Q. Tyroller, R.-D. Buhai, L. Marceau, B. Deml, and J. A. Shah, "Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2394–2401, Jul. 2018, doi: 10.1109/LRA.2018.2812906.

[2] P. Schydlo, M. Rakovic, L. Jamone, and J. Santos-Victor, "Anticipation in human–robot cooperation: A recurrent neural network approach for multiple action sequences prediction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Brisbane, QLD, Australia, Jul. 2018, pp. 5909–5914, doi: 10.1109/ICRA.2018.8460924.

[3] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human–robot collaboration," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Christchurch, New Zealand, Mar. 2016, pp. 83–90, doi: 10.1109/HRI.2016.7451737.

[4] H. Admoni and S. Srinivasa, "Predicting user intent through eye gaze for shared autonomy," in *Proc. AAAI Fall Symp. Ser.*, 2016, pp. 298–303.
[5] T. Yonezawa, H. Yamazoe, A. Utsumi, and S. Abe, "Attractive, informative, and communicative robot system on guide plate as an attendant with awareness of user's gaze," *J. Paladyn Behav. Robot.*, vol. 4, no. 2, pp. 113–122, Jan. 2013.
[6] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Stockholm, Sweden, May 2016, pp. 3118–3125, doi: 10.1109/ICRA.2016.7487478.
[7] D. Pavllo, D. Grangier, and M. Auli, "QuaterNet: A quaternion-based recurrent model for human motion," 2018, *arXiv:1805.06485*.
[8] Y. Wang, X. Wang, P. Jiang, and F. Wang, "RNN-based human motion prediction via differential sequence representation," in *Proc. IEEE 6th Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Singapore, Jul. 2019, pp. 138–143, doi: 10.1109/CCIS48116.2019.9073734.
[9] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12116–12125.
[10] S. B. Abdullahi and K. Chamnongthai, "American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach," *IEEE Access*, vol. 10, pp. 15911–15923, 2022, doi: 10.1109/ACCESS.2022.3148132.
[11] S. B. Abdullahi and K. Chamnongthai, "American sign language words recognition of skeletal videos using processed video driven multi-stacked deep LSTM," *Sensors*, vol. 22, p. 1406, Feb. 2022, doi: 10.3390/s22041406.
[12] A. Hussain, S. U. Khan, N. Khan, I. Rida, M. Alharbi, and S. W. Baik, "Low-light aware framework for human activity recognition via optimized dual stream parallel network," *Alexandria Eng. J.*, vol. 74, pp. 569–583, Jul. 2023.
[13] A. Hussain, K. Muhammad, H. Ullah, A. Ullah, A. S. Imran, M. Y. Lee, S. Rho, and M. Sajjad, "Anomaly based camera prioritization in large scale surveillance networks," *Comput., Mater. Continua*, vol. 70, no. 2, pp. 2171–2190, 2022.
[14] A. Hussain, T. Hussain, W. Ullah, and S. W. Baik, "Vision transformer and deep sequence learning for human activity recognition in surveillance videos," *Comput. Intell. Neurosci.*, vol. 2022, Apr. 2022, Art. no. 3454167, doi: 10.1155/2022/345416.
[15] F. U. M. Ullah, M. S. Obaidat, A. Ullah, K. Muhammad, M. Hijji, and S. W. Baik, "A comprehensive review on vision-based violence detection in surveillance videos," *ACM Comput. Surveys*, vol. 55, no. 10, pp. 1–44, Oct. 2023, doi: 10.1145/3561971.
[16] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–9.
[17] W. Peng, J. Shi, and G. Zhao, "Spatial temporal graph deconvolutional network for skeleton-based human action recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 244–248, 2021, doi: 10.1109/LSP.2021.3049691.
[18] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1113–1122.
[19] Q. Cui, H. Sun, and F. Yang, "Learning dynamic relationships for 3D human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 6518–6526, doi: 10.1109/CVPR42600.2020.00655.
[20] Q. Cui and H. Sun, "Towards accurate 3D human motion prediction from incomplete observations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 4799–4808, doi: 10.1109/CVPR46437.2021.00477.
[21] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 211–220, doi: 10.1109/CVPR42600.2020.00029.
[22] H. Zhou, C. Guo, H. Zhang, and Y. Wang, "Learning multiscale correlations for human motion prediction," in *Proc. IEEE Int. Conf. Develop. Learn. (ICDL)*, Beijing, China, Aug. 2021, pp. 1–7, doi: 10.1109/ICDL49984.2021.9515609.
[23] Q. Li, G. Chalvatzaki, J. Peters, and Y. Wang, "Directed acyclic graph neural network for human motion prediction," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Xi'an, China, Jul. 2021, pp. 3197–3204, doi: 10.1109/ICRA48506.2021.9561540.
[24] Q. Zhang, T. Wang, H.-N. Wu, M. Li, J. Zhu, and H. Snoussi, "Human action prediction based on skeleton data," in *Proc. 39th Chin. Control Conf. (CCC)*, Shenyang, China, 2020, pp. 6608–6612, doi: 10.23919/CCC50068.2020.9189122.
[25] C. Yujun, H. Lin, W. Yiwei, C. Tat-Jen, C. Jianfei, and Y. Junsong, "Learning progressive joint propagation for human motion prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 226–242.
[26] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to MLP: A simple baseline for human motion prediction," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2023, pp. 4798–4808, doi: 10.1109/WACV56688.2023.00479.
[27] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A spatio-temporal transformer for 3D human motion prediction," in *Proc. Int. Conf. 3D Vis. (3DV)*, London, U.K., Dec. 2021, pp. 565–574, doi: 10.1109/3DV53792.2021.00066.
[28] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 11447–11456, doi: 10.1109/ICCV48922.2021.01127.
[29] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Multiscale spatio-temporal graph neural networks for 3D skeleton-based motion prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 7760–7775, 2021, doi: 10.1109/TIP.2021.3108708.
[30] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso, "Space-time-separable graph convolutional network for pose forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 11189–11198, doi: 10.1109/ICCV48922.2021.01102.
[31] A. A. Ballakur and A. Arya, "Empirical evaluation of gated recurrent neural network architectures in aviation delay prediction," in *Proc. 5th Int. Conf. Comput., Commun. Secur. (ICCCS)*, Patna, India, Oct. 2020, pp. 1–7, doi: 10.1109/ICCCS49678.2020.9276855.
[32] A. Sampieri, G. M. D. A. di Melendugno, A. Avogaro, F. Cunico, F. Setti, G. Skenderi, M. Cristani, and F. Galasso, "Pose forecasting in industrial human–robot collaboration," in *Proc. Comput. Vis. ECCV 17th Eur. Conf.*, Tel Aviv, Israel, Oct. 2022, pp. 51–69.
[33] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 6427–6436, doi: 10.1109/CVPR52688.2022.00633.
[34] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014, doi: 10.1109/TPAMI.2013.248.
[35] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 5441–5450, doi: 10.1109/ICCV.2019.00554.
[36] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pos-Moll, "Recovering accurate 3D human pose in the wild using IMUS and a moving camera," in *Proc. Eur. Conf. comput. Vis. (ECCV)*, 2018, pp. 601–617.
[37] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5226–5234, doi: 10.1109/CVPR.2018.00548.
[38] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 9488–9496, doi: 10.1109/ICCV.2019.00958.
[39] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Proc. Comput. Vis.-ECCV 16th Eur. Conf.*, Glasgow, U.K.: Springer, Aug. 2020, pp. 474–489.
[40] B. Wang, E. Adeli, H.-K. Chiu, D.-A. Huang, and J. C. Niebles, "Imitation learning for human pose prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Jul. 2019, pp. 7123–7132, doi: 10.1109/ICCV.2019.00722.



RUI LI received the M.S. and Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, in 2013 and 2019, respectively. Currently, she is with the Xi'an University of technology. Her research interests include intelligent robotics, brain–computer interface, EEG decoding method, and brain control prosthesis.

DUO HE received the B.Eng. degree in mechanical design, manufacturing and automation major from the Xi'an University of Technology, in 2020, where she is currently pursuing the M.S. degree in mechanical engineering. Her current research interests include image recognition, behavior recognition, and behavior prediction.

SHIQIANG YANG received the Ph.D. degree in mechanical engineering from the Xi'an University of Technology, in 2010. From 2005 to 2018, he was with the Xi'an University of Technology, where he has been an Associate Professor with the School of Mechanical and Precision Instrument Engineering, since 2009. His current research interests include intelligent robot control, image recognition, behavior detection, and recognition.

AN YAN received the B.Eng. degree in mechanical design, manufacturing and automation from Henan Polytechnic University, in 2020. He is currently pursuing the M.S. degree in mechanical engineering with the Xi'an University of Technology. His research interests include human motion detection and recognition, and human motion prediction.

XIN ZENG received the B.Eng. degree in mechanical design, manufacturing and automation from the Southwest University of Science and Technology, in 2021. She is currently pursuing the M.S. degree in mechanical engineering with the Xi'an University of Technology. Her current research interests include image recognition, behavior detection, and recognition.

DEXIN LI received the Ph.D. degree in mechanics from Xi'an Jiaotong University, Xi'an, China, in 2003. From 1992 to 2019, he was with the Xi'an University of Technology, Xi'an, where he has been an Associate Professor with the School of Mechanical and Precision Instrument Engineering, since 2002. His current research interests include mechanical computer aided design and manufacturing.