3D Skeleton-Based Human Motion Prediction Using Dynamic Multi-Scale Spatiotemporal Graph Recurrent Neural Networks

Mayank Lovanshi, *Student Member, IEEE*, Vivek Tiwari, *Member, IEEE*, and Swati Jain, *Member, IEEE*

***Abstract*—A dynamic multi-scale spatiotemporal graph recurrent neural network (DMST-GRNN) model has been introduced, which is leveraged to use human motion prediction on a 3D skeleton-based human activity dataset. It offers a multi-scale approach to spatial & temporal graphs using multi-scale graph convolution units (MGCUs) to describe the human body's semantic interconnection. The proposed DMST-GRNN is an encoder-decoder framework where a series of MGCUs are used as encoders to learn spatiotemporal features and a novel graph-gated recurrent unit lite (GGRU-L) for the decoder to predict human pose. Extensive experiments have been carried out with two datasets, Human3.6M and CMU Mocap, where both short and long videos were considered to validate the performance of the proposed model. The DMST-GRNN model outperforms the existing baseline on the Human3.6M datasets by 11.95% and 7.74% of average mean angle errors (avg MAE) for short and long-term motion prediction, respectively. Similarly, CMU Mocap datasets, the DMST-GRNN model predicts future posture more accurately than the previous best approaches by 2.77% and 5.51% of average mean angle errors (avg MAE) for short and long-term motion prediction, respectively. A comparison analysis was also presented with other measures like mean angle error, prediction loss and standard deviation. A separate discussion has been included to analyze the effect of different multiscale on spatial and temporal graphs, along with the impact of MGCU unit counts.**

***Index Terms*—CMU Mocap, DMST-GRNN, graph GRU lite, human3.6M, human activity, multi-scale, skeleton, spatiotemporal.**

I. Introduction

NOWADAYS human body 3D skeleton is used to estimate the pose of the human body based on the previous activity. Human motion prediction is becoming increasingly popular since it enables computers to comprehend human activity. Such a strategy (skeleton-based human motion prediction) might be used in various computer vision and robotics applications, including human-computer interaction, autonomous driving, and pedestrian tracking [1], [2]. Real-world human motion prediction is challenging since the movement is unpredictable, flexible, and non-linear in nature. Many approaches have been put forward to address this problem, such as the traditional state-based and deep-network approaches, which provide group posture features to learn the motion sequence [3]. Although these approaches came up with satisfactory results and tried to handle inherent challenges, studying spatial or temporal relationships among the human body joints in this domain is still an opportunity. The temporal relationships represent inter-frame interaction between displaying the continuous movements, while the spatial relations record the underlying pose. Spatial features are to be calculated by determining the spatial deviation of each frame of the activity sequence. The temporal feature is associated with or changes over time. So, it seems promising to estimate the temporal change of each frame of the activity sequence.

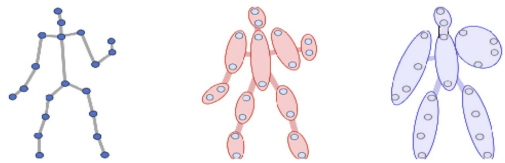

Fig. 1. Three body scales on H3.6M with 20, 10, and 5 scales, respectively.

Traditional approaches are a single-scale model that fails to represent a functional set of joints or high-order connections. Discussing human motion, like running, involves the large limbs moving together, while other acts, like smoking, include the wrist moving very little or the elbow could result differently in future poses. This is attracted to depict human motions by its scalable attention and then multi-scale correlations used for human motion prediction. Multi-scale has a huge advantage over solving problems with important features at multiple scales of time or space. Multi-scale shows the various scale of body joints depending upon the joints/key points value of the human body. Fig. 1 depicts the multi-scale body parts of the H3.6M dataset. In this view, a spatial network comes up as an appealing solution which covers each frame's body joints and internal interactions [4].

In this research study, a multi-scale spatiotemporal graph recurrent neural network (DMST-GRNN) has been introduced. The fundamental principle of the DMST-GRNN is to

Manuscript received 18 January 2023; revised 8 June 2023 and 6 July 2023; accepted 8 August 2023. *(Corresponding author: Vivek Tiwari.)*

Mayank Lovanshi is with the Department of Computer Science & Engineering, International Institute of Information Technology, Naya Raipur 493661, India (e-mail: mayank@iiitnr.edu.in).

Vivek Tiwari is with the Department of Computer Science & Engineering, ABV-Indian Institute of Information Technology & Management (ABV-IIITM), Gwalior 474015, India (e-mail: viveknitbpl@gmail.com).

Swati Jain is with the Govt. J. Yoganandam Chhattisgarh College, Raipur 492001, India (e-mail: sjcscghed@gmail.com).

Digital Object Identifier 10.1109/TETCI.2023.3318985

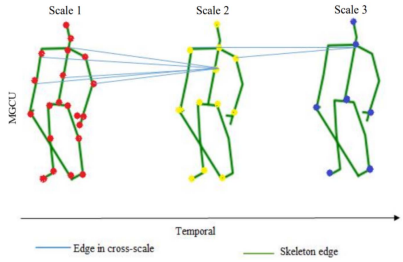

Fig. 2. Two multi-scale graphs temporal variation from one scale to another scale, capturing local and distinct relations.

use multi-scale spatial and temporal graphs to learn the semantic interconnection of a human body for kinematic feature extraction. The spatial & temporal representation of the skeleton image. The multi-scale graph convolution unit (MGCU) is developed to extract spatial and temporal features at each scale, i.e. discussed in Fig. 2. A detailed discussion has been made in Section IV. However, the major contributions of the article are as follows:

- The DMST-GRNN model introduces an encoder-decoder architecture, where the encoder incorporates MGCU and the decoder incorporates GGRU-L as crucial components. A decoder with a GGRU-L unit has been employed for the first time for human motion prediction.
- The proposed DMST-GRNN model employs a minimum of four MGCU units in the encoder, ensuring fewer parameters in the decoder without compromising overall performance.
- The encoder of the proposed model utilises various numbers of multiscale graph convolution units (MGCU) to effectively capture the spatio-temporal feature relationship of human motion, resulting in optimal performance description.
- Two large datasets, i.e. Human3.6M [5], [6] & CMU Mocap [7], [8] employed to validate proposed DMST-GRNN model for human activity prediction.
- The model is validated with various measures like mean angle error (MAE), average MAE and prediction loss.
- A separate discussion is included to analyse the effect of different multiscale graphs on spatial and temporal aspects and the impact of MGCU unit counts.

The remaining article is organized as follows: Section II discussed a brief review of related work for the graph's human activity prediction and visualization learning. Section III presents a methodology and the statistical underpinnings of our approach. Section IV illustrates a detailed study of the DMST-GRNN model: a multi-scale spatiotemporal graph computational unit. Section V is about the datasets used for the experiment. Finally, the model's experiment results & validations are presented and followed by the conclusion in Sections VI and VII, respectively.

II. Related Work

Nowadays, visual information-based human motion prediction is a hot topic among researchers. 3D human motion prediction comprises estimating future postures from a previous movement is a problem in computer vision and artificial intelligence, which can aid computers in understanding human actions. It aims to develop temporal representations that can perform short and long-term human motion prediction tasks. In recent research, deep recurrent neural networks (RNNs) have been utilised to simulate motion [9]. Here, the author focuses on the issues like the apparent discrepancy between the actual training result and the first predicted frame during the continuous motion of a human body. Therefore, a CNN-based sequence-to-sequence (seq2seq) architecture [10] was employed to resolve this issue. It utilises a fixed-size context window, which proves inadequate for capturing the short-term dynamics of human body joint motion. Given human joint motion's highly dynamic and variable nature, accurately capturing these motions within a short time interval presents a challenge for any CNN-based model [11]. It is important to note that CNN treats each input independently and does not retain information from previous inputs or values, resulting in significant memory loss [12]. Furthermore, the model presented in [10] requires actual training data to effectively learn complex patterns in the input sequence, which was unavailable in their research. In conclusion, the model must capture spatial-temporal information in short-term intervals, leading to poor performance in short-term prediction tasks.

A hypothesis discussed in [13] proposed a motion context modelling by summarising the historical human motion concerning the current prediction. A modified highway unit (MHU) [13] is proposed to eliminate motionless joints and estimate the next pose, given the motion context. In this research, many activities are highly uncertain with different activity classes because of the Restricted Boltzmann Machines (RBMs) limitation. Therefore, the observed information cannot provide enough evidence for modelling and predicting. [14] suggested replacing it with Long short-term memory (LSTM), Gated recurrent unit (GRU) & Bi-LSTM [14] based encoder-recurrent-decoder network. A non-linear transformation has been employed to encode the posture feature and decode the LSTM output. The results claim it's a better performance in understanding the temporal dynamics of human motion. RNN-based models like LSTM and GRU struggle with learning long-term sequences and capturing complex human behaviour [15]. They treat all inputs equally and fail to determine the significance within a context. Additionally, these models have difficulty predicting future movements based on short-term data and capturing non-linear patterns in human motion [13]. This limits their accuracy in identifying short-term and long-term motion patterns.. Their article doesn't validate their experimental result on the large state-of-the-art dataset. Therefore, the result of models is not capable of identifying short-term & long-term motion prediction.

Deep learning research aims to generalize graph neural networks (GNN) for motion prediction. In research work [16], Graph convolution networks (GCNs) were introduced, combining CNNs with GNN. GCN is classified into temporal and spatial perspectives, addressing locality and incorporating spectrum analysis. Then ST-GCN model is introduced in [17], [18] while applying CNN filters to the spatial domain on 1-neighbour nodes. The input data scale remains constant across joints due to

weight sharing. However, the ST-GCN model couldn't validate results on state-of-the-art datasets, excluding the discussion on short and long-term human motion prediction.

A research article [19] discussed a recurrent neural network with skeleton joint co-attention that captures spatial coherence and temporal evolution in spatio-temporal space. However, it struggles with generalizing to new or different motion patterns. In [20], a self-renewing convGRU architecture predicts skeleton motion by capturing temporal and spatial dependencies. The model incorporates a long-term semantic vector to improve the accuracy of generated motion sequences, but it may face challenges in capturing long-term dependencies in complex motions.

The study [21] introduced a hypothesis that not all body joints are equally crucial for activity detection. Using recurrent temporal encoding, an end-to-end action-attending network can describe irregular skeletons as undirected attribute graphs. The model is powerful enough to adaptively predict separate incoherent action units for various activities. The effectiveness of the author's model was assessed using medium and large-scale datasets. Different joints contribute to human activity recognition for the skeleton-based system in different ways. The human posture couldn't be scaled up excessively in this model because it only represented a single spatial and temporal scale in a network.

The article in [4] proposed dynamic multi-scale graphs to represent the human body and introduce dynamic multi-scale graph neural networks (DMGNN) with an encoder-decoder architecture for human motion prediction on 3D skeleton-based data. It develops a multi-scale graph computational unit (MGCU) in the encoder for feature extraction and a graph-based GRU (G-GRU) in the decoder for future posture generation. A multi-scale network with nodes representing body parts at different scales and edges representing pairwise relationships between body parts is addressed in [4]. Human motion prediction is a sequence-based method that needs an attention model, a limitation of this research.

Therefore, a hypothesis [1] discussed human motion prediction on 3D skeleton data with an attention model. This article introduces a multi-scale spatiotemporal graph neural network (MST-GNN) with an encoder-decoder architecture. It also creates multi-scale spatiotemporal graphs representing the performance of different human body movements. Multi-scale spatiotemporal graph computational units (MST-GCU) develop features in the encoder, while graph attention-based GRU (GA-GRU) generates poses in the decoder. A multi-scale spatiotemporal graph represents the spatiotemporal relationships among human motion based on 3D skeleton data. There are three distinct intuitions. The first is that while moving, human bodies are subject to a few limitations; the second is that locations obtained at several timestamps are inertial and correlated; the third is that many motions include a group's joints at temporal intervals. These three findings point to the requirement for a multi-scale spatiotemporal graph to be developed to analyse the dynamics of human activity. This research created an issue in updating weights for spatiotemporal modelling because of the manual modification in weight.

In light of various literature surveys, several limitations have been identified that can be addressed by employing the proposed model. These limitations include the underutilisation of spatio-temporal information, the complexity of the model used to learn spatio-temporal features, limited opportunities for learning more comprehensive features, and a weaker ability to handle long-term context dependencies. To overcome these shortcomings, the proposed model utilises an encoder-decoder architecture. The encoder facilitates the effective extraction of spatial-temporal features, while the decoder enables accurate human motion prediction with fewer parameters compared to state-of-the-art models. The proposed encoder-decoder model has demonstrated its ability to learn non-linear short and long-term sequences.

III. Problem Foundation & Mathematical Formulation

This section helps to understand the mathematics behind predicting human motion pose based on historical motion. 3D skeleton-based human action prediction aims to create a sequence of observation-generated future postures. The body joints of the human body help to estimate human motion, i.e. mathematically described here, $P^t \in R^{J*3}$ = posture matrix that logs J body joint's three-dimensional (3D) positions at time t. $P = [P^1, P^2 \ldots\ldots P^N] \in R^{N*J*3}$ = Pose matrices are concatenated into a three-mode tensor using a series of N timestamps. $P^{[t,i,c]} = i$th body joint's cth coordinate value at timestamp t.

Therefore, $P^{Hist} = [P^{(-N+1)}, \ldots\ldots\ldots P^{(0)}] \in R^{N*J*3}$ is tensor that represents N historical poses, $P^{Fut} = [P^{(1)}, \ldots\ldots\ldots P^{(N)}] \in R^{N*J*3}$ is tensor that represents N future poses. So $P^{Fut} = F_{pred}(P^{hist})$. To approach the ground truth P^{Fut} in motion prediction, a trainable predictor F_{pred} is presented, which creates a series of predicted poses $P^{Fut} = F_{pred}(P^{hist})$.

A. Multi-Scale Spatial & Temporal Graph

A multi-scale spatiotemporal network is based on a skeleton representing human motion's spatiotemporal relationships. There are three different intuitions; first, human bodies are regularised by a few spatial restrictions while moving; second, positions taken at different timestamps are inertial and correlated; third, many motions include a functional set of joints and some temporal segments. These three findings suggest the potential introduction of a multi-scale spatiotemporal graph to study the dynamics of human activities.

A spatiotemporal graph at the first joint scale should be presented initially to define the multi-scale graph. $G_0(V_0, E_0, A_0)$ is a spatiotemporal graph that simulates the inter-joint relations, where $A_0 \in R^{(NJ)(NJ)}$ is the graph adjacency matrix, E_0 is the edge set carrying the spatiotemporal relations, and V_0 is the vertex set with $|V_0| = NJ$ joints. Then rearrange P and combine the first two dimensions to form a posture matrix supported on this spatiotemporal graph, $P \in R^{(NJ)*3}$ [1], [22].

A multi-scale spatiotemporal graph consists of $R+1$ layered graphs based on the spatiotemporal graph at the joint scale. We abstract additional R graphs, in addition to the initial as $G_0(V_0, E_0, A_0)$, $G_1(V_1, E_1, A_1), \ldots G_R(V_R, E_R, A_R)$

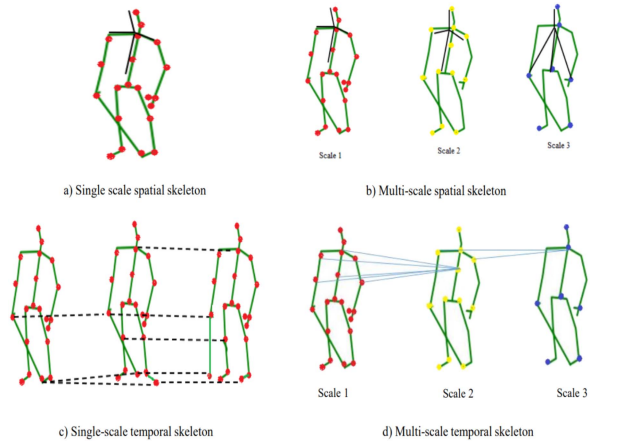

Fig. 3. (a) Single-scale spatial skeleton. (b) Multi-scale spatial skeleton. (c) Single-scale temporal skeleton. (d) Multi-scale temporal skeleton.

includes $N_1 J_1, \ldots, N_R J_R$ vertices. In the rth scale, every vertex in V_r represents a collection of body joints.

Initially, two challenges were identified for building the proposed DMST-GRNN structure: 1) Fixed spatiotemporal graphs cannot accommodate various human motions because of implicit and action-related limits on bodies. 2) The links between space and time could produce a vast, difficult graph to store and interpret in real-time prediction. The first issue may be handled using trainable multi-scale spatiotemporal graphs, which capture flexible connections in both the spatial and temporal domains. In this regard, the graph adjacency matrix at every scale is modified, which best matches the implicit associations in movements. The second issue may be solved by assuming that the spatiotemporal graph is decomposable. The hybrid spatial-temporal graph was converted into a spatial and temporal graph. The adjacency matrix A_r is the Cartesian product of the temporal and spatial graphs: $A_r = S_r \otimes T_r$, where A_r is the Cartesian product of $S_r \in R^{J_r * J_r}$, and $T_r \in R^{T_r * T_r}$ are the adjacency matrices of spatial and temporal graphs. In comparison, T_r indicates the temporal dependencies across time at a joint, and S_r displays the spatial relationships between joints at a frame. Fig. 3 discussed the single-scale & multi-scale spatial skeleton and single-scale & multi-scale temporal skeleton.

B. Spatial-Temporal Graph Convolution

A spatiotemporal graph convolution obtains features from a spatial-temporal network at a different scale. So a spatiotemporal graph $A = S \otimes T \in R^{(NJ)*(NJ)}$ at a single scale and $M \in R^{(NJ)*D}$ is a spatiotemporal matrix. Furthermore, (1) represents spatiotemporal graph convolution (M'), which is being generated by convolving Graph filter (G) with input M, spatiotemporal graph A.

$$M' = G *_A M = G *_{S \otimes T} M \tag{1}$$

where, M= Spatiotemporal matrix, M' = Spatiotemporal graph convolution, G = Graph filter fuse the input M on the spatiotemporal graph A.

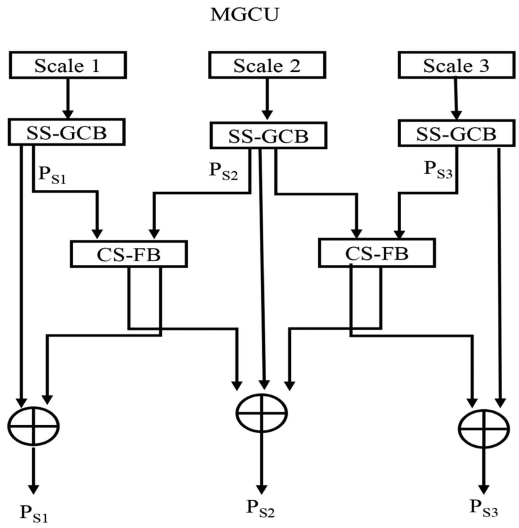

Fig. 4. MGCU: single-scale graph convolution blocks (SS-GCB) and cross-scale fusion blocks(CS-FB).

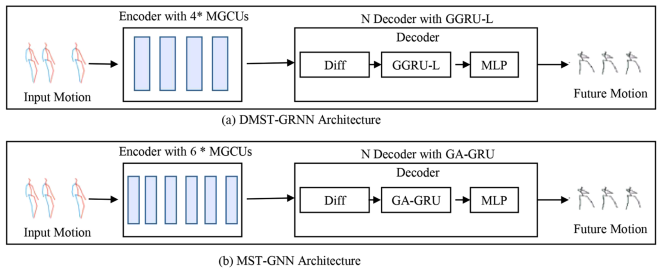

Fig. 5. Architectural difference between (a) The proposed DMST-GRNN model; (b) the existing MST-GNN model.

IV. DMST-GRNN Framework

In this research, a novel dynamic multi-scale spatiotemporal graph recurrent neural network (DMST-GRNN) was proposed on the multi-scale spatiotemporal graph to estimate future 3D postures through activity class [3]. Primarily, it consists of two phases, an encoder and a decoder. Furthermore, the encoder offers two components: single-scale graph convolution block (SS-GCB) and cross-scale fusion block (CS-FB), as depicted in Fig. 4. At the same time, Graph Gated Recurrent Unit Lite (GGRU-L) serves as the decoder's core part. The DMST-GRNN is an encoder-decoder architecture in which the encoder captures prominent history information, and the decoder is used to predict future poses. The key component of the encoder is a multi-scale graph computational unit (MGCU) [1] that estimates movement characteristics passed to the multi-scale spatiotemporal graph. Fig. 5 depicts the architectural differences that include the number of MCGU units utilized (4 in DMST-GRNN and 6 in MST-GNN) and the type of GRU used in the decoder (GGRU-L in DMST-GRNN and GA-GRU in MST-GNN). So following are the different modules involved in DMST-GRNN.

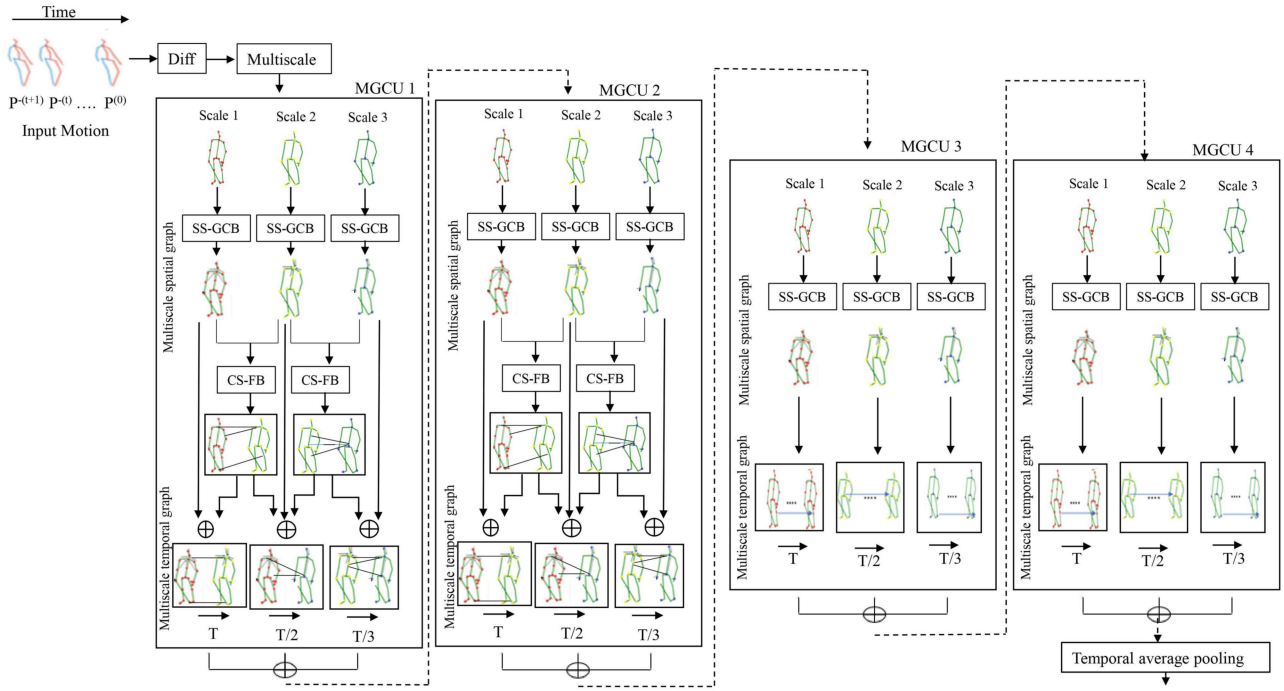

Fig. 6. Encoder framework of DMST-GRNN model. In the encoder, 4 MGCU is used for extracting spatiotemporal features.

A. *Encoder*

The encoder attempts to give relevant motion states by extracting semantics from observed motions. As Encode and decoder work in sequence, the encoder's output becomes an input for a decoder. An Encoder is a combination of various MGCUs. Each MGCU fuses the multi-scale graph creation feature with a single-scale graph convolution block and cross-scale fusion block. Fig. 6 discussed the complete architecture of the encoder part of DMST-GRNN.

In MGCU, concatenate the motion sample of each scale, i.e. scale 1, scale 2 & scale 3, and extract spatiotemporal characteristics using a cascade of Multi-scale graph convolutional units (MGCUs). The multi-scale network of each MGCU is trained individually. As a result, the topology of a graph may change dynamically between MGCUs. The output characteristics are weighted and summed to integrate the three scales for comprehensive semantics. Then broadcast the lower body components to coincide with their spatially relevant joints since the number of body components varies between sizes. The three scale's broadcast output includes H_{s1}, H_{s2}, $H_{s3} \in R^{T' \times J \times D_h}$; so the summed feature is represented in (2).

$$H = H_{s1} + \lambda(H_{s2} + H_{s3}) \tag{2}$$

where, λ = hyper-parameter to balance different scales. Equation (2) helps to find an average temporal pooling to remove the time dimension of H and obtain $H \in R^{J \times D_h}$, which joins historical information as the initial state of the decoder.

The final architecture of the encoder combines the SS-GCBs & CS-FB. The four MGCUs were included in the architecture of the encoder. Extensive experiments were performed to decide the optimal number of MGCUs (four), which has been discussed in Section VI (Result and Experiments). The first two contain the SS-GCBs & CS-FBs, while the last two MGCUs use only the SS-GCB [4]. The encoder modules are as follows:

1) Single-Scale Graph Convolution Block (SS-GCB): It represents the graph convolution and temporal convolution. The proposed model uses four cascaded SS-GCBs employed to obtain spatiotemporal features. Each SS-GCB unit includes a ReLU activation function, batch normalization layer, and dropout operation. It is denoted as ReLU $\rightarrow$ BN $\rightarrow$ Dropout.

2) Cross-Scale Fusion Block (CS-FB): The study focused on multi-scale features and their contribution. In this view, CS-FB has been adopted to fuse multi-scale features. It is used to minimise the dimension & feature vector of the body component through temporal convolution. These feature vectors are passed from the four multi-layer perceptrons (MLP) to learn feature embedding for two body scales & then employ the softmax layer to calculate edge weight in a corresponding cross-scale graph.

B. *Decoder*

Future postures are predicted successively by the decoder. The decoder of the architecture includes three modules, i.e. Multi-layer perceptron (MLP), graph-gated recurrent unit lite (GGRU-L) and Difference operator (Diff). The core part of a decoder is our proposed GGRU-L, which reproduces motion state for sequential motions. MLP is used to learn features from the encoder part, and the output is treated as the input for GGRU-L. The difference operator is used to update hidden state motion. Fig. 7 depicts the complete framework of the decoder part of DMST-GRNN. Then use an output function (fpred()) to

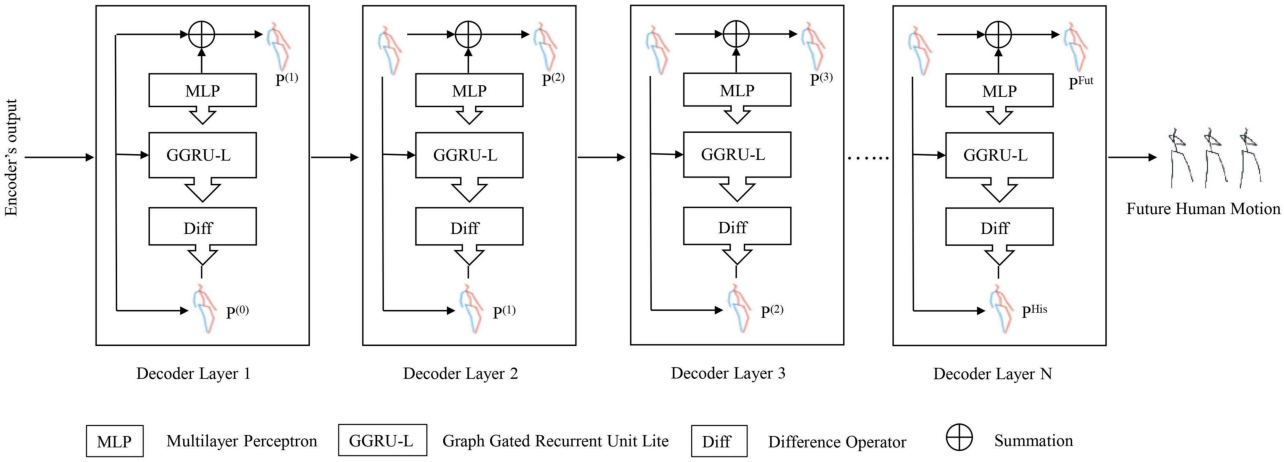

Fig. 7. Decoder framework of DMST-GRNN model. In the decoder, graph GRU-Lite is used to predict human motion.

produce future posture displacement. To forecast the next frame and apply the displacement to the input posture. The decoder works at frames t are mentioned in (3).

$$\widehat{\mathbf{X}}^{(t+1)} = \widehat{\mathbf{X}}^{(t)} + f_{\text{pred}}\left(\text{GGRU}-\text{L}\left(\text{diff}_2\left(\widehat{\mathbf{X}}^{(t)}\right), \mathbf{H}^{(t)}\right)\right) \tag{3}$$

where, $f_{pred}(.)$ = output function, $H^{(0)}$ = Initial hidden state, $H^{(t)}$ = Updated hidden state

1) Graph Gated Recurrent Unit Lite(GGRU-L): GGRU-L is the main key component of the decoder used for future poses. A GGRU-L's prominent role is to modify hidden states using a trainable graph as a guide. Using a trainable graph, the key is to regularise the states to generate future postures [25]. Let $A^H \in R^{J\times J}$ = adjacent matrix of the built-in graph, which is developed to create dynamic edges and is initially initialized with the skeleton graph, and $H^0 \in R^{M\times D_h}$ = initial state of GGRU-L At the time $t > 0$, GGRU-L takes input as the initial hidden state H^0, Hidden state at t $H^{(t)}$, and 3D skeleton-based information input: $I(t) \in R^{J\times d}$. Then, GGRU-L $(I^{(t)}, H^{(t)})$ discussed in (4), (5), (6), and (7).

$$\mathbf{r}^{(t)} = \sigma\left(r_{\text{hid}}\left(\mathbf{A_H}\mathbf{H}^{(t)}\mathbf{W_H}\right)\right) \tag{4}$$

$$\mathbf{u}^{(t)} = \sigma\left(u_{\text{hid}}\left(\mathbf{A_H}\mathbf{H}^{(t)}\mathbf{W_H}\right)\right) \tag{5}$$

$$\mathbf{c}^{(t)} = \tanh\left(\mathbf{r}^{(t)} \odot c_{\text{hid}}\left(\mathbf{A_H}\mathbf{H}^{(t)}\mathbf{W_H}\right)\right) \tag{6}$$

$$\mathbf{H}^{(t+1)} = \mathbf{u}^{(t)} \odot \mathbf{H}^{(t)} + \left(1 - \mathbf{u}^{(t)}\right) \odot \mathbf{c}^{(t)} \tag{7}$$

where, $r_{hid}(.), u_{hid}(.), c_{hid}(.)$ = trainable linear mapping, W_H = trainable weight, $r^{(t)}$ = reset gate, $u^{(t)}$ = update gate, $H^{(t+1)}$= updated candidate hidden state.

Each GGRU-L cell creates the state for the next frame by applying a graph convolution to the hidden state for information propagation. The GGRU-L & MLP produced output functions are combined in the decoder's final architecture. It helps to predict future postures and estimate the motion between two consecutive frames.

V. Dataset Used

Extensive tests are performed on two large datasets, i.e., Human 3.6M [5], [6], CMU Mocap [7], [8], to validate our DMST-GRNN model for human activity prediction. The results of the experiments demonstrate that DMST-GRNN outperforms most existing techniques for short and long-term human motion prediction [1]. So following is the description of the dataset:

A. Human 3.6m (H3.6M)

In the H3.6M dataset, seven participants conduct fifteen different activities. Each subject has 32 joints and utilizes those with non-zero values after transforming the joint locations into exponential maps (20 joints remain). Downsample all sequences by two along the time axis. The model training was done on six individuals and evaluated on a particular video from the fifth subject, following the next paradigms [5], [6].

B. CMU Motion Capture (CMU Mocap)

The five general action classes in CMU Mocap are human interaction, environment interaction, locomotion, physical activities & sports, and situations & scenarios. Each action has thirty-eight joints and preserves twenty-six joints. The eight actions selected by this dataset are "basketball", "basketball signal", "directing traffic", "jumping", "running", "soccer", "walking", and "washing windows" to preserve consistency [7], [8].

VI. Result & Experiment

Extensive experiments were carried out to validate the model using two state-of-the-art datasets, i.e., Human3.6m and CMU Mocap. Furthermore, two sorts of videos (short-term and long-term motion) were used to test the learning capability of the model. The following sections have discussed the underlying protocol of the experiment setup.

A. Short-Term & Long-Term Motion Prediction

In the light of existing literature, videos of lengths up to 400 milliseconds and 1000 milliseconds have been referenced as

TABLE I
MEAN ANGLE ERRORS (MAE) OF STATE-OF-THE-ART METHODS FOR SHORT-TERM & LONG-TERM MOTION PREDICTION ON 15 ACTIONS OF H3.6M

Motion	Walking						Eating						Smoking					
Milliseconds	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
Res-sup [9]	0.27	0.46	0.67	0.75	0.93	1.03	0.23	0.37	0.59	0.73	0.95	1.08	0.32	0.59	1.01	1.10	1.25	1.60
CSM [10]	0.33	0.54	0.68	0.73	0.88	0.92	0.22	0.36	0.58	0.71	1.01	1.24	0.26	0.49	0.96	0.92	0.97	1.62
Traj-GCN [23]	0.18	0.32	0.49	0.56	0.65	0.67	0.17	0.31	0.52	0.62	0.76	1.12	0.22	0.41	0.84	0.79	0.87	1.57
SC-RNN [19]	0.32	0.47	0.54	0.63	0.72	0.88	0.25	0.42	0.56	0.68	0.67	0.87	0.35	0.41	0.74	1.06	1.10	1.28
SAED [20]	0.25	0.41	0.65	0.75	0.72	0.91	0.21	0.34	0.34	0.66	0.67	1.20	0.26	0.49	0.92	0.91	1.10	1.65
DMGNN [4]	0.18	0.31	0.49	0.58	0.66	0.75	0.17	0.30	0.49	0.59	0.74	1.14	0.21	0.40	0.81	0.78	0.84	1.52
Hisrep [24]	0.18	0.30	0.46	**0.51**	**0.59**	**0.64**	0.16	0.28	0.47	0.55	0.74	**1.10**	0.21	0.39	0.78	0.77	0.86	1.58
MST-GNN [1]	0.18	0.31	0.49	0.57	0.62	0.73	0.16	0.28	0.47	0.55	0.74	1.13	0.21	0.39	0.78	0.77	0.83	1.51
DMST-GRNN	**0.18**	**0.30**	**0.46**	0.52	0.62	0.73	**0.15**	**0.26**	**0.44**	**0.54**	**0.73**	1.12	**0.21**	**0.38**	**0.77**	**0.76**	**0.81**	**1.50**
Motion	Discussion						Directions						Greeting					
Milliseconds	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
Res-sup [4]	0.30	0.67	0.98	1.06	1.43	1.69	0.41	0.64	0.80	0.92	1.26	1.48	0.57	0.83	1.45	1.60	1.84	1.96
CSM [10]	0.32	0.67	0.94	1.01	1.56	1.86	0.39	0.60	0.80	0.91	0.98	1.37	0.51	0.82	1.21	1.38	1.72	1.79
Traj-GCN [23]	0.20	**0.51**	**0.79**	**0.86**	1.33	1.70	0.26	0.45	0.70	0.79	0.87	1.29	0.35	0.61	0.96	1.13	1.54	1.59
SC-RNN [19]	0.33	0.67	0.94	1.05	1.27	1.86	0.30	0.52	0.69	0.79	0.87	1.07	0.54	0.79	1.13	1.31	1.49	1.64
SAED [20]	0.30	0.66	0.93	0.98	1.27	1.73	0.39	0.54	0.75	0.84	1.14	1.38	0.54	0.87	1.24	1.41	1.49	1.75
DMGNN [4]	0.26	0.65	0.92	0.99	1.33	1.45	0.25	0.44	0.65	0.71	0.86	1.30	0.36	0.61	**0.94**	**1.12**	1.57	1.63
Hisrep [24]	0.20	0.52	0.78	0.87	1.29	1.63	0.25	0.43	0.60	0.69	0.81	1.27	0.35	0.60	0.95	1.14	1.47	**1.57**
MST-GNN [1]	0.22	0.56	0.83	0.89	1.30	1.58	0.24	0.43	0.60	0.67	0.82	1.26	**0.34**	0.58	0.95	1.12	1.55	1.61
DMST-GRNN	**0.20**	0.56	0.81	0.88	**1.21**	**1.24**	**0.23**	**0.42**	**0.58**	**0.64**	**0.73**	**1.18**	0.36	**0.58**	0.95	1.13	**1.51**	1.59
Motion	Phoning						Posing						Purchases					
Milliseconds	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
Res-sup [9]	0.59	1.06	1.45	1.60	1.75	2.01	0.45	0.85	1.34	1.56	2.03	2.55	0.58	0.79	1.08	1.15	1.76	2.52
CSM [10]	0.59	1.13	1.51	1.65	1.68	1.92	0.29	0.60	1.12	1.37	1.93	2.62	0.63	0.91	1.19	1.29	1.64	2.42
Traj-GCN [23]	0.53	1.02	1.32	1.45	1.47	1.66	0.23	0.54	1.26	1.38	1.57	2.37	0.42	0.66	1.04	1.12	1.52	2.28
SC-RNN [19]	0.59	1.13	1.48	1.61	1.47	1.78	0.28	0.56	1.15	1.40	1.57	2.46	0.60	0.85	1.16	1.24	1.52	2.27
SAED [20]	0.27	0.43	0.62	0.68	0.79	1.08	0.34	0.67	1.20	1.45	1.86	2.53	0.47	0.93	1.07	1.07	1.13	1.30
DMGNN [4]	0.52	0.97	1.29	1.43	1.44	1.64	0.20	0.46	1.06	1.34	1.49	2.17	0.41	0.61	1.05	1.14	1.46	2.26
Hisrep [24]	0.53	1.01	1.31	1.43	1.41	1.68	0.19	0.46	1.09	1.35	1.60	2.32	0.42	0.65	1.00	1.07	1.39	2.13
MST-GNN [1]	0.52	**0.83**	1.25	1.38	1.36	1.60	0.18	0.44	0.98	1.20	1.51	2.15	0.40	0.60	0.97	1.04	1.43	2.22
DMST-GRNN	**0.50**	0.84	**1.24**	**1.38**	**1.27**	**1.52**	**0.17**	**0.42**	**0.93**	**1.12**	**1.25**	**1.78**	**0.40**	**0.60**	**0.96**	**1.04**	**1.31**	**2.05**
Motion	Sitting						sitting Down						Taking Photo					
Milliseconds	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
Res-sup [9]	0.41	0.68	1.12	1.33	1.54	1.85	0.47	0.88	1.37	1.54	1.60	2.17	0.28	0.57	0.90	1.02	1.36	1.59
CSM [10]	0.39	0.61	1.02	1.18	1.32	1.68	0.41	0.78	1.16	1.31	1.48	1.98	0.23	0.49	0.88	1.06	1.05	1.25
Traj-GCN [23]	0.29	0.45	0.82	0.97	1.15	1.50	0.30	0.63	0.89	1.01	1.20	1.72	0.15	0.36	0.59	0.72	0.86	1.08
SC-RNN [19]	0.43	0.89	1.12	1.34	1.50	1.81	0.34	0.72	1.37	1.51	1.65	1.92	0.25	0.53	0.81	0.90	0.97	1.27
SAED [20]	0.40	0.63	1.01	1.18	1.50	1.66	0.40	0.75	1.08	1.21	1.65	1.89	0.26	0.51	0.77	0.88	0.97	1.17
DMGNN [4]	0.26	0.42	0.76	0.97	1.12	1.51	0.32	0.65	0.93	1.05	1.30	1.74	0.15	0.34	0.58	0.71	0.83	1.06
Hisrep [24]	0.29	0.47	0.83	1.01	1.16	1.55	0.30	0.63	0.92	1.04	1.18	1.70	0.16	0.36	0.58	0.70	0.82	1.08
MST-GNN [1]	0.26	0.41	0.75	0.92	1.09	1.47	0.30	0.62	0.88	0.99	1.23	1.72	0.15	0.35	0.57	0.69	0.82	1.03
DMST-GRNN	**0.24**	**0.39**	**0.73**	**0.89**	**1.06**	**1.45**	**0.29**	**0.59**	**0.86**	**0.96**	**1.08**	**1.46**	**0.15**	**0.33**	**0.56**	**0.67**	**0.76**	**0.93**
Motion	Waiting						Walking Dog						Walking Together					
Milliseconds	80	160	320	400	560	1000	80	160	320	400	560	1000	80	160	320	400	560	1000
Res-sup [9]	0.32	0.63	1.07	1.26	1.73	2.43	0.52	0.89	1.25	1.40	1.82	2.36	0.27	0.53	0.74	0.79	1.03	1.48
CSM [10]	0.30	0.62	1.09	1.30	1.65	2.39	0.59	1.00	1.32	1.44	1.70	2.01	0.27	0.52	0.71	0.74	0.87	1.32
Traj-GCN [23]	0.23	0.50	0.92	1.15	1.58	2.32	0.46	0.80	1.12	1.30	1.55	1.79	0.15	0.35	0.52	0.57	0.61	1.17
SC-RNN [19]	0.30	0.54	1.11	1.23	1.31	1.63	0.46	0.79	1.12	1.34	1.68	1.80	0.41	0.43	0.50	0.58	0.67	1.02
SAED [20]	0.30	0.61	1.08	1.31	1.31	2.43	0.55	0.91	1.22	1.37	1.68	1.94	0.23	0.44	0.63	0.69	0.87	1.26
DMGNN [4]	0.22	0.49	0.88	1.10	1.46	2.12	0.42	0.72	1.16	1.34	1.57	1.79	0.27	0.52	0.83	0.95	0.70	1.24
Hisrep [24]	0.22	0.49	0.92	1.14	1.54	2.30	0.46	0.78	**1.05**	**1.23**	1.57	1.82	0.27	0.52	0.82	0.94	**0.63**	**1.16**
MST-GNN [1]	**0.21**	0.49	0.88	1.08	1.40	2.05	0.41	0.73	1.11	1.25	1.56	1.74	0.26	0.50	0.79	0.92	0.72	1.23
DMST-GRNN	0.22	**0.48**	**0.83**	**0.98**	**1.31**	**1.96**	**0.40**	**0.71**	1.08	1.24	**1.53**	**1.73**	**0.13**	**0.30**	**0.46**	**0.51**	0.68	1.21

short-term and long-term. There are 15 and 8 activities from the Human 3.6M and CMU Mocap datasets, respectively, used for short-term motion prediction. Similarly, the same protocol setup was employed for long-term motion prediction (15 and 8 activities from Human 3.6M and CMU Mocap dataset). However, the model was also validated with videos of other different milliseconds, such as 80 ms, 160 ms, and 320 ms, that are treated as short-term.

Tables I and IV represent the mean angle error (MAE) for various state-of-the-art methods on the 15 and 8 activities of the Human3.6M and CMU Mocap, respectively. Table III depicts the standard deviation error for the 15 activities of DMST-GRNN on a human 3.6m dataset. Furthermore, Tables II and V show the Average MAE with different video lengths (80 ms, 160 ms, 320 ms, 400 ms, 1000 ms) on Human3.6M and CMU Mocap respectively. The proposed model has been evaluated with various spatial and temporal scale values, and it has been discovered that it performs best with scales I, II, and III. The underlying setup and result of the validation are explained in the following section with Tables VI and VII.

TABLE II
AVERAGE MEAN ANGLE ERRORS (MAE) OF STATE-OF-THE-ART METHODS FOR SHORT-TERM & LONG-TERM MOTION PREDICTION ON 15 ACTIONS OF H3.6M

Motion	Average MAE					
Milliseconds	80	160	320	400	560	1000
Res-sup [9]	0.40	0.69	1.04	1.18	1.48	1.75
CSM [10]	0.38	0.68	1.01	1.13	1.36	1.76
Traj-GCN [24]	0.27	0.53	0.85	0.96	1.17	1.59
SC-RNN [23]	0.37	0.62	0.92	1.04	1.14	1.46
SAED [20]	0.37	0.65	0.97	1.14	1.19	1.69
DMGNN [4]	0.27	0.52	0.83	0.95	1.17	1.57
Hisrep [24]	0.27	0.52	0.82	0.94	1.14	1.57
MST-GNN [1]	0.26	0.50	0.79	0.92	1.15	1.55
DMST-GRNN	**0.25**	**0.47**	**0.77**	**0.81**	**1.05**	**1.43**

These experimental results favour the proposed DMST-GRNN model. It performs exceptionally well on both datasets compared to other state-of-the-art approaches. The Human3.6

TABLE III
STANDARD DEVIATION ANGLE ERRORS OF DMST-GRNN FOR SHORT-TERM & LONG-TERM MOTION PREDICTION ON 15 ACTIONS OF H3.6M

Motion	Standard deviation angle error of DMST-GRNN					
Milliseconds	80	160	320	400	560	1000
Walking	0.008	0.016	0.028	0.034	0.041	0.054
Eating	0.007	0.014	0.026	0.036	0.049	0.112
Smoking	0.012	0.024	0.036	0.035	0.056	0.180
Discussion	0.013	0.038	0.057	0.064	0.132	0.141
Direction	0.015	0.031	0.034	0.036	0.037	0.138
Greeting	0.021	0.037	0.068	0.123	0.181	0.152
Phoning	0.031	0.042	0.142	0.162	0.152	0.179
Posing	0.012	0.031	0.067	0.131	0.149	0.192
Purchase	0.028	0.040	0.069	0.112	0.141	0.263
Sitting	0.016	0.024	0.037	0.046	0.121	0.171
Sitting Down	0.018	0.038	0.042	0.069	0.131	0.146
Taking Photo	0.007	0.037	0.039	0.041	0.044	0.069
Waiting	0.015	0.032	0.042	0.070	0.162	0.210
Walking Dog	0.031	0.036	0.121	0.140	0.163	0.192
Walking Together	0.007	0.021	0.032	0.034	0.036	0.129

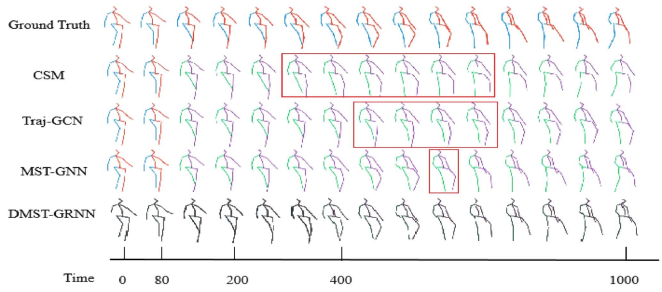

Fig. 8. Walking motion prediction on the short & long-term video with H3.6M.

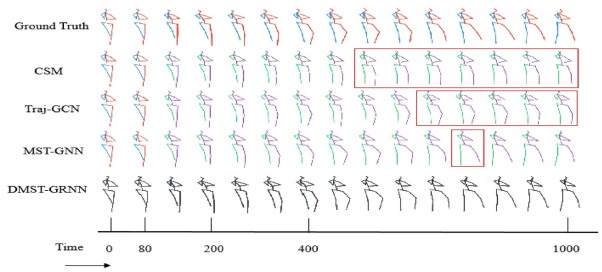

Fig. 9. Smoking motion prediction on the short & long-term video with H3.6M.

million dataset improved performance by 11.95% at 400 ms and 7.74% at 1000 ms, a higher increase in performance (avg MAE) than the best available method [1] discussed in Table II. Similarly, the CMU Mocap dataset yields 3% at 400 ms and 6% at 1000 ms using an average MAE, and the best available method [1] is discussed in Table V.

B. Motion Prediction Visualization

Figs. 8 and 9 show future poses for "walking" and "smoking" motion respectively, on short-term and long-term videos of H3.6M. It clearly visualizes that the DMST-GRNN model forecast is substantially better than others. This visualization result was constructed based on the MAE with the H3.6M dataset. Figs. 8 and 9 illustrate some red boxes over human poses in baseline methods that indicate a slightly high deviation from the ground truth of human poses. These deviations were discovered during prediction using the MAE values for different video lengths on the H3.6M dataset.

C. Effect of Multiple Scales on Spatial & Temporal Graphs

Spatial graphs are classified into scales like 20, 10, 5, 3, 2 to validate multi-scale graph representation. The Average MAE is depicted in Table VI based on various spatial scales. The DMST-GRNN model offers the best results with 20, 10, 5 nodes. For example, spatial scales use 20,10,5 vertices means the 20, 10, 5 nodes (key points) are extracted from the human body for generating spatial scale features.

Similarly, temporal graphs are classified into scales like 1, 1/2, 1/3, 1/4, 1/5 to validate multi-scale graph representation. The Average MAE is depicted in Table VII with different temporal scales. The DMST-GRNN model offers the best results with 1, 1/2, 1/3 length size. For example, temporal scales use 1, 1/2, 1/3 times of lengths of the original video for the temporal scale features. It is worth noting that spatial and temporal scale features must pass together in the model. In this regard, the best results have been observed with three scales (I, II, III) for both temporal and spatial, as depicted in Tables VI and VII.

D. Effect of the MGCU's Count

The MGCU unit is an integral part of the proposed framework and plays an essential role in overall performance. The proposed framework has offered four MGCUs, as presented in Section IV and Fig. 6. The study has included an extensive experiment with various combinations of MGCU units, and results are summarized in Table VIII. It is very much evident a better performance with four M-GCU units.

The study conducted experiments on encoders with varying numbers of MGCU units and found that using four units yielded the best performance for accurate human motion prediction. Employing fewer units resulted in underutilising features, while employing more units led to overfitting. It is important to note that these findings are domain-specific and may not be generalizable to other applications.

E. Perfomance With Prediction Loss

The DMST-GRNN evaluated on prediction loss, which states the error for the model's current state over different iterations. Fig. 10 depicts the prediction loss of short-term & long-term motion prediction of the H3.6M and CMU Mocap.

The DMST-GRNN model utilizes four MGCU units in the encoder, maintaining performance while reducing parameters in the decoder, resulting in lower computational requirements. The computational complexity can be quantitatively measured using Giga floating-point operations (GFLOPs) [26], which several factors into account, such as input size and model implementation details. The DMST-GRNN model's GFLOPs value is 236.7, indicating its computational efficiency.

It is worth mentioning that the GGRU-L employed in the decoder, as discussed in Section IV, is responsible for improved performance. Because it transmits fewer parameters but is more

TABLE IV
MEAN ANGLE ERRORS (MAE) OF STATE-OF-THE-ART METHODS FOR SHORT-TERM & LONG-TERM MOTION PREDICTION ON 8 ACTIONS OF CMU MOCAP

Motion	Basketball					Basketball Signal					Directing Traffic					Jumping				
Milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res-sup [9]	0.49	0.77	1.26	1.45	1.77	0.42	0.76	1.33	1.54	2.17	0.31	0.58	0.94	1.10	2.26	0.57	0.86	1.76	2.03	2.42
SAED [20]	0.40	0.69	1.20	1.39	1.84	0.30	0.54	0.93	1.08	1.48	0.29	0.52	0.83	0.96	1.90	0.34	0.54	1.26	1.45	1.88
CSM [10]	0.37	0.62	1.07	1.18	1.95	0.32	0.59	1.04	1.24	1.96	0.25	0.56	0.89	1.10	2.06	0.39	0.60	1.36	1.56	2.01
Traj-GCN [23]	0.33	0.52	0.89	1.01	1.71	0.11	0.20	0.41	0.53	1.00	0.15	0.32	0.52	**0.60**	2.00	0.31	**0.49**	**1.23**	**1.39**	1.80
DMGNN [4]	0.33	0.46	0.89	1.11	1.66	0.10	0.17	0.31	0.41	1.26	0.15	0.30	0.57	0.72	1.98	0.37	0.65	1.49	1.71	1.79
MST-GNN [1]	0.28	0.49	0.88	1.01	1.68	0.10	0.18	0.29	0.38	1.04	0.15	**0.29**	0.51	0.62	1.95	0.31	0.53	1.29	1.46	1.82
DMST-GRNN	**0.25**	**0.41**	**0.80**	**0.97**	**1.63**	**0.08**	**0.14**	**0.26**	**0.34**	**0.97**	**0.15**	0.30	**0.50**	0.62	**1.91**	**0.31**	0.52	1.29	1.46	**1.74**
Motion	Running					Soccer					Walking					Washing Window				
Milliseconds	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
Res-sup [9]	0.32	0.48	0.65	0.74	1.00	0.29	0.50	0.87	0.98	1.73	0.35	0.45	0.59	0.64	0.88	0.32	0.47	0.74	0.93	1.37
SAED [20]	0.30	0.46	0.55	0.58	0.64	0.24	0.41	0.74	0.84	1.36	0.33	0.40	0.43	0.47	0.67	0.33	0.54	0.85	1.03	1.43
CSM [10]	0.28	0.41	0.52	0.57	0.67	0.26	0.44	0.75	0.87	1.56	0.35	0.44	0.45	0.50	0.78	0.30	0.47	0.80	1.01	1.39
Traj-GCN [23]	0.33	0.55	0.73	0.74	0.95	0.18	0.29	**0.61**	0.71	1.40	0.33	0.45	0.49	0.53	0.61	0.22	0.33	0.57	0.75	1.20
DMGNN [4]	0.19	0.31	0.47	0.49	0.64	0.22	0.32	0.79	0.91	1.54	0.30	0.34	0.38	0.43	0.60	0.20	0.27	0.62	0.81	1.09
MST-GNN [1]	0.18	0.28	0.43	0.48	0.66	0.18	0.28	0.63	0.71	**1.38**	0.26	0.32	0.37	0.41	0.54	0.21	0.31	0.54	**0.69**	1.07
DMST-GRNN	**0.18**	**0.25**	**0.36**	**0.40**	**0.47**	**0.18**	**0.28**	0.62	**0.70**	1.40	**0.25**	**0.32**	**0.35**	**0.40**	**0.51**	**0.18**	**0.24**	**0.52**	0.71	**1.03**

TABLE V
AVERAGE MEAN ANGLE ERRORS (MAE) OF STATE-OF-THE-ART METHODS FOR SHORT-TERM & LONG-TERM MOTION PREDICTION ON 8 ACTIONS OF CMU MOCAP

Motion	Average MAE				
Milliseconds	80	160	320	400	1000
Res-sup [9]	0.38	0.61	1.02	1.18	1.68
SAED [23]	0.31	0.51	0.85	0.97	1.41
CSM [10]	0.32	0.52	0.88	0.99	1.55
Traj-GCN [24]	0.24	0.39	0.68	0.78	1.33
DMGNN [4]	0.23	0.35	0.69	0.82	1.32
MST-GNN [1]	0.21	0.33	0.62	0.72	1.27
DMST-GRNN	**0.19**	**0.30**	**0.58**	**0.70**	**1.20**

TABLE VI
AVERAGE MEAN ANGLE ERRORS OF DMST-GRNN MODEL AT VARIOUS SPATIAL SCALES FOR SHORT-TERM & LONG-TERM MOTION PREDICTION ON H3.6M DATASET

	Spatial Scale					Average MAE						
	I	II	III	IV	V							
Scale index	20	10	5	3	2	80	160	320	400	560	1000	Avg
I	✓					0.261	0.493	0.811	0.834	1.071	1.450	0.820
I,II	✓	✓				0.254	0.480	0.796	0.820	1.069	1.442	0.815
I,III	✓		✓			0.263	0.477	0.798	0.825	1.064	1.441	0.812
I,IV	✓			✓		0.256	0.486	0.796	0.829	1.059	1.433	0.810
I,V	✓				✓	0.258	0.486	0.796	0.829	1.058	1.433	0.810
I,II,III	✓	✓	✓			**0.251**	**0.474**	**0.773**	**0.819**	**1.056**	**1.431**	**0.800**
I,II,IV	✓	✓		✓		0.253	0.479	0.795	0.819	1.059	1.435	0.806
I,II,V	✓	✓			✓	0.254	0.479	0.795	0.819	1.059	1.436	0.807
I,II,III,IV	✓	✓	✓	✓		0.251	0.474	0.794	0.815	1.064	1.436	0.806
I,II,III,V	✓	✓	✓		✓	0.253	0.477	0.797	0.816	1.063	1.431	0.806
I,II,III,IV,V	✓	✓	✓	✓	✓	0.252	0.474	0.797	0.815	1.064	1.441	0.807

TABLE VII
AVERAGE MEAN ANGLE ERRORS OF DMST-GRNN MODEL AT VARIOUS TEMPORAL SCALES FOR SHORT-TERM & LONG-TERM MOTION PREDICTION ON H3.6M DATASET

	Temporal Scale					Average MAE						
	I	II	III	IV	V							
Scale Index	1	1 / 2	1 / 3	1 / 4	1 / 5	80	160	320	400	560	1000	Avg
I	✓					0.257	0.487	0.793	0.841	1.070	1.512	0.826
I,II	✓	✓				0.255	0.488	0.787	0.855	1.064	1.508	0.826
I,III	✓		✓			0.255	0.484	0.790	0.835	1.070	1.501	0.823
I,IV	✓			✓		0.255	0.479	0.786	0.831	1.064	1.478	0.816
I,V	✓				✓	0.257	0.480	0.797	0.833	1.062	1.461	0.815
I,II,III	✓	✓	✓			**0.251**	**0.474**	**0.773**	**0.819**	**1.056**	**1.431**	**0.800**
I,II,IV	✓	✓		✓		0.253	0.479	0.793	0.822	1.058	1.500	0.818
I,II,V	✓	✓			✓	0.252	0.477	0.794	0.821	1.059	1.510	0.819
I,II,III,IV	✓	✓	✓	✓		0.253	0.479	0.795	0.825	1.059	1.511	0.820
I,II,III,V	✓	✓	✓		✓	0.253	0.477	0.795	0.820	1.062	1.510	0.820
I,II,III,IV,V	✓	✓	✓	✓	✓	0.254	0.485	0.798	0.825	1.061	1.520	0.824

TABLE VIII
AVERAGE MEAN ANGLE ERRORS OF DMST-GRNN MODEL WITH DIFFERENT NUMBERS OF MGCUS FOR SHORT-TERM & LONG-TERM MOTION PREDICTION

MGCU	Average MAE at different timestamp						
	80	160	320	400	560	1000	Avg
1	0.261	0.491	0.798	0.829	1.069	1.444	0.815
2	0.262	0.484	0.794	0.830	1.062	1.442	0.812
3	0.257	0.484	0.784	0.825	1.059	1.438	0.808
4	**0.251**	**0.474**	**0.773**	**0.819**	**1.056**	**1.431**	**0.800**
5	0.258	0.479	0.797	0.824	1.061	1.440	0.810
6	0.259	0.482	0.799	0.825	1.064	1.439	0.811

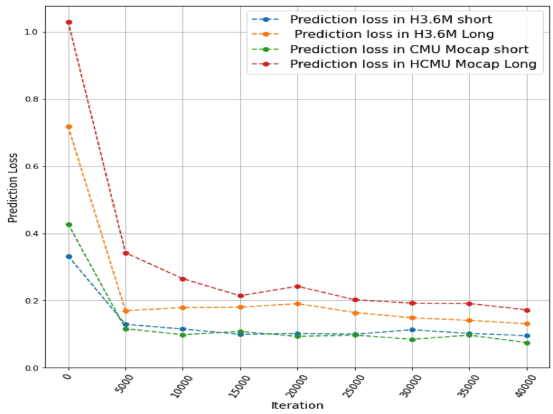

Fig. 10. Prediction Loss in short-term & long-term motion prediction on H3.6M dataset and CMU Mocap.

concrete, robust, and precise to MLP, further, it is helpful in the training to optimize results on the complex function used in the DMST-GRNN model.

VII. CONCLUSION AND FUTURE WORK

Dynamic multi-scale spatiotemporal graph recurrent neural network (DMST-GRNN) is an encoder-decoder architecture that describes the human body and is further employed to predict human motion based on 3D skeletons. This architecture was proposed with four multi-scale graph computational units (MGCU) in the encoder to obtain features & graph GRU lite (GGRU-L) in the decoder to estimate poses. In light of the findings, the suggested model performs better than most of

the existing short and long-term prediction techniques. Two cutting-edge datasets, Human3.6m and CMU Mocap, were used to validate the DMST-GRNN model with various evaluation metrics like Mean Angle Error (MAE), Average MAE, and Prediction Loss. The DMST-GRNN model outperforms the current best available baseline on the Human3.6m datasets by 11.95% and 7.74%, in terms of average mean angle errors for short- and long-term motion prediction, respectively. Similar results were observed with CMU Mocap datasets. This work may help in many promising areas like generating realistic animations of human motion, anticipating a user's intentions with the machine, controlling the motion of robotic systems, predicting the motion of sports person, and predicting imbalance motion in the context of the medical field and generate realistic animations in movies & videos.

The DMST-GRNN model relies on large annotated datasets for effective training and may struggle to generalise to unfamiliar human motion scenarios due to limited underutilized data. Additionally, the quality of automatically extracted features plays a crucial role in the model's performance, but incorporating hand-crafted or domain-specific features could enhance its capabilities.

It is worth to suggest to employ additional evaluation metrics such as the percentage of correct keypoint (PCKh), mean per joint positional error (MPJPE), and loss functions like cross-entropy loss and objective loss for robust evaluation of human pose. Furthermore, semantic segmentation techniques can also extract human poses and may help in human motion prediction. To make the DMST-GRNN model quick and memory-efficient, the attention-based transformer technique and hierarchical temporal memory idea can be applied.

REFERENCES

[1] 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.

[2] Y. Tang, L. Zhang, Q. Teng, F. Min, and A. Song, "Triple cross-domain attention on human activity recognition using wearable sensors," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1167–1176, Oct. 2022.

[3] 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.*, 2021, pp. 3197–3204.

[4] 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.*, 2020, pp. 214–223.

[5] 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.

[6] C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2220–2227.

[7] G. Rogez and C. Schmid, "MoCap-guided data augmentation for 3D pose estimation in the wild," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3116–3124.

[8] H. Kadu and C.-C. J. Kuo, "Automatic human mocap data classification," *IEEE Trans. Multimedia*, vol. 16, pp. 2191–2202, 2014.

[9] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2891–2900.

[10] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5226–5234.

[11] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. IEEE 3rd IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 579–583.

[12] B. Hu et al., "Adaptation supports short-term memory in a visual change detection task," *PLoS Comput. Biol.*, vol. 17, no. 9, 2021, Art. no. e1009246.

[13] Y. Tang, L. Ma, W. Liu, and W. Zheng, "Long-term human motion prediction by modeling motion context and enhancing motion dynamic," 2018, *arXiv:1805.02513*.

[14] C. Taramasco et al., "A novel monitoring system for fall detection in older people," *IEEE Access*, vol. 6, pp. 43563–43574, 2018.

[15] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 2, pp. 361–370, 2017.

[16] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang, "Action-attending graphic neural network," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3657–3670, Jul. 2018.

[17] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, Art. no. 912.

[18] R. Liu, C. Xu, T. Zhang, W. Zhao, Z. Cui, and J. Yang, "Si-GCN: Structure-induced graph convolution network for skeleton-based action recognition," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[19] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3300–3315, Jun. 2022.

[20] R. Zhang, X. Shu, R. Yan, J. Zhang, and Y. Song, "Skip-attention encoder–decoder framework for human motion prediction," in *Multimedia Syst.*, vol. 28, pp. 413–422, 2022.

[21] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 128–145, Apr. 2021.

[22] Y. Tang, J. Lu, Z. Wang, M. Yang, and J. Zhou, "Learning semantics-preserving attention and contextual interaction for group activity recognition," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4997–5012, Oct. 2019.

[23] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9489–9497.

[24] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 474–489.

[25] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst.*, 2017, pp. 1597–1600.

[26] N. T. H. Thu and D. S. Han, "HiHAR: A hierarchical hybrid deep learning architecture for wearable sensor-based human activity recognition," *IEEE Access*, vol. 9, pp. 145271–145281, 2021.

Mayank Lovanshi (Student Member, IEEE) is a Doctoral Research Scholar with the Department of Computer Science Engineering, International Institute of Information Technology, Naya Raipur, India. His research interests include human activity recognition, computer vision, and semantic segmentation.

Vivek Tiwari (Member, IEEE) is an Associate Professor with the Department of Computer Science Engineering, ABV-Indian Institute of Information Technology & Management (ABV-IIITM), Gwalior, India. He has successfully supervised many doctoral and PG scholars. His research interests include machine learning, data mining, computer vision (bioinformatics), business analytics, and data warehousing.

Swati Jain (Member, IEEE) is an Assistant Professor and the Head of the Department of Computer Science with Govt. J. Yoganandam Chhattisgarh College, Raipur (CG), India. A significant number of research work is her credits, including journals, conferences, and book chapters. Her research interests include Data Science, Image Processing, and optimization.