Graph Convolutional Networks for Assessment of Physical Rehabilitation Exercises

Swakshar Deb, Md Fokhrul Islam, Shafin Rahman, and Sejuti Rahman

***Abstract*— Health professionals often prescribe patients to perform specific exercises for rehabilitation of several diseases (e.g., stroke, Parkinson, backpain). When patients perform those exercises in the absence of an expert (e.g., physicians/therapists), they cannot assess the correctness of the performance. Automatic assessment of physical rehabilitation exercises aims to assign a quality score given an RGBD video of the body movement as input. Recent deep learning approaches address this problem by extracting CNN features from co-ordinate grids of skeleton data (body-joints) obtained from videos. However, they could not extract rich spatio-temporal features from variable-length inputs. To address this issue, we investigate Graph Convolutional Networks (GCNs) for this task. We adapt spatio-temporal GCN to predict continuous scores(assessment) instead of discrete class labels. Our model can process variable-length inputs so that users can perform any number of repetitions of the prescribed exercise. Moreover, our novel design also provides self-attention of body-joints, indicating their role in predicting assessment scores. It guides the user to achieve a better score in future trials by matching the same attention weights of expert users. Our model successfully outperforms existing exercise assessment methods on KIMORE and UI-PRMD datasets.**

***Index Terms*— Automated assessment, dynamically changing attention, graph convolution network, performance metrics, physical rehabilitation.**

I. Introduction

PHYSICAL therapy intervention by exercises on given tasks is one of the most effective ways to assess musculoskeletal conditions and rehabilitate post-stroke patients. Patients mostly perform those exercises in a home environment without the presence of experts/therapists. Consequently, patients cannot get adequate guidance and evaluation for the prescribed exercise. It motivates researchers to build models for automatic assessment of physical rehabilitation exercises given RGBD camera data of exercises as input [1], [2]. Such models ensure a economical solution for patients and assist health professionals in monitoring patients' progress. This paper proposes a novel end-to-end model that assesses rehabilitation exercises and provides explicit guidance about achieving a better assessment score. We identify the following drawbacks of existing approaches. ***(1)*** Initial studies solve a classification problem (correct or incorrect exercise) instead of predicting a continuous assessment score [3], [4]. They cannot monitor the subtle improvement of performance. ***(2)*** Many approaches address the regression problem by predicting a numerical score, but they mostly rely on hand-crafted features (e.g., Relative Trajectory, Projected Trajectory, and Jerk) [2]. It requirescostly pre-processing (e.g., PCA, dynamic time warping) and expert knowledge that hampers end-to-end processing. ***(3)*** Recent approaches employ deep learning techniques (CNN) for feature extraction [1]. For this, they convert all input RGBD videos to a fixed-length before feeding them to the network. As a wide variety of users (patients/experts) will perform the same exercises in

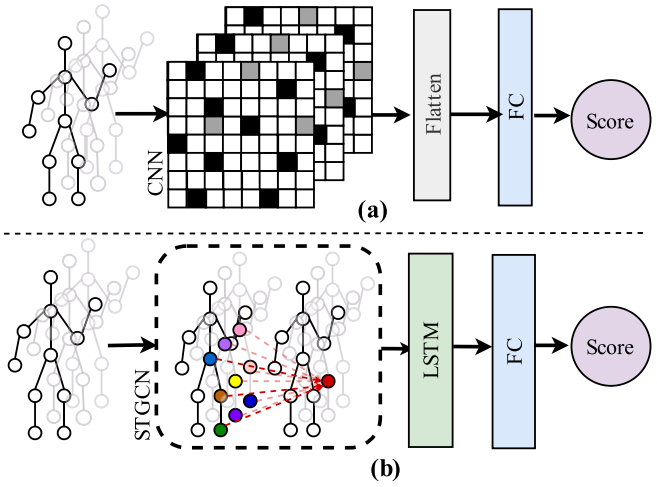

Fig. 1. Overview of existing vs. the proposed method. (a) The existing deep learning method [1] applies CNN to the grid structure of stacked skeleton (body-joints) data. It performs consistently only with fixed-length input and ignores spatio-temporal topological structure from interaction among neighborhood joints. (b) Our proposed method employs STGCN to address the issues mentioned above. We offer extensions to STGCN using LSTM to extract rich spatio-temporal features and attend to different body-joints (as illustrated in colored joints) based on their role in the given exercise. It enables our method to guide users for better assessment scores.

Manuscript received August 14, 2021; revised December 31, 2021; accepted January 29, 2022. Date of publication February 9, 2022; date of current version February 23, 2022. This work was supported in part by the Information and Communication Technology (ICT) Division, Ministry of Posts, Telecommunications and Information Technology of the Government of Bangladesh; and in part by the University of Dhaka. *(Swakshar Deb and Md Fokhrul Islam contributed equally to this work.) (Corresponding author: Sejuti Rahman.)*

Swakshar Deb, Md Fokhrul Islam, and Sejuti Rahman are with the Department of Robotics and Mechatronics Engineering, University of Dhaka, Dhaka 1000, Bangladesh (e-mail: sejuti.rahman@du.ac.bd).

Shafin Rahman is with the Department of Electrical and Computer Engineering, North South University, Dhaka 1229, Bangladesh.

Digital Object Identifier 10.1109/TNSRE.2022.3150392

diverse environments (indoor/outdoor/lab/home) with a different number of repetitions, the assessment network should have the ability to process variable-length input. Moreover, because of depending on CNN, they ignore topological structure information from interaction among neighborhood joints (see Fig. 1(a)). ***(4)*** Existing approaches predict an evaluation score for an input exercise video but fail to explicitly guide the subject by pointing out which body movements/joints to focus on for a better score. This work attempts to solve the mentioned problems. We investigate Graph Convolutional Networks (GCNs) to assess physical rehabilitation exercises. Spatio-temporal GCN (STGCN) was first introduced in traffic forecasting problems [5]. Because of its ability to explore the higher-order topological structure from body joints during human movements, it has been widely used in action/gesture/emotion classification problems since then [6]–[8]. Primarily, STGCN methods use skeleton data (body-joint co-ordinates represented as graphs) of human movements as input and predict discrete class levels [6], [9], [10]. The development of cost-effective depth cameras [11] and pose estimation technology [12]–[15] has made skeleton data considerably more accessible. In this paper, we extend vanilla STGCN to predict a continuous assessment score for physical rehabilitation exercises. Our proposed extension is two-fold (see Fig. 1(b)). Firstly, we incorporate an LSTM after the STGCN architecture (instead of global-pooling) to extract temporal features from variable-length input. It significantly helps to adapt STGCN to regression-based problems. Secondly, we propose a self-attention mechanism operated on the adjacency matrix of body-joints by using ConvLSTM layers. Different body-joints play a different role in the exercise assessment process. By analyzing the attention quantity, we determine how body-joints contribute to the final score, which eventually guides the user to perform better in later trials. We experiment on two established rehabilitation exercise datasets, namely KIMORE [16] and UI-PRMD [17]. Our method outperforms existing methods across several evaluation metrics. We summarize the contributions of this paper as follows:

- We extend the popular STGCN to adapt it for the assessment of rehabilitation exercises in an end-to-end manner.
- Our proposed model supports variable-length exercise input considering any number of repetitions of a given exercise during training and testing. We also offer a self-attention mechanism to guide users by highlighting body-joints contributing more to the prediction.
- We provide extensive experiments on two rehabilitation datasets (KIMORE and UI-PRMD) and establish a new state-of-the-art performance.

II. Related Works

A. Exercise Classification

To replace the costly and subjective evaluation of human expert with an automated system, several works investigate exercise classification for body motion tracking [12], [18], activity classification [6], [19], gait analysis [20], robot programming [21], virtual reality [22], and rehabilitation therapy [1]. To capture motion data, researchers employ RGBD cameras [23], Accelerometers [24], Kinect [2], Gyroscopes [25], Vicon [23], IR cameras [4] and so on. Existing methods generally extract hand-crafted features [3], [4], [23], [26], [27] from motion data to represent human body motion and classify using K-nearest neighbors, SVM, Random Forest, and Logistic Regression, all of which need extensive domain knowledge and lack end-to-end learning intuition. Authors in [26] used machine learning algorithms to detect compensation in stroke survivors based on muscle pressure distribution during exercise. To improve the classification results with the help of more discriminative hand-crafted features, [4] used features across spatial and temporal domains (elbow angle, maximum and minimum knee angle, distance between shoulders, velocity, etc.) to classify different types of disability (e.g., Parkinson's diseases, hemiplegia, etc.). [23] used spectral features (e.g., frequency-domain entropy, high frequency energy content, etc.) to incorporate more useful information and used a support vector machine (SVM) classifier, similar to [4]. These methods, however, rely solely on hand-crafted features and do not fully harness the power of modern deep learning architectures. On the other hand, due to its ability to learn meaningful features and impressive performance, deep learning-based exercise classification [24], [28] has recently received much attention. In [24], a CNN based classifier was implemented based on the wearable sensor data (accelerometer). To improve performance, [28] proposed an RNN-based compensation classification based on 3D joint locations with noise reduction steps, in which they used a Savitzky Golay filter with a fixed-length window size. These studies, however, overlook the interconnectivity of the human body and fail to capture important spatio-temporal features of the body's natural topological structure.

B. Exercise Assessment

Instead of predicting discrete class labels, exercise assessment aims to assign a continuous value measuring the quality of exercise in comparison to a prescribed version. Previous works in this area usually learn a distance function to judge the quality of performed and prescribed exercise. Example works include [29] using Mahalanobis distance, [30]–[32] using dynamic time warping algorithm [33]. For not being exercise specific models, those methods can compare two arbitrary exercises but cannot model task-specific targeted exercise. To address this problem, another stream of works focuses on probabilistic approaches like Hidden Markov models [34], [35] and mixtures of Gaussian distributions [36] for assessing exercises. Those approaches require several preprocessing stages, which hampers the end-to-end processing of the system. Considering the recent success of deep learning, we attempt to assess exercises using deep end-to-end models.

C. Assessment of Rehabilitation Exercise

Not enough work has addressed this topic. Initial results on the assessment of rehabilitation exercise have been reported by [1], [2]. Among them, Lee *et al.* [2] used hand-crafted features to classify a range of motions, smoothness and detect the occurrence of correct and incorrect movements. In another

work, Liao *et al.* [1] proposed a Spatio-temporal network that can assess an exercise. They combined temporal pyramid, multi-branch convolution, and recurrent layers to improve the performance. Both of the methods discussed above employ skeleton data (consisting of 22-39 body joints) to model human users. They applied convolutions on tensors of joint data that break the subtle spatial organization of the natural graph structure inside the human body. Moreover, these methods could not process variable-length input sequences. Also, methods become dependent on many pre-processing steps (PCA, Autoencoder, dynamic time warping and so on) that hampers the end-to-end processing of the system. In line with the recent success of GCNs for modeling human motion [6], [37], we employ GCN to assess rehabilitation exercise. GCNs are good at processing graph structure, useful to extract features from variable-length input and to provide end-to-end solutions.

III. Method

Problem Formulation: Suppose, ith RGBD video of an arbitrary exercise $\mathbf{e}$ is represented by $\boldsymbol{V_i} = \langle \mathbf{X}_t | 1 \dots n_i \rangle$ where, $\mathbf{X}_t$ and n_i denote frame t and number of frames, respectively. We consider n_i to be variable across videos. Each video has a ground-truth score annotation $y_i \in [0, 1]$ representing the assessment/quality of the performed exercise. A higher y_i score indicates better performance by the user. The training dataset includes a set of tuples $\{(\boldsymbol{V_i}, y_i) : i \in [0, \mathcal{T}]\}$ where $\mathcal{T}$ represents the total number of training videos associated with the exercise $\mathbf{e}$. We train an end-to-end θ_e parameterized model, $\mathcal{F}$, for exercise $\mathbf{e}$ that can predict a continuous score, $\hat{y}_j$ close to the ground-truth assessment score, y_j^* for a given test video V_j. We formulae $\hat{y}_j$ as follows:

$$\hat{y}_j = \mathcal{F}(V_j; \theta_{\mathbf{e}}), \quad s.t.\ \hat{y}_j \approx y_j^* \tag{1}$$

We assume that each RGBD frame presents a single prominent human subject. A well-known way to model human is to use skeleton data of different joint co-ordinates [1], [2]. In this paper, we adopt skeleton-based human modeling because it is easy to obtain. Advanced RGBD cameras like Kinects can provide automatic skeleton joint co-ordinates in real-time [16], [17]. Other possible choices include the BlazePose [12] and VideoPose3D [13] algorithms. According to studies [38], [39], Kinectv2 [11] and Vicon are reliable sources of skeleton data for motion-based tasks because the output is closely aligned with the ground truth system (i.e., the stereophotogrammetric system).

Suppose, there are N number of joints and each joint has C dimensional co-ordinate vector that depend on the pose estimation algorithm (2D, 3D or 6D [13], [40], [41]). The dimension of tth frame and ith video become $\mathbf{X}_t \in \mathbb{R}^{N \times C}$ and $\boldsymbol{V}_i \in \mathbb{R}^{\mathcal{T} \times N \times C}$, respectively. Different body joints play essential roles while assessing the qualify of the performed exercise. In our experiments, we notice that such roles vary from one exercise to other. Suppose, $\mathbf{M}_j \in \mathbb{R}^{\mathcal{T} \times N \times N}$ is self-attention map of all joints representing their roles/importance for jth video. By analyzing $\mathbf{M}_j$ of many users (patients and experts), we can get meaningful insight about final predicted score.

TABLE I
Commonly Used Notations

Notations	Descriptions
$\otimes$	Temporal convolution operation.
$\oplus$	Concatenation operation.
$\odot$	Element-wise product.
$*$	Convolution operation.
G	A graph.
$\mathcal{V}$	The set of nodes in a graph.
v	A node $v \in \mathcal{V}$.
N	The number of nodes, $N = \lvert\mathcal{V}\rvert$.
$\mathcal{E}$	The set of edges in a graph.
e_{ij}	An edge $e_{ij} \in \mathcal{E}$.
$\mathcal{N}(v)$	The neighbors of a node v.
$\mathcal{A}_k$	The k hop graph adjacency matrix.
$\tilde{\mathcal{A}}_k$	The normalize k hop graph adjacency matrix.
D_k	The k hop degree matrix of $\mathcal{A}_k$.
$\mathcal{T}$	The total number of timesteps.
C	The number of input channels.
C_{out}	The number of output channels.
$\mathbf{X} \in \mathbb{R}^{\mathcal{T} \times N \times C}$	The input feature matrix.
$\mathbf{G}_k \in \mathbb{R}^{\mathcal{T} \times N \times C'}$	The output of k hop graph convolution.
Γ_i^l	The i^{th} temporal kernal of l^{th} STGCN block.
$\mathbf{Y}^l \in \mathbb{R}^{\mathcal{T} \times N \times C_{out}}$	The output from the l^{th} STGCN block.
$\mathbf{Y}_r^t \in \mathbb{R}^{NC_{out}}$	The tth sequence of Y_r.
$\mathbf{M}_k \in \mathbb{R}^{\mathcal{T} \times N \times N}$	The learned self attention map at k hop.
l	The layer index
t	The time step/iteration index
$\sigma(\cdot)$	The nonlinear activation function.
$\phi(\cdot)$	The normalizing factor.
$\mathbf{W}_k, W_i, W_o, W_f$	Learnable model parameters.

Moreover, it can provide guidance to the user about where to emphasize to achieve better performance. With this motivation, given an RGBD video $\boldsymbol{V_j}$ as input, our overall goals are: ***(a)*** Exercise assessment: to predict the score $\hat{y}_j$, ***(b)*** Role of body-joints: to calculate self-attention map, $\mathbf{M}_j$.

A. Solution Framework

We investigate Graph Convolutional Networks (GCNs) [42] as our core solution framework for assessing physical rehabilitation exercises. GCN is a natural choice because GCN represents human skeleton data as graphs to extract features from the topological structure of neighborhood body-joints. Our final design incorporates a recent spatio-temporal GCN (STGCN) to extract both spatial and temporal feature together as a unified network. Due to the nature of rehabilitation exercise data, our extension addresses the following challenges. ***(1)*** *Variable-length processing:* Unlike related problems with sequential data (action/motion classification in general), rehabilitation exercise data exhibit significant variations within variable-length data. A notable reason could be the exercise actors are largely diverse people, from expert therapists to patients with different disabilities and diseases. Moreover, collected rehabilitation data may originate from constrained lab/gymnasium environments, indoor/outdoor, and home-based settings. Furthermore, based on the therapist's prescription, rehabilitation exercises may need to perform with a variable number of repetitions. As a result, different users task a different amount of time to perform the same exercise with the same number of repetitions. ***(2)*** *Explicit guidance:*

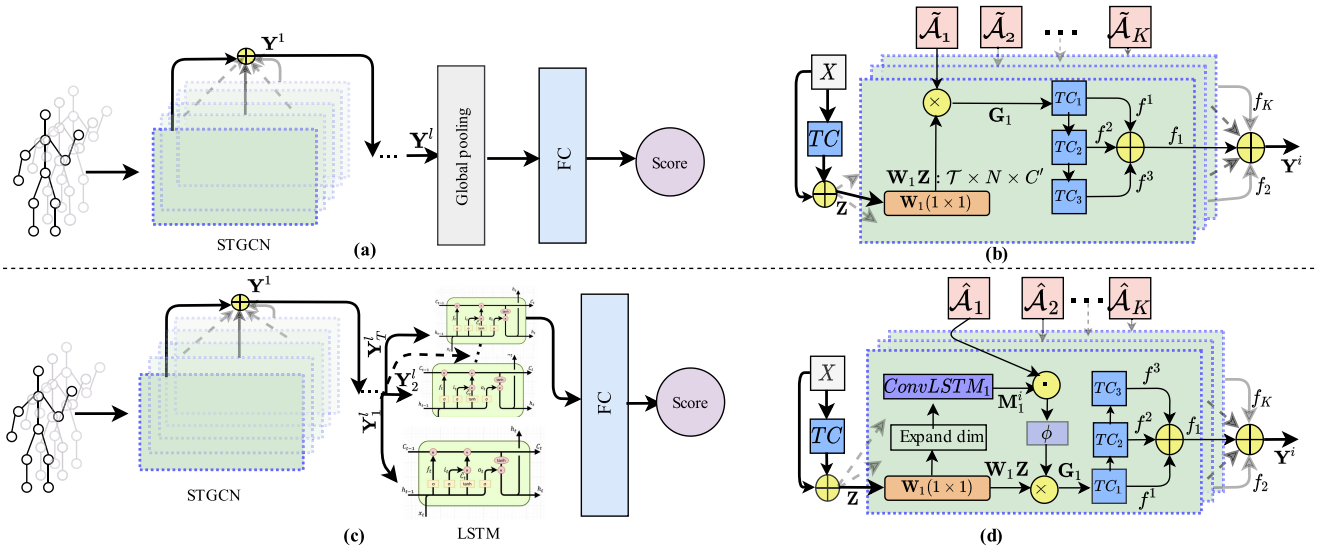

Fig. 2. GCN based end-to-end models using (a-b) vanilla STGCN and (c-d) extended STGCN for rehabilitation exercise assessment. 'TC,' $\oplus$ and $\odot$ denote temporal convolution, concatenation and element-wise multiplication, respectively. (b) and (d) illustrate the detailed components of the green STGCN block of (a) and (c), respectively.

Existing exercise assessment approaches converts this task as a classification [43], [44] or regression [1], [2] problem. However, they do not provide any explicit guidance about where (body-joints) to emphasize or attend to improve the assessment quality.

We first describe the preliminaries of a vanilla STGCN adapted to address rehabilitation exercise assessment task. Then, we show that it can predict a baseline assessment score, but it cannot fully address the challenges mentioned earlier.

1) Exercise Assessment With Vanilla STGCN: We illustrate the network in Fig. 2(a) and (b). We represent human body-joints at each video frame as graph, $G = (\mathcal{A}, \mathcal{V}, \mathcal{E})$ where, $\mathcal{V}$ is set of vertex (body-joints), $\mathcal{E}$ is a set of edges (connection of body-joints) and $\mathcal{A} \in \mathbb{R}^{N\times N}$ is the adjacency matrix of the graph G. For spatial configuration partitioning, we dismantle $\mathcal{A}$ into several matrixes $\mathcal{A}_k$, where $\mathcal{A} + \mathbf{I} = \sum_k \mathcal{A}_k$, representing the adjacency matrix of hop k. The initial conditions are $\mathcal{A}_0 = \mathbf{I}$ and $\mathcal{A}_1 = \mathcal{A}$. To extract dependencies from body-joints, we forward the skeleton data of a video, $\boldsymbol{V} \in \mathbb{R}^{\mathcal{T}\times N\times C}$ to STGCN block. Initially, we perform temporal convolution over the input skeleton sequences using kernel Γ^u. Then, we concatenate the input and the temporal features to extract spatial feature. All he initial processing can be written as:

$$\mathbf{Z} = \mathbf{X} \oplus (\Gamma^u \otimes \boldsymbol{V}) \tag{2}$$

Here, $\mathbf{Z} \in \mathbb{R}^{\mathcal{T}\times N\times (C+C^u)}$ is processed video representation, where, C^μ is the number of filter used in temporal convolution, $\otimes$ and $\oplus$ denote the temporal convolution and concatenation respectively. To extract spatial features from the topological structure of human skeleton, we perform graph convolution on $\mathbf{Z}$ and for the kth hop adjacency matrix, $\mathcal{A}_k$ using the update rule of GCN [42] is:

$$\mathbf{G}_k = \sigma(\tilde{\mathcal{A}}_k \mathbf{Z} \mathbf{W}_k) \tag{3}$$

where, $\tilde{\mathcal{A}}_k = D_k^{-\frac{1}{2}}(\mathcal{A}_k + I)D_k^{-\frac{1}{2}}$, D is a diagonal degree matrix, $\mathbf{W}_k$ is the learnable weight matrix and σ is an nonlinear activation function. This equation performs linear transformation on the feature space and then aggregate the neighbour information using the normalize adjacency matrix. Then, we implement three Temporal Convolutional Layers (TCNs) with same padding and kernel Γ_1^l, Γ_2^l and Γ_3^l correspondingly to extract different level of temporal features. To recognize movement patterns at different levels of abstraction, we concatenate both higher and lower level features. The operations involve: $f^1 = \Gamma_1^l \otimes \mathbf{G}_k$, $f^2 = \Gamma_2^l \otimes f^1$, $f^3 = \Gamma_3^l \otimes f^2$ and $f_k = f^1 \oplus f^2 \oplus f^3$, where f_k represents the spatio-temporal features extracted from the kth hop. Finally, we concatenate outputs from each hop.

$$\mathbf{Y} = f_1 \oplus f_2 \oplus \ldots \oplus f_k \tag{4}$$

The output of the STGC block, $\mathbf{Y} \in \mathbb{R}^{\mathcal{T}\times N\times C_{out}}$, where, $C_{out} = \sum_{i=1}^{3} kC_i^l$ is a 3D tensor. In experiment, we stack multiple STGCN blocks to extract more complex features. Next, we apply a global average pooling to the output of the last STGCN, $\mathbf{Y}^l$, and calculate the feature vector $Y_{pool} \in \mathbb{R}^{C_{out}}$ is further processed by a series of Fully Connected (FC) layers to predict a continuous assessment score.

2) Issues With Vanilla STGCN: The extracted spatio-temporal features, $\mathbf{Y}_{pool}$ can provide a baseline assessment score, but there is scope for improvement. In Fig. 2(a), one can notice that the global pooling layer before the FC layers ignores sequential dependencies resided among the spatio-temporal features across frames/body movements. Thus, users performing the same exercise at different pace (slow or fast) extract different spatio-temporal features. This issue becomes more critical in regression-based learning. Furthermore, different exercises pay particular emphasis to specific joints. But, vanilla STGCN treats all body-joints equally.

It cannot provide the role/importance of joints in predicting the assessment score.

B. Our Extension

In this subsection, we extend the vanilla STGCN for the rehabilitation exercise assessment task to address the issues mentioned above. We include two components to address variable-length processing and calculating the role of body-joints to guide users explicitly.

1) Variable-Length Processing: We employ an LSTM instead of global pooling layers (see Fig. 2(c)). It has some benefits over pooling. **(1)** LSTM captures sequential dependencies presented in spatio-temporal feature vectors. It is inherently designed to extract discriminative features that have accumulated over time. This subtle information plays an important role in predicting the correctness score of an exercise. Smoothness, for example, is an important criterion for scoring a given exercise. We must examine the temporal features (velocity, acceleration) overall conjugative time frames to discern the smoothness of a movement (see Table VI for experimental evidence). Using a pooling layer instead of an LSTM may fail to extract the smoothness information. **(2)** Pooling layers take average or maximum response for predefined window size. It may lose some subtle features necessary for predicting correctness score. Instead, LSTM searches for more meaningful regions from these extracted features. We reshape, $\mathbf{Y}^l$ to $\mathbf{Y}_r \in \mathbb{R}^{\mathcal{T}\times NC_{out}}$ where $\mathbf{Y}^l$ is the output from the last STGCN. The input to the LSTM is $\mathbf{Y}_r^t \in \mathbb{R}^{NC_{out}}$, the tth sequence of Y_r. Considering the characteristics of the rehabilitation exercise data and regression-based problem instead of classification, LSTM fits better instead of global pooling. LSTM considers the variation of spatio-temporal features over the temporal dimension, whereas global pooling breaks the sequential nature of data. Therefore, it helps variable-length processing and allows users to perform the same exercise at different paces.

2) Role of Body-Joints: The role of each joint is different in each exercise. For example, in the squatting exercise (Ex 5), the ankle, knee, spine, and shoulder play an important role, whereas, in the lifting arms exercise (Ex 1), the elbow, spine, thumb, and wrist joint are more significant than the rest of the joints. We want to emphasize that capturing this joint role is crucial in determining exercise quality (Table VI). However, vanilla STGCN (in Eq 3) treats all body joints equally. This joint role should change depending on temporal and spatial context while assessing rehabilitation exercise. It motivates us to implement the self-attention module where we treat each joint differently depending on its role in a given exercise. To calculate the role of individual body-joints depending on the both temporal and spatial context, we replace the $\tilde{\mathcal{A}}_k$ with a trainable self-attention map, ($\mathbf{M}_k \in \mathbb{R}^{\mathcal{T}\times N\times N}$) as follows:

$$\mathbf{G}_k = \sigma(\phi(\hat{\mathcal{A}}_k \odot \mathbf{M}_k)\mathbf{Z}\mathbf{W}_k) \tag{5}$$

where, $\odot$ represents the Hadamard product, $\hat{\mathcal{A}}_k = \mathcal{A}_k + I$ and $\phi(\cdot)$ is the normalizing factor. Fig. 2(d) illustrates the attention-guided STGC block. We improve STGCN by modifying the adjacency matrix dynamically with a self-attention map calculated from ConvLSTM layers. Let, $\mathbf{Q}_k = ZW_k \in$

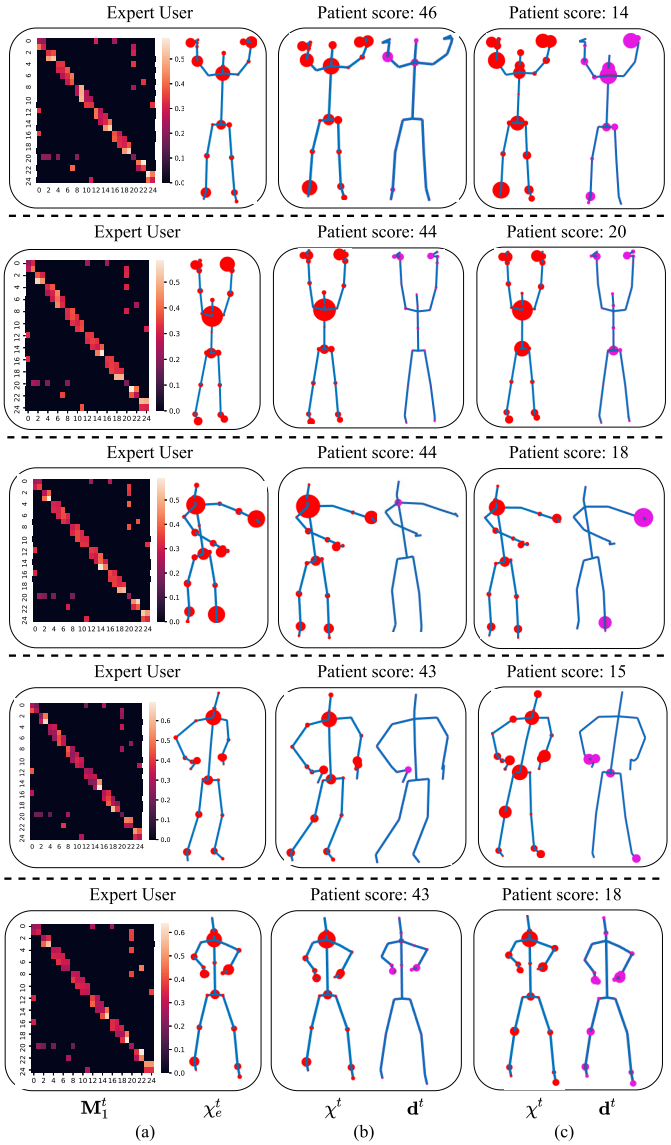

Fig. 3. Visualization of attention maps (red circles) and the role of body joints for five exercises. The larger circle represents the higher role of that joints. (a) Average attention map ($\mathbf{M}_1^t$) and joint role (χ_e^t) of expert users. (b) and (c) *left*: Role (χ^t) of joints when patients score high and low respectively, *right*: the role difference (d^t) from the expert representing where (violet circles) to emphasize to get better assessment score.

$\mathbb{R}^{\mathcal{T}\times N\times C'}$ expanded to $\mathbf{Q}_k$ to $\mathbb{R}^{\mathcal{T}\times N\times 1\times C'}$, where, C' = no of convolutional filters, is the input to the ConvLSTM. The operations at time t are as follows:

$$\begin{aligned}
i_k^t &= \sigma(W_i * \mathbf{Q}_k^t + U_i * h_k^{t-1} + b_i)\\
f_k^t &= \sigma(W_f * \mathbf{Q}_k^t + U_f * h_k^{t-1} + b_f)\\
o_k^t &= \sigma(W_o * \mathbf{Q}_k^t + U_o * h_k^{t-1} + b_o)\\
g_k^t &= \tanh(W_c * \mathbf{Q}_k^t + U_c * h_k^{t-1} + b_c)\\
c_k^t &= f_k^t \odot c_k^{t-1} + i_k^t \odot g_k^t\\
h_k^t &= o_k^t \odot \tanh(c_k^t)
\end{aligned}$$

where, $*$ is the convolution operation, σ represents the sigmoid function, $W_i, W_f, W_o, W_c, U_i, U_f, U_o, U_c \in \mathbb{R}^{1\times 1}$ represent the conv. kernels, b_i, b_f, b_o, b_c are the bias parameter. The number of kernels is the same as the number of joints

in the skeleton graph. The final output from convLSTM does not encapsulate any structural information. We inject the graph structure by elementwise multiplication with the adjacency matrix (Eq. 5). Then, we apply a normalizing factor $\phi(\cdot)$. $\mathbf{M_k} = h_k \in \mathbb{R}^{\mathcal{T}\times N\times N}$, is the self-attention map where each row indicate the attention weights for a body-joints with its neighbors. We can calculate the role of the body-joints in assessing rehabilitation exercises. Fig. 3(a) shows a self-attention map where we highlight the role of body joints. The greater attention value demonstrates the higher emphasis/role on that joint. Some joints may get the same importance but contribute to both high and low scores across trials. It means that those joints have little influence on the overall score because subjects have moved them ideally, and there is nothing much to improve scores by focusing on those joints. On the other hand, some joints are more prominent in determining low scores. Such joints are the ones on which the patient should focus. We can calculate such roles of joints for both expert and inexpert (possibly patient) users. We notice that when a patient gets a low assessment score (0-20), his/her joint-role pattern varies significantly in comparison to the expert's pattern (35-40). The joint role, $\chi \in \mathbb{R}^{\mathcal{T}\times N}$ is computed by column-wise summation over $\mathbf{M}_1 \odot A_1$, where M_1 denotes the first hop attention map. In Fig. 3(b) and (c), we calculate $d^t = \boldsymbol{\chi}_e^t - \boldsymbol{\chi}^t$ to find this pattern variation at time t. $\boldsymbol{\chi}_e$ is calculated by averaging the joint role of the expert therapists. Visualizing d^t, a user can know where to emphasize to achieve a better score in future trails.

We train our proposed network using Huber loss because of less sensitivity to outliers [45]. We simply forward a test sequence of skeleton data (body-joints) to calculate a continuous assessment score during inference. We employ the Huber loss function which is defined as follows:

$$L(y_i - \hat{y}_i) = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2; & |y_i - \hat{y}_i| \leq \delta \\ \delta(|y_i - \hat{y}_i| - \frac{\delta}{2}); & otherwise \end{cases} \quad (6)$$

where y is the true label and $\hat{y}$ is the predicted label. This Huber loss uses the properties of both the Mean squared loss and the Mean absolute deviation loss depending on a parameter δ. When the error (i.e., the difference between actual and the estimated value) is less than a small value δ, it acts as the Mean squared loss. When the error is greater than δ, it approaches the mean absolute loss, which is less sensitive to the outliers. The skeleton data captured by various pose estimation algorithms (Kinectv2, Vicon, BlazePose, VideoPose3D) may contain incorporate incomplete, noisy or redundant information of joint positions [46], [47]. This incomplete or noisy skeleton data can deteriorate the performance, especially when some prominent joints are disturbed. They act as outliers in the sample space. The Huber loss is less sensitive to those outliers.

3) Inference: In inference, we utilize the same architecture as shown in Fig. 2. For a test video $\mathbf{X} \in \mathbb{R}^{\mathcal{T}\times N\times C_{in}}$ where $\mathcal{T}$ can be of any length, first we construct the skeleton graph. The skeleton data can be extracted from pose estimation algorithm or any motion-capture devices. For STGC block we use skeleton graph as the input. Inside each STGC block, first we perform the temporal convolution operation and concatenate with the original input namely $\mathbf{Z} \in \mathbb{R}^{\mathcal{T}\times N\times(C+C^\gamma)}$. Second, inside each graph convolutional network we perform attention guided graph convolution operation over $\mathbf{Z}$. Finally there are three temporal convolutional layers concatenated together. There can be multiple stacked STGC blocks and after the last STGC block there is multiple LSTM layers followed by a dense layer to predict the final score.

TABLE II
SUMMARY OF THE BOTH KIMORE AND UI-PRMD DATASETS

Feature	UI-PRMD dataset [17]	KIMORE dataset [16]
Reference	Vakanski *et al.* (2018)	Capecci *et al.* (2019)
Year	2018	2019
Sensor	Kinect v2 + Vicon	Kinect v2
Modality	Skeleton data	RGB-D and skeleton data
No. of Subjects	10	78
No. of Exercises	10	5
Score range	0 - 1	0 - 50

TABLE III
COMPUTATIONAL COST FOR KIMORE DATASET

Stage	# of Videos	# of Parameters	Execution Time	Avg. time per video
Train	460	0.722 million	50.13 minutes	–
Test	116	0.722 million	2.64 seconds	22.7 milliseconds

IV. EXPERIMENT

A. Setup

1) Dataset: We experiment on two rehabilitation exercise datasets. ***(1)*** Kinematic assessment of Movement for remote monitoring of physical Rehabilitation (KIMORE) [16] dataset includes RGBD videos and the ground-truth score annotations of five types of exercises. It has two groups, control groups (expert and non-expert) and a group with pain and postural disorder (Parkinson, back-pain, stroke). The control group includes 44 healthy subjects, from which 12 subjects were physiotherapists and experts in the rehabilitation of back pain and postural disorders, and the remaining 32 were non-expert healthy subjects. The pain and postural disorder group contain 34 subjects suffering from chronic motor disabilities. ***(2)*** UI-PRMD [17] dataset contains ten rehabilitation exercises collected from 10 healthy subjects using Kinect and Vicon sensors. Each subject performed ten repetitions of the same activity. The data includes positions and angles of full-body joints. A summary of these two datasets is given in Table II.

2) Computational Cost: In Table III, we report training/testing time for the KIMORE dataset. The timing suggests that our proposed method can provide results in real-time, considering real-time skeleton data generation is available. Note that the computational cost is measured using a single Tesla K80 GPU.

3) Evaluation Process: Similar to [1], [2], [5], we evaluate our model using mean absolute deviation (MAD), mean absolute percentage Error (MAPE) and root mean square error (RMSE) scores. The lower the score, the more accurate is the predicted score. Mean Absolute deviation (MAD) is an average of the absolute deviation between true values and

predicted values. MAD, being a scale-dependent measure, cannot be applied to compare methods that are applied to data with different scales. On the other hand, Mean Absolute Percentage Error (MAPE), the percentage equivalent of MAD, is a scale-independent metric. However, MAPE tends to infinity or becomes undefined if the ground truth value equals 0 for any sample in the data. Another widely used evaluation metric is Root Mean Squared Error (RMSE) which is defined as the square root of the squared error (no absolute deviation). RMSE penalizes large errors due to the squared term. It is also a scale-dependent metric like MAD. The equations of MAD, MAPE and RMSE are given as follows:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y - \hat{y}}{y}| \times 100$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2}$$

Here, n is the sample size and y and $\hat{y}$ are the ground-truth and predicted value respectively.

4) Implementation Details[1]*:* We train the model using Adam optimizer for 1500 epochs with a learning rate of 0.0001. The batch size is 3 and 10 for UI-PRMD and KIMORE dataset, respectively. We use cross-validated hop size, $k = 2$. The output space dimensionality of LSTM layers is 80, 40, 40, 80 followed by a fully connected layer with linear activation. According to the validation set, we choose the best model to evaluate the model performance on the test set. We apply a dropout mechanism with a dropout probability of 0.25. Similar to [1], [42], we also report the 10-run result to fairly evaluate the performance of our model against existing works. We perform both training and testing ten times. After each run, we store the performance metrics (MAD, RMSE, MAPE) and finally take the average of stored results, ensuring the reliability of our results. We implement our model using tensorflow2.0. Similar to [6], we apply the Resnet mechanism on each STGC block.

5) Hyper-Parameter Validation: We conduct hyperparameter tuning on a separate validation set. We set the validation split = 0.2 from the total training set inside the `fit()` method of tensorflow 2.0. The train-test-valadiation split is shown in Fig. 4 (a). We validate the model for hop size k within $\{1, 2, 3\}$, the number of STGCN blocks within $\{1, 2, 3, 4, 5\}$, the number of stacked LSTM layers within $\{2, 3, 4, 5\}$, the learning rate within $\{10^{-3}, 10^{-4}, 10^{-5}\}$. When the model suffers from high variance, we can decrease the hop size (k), STGCN, and LSTM layers, reducing model complexity. In contrast, we can increase the values of those hyper-parameters to lessen the bias problem. Based on those validation experiments, we use $k = 2$, three STGC blocks, four LSTM layers and learning rate 10^{-4} for our final model. After finding the optimal model, we conduct experiment on δ sensitivity in Fig. 4(b). Fig. 4(c) shows the validation results (MAD) of our method using different numbers of hop size, STGCN blocks, and stacked LSTM layers for KIMORE exercise 5. It shows that with the increase of STGCN block number from 2 to 3, the performance is boosted, while further increasing the block number leads to no significant improvement. As we increase hop size from 1 to 2, the performance is improved, but for hop size 3 the overall performance deteriorates since there is an influence of unnecessary joints disrupting the informative features for larger neighborhood size. Moreover, increasing the stacked LSTM layers, the model's overall performance continues to improve since LSTM is better suited for extracting meaningful features from sequential dependencies. As we increase the number of LSTM, the performance increases significantly, but after 4 LSTM layers, there is a slight improvement. Considering the computational cost, we select LSTM layers to be 4.

[1]Code and evaluation are available at: https://github.com/fokhruli/STGCN-rehab

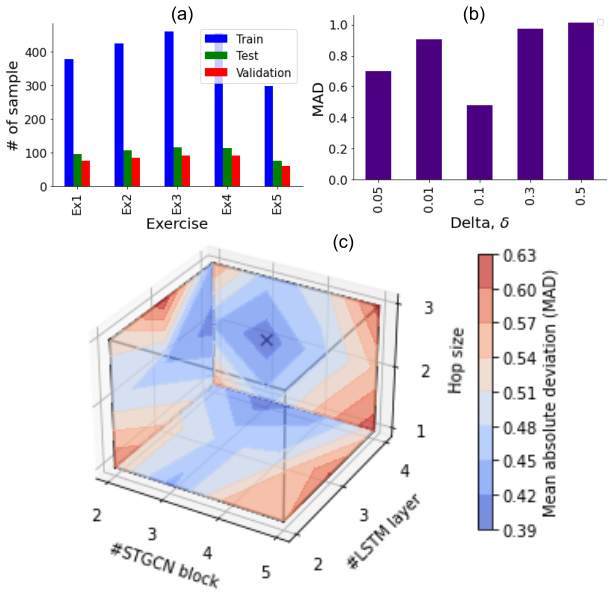

Fig. 4. (a) Train-test-val split on KIMORE dataset. (b) Performance of our proposed model on different values of δ. The optimal δ value is 0.1. (c) Hyper-parameter validation on KIMORE exercise 5 (in MAD). The cross ($\times$) indicates the optimum point.

B. Overall Result

We report our results on UI-PRMD and KIMORE datasets in Table IV and V, respectively. Following the work [1], the table mentions MAD performance on each of ten different exercises included in UI-PRMD dataset. Similarly, we present MAD, RMS, and MAPE results on five exercises of the KIMORE dataset. Liao *et al.* [1] proposed a temporal pyramid network to process the multiple-scale version of the movement repetitions. The initial hierarchical layers in the model employ for learning spatial dependencies in human movements and are followed by a series of LSTM recurrent layers for modeling temporal correlations in learned representations. Still, they ignored the human skeleton's topology information, failing to extract the expressive power residing in spatial features. On the other hand, Yan *et al.* [6] built a spatio-temporal graph neural network (STGCN) while neglecting the sequential

TABLE IV
RESULTS OF TEN EXERCISES (EX) ON THE UI-PRMD DATASET USING THE EVALUATION METRIC MAD (LOWER VALUES INDICATE BETTER RESULTS)

Ex	Ours	Song *et al.* [46]	Zhang *et al.* [48]	Liao *et al.* [1]	Li *et al.* [49]	Shahroudy *et al.* [50]	Du *et al.* [51]
Ex1	**0.009**	0.011	0.022	0.011	0.011	0.018	0.030
Ex2	**0.006**	**0.006**	0.008	0.028	0.029	0.044	0.077
Ex3	0.013	**0.010**	0.016	0.039	0.056	0.081	0.137
Ex4	**0.006**	0.014	0.016	0.012	0.014	0.024	0.036
Ex5	**0.008**	0.013	0.008	0.019	0.017	0.032	0.064
Ex6	**0.006**	0.009	0.008	0.018	0.019	0.034	0.047
Ex7	**0.011**	0.017	0.021	0.038	0.027	0.049	0.193
Ex8	**0.016**	0.017	0.025	0.023	0.025	0.051	0.073
Ex9	**0.008**	**0.008**	0.027	0.023	0.027	0.043	0.065
Ex10	**0.031**	0.038	0.066	0.042	0.047	0.077	0.160

TABLE V
RESULTS OF FIVE EXERCISES (EX) ON KIMORE USING THE EVALUATION METRICS MAD, RMS, AND MAPE (LOWER VALUES INDICATE BETTER RESULTS)

Metric	Ex	Ours	Song *et al.* [46]	Zhang *et al.* [48]	Liao *et al.* [1]	Yan *et al.* [6]	Li *et al.* [49]	Du *et al.* [51]
MAD	Ex1	**0.799**	0.977	1.757	1.141	0.889	1.378	1.271
	Ex2	**0.774**	1.282	3.139	1.528	2.096	1.877	2.199
	Ex3	**0.369**	1.105	1.737	0.845	0.604	1.452	1.123
	Ex4	**0.347**	0.715	1.202	0.468	0.842	0.675	0.880
	Ex5	**0.621**	1.536	1.853	0.847	1.2184	1.662	1.864
RMS	Ex1	2.024	2.165	2.916	2.534	**2.017**	2.344	2.440
	Ex2	**2.120**	3.345	4.140	3.738	3.262	2.823	4.297
	Ex3	**0.556**	1.929	2.615	1.561	0.799	2.004	1.925
	Ex4	**0.644**	2.018	1.836	0.792	1.331	1.078	1.676
	Ex5	**1.181**	3.198	2.916	1.914	1.951	2.575	3.158
MAPE	Ex1	**1.926**	2.605	5.054	2.589	2.339	3.491	3.228
	Ex2	**1.272**	3.296	10.436	3.976	6.136	5.298	6.001
	Ex3	**0.728**	2.968	5.774	2.023	1.727	4.188	3.421
	Ex4	**0.824**	2.152	3.901	2.333	2.325	1.976	2.584
	Ex5	**1.591**	4.959	6.531	2.312	3.802	5.752	5.620

nature of the spatio-temporal features because of using global average pooling. Moreover, Song *et al.* [46] also proposed a GCN to explore discriminative features that spread over all skeleton joints using multi-stream information (occlusion, jittering, etc.) and focus on the spatial context to learn the joint attention. Zhang *et al.* [48] incorporated high-level semantics of joints (joint type and frame index) into the network with the help of joint- and frame-level modules to hierarchically exploit the joint relationship. However, they fail to capture strong sequential dependencies between consecutive frames through several spatial and temporal maxpooling layers. Du *et al.* [51] did use temporal information but ignored the spatial information. Li *et al.* [49] ignored both the topological structure and the sequential nature of human body-joint features. Our model successfully outperforms existing methods in the rehabilitation exercise assessment task on both datasets. This success becomes possible because of the spatio temporal graph network with learned anisotropic filters by the self-attention mechanism that separately considers spatial and temporal directions. Moreover, our model fully utilizes the benefit of deep learning techniques, ensuring end-to-end learning. For [6], [46], [48], [49], [51], we replaced the last softmax layer with a fully-connected layer with linear

TABLE VI
ABLATION STUDY ON EX5 (EXERCISE 5) OF KIMORE DATASET. WE INCREMENTALLY ADD MORE COMPONENTS TO COMPARE THE PERFORMANCE (LOWER VALUES INDICATE BETTER RESULT)

Is-Stacked	Aggregation Style	Has-TCN Concatenated	Has-self attention	MAD	RMSE	MAPE
No	Global Pool	No	No	2.585	3.795	8.920
Yes	Global Pool	No	No	1.472	2.560	4.878
Yes	Global Pool	Yes	No	1.365	2.184	4.320
Yes	LSTM	Yes	No	0.767	1.484	2.340
Yes	LSTM	Yes	Yes	**0.478**	**0.981**	**1.516**

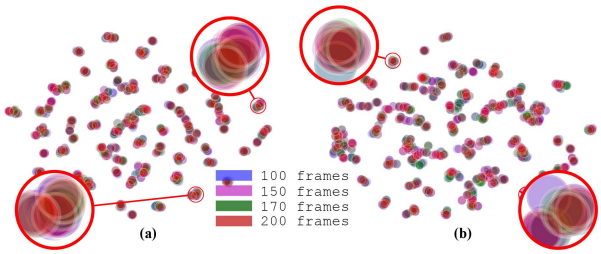

Fig. 5. t-SNE [52] visualization of features for two exercise videos (Ex2 and Ex3) from KIMORE dataset by representing exercises in different lengths. Our model extracts similar features from different lengths of the same video.

activations. Other than that, we closely followed the proposed implementation as described by the authors in the respective papers.

1) Ablation Study: In Table VI, we report different variations of GCN assessing rehabilitation exercises. We include or replace different components to estimate the contribution of them. Here, we experiment with and without stacked STGCN, spatio-temporal feature aggregation style (global pool/LSTM), the inclusion of concatenated TCN outputs, and self-attention (ConvLSTM based joint role) components. First, we compare a plain (without stacked) GCN vs. stacked GCN using global pooling as feature aggregation and without using any other parts. We notice performance improved in the stacked case because of extracting more complex features. On top of the stacked version, we add TCN feature concatenations mentioned in Sec. III-A. It helps to improve performance because of considering temporal information. Results further improve while replacing the global pool with LSTM (see Sec. III-B), since it augments more spatio-temporal information. Finally, we include the self-attention mechanism using ConvLSTM, which is our final recommendation. We achieve the best results in this configuration because of dynamically calculating the role of different joints while assessing exercises.

2) Feature Visualization: After training our final model with variable-length data, we also test our model with fixed-length input. We create four fixed-length versions (100, 150, 170 and 200 frames) of the same test data. These fixed-length versions represent rehabilitation exercises performed at different (slow/fast) pace and varying repetitions. We extract the latent feature representation after the LSTM layer for both variable-length and fixed-length test data as input. Fig. 5 shows those features using the 2D t-SNE plot of several

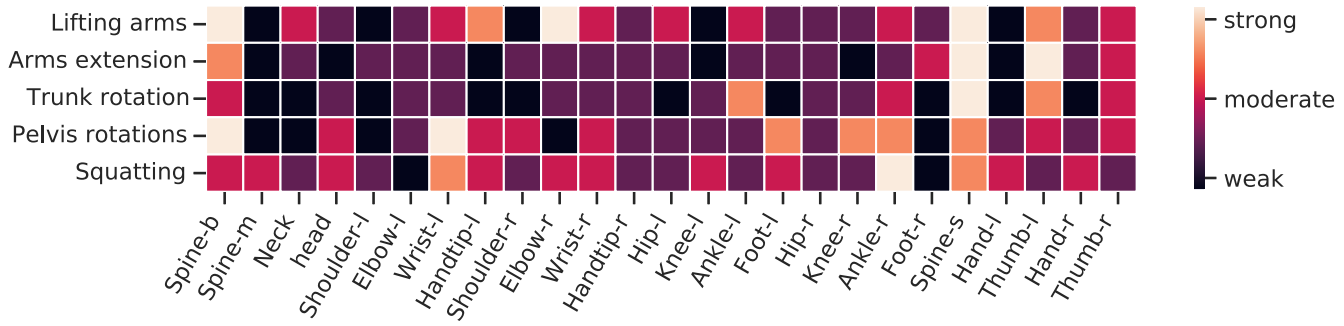

Fig. 6. A visualization of how the role of joints varies with different exercises as determined by the attention-value computed by our method. For example, in the lifting arm exercise (Ex 1), the elbow, spine, thumb, and wrist play a more important (strong/moderate) role than the rest of the joints. Similarly, in pelvis rotation (Ex 4), spine base and wrist contribute more than other joints. These findings closely align with the KIMORE dataset paper [16] that discusses the variation of role of joints in individual exercises as suggested by medical professionals.

TABLE VII
RESULTS FOR EVALUATING EXERCISE 5 OF THE KIMORE DATASET WITH MAD, RMS, AND MAPE WHILE COLLECTING JOINT/SKELETON DATA USING VARIOUS POSE ESTIMATION ALGORITHMS

Metric	Algorithm	Ours	Liao *et al.* [1]	Yan *et al.* [6]	Li *et al.* [49]	Du *et al.* [51]
MAD	BlazePose [12]	**0.971**	4.043	3.709	4.548	6.309
	VideoPose3D [13]	**1.855**	2.554	3.084	3.546	4.669
	Kinectv2 [11]	**0.621**	0.847	1.218	1.663	1.864
RMS	BlazePose [12]	**1.993**	5.991	5.657	7.194	8.681
	VideoPose3D [13]	**3.822**	3.908	4.943	5.202	6.012
	Kinectv2 [11]	**1.180**	1.914	1.951	2.575	3.158
MAPE	BlazePose [12]	**3.081**	15.618	15.917	20.897	25.816
	VideoPose3D [13]	**6.810**	8.102	10.790	11.964	14.750
	Kinectv2 [11]	**1.591**	2.312	3.802	5.752	5.620

exercises from the KIMORE dataset. One can notice that our model provides similar feature representations for different input lengths. It tells that our model can successfully assess physical rehabilitation exercises no matter how many repetitions or how slowly users perform the movement.

3) Effect of Joints in Different Exercises: By analyzing the natural topological structure of the human body and extracting spatial information using attention-guided graph convolution, we treat each joint differently based on spatial and temporal context. It produces anisotropic filters that are more powerful than isotropic filters used in vanilla STGCN. In Fig. 6, we show the effect of different joints (from expert users) on individual exercises belonging to the KIMORE dataset.

4) Result Using RGB Camera: The KIMORE and UI-PRMD datasets collect joint positions using Microsoft Kinectv2 [11] and Vicon sensors. However, rather than solely based on RGBD sensors, we also experiment with economically available RGB videos. We use pre-trained pose estimation algorithms to extract the 3D pose of a human being. The KIMORE dataset [16] provides RGB videos of patients performing rehabilitation exercises. We validate the performance of our model with two other pose estimator methods, namely BlazePose [12] and VideoPose3D [13] trained on MS Coco [53] and Human3.6M [54] datasets, respectively. Both ways use RGB information to estimate the 3D pose of a movement. In Table VII, we compare the performance of our model with different pose estimation algorithms. Since the Microsoft Kinectv2 sensor uses RGBD information to detect human poses, it outperforms the other two pose estimation algorithms based on RGB cameras.

V. CONCLUSION

In this paper, we propose attention guided GCN for assessing physical rehabilitation exercises. Our model takes skeleton data of human movement (represented as graph) as input and predicts an assessment score indicating the quality of the performed exercise compared to the prescribed version. We modify the popular STGCN architecture to adapt it to our regression-based problem setting. Our proposed network carefully extracts discriminative spatio-temporal features to facilitate variable-length exercise data. Besides, we propose a self-attention mechanism to attend body joints differently for various exercises. It also helps to guide users on which body joints to emphasize to achieve a better assessment score. Our model provides state-of-the-art performances on two well-known physical rehabilitation datasets, KIMORE and UI-PRMD. In addition to quantitative results, we present qualitative illustrations (as guidance) to visualize the reasoning about the predicted assessment score.

REFERENCES

[1] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 468–477, Feb. 2020.
[2] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. I. Badia, "Learning to assess the quality of stroke rehabilitation exercises," in *Proc. 24th Int. Conf. Intell. User Interface*, Mar. 2019, pp. 218–228.
[3] T. Hamaguchi *et al.*, "Support vector machine-based classifier for the assessment of finger movement of stroke patients undergoing rehabilitation," *J. Med. Biol. Eng.*, vol. 40, no. 1, pp. 91–100, Feb. 2020.
[4] B. Pogorelc, Z. Bosnić, and M. Gams, "Automatic recognition of gait-related health problems in the elderly using machine learning," *Multimedia Tools Appl.*, vol. 58, no. 2, pp. 333–354, May 2012.
[5] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018, pp. 3634–3640.
[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–10.

[7] J.-H. Kim, G.-S. Hong, B.-G. Kim, and D. P. Dogra, "DeepGesture: Deep learning-based gesture recognition scheme using motion sensors," *Displays*, vol. 55, pp. 38–45, Dec. 2018.

[8] Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, "Residual-based graph convolutional network for emotion recognition in conversation for smart Internet of Things," *Big Data*, vol. 9, no. 4, pp. 279–288, Aug. 2021.

[9] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 9532–9545, 2020.

[10] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3289–3299.

[11] J. Shotton *et al.*, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

[12] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," in *Proc. CVPR Workshop Comput. Vis. Augmented Virtual Reality*, 2020, pp. 1–4.

[13] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7753–7762.

[14] Y. Gu, H. Zhang, and S. Kamijo, "Multi-person pose estimation using an orientation and occlusion aware deep learning network," *Sensors*, vol. 20, no. 6, p. 1593, Mar. 2020.

[15] J. Wang *et al.*, "Deep 3D human pose estimation: A review," *Comput. Vis. Image Understand.*, vol. 210, Sep. 2021, Art. no. 103225.

[16] M. Capecci *et al.*, "The KIMORE dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 7, pp. 1436–1448, Jul. 2019.

[17] A. Vakanski, H.-P. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, p. 2, Jan. 2018.

[18] B. Artacho and A. Savakis, "UniPose: Unified human pose estimation in single images and videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7035–7044.

[19] K. Thakkar and P. J. Narayanan, "Part-based graph convolutional network for action recognition," 2018, *arXiv:1809.04983*.

[20] J. Marin, J. J. Marin, T. Blanco, J. de la Torre, I. Salcedo, and E. Martitegui, "Is my patient improving? Individualized gait analysis in rehabilitation," *Appl. Sci.*, vol. 10, no. 23, p. 8558, Nov. 2020.

[21] F. M. Carrillo *et al.*, "Adapting a general-purpose social robot for paediatric rehabilitation through *in situ* design," *ACM Trans. Hum.-Robot Interact.*, vol. 7, no. 1, pp. 1–30, May 2018.

[22] K. J. Bower, J. Louie, Y. Landesrocha, P. Seedy, A. Gorelik, and J. Bernhardt, "Clinical feasibility of interactive motion-controlled games for stroke rehabilitation," *J. neuroeng. Rehabil.*, vol. 12, no. 1, p. 63, Dec. 2015.

[23] S. Das *et al.*, "Quantitative measurement of motor symptoms in Parkinson's disease: A study with full-body motion capture data," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 6789–6792.

[24] T. T. Um, V. Babakeshizadeh, and D. Kulic, "Exercise motion classification from large-scale wearable sensor data using convolutional neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 2385–2390.

[25] T. Hussain, H. F. Maqbool, N. Iqbal, M. Khan, Salman, and A. A. Dehghani-Sanij, "Computational model for the recognition of lower limb movement using wearable gyroscope sensor," *Int. J. Sensor Netw.*, vol. 30, no. 1, pp. 35–45, 2019.

[26] S. Cai *et al.*, "Detecting compensatory movements of stroke survivors using pressure distribution data and machine learning algorithms," *J. Neuroeng. Rehabil.*, vol. 16, no. 1, pp. 1–11, Dec. 2019.

[27] S. Patel *et al.*, "A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology," *Proc. IEEE*, vol. 98, no. 3, pp. 450–461, Mar. 2010.

[28] X. Z. Ying, M. Lukasik, M. H. Li, E. Dolatabadi, R. H. Wang, and B. Taati, "Automatic detection of compensation during robotic stroke rehabilitation therapy," *IEEE J. Transl. Eng. Health Med.*, vol. 6, pp. 1–7, 2018.

[29] R. Houmanfar, M. Karg, and D. Kulic, "Movement analysis of rehabilitation exercises: Distance metrics for measuring patient progress," *IEEE Syst. J.*, vol. 10, no. 3, pp. 1014–1025, Sep. 2016.

[30] C.-J. Su, C.-Y. Chiang, and J.-Y. Huang, "Kinect-enabled home-based rehabilitation system using dynamic time warping and fuzzy logic," *Appl. Soft Comput.*, vol. 22, pp. 652–666, Sep. 2014.

[31] Z. Zhang, Q. Fang, and X. Gu, "Objective assessment of upper-limb mobility for poststroke rehabilitation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 859–868, Apr. 2016.

[32] A. Goñi, A. Illarramendi, and D. Antón, "Exercise recognition for Kinect-based telerehabilitation," *Methods Inf. Med.*, vol. 54, no. 2, pp. 145–155, 2015.

[33] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 1, pp. 43–49, Feb. 1978.

[34] M. Capecci *et al.*, "A hidden semi-Markov model based approach for rehabilitation exercise assessment," *J. Biomed. Inform.*, vol. 78, pp. 1–11, Feb. 2018.

[35] J. F.-S. Lin, M. Karg, and D. Kulić, "Movement primitive segmentation for human motion modeling: A framework for analysis," *IEEE Trans. Hum.-Mach. Syst.*, vol. 46, no. 3, pp. 325–339, Jun. 2016.

[36] A. Vakanski, J. M. Ferguson, and S. Lee, "Mathematical modeling and evaluation of human motions in physical therapy using mixture density neural networks," *J. Physiotherapy Phys. Rehabil.*, vol. 1, no. 4, p. 118, 2016.

[37] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 103–118.

[38] M. Capecci *et al.*, "Accuracy evaluation of the Kinect v2 sensor during dynamic movements in a rehabilitation scenario," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 5409–5412.

[39] P. Merriaux, Y. Dupuis, R. Boutteau, P. Vasseur, and X. Savatier, "A study of Vicon system positioning performance," *Sensors*, vol. 17, no. 7, p. 1591, Jul. 2017.

[40] R. Khirodkar, V. Chari, A. Agrawal, and A. Tyagi, "Multi-instance pose networks: Rethinking top-down pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3122–3131.

[41] M. A. Fisch and R. Clark, "Orientation keypoints for 6D human pose estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 16, 2021, doi: 10.1109/TPAMI.2021.3136136.

[42] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[43] Z. Zhang, Q. Fang, L. Wang, and P. Barrett, "Template matching based motion classification for unsupervised post-stroke rehabilitation," in *Proc. Int. Symp. Bioelectron. Bioinf. (ISBB)*, Nov. 2011, pp. 199–202.

[44] I. Ar and Y. S. Akgul, "A computerized recognition system for the home-based physiotherapy exercises using an RGBD camera," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 6, pp. 1160–1171, Nov. 2014.

[45] J. Cavazza and V. Murino, "Active regression with adaptive Huber loss," 2016, *arXiv:1606.01568*.

[46] Y. F. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1915–1925, May 2021.

[47] J. Wang, S. Jin, W. Liu, W. Liu, C. Qian, and P. Luo, "When human pose estimation meets robustness: Adversarial algorithms and benchmarks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11855–11864.

[48] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1112–1121.

[49] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 786–792.

[50] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[51] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[52] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Jan. 2014.

[53] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[54] 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, Dec. 2013.