Sparse Adaptive Graph Convolutional Network for Leg Agility Assessment in Parkinson's Disease

Rui Guo, Xiangxin Shao, Chencheng Zhang, and Xiaohua Qian

***Abstract*— Motor disorder is a typical symptom of Parkinson's disease (PD). Neurologists assess the severity of PD motor symptoms using the clinical rating scale, i.e., MDS-UPDRS. However, this assessment method is time-consuming and easily affected by the perception difference of assessors. In the recent outbreak of coronavirus disease 2019, telemedicine for PD has become extremely urgent for clinical practice. To solve these problems, we developed an automated and objective assessment method of the leg agility task in the MDS-UPDRS using videos and a graph neural network. In this study, a sparse adaptive graph convolutional network (SA-GCN) was proposed to achieve fine-grained quantitative assessment of skeleton sequences extracted from videos. Specifically, the sparse adaptive graph convolutional unit with a prior knowledge constraint was proposed to perform adaptive spatial modeling of physical and logical dependency for skeleton sequences, thus achieving the sparse modeling of the discriminative spatial relationships. Subsequently, a temporal context module was introduced to construct the remote context dependency in the temporal dimension, hence determining the global changes of the task. A multi-domain attention learning module was also developed to integrate the static spatial features and dynamic temporal features, and then to emphasize the salient feature selection in the channel domain, thereby capturing the multi-domain fine-grained information. Finally, the evaluation results using a dataset with 148 patients and 870 samples confirmed the effectiveness and reliability of our scheme, and the method outperformed other related state-of-the-art methods. Our contactless method provides a new potential tool for automated PD assessment and telemedicine.**

***Index Terms*— Parkinson's disease, leg agility, video-based assessment, multi-domain attention learning, sparse adaptive graph convolution.**

I. Introduction

PARKINSON'S disease (PD) is the second most common chronic neurodegenerative disorder in the world. With the increasing aging population, PD imposes a huge socioeconomic burden [1]. PD leads to the gradual decline of patients' motor ability, characterized by tremor at rest, rigidity, akinesia (or bradykinesia), and postural instability [2]. The motor ability assessment of PD patients is essential for clinical diagnosis and treatment. Neurologists generally utilize the Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [3] to evaluate PD patients' motor symptoms. Leg agility (LA) assessment is one of the vital parts of the MDS-UPDRS. In this assessment, the neurologist requires the patient to raise each leg from the ground with the maximum amplitude and speed and then scores them 0–4 according to the patient's movement speed, amplitude, hesitation and pause, gradual reduction of amplitude, and other factors. A previous study [4] has shown that objective LA action parameters could predict the severity of PD well. Thus, LA is crucial in the assessment of PD, especially in the assessment of the severity of the motor disorder in lower limbs.

However, there are many limitations to the clinical practice of the MDS-UPDRS. First, although the action assessment in the MDS-UPDRS is direct and detailed, the quantitative assessment still depends on the clinical experience of neurologists, so subjectivity and inter-rater variability inevitably exist. In addition, a complete assessment usually takes more than half an hour to finish. This time-consuming assessment imposes a substantial burden on clinicians as the number of PD patients increases. Second, the scarcity of neurologists in the movement disorder and neurodegenerative disorder fields make it difficult for performing regular follow-ups to monitor PD patients' disease status.

The automated assessment of PD motor symptoms provides a new solution to tackle the problems mentioned above. It not only achieves objective assessment to avoid inconsistency of the ratings of neurologists but also is conducive to realize PD telemedicine. The automated assessment of PD is notably essential in public health emergencies. For example, to prevent further outbreak and development of the coronavirus disease 2019 (COVID-19), many regions in the world have implemented traffic control policies; this makes it difficult for PD patients to receive timely diagnosis or follow-up. This case underlines the urgency of automated PD motor assessment systems. In this study, we plan to realize the automated assessment of LA in PD patients.

The automated PD motor assessment is mainly based on the strong correlation between kinematic features and symptom

Manuscript received June 21, 2020; revised November 11, 2020; accepted November 11, 2020. Date of publication November 19, 2020; date of current version January 29, 2021. *(Corresponding author: Xiaohua Qian.)*

Rui Guo and Xiaohua Qian are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: xiaohua.qian@sjtu.edu.cn).

Xiangxin Shao is with the School of Electrical and Electronic Engineering, Changchun University of Technology, Changchun 130012, China.

Chencheng Zhang is with the Department of Functional Neurosurgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China.

Digital Object Identifier 10.1109/TNSRE.2020.3039297

1558-0210 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

severity [5]. Currently, most studies mainly applied wearable sensors and optical devices to extract kinematic parameters [6]–[9]. Patel *et al.* [6] used the accelerometer data features and support vector machine (SVM) to assess the severity of tremor, bradykinesia, and dyskinesia. Ramsperger *et al.* [7] utilized the data measured from different movement stages using a single inertial sensor on the patient's ankle for the evaluation of the severity of the patient's leg dyskinesia. Bank *et al.* [8] applied the optical hand tracking technology to obtain the real-time coordinates of the wrist, finger joints, and fingertips for calculating the kinematic variables. They then performed the intra-class correlation analysis, showing the potential of quantifying the components of bradykinesia in PD. Ferraris *et al.* [9] tracked the movements of hands and the body through the skeleton tracking function of an optical RGB depth device, and calculated the relevant kinematic parameters, followed by the quantitative assessment of three upper limb tasks and two lower limb tasks by SVM.

With the rapid development of deep learning technology, some researchers have explored the application of the deep learning-based pose estimation model in the assessment of PD videos. The joint coordinates are extracted to calculate the kinematic parameters from the motion trajectories. For example, Li *et al.* [4] applied the convolutional pose machines (CPM) for four different scale tasks to extract the motion trajectories of joints from videos and calculated the relevant feature parameters. Then, the random forest method was used to detect the pathological motion and predict the clinical grade of PD or levodopa-induced dyskinesia (LID). After that, Li *et al.* [10] proposed a video-based system to quantify the important changes of patients in dyskinesia during acute levodopa infusions, and finally concluded that video-based features may show good responsiveness of patient reported changes in dyskinesia. Liu *et al.* [11] estimated the hand poses and extracted the kinematic features and then applied the SVM to generate the score level of three tasks related to the upper limb bradykinesia. Recently, Sabo *et al.* [12] utilized 3D joint coordinates provided by the Microsoft Kinect sensor and 2D joint coordinates extracted from recorded videos to calculate gait features. Subsequently, through regression analysis, they concluded that the extracted gait features were correlated to the severity scores of PD gait.

In the automated assessment of the LA task in PD patients, existing studies also mainly use the three types of methods described above; that is, methods based on wearable sensors, optical devices, and videos. Giuberti *et al.* [13], [14] utilized the human body sensor network (BSN) composed of wearable wireless inertial nodes to extract the relevant kinematic features in the time-frequency domain. They obtained accuracies of approximately 40% based on time-domain features and 50% based on time- and frequency-domain features through the k-nearest neighbor (kNN) classifier. Paris *et al.* [15] extracted the kinematic feature set in the time-frequency domain based on the BSN. Then, principal component analysis and the kNN were used to achieve an accuracy of 43%. Ferraris *et al.* [9] calculated kinematic parameters such as angle, speed, and time through the joint coordinates of the hip, knee, and ankle provided by the optical RGB depth device Microsoft Kinect SDK, and finally achieved 60% assessment accuracy by SVM. Li *et al.* [4] used the CPM to extract joint coordinates from assessment videos and calculated 32 features, that is, 15 kinematic features, 16 spectrum features, and a convex hull for each joint trajectory. Regression analysis was performed through a random forest to predict the clinical grade of symptom severity, and a Pearson correlation coefficient of 0.618 was obtained.

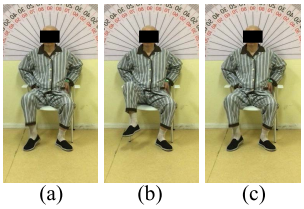

Fig. 1. Process diagram of completing an LA action.

However, these methods for PD LA assessment based on wearable sensors or optical devices are still problematic to be popularized in clinical practice, especially in telemedicine. The wearable sensors make physical contact with the patients' bodies, thus unavoidably interfering with the patients' movements. To ensure the continuous accuracy and uniformity of measurement, the sensors and optical devices must be calibrated regularly, which causes trouble for PD patients with mobility difficulties. However, as shown in Fig. 1, the vision-based assessment requires a device with an ordinary camera, such as a smartphone, to capture the video data. This scheme is contactless and is easy to be propagated in clinical application and telemedicine. Although the LA task assessment based on videos [4] was proposed, the entire discriminative information is challenging to be obtained by manual feature engineering. Deep learning technology [16] allows the model to learn data representation with multiple levels and can extract sophisticated discriminative features, thus providing a new solution for the vision-based automated quantitative assessment of the PD LA task from videos.

In human action recognition, such as in the PD LA task assessment, dynamic human skeleton conveys critical action features, effectively improving the performance and robustness of the action recognition system. Previous modeling methods [17]–[19] mostly relied on components or rules defined manually to analyze the spatial patterns of the dynamic skeleton, whereas the skeleton shape is not regular grids, hence making it challenging to capture the features embedded in the skeleton spatial topology graph automatically. Yan *et al.* [20] first applied a graph convolutional network (GCN) to skeleton-based action recognition, and extended the GCN to a spatial-temporal graph convolutional network (ST-GCN), providing a paradigm of modeling the dynamic skeleton based on the GCN. Therefore, in this study, we will extend the ST-GCN framework for automated quantitative assessment of the LA task in PD patients based on videos.

However, there are still three challenges in the quantitative assessment of the LA task through deep learning technology based on the GCN:

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

1) In the LA task, it is necessary to consider the physical and logical dependency of human joints simultaneously. The logical dependency between the left and right legs is very important, especially the dependency between the left and right ankles, which is the most important spatial structure relationship in the LA task. However, the existing skeleton-based action recognition methods mainly focus on the adaptability of physical dependency and ignore the adaptive modeling of logical dependency; moreover, they lack the exploration of discriminative features.

2) Neurologists provide scores by considering all the changes in patients' LA action execution, that is, assessment globally depends on the remote dependency between different time-points in videos. Nevertheless, most of the previous action recognition studies focus on the selection of discriminative periods, while few consider the contextual relationship of the time series.

3) In the video analysis of the LA task, the importance of coordinate changes in different directions is diverse. However, the conventional feature extraction strategies apply the attention mechanism to enhance the saliency regions in space and time domain, thus neglecting the enhancement of important features in the channel (i.e., feature) domain.

Therefore, we designed a novel graph convolutional unit to aggregate physical and logical dependency adaptively and combined it with a sparsification strategy to extract the discriminative features from the graph. The significant feature regions were further enhanced by constructing the context dependency of time series and integrating the multi-domain salient features. Overall, we proposed a sparse adaptive graph convolutional network (SA-GCN) to realize automated quantitative assessment of the LA task in PD patients. Specifically, 1) The joint sequences of PD patients were extracted from videos by a state-of-the-art human pose estimation model. 2) A sparse adaptive graph convolutional unit (SAGCU) was developed to realize the spatial modeling of joint and joint-motion spatial-temporal graphs. Two types of spatial structure relationships, i.e., physical and logical dependency, were proposed to encode the joint connections adaptively. The sparsification constraint was then embedded into the cost function to determine the discriminative features. 3) A temporal context module (TCM) was introduced to construct the context dependency of video sequences by calculating the correlation of time positions, with the purpose of capturing the remote dependency in the temporal dimension. 4) A multi-domain attention learning module (MDALM) was developed. The high-level spatial-temporal features were used to guide low-level features to enhance the salient features in the channel domain, hence achieving the feature integration of time, space, and channel.

Further, the contributions of this study are summarized as follows:

1) A sparse adaptive graph convolutional unit (SAGCU) was proposed to combine the physical dependency and logical dependency for exploring the most important spatial structure relationships in the LA action.

2) A temporal context module (TCM) was introduced to capture the remote dependency in the temporal dimension for obtaining the global changes of the same joint in the process of action execution, thereby avoiding the localization of temporal features.

3) A multi-domain (i.e., time, space, and channel) attention learning module (MDALM) was designed to integrate the critical information in the time, space, and channel domains, thereby mining the significant features in multi-domains.

In a word, in the proposed method, the LA task can be assessed using videos recorded by ordinary cameras, thus providing a potential tool for the PD automated assessment system and telemedicine. In addition, the SAGCU, TCM, and MDALM can not only solve the challenges in our task but also provide a paradigm for GCNs, video analysis, and skeleton-based action recognition.

II. Related Work

A. Application of the Graph Neural Network in Skeleton-Based Human Action Recognition

The dynamic skeleton sequence represented by joint coordinates can be obtained through sensors or pose estimation algorithms; then, human action recognition can be realized by analyzing the sequence.

The previous skeleton-based action recognition [17]–[19] mostly relied on the analysis of human spatial patterns through pre-defined rules. For example, Du *et al.* [17] proposed an end-to-end hierarchical recurrent neural network (RNN). The physical structure of human body was divided into five parts, which were fed into five subnets, respectively, thus achieving hierarchical fusion and obtaining a higher-level input representation. Liu *et al.* [18] proposed a tree-structure-based traversal approach and then introduced a novel gating mechanism with a long short-term memory (LSTM) network.

Human skeleton can be simplified as a graph composing points (i.e., joints) and edges (i.e., bones), which is an irregular grid structure. The above methods are difficult to model the natural spatial connectivity of the topological graph of the human skeleton dynamically, such as the information exchange between different joints. For this reason, Yan *et al.* [20] first used the GCN to construct the human skeleton sequence as a spatial-temporal graph. They proposed the spatial interaction of graph convolution to implicitly learn the spatial connectivity information of the human skeleton. Further, the temporal dynamic information was overlaid through conventional convolutional operation; thus, better action representation was learned effectively. Tang *et al.* [21] extracted keyframes through a deep progressive reinforcement learning method. A pre-defined weighted adjacency matrix was used to present the relationships between joints, including physical connections and physical disconnections. Then, the spatial dependency between joints was constructed by graph convolution. Thakkar and Narayanan [22] divided the human skeleton into four subgraphs. Then, the part-based GCN was developed, and the relative coordinates and temporal displacements of joints were introduced to improve the performance. Shi *et al.* [23] designed a new adjacency matrix strategy to learn the graph structure of different network layers and skeleton samples adaptively. This data-driven method improved the flexibility

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

of graph construction. Finally, the joint and bone information was modeled simultaneously by a two-stream network. Shi *et al.* [24] expressed the human skeleton as a directed acyclic graph based on the motion dependency of human joints and bones, and proposed a novel directed graph neural network to model spatial information and motion information. Wen *et al.* [25] designed a GCN based on motif to encode the hierarchical spatial structure, which allowed the modeling of the physical connection and disconnection of joints simultaneously.

The methods mentioned above achieved good performance in skeleton-based action recognition; however, the logical dependency between joints is fixed in these methods, and graph features are redundant for action recognition. In this paper, we proposed an adaptive graph convolution to model the physical and logical dependency of joints adaptively and introduced a sparsification strategy in the cost function to extract discriminative graph features.

B. Spatial-Temporal Relationship Modeling for Video-Based Action Recognition

In the video-based human action recognition, modeling the spatial-temporal relationship and extracting discriminative spatial-temporal features play a pivotal role. Most methods [26]–[28] enhanced the spatial-temporal relationship by designing attention modules in the action recognition task based on RGB frames or optical flow. Du *et al.* [26] enhanced the LSTM network with a novel spatial-temporal attention module to leverage the global video context for closely learning the significant features related to the current frame. Li *et al.* [27] designed a general attention neuron to estimate the attention probability of both spatial location and video clips in the time series. Yu *et al.* [28] proposed a spatial attention module to focus on the spatial salient features and introduced a bidirectional LSTM-based temporal attention module to focus on vital video cubes.

In the skeleton-based action recognition tasks, some studies [29], [30] used an attention module to improve the original LSTM network to focus on the discriminative joints in each frame; by contrast, most studies [20], [22], [23] adopted the ST-GCN as the basic framework. Graph convolution was introduced to model spatial features, and conventional convolution was used to capture dynamic temporal information. Based on the ST-GCN, Wu *et al.* [31] further proposed cross-domain spatial residual layers to effectively capture spatial-temporal information, and introduced dense connection blocks to improve the robustness of features. Wen *et al.* [25] proposed variable temporal dense blocks with different kernel sizes to extract temporal features in different ranges.

Although these studies proposed various modeling methods for the spatial-temporal relationship of video sequences, they always extracted the discriminative features in the spatial and temporal domains and ignored the importance of different coordinate directions in the channel domain to the task of action recognition. Furthermore, they often focused on the selection of significant periods in the time domain, without too much consideration of global dependency. Therefore, we proposed the MDALM to integrate the salient feature information from the three domains (i.e., time, space, and channel), emphasizing the feature selection in the channel domain and the global dependency in the time domain.

III. Method

Fig. 2 presents the architecture of our proposed SA-GCN. First, the skeleton sequences of the human body are extracted from the assessment videos of the LA task through the state-of-the-art human pose estimation model, and the joint and joint-motion spatial-temporal graphs are constructed. Joint and joint-motion coordinates serve as the input to the two-stream network, and these initial coordinate features are first fused through a MDALM without the residual mechanism before entering the corresponding stream. Each stream is composed of 9 basic MDALMs with the residual mechanism. Each MDALM consists of three branches: the spatial-temporal feature extraction branch, composed of the SAGCU and TCM; the spatial learning branch, i.e., the SAGCU; and the adaptive channel-wise attention learning branch. The weight function in the last SAGCU is constrained through the sparsification term in the cost function to realize the sparsification of the graph features. Finally, the output scores of the fully connected layers in the two-stream network are added with equal weights to obtain the assessment categories of the input videos. In the following subsections, we present a detailed description of our proposed approach.

A. Sparse Adaptive Graph Convolutional Unit (SAGCU)

1) Graph Construction: The physical connections between human joints, defined as physical dependency, are very important. Additionally, the relationships between joints without physical connections are vital for the LA task recognition. For example, the relationship between the left and right ankles is closely related to the dynamic characteristics of the LA task. These relationships are defined as logical dependency. Subsequently, the physical- and logical-dependency graphs are constructed simultaneously to model the human body's spatial structure. According to the motion characteristics of different body regions during the LA task and inspired by some previous skeleton-based action recognition work based on the division of human body parts [22], [32], the physical-dependency graph, consisting of three node types, i.e., the joints themselves, upper body region, and lower body region, is used to encode the physical connections of human joints, as shown in Fig. 3. This graph can be formally expressed by

$$G^{(1)} = \bigcup\nolimits_{i\in\{1,2,3\}} \mathcal{G}_i \,|\, \mathcal{G}_i = \left(\mathcal{V}_i, \mathcal{E}_i, X_i\right) \tag{1}$$

where $\mathcal{V}_i = \{v_1, v_2, \ldots, v_V\}$, and $\mathcal{E}_\mathrm{i}$ are the set of V nodes and the set of edges, respectively. $X \in \mathbb{R}^{V\times C}$ is the attribute matrix (i.e., the feature matrix) of V nodes. i represents different node types, which can be mapped through a type mapping function $\xi{:}\mathcal{V}_i \rightarrow \Theta_i, i = 1, 2, 3$. Θ_i is the set of node types. Inspired by [25], the logical-dependency graph is used to encode the logical connections between different joints by constructing a weighted adjacency matrix. Larger weight values are assigned to the connections formed by the nodes with a closer spatial

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

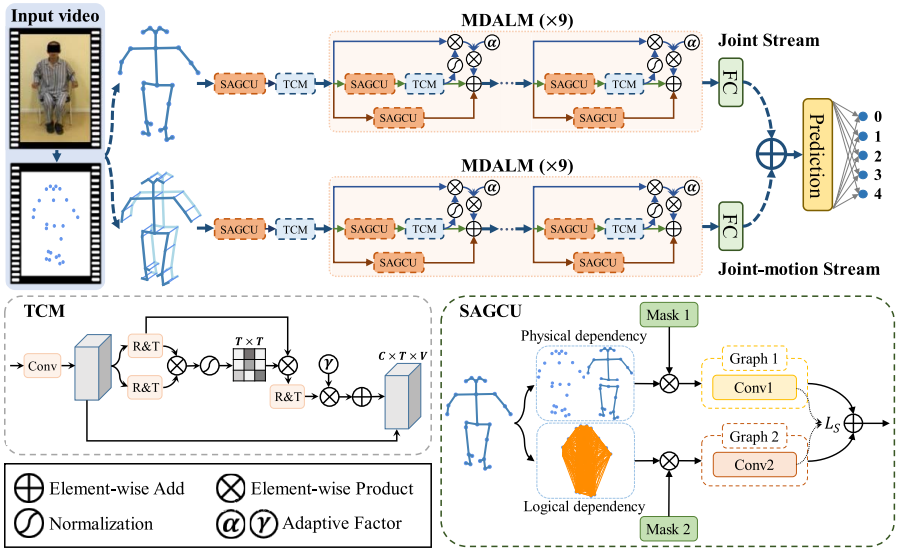

Fig. 2. Architecture of the proposed SA-GCN.

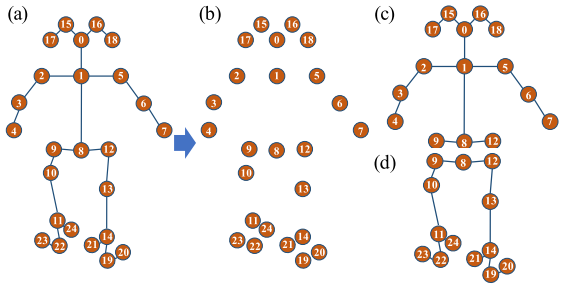

Fig. 3. Construction of physical dependency graph. (a) 25 joints of human body. (b)–(d) Joints themselves, upper body region, and lower body region, respectively, together constituting the physical dependency graph.

distance. The weight value between node i and j is defined as $\alpha_{i,j} = \max\{d\} - d_{i,j}$, where d is the average Euclidean distance of two nodes in the temporal dimension. Finally, the weighted adjacency matrix is normalized to reduce the deviation caused by nodes with high connections. Therefore, the logical-dependency graph can be formally expressed as

$$G^{(2)} = \mathcal{G}|\mathcal{G} = (\mathcal{V}, \mathcal{E}, X) \tag{2}$$

2) Sparse-Adaptive Graph Convolutional Operation: Inspired by the application of graph convolution in action recognition [20], we follow the method of spatial perspective to construct graph convolution on graphs. In the spatial dimension, the graph convolutional operation on node v_i can be expressed as follows:

$$f(v_i) = \sum_{j=1}^{V} \frac{1}{\Lambda_{ii}} (A_{ij} + I_{ij}) x_j w \tag{3}$$

where x is the input feature map, v is the node on the graph, and w is the weight function. $\Lambda_{ii} = \sum_j (A_{ij} + I_{ij})$ is the diagonal matrix to realize normalization, where A is the $V \times V$ adjacency matrix that defines the connection relationship between nodes, and I is the identity matrix that defines the self-connection of nodes. $f(v_i)$ is the graph convolutional operation result at node v_i. Then, Yan *et al.* [20] adopted the graph convolutional operation similar to that proposed in [33], and extended (3) to the input feature $X \in \mathbb{R}^{V \times C}$ with V nodes and C channels, which can be further expressed as

$$F = \Lambda^{-1/2}(A + I)\Lambda^{-1/2} X W \tag{4}$$

Given the two graphs $G^{(1)}$ and $G^{(2)}$ defined in the previous section, the graph convolutional operation in (4) is obviously not applicable to the case of multiple spatial structures and node types. Therefore, on the basis of [20], [34], we propose the adaptive graph convolutional unit (AGCU), which can be expressed mathematically as

$$f^M(v_i) = \sum_{j=1}^{V} \sum_{k=1}^{K_M} \left(\frac{1}{\Lambda_{kii}^M} A_{kij}^M\right) \otimes \mathcal{M}_{kij}^M x_j w_k^M \tag{5}$$

where $M \in \{1, 2\}$ represents the physical-dependency graph and logical-dependency graph, respectively, and K_M is the number of node types in graph M. $\mathcal{M}$ is the adaptive mask that is used to adaptively scale the importance of connections between different nodes under each node type. Consequently, the graph convolutional operation for graph M generates new node representations by assigning different adaptive weights to different node connections. Further, (5) can be extended to the feature map $X \in \mathbb{R}^{V \times C}$ with V nodes, which can be formulated as

$$F^M = \sum_{k=1}^{K_M} \left((\Lambda_k^M)^{-1/2} A_k^M (\Lambda_k^M)^{-1/2}\right) \otimes \mathcal{M}_k^M X W_k^M \tag{6}$$

The importance of connections between nodes in two different graphs also differs. In our implementation, since physical dependency is an inherent attribute of the human body, the adaptive mask $\mathcal{M}^1$ in the physical-dependency graph is initialized as an all-one matrix, and $\mathcal{M}^2$ in the logical-dependency graph is initialized as a matrix with values of 10^{-6}, such that the network can pay more attention to the basic physical dependency between joints. As training progresses, the network increases attention to the logical

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

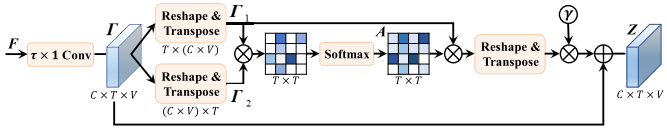

Fig. 4. Implementation of the TCM. γ is an adaptive weight factor.

dependency of the joints adaptively. Finally, the output of AGCU is calculated as follows:

$$F^M = \sum_{i=1}^{M} F^M = \sum_{i=1}^{M} \sum_{k=1}^{K_i} ((\Lambda_k^i)^{-1/2} A_k^i (\Lambda_k^i)^{-1/2}) \otimes \mathcal{M}_k^i X W_k^i \quad (7)$$

In addition, the graphs have corresponding attribute features of all nodes and adjacency matrices between nodes, but not all features and defined connections are meaningful. Thus, we propose the SAGCU. An L1 regularization term is added to the cost function to constrain the weight function W^M in (7) in the last SAGCU; thus, only meaningful sparse features in the graphs are selected, and the mining of discriminative spatial features of the graphs is effectively realized. This process can be expressed as

$$L_{sparse} = \sum_{i=1}^{M} \sum_{k=1}^{K_i} \left\| W_k^i \right\|_1 = \sum_{i=1}^{M} \sum_{k=1}^{K_i} \sum_{j=1}^{n_k} \left| w_{kj}^i \right| \quad (8)$$

B. Temporal Context Module (TCM)

To model the remote context dependency in the temporal dimension, inspired by dependency modeling in the semantic segmentation task in [35], we introduce the TCM. It encodes remote context dependency into local features, hence enhancing the feature representation ability. As shown in Fig. 4, the output F of the SAGCU is fed into the conventional $\tau \times 1$ temporal convolutional operation to obtain the local feature map $\Gamma \in \mathbb{R}^{C \times T \times V}$; then, Γ is reshaped and transposed as $\Gamma_1 \in \mathbb{R}^{T \times (C \times V)}$ and $\Gamma_2 \in \mathbb{R}^{(C \times V) \times T}$, respectively. Subsequently, the context dependency in the temporal dimension is acquired through matrix multiplication. Finally, the softmax is applied to gain the temporal context map $A \in \mathbb{R}^{T \times T}$ as follows:

$$a_{ji} = \exp(\Gamma_{1i} \cdot \Gamma_{2j}) \Big/ \sum_{i=1}^{T} \exp(\Gamma_{1i} \cdot \Gamma_{2j}) \quad (9)$$

where a_{ji} represents the correlation between the i^{th} and j^{th} time points. Afterwards, A is multiplied by the original local feature Γ_1; then, the result is scaled through the weight factor γ, which is initialized to 0, thereby acquiring the feature map with the temporal context response. Finally, the scaling result and Γ are summed to obtain the final output $Z \in \mathbb{R}^{C \times T \times V}$, as follows:

$$Z_j = \gamma \sum_{i=1}^{T} (a_{ji} \Gamma_{1i}) + \Gamma_j \quad (10)$$

Therefore, the TCM aggregates the remote context dependency adaptively according to the temporal context map; thus, the final output feature is the weighted sum of the original feature and the features of all the time points.

C. Multi-Domain Attention Learning Module (MDALM)

The conventional spatial-temporal graph convolutional unit has been proved to have the capability of internal self-attention, and the graph convolutional operation has a good ability of spatial feature extraction [20]. Therefore, we propose the MDALM. As shown in Fig. 2, it has three branches and can be regarded as a special three-stream structure. The middle branch extracts the dynamic spatial-temporal features by the SAGCU, followed by the TCM. The upper branch, that is, the adaptive channel-wise attention learning branch, uses the high-level spatial-temporal feature map as the attention mask of the low-level layer without any additional attention layer, hence enhancing the salient features in the channel domain. Inspired by [31], the lower branch, i.e., the spatial learning branch, learns the static spatial features through the SAGCU.

In the process of adaptive channel-wise attention learning, first, the output feature map of the high-level layer is normalized through softmax at every spatial-temporal position as follows:

$$B = \mathcal{N}(Z) = \{ b | b_{i,j}^c = \frac{\exp\left(Z_{i,j}^c\right)}{\sum_{c'} \exp\left(Z_{i,j}^{c'}\right)}, \quad i = 1, \ldots, T, \; j = 1, \ldots, V \} \quad (11)$$

where i and j represent the location index in the spatial-temporal dimension, and c represents the channel index of the high-level feature map Z. Thus, the salient features in the channel domain are enhanced. Then, B is used as the attention mask and the channel salient map is derived by multiplying B with the original feature X. Subsequently, the salient map is multiplied with a learnable adaptive factor α, which is initialized to 0 for avoiding the effect of negative attention at the beginning of training. During network training, the contribution of the channel salient map is constantly adjusted to achieve the balance between the attention branch and the other two branches.

Finally, the output of MDALM is represented through the element-wise addition result of three branches as follows:

$$O = \mathcal{T}\left(\sum_{i=1}^{M} F^M\right) + \sum_{i=1}^{M} F^M + \alpha \cdot \mathcal{N}\left(\mathcal{T}\left(\sum_{i=1}^{M} F^M\right)\right) \cdot X = \mathcal{T}(\mathcal{S}(X)) + \mathcal{S}(X) + \alpha \cdot \mathcal{N}(\mathcal{T}(\mathcal{S}(X))) \cdot X \quad (12)$$

where $\mathcal{S}(\cdot)$ represents the SAGCU, and $\mathcal{T}(\cdot)$ represents the TCM.

D. Model Architecture

Based on the skeleton sequences extracted from videos, joint and joint-motion spatial-temporal graphs are constructed, respectively, and can be represented as $\Omega_J = (V_J, E_J)$ and $\Omega_{Jm} = (V_{Jm}, E_{Jm})$, where $V_J = \{v_1, v_2, \ldots, v_V\}$ and $V_{Jm} = \{v_{t+1,1} - v_{t,1}, v_{t+1,2} - v_{t,2}, \ldots, v_{t+1,V} - v_{t,V} | t = 1, \ldots, T\}$ are the joints of the human body in all video frames. E_J and E_{Jm} contain not only the connections between the joints

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

defined by the two graphs in the spatial dimension but also the connections of the same joints in the temporal dimension. Then, the whole model is composed of the joint stream and joint-motion stream. Each stream is stacked with 9 MDALMs with the residual mechanism. Well-trained models often have stronger ability of feature extraction. Therefore, two adaptive masks of adjacency matrices and two adaptive factors of context or salient maps are introduced to guide the adaptive learning of the model.

The cost function of the model consists of three parts: L_{ce} is the cross-entropy term to minimize the classification error of the network; L_{sparse} is the L1 regularization term in the SAGCU to drive the model to learn the sparse discriminative features in the spatial dimension; L_{param} is the weight decay term to prevent the over-fitting problem. The cost function can be expressed as

$$\begin{aligned} L &= L_{ce} + L_{sparse} + L_{param} \\ &= -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} log\hat{y}_{i,k} \\ &\quad +\lambda_1 \sum_{i=1}^{M}\sum_{k=1}^{K_i}\left\|W_k^i\right\|_1 + \lambda_2 \|\theta\|_2 \end{aligned} \tag{13}$$

where N is the batch size. K is the number of categories in the dataset. y is the true label, and $\hat{y}_{i,k}$ indicates the probability that the i^{th} sample is predicted as the k^{th} category. λ_1 and λ_2 are the trade-off factors.

When using the gradient descent method to train the neural network, the back propagation process is needed to calculate the partial derivative of the cost function to the weight parameters. The cross-entropy loss function is differentiable and has been widely used in the classification task of deep learning [36]. Therefore, the following will prove that L_{sparse} and L_{param} are differentiable.

The partial derivatives of L_{sparse} and L_{param} with respect to W and θ are calculated as follows:

$$\begin{aligned} \frac{\partial L_{sparse}}{\partial W} &= \frac{\partial(\lambda_1 \sum_{i=1}^{M}\sum_{k=1}^{K_i}\left\|W_k^i\right\|_1)}{\partial W} \\ &= \lambda_1 \frac{\partial \sum_W |W|}{\partial W} = \lambda_1 sgn(W) \end{aligned} \tag{14}$$

$$\begin{aligned} \frac{\partial L_{param}}{\partial \theta} &= \frac{\partial(\lambda_2 \|\theta\|_2)}{\partial \theta} = \lambda_2 \frac{\partial(\sqrt{\sum_\theta \theta^2})}{\partial \theta} \\ &= 2\lambda_2 \frac{\theta}{\sqrt{\sum_\theta \theta^2}} \end{aligned} \tag{15}$$

In (14), the derivative of L_{sparse} is undefined at 0. However, if L_{sparse} is 0, the weights need not be updated, so the convention of $sgn\,(0) = 0$ is adopted. After calculating the derivatives of L_{sparse} and L_{param}, the rest of the back propagation process is the same as the conventional classification network.

To show the parameter updating process of the cost function, two regularization terms L_{sparse} and L_{param} are added to the original cross-entropy term, respectively, thus obtaining L_1 and L_2 as follows:

$$L_1 = L_{ce} + \lambda_1 \sum_W |W| \tag{16}$$

$$L_2 = L_{ce} + \lambda_2 \sqrt{\sum_\theta \theta^2} \tag{17}$$

TABLE I
CLINICAL CHARACTERISTICS OF STUDY PARTICIPANTS

Characteristics			Value
Age at surgery, years			63.8 ± 8.7
Gender			103 M / 54 F
Disease duration, years			9.4 ± 3.9
MDS-UPDRS motor score	Surgery	Before	44.9 ± 17.9
		After	36.9 ± 17.5
	Medication	ON	35.9 ± 13.6
		OFF	55.4 ± 15.9
	DBS	ON	31.4 ± 15.8
		OFF	44.6 ± 16.5
Modified Hoehn and Yahr Scale [median (IQR)]			2.5 (2.0-3.0)

Partial derivatives of (16) and (17) are taken with respect to parameters W and θ, respectively, as follows:

$$\begin{aligned} \frac{\partial L_1}{\partial W} &= \frac{\partial L_{ce}}{\partial W} + \frac{\partial\left(\lambda_1 \sum_W |W|\right)}{\partial W} \\ &= \nabla L_{ce(W)} + \lambda_1 sgn\,(W), \\ sgn\,(W) &= \begin{cases} +1, & W > 0 \\ 0, & W = 0 \\ -1, & W < 0 \end{cases} \end{aligned} \tag{18}$$

$$\begin{aligned} \frac{\partial L_2}{\partial \theta} &= \frac{\partial L_{ce}}{\partial \theta} + \frac{\partial(\lambda_2\sqrt{\sum_\theta \theta^2})}{\partial \theta} \\ &= \nabla L_{ce(\theta)} + 2\lambda_2 \frac{\theta}{\sqrt{\sum_\theta \theta^2}} \end{aligned} \tag{19}$$

Finally, the rules for updating parameters W and θ can be expressed as follows:

$$W \mapsto W' = W - \eta \nabla L_{ce(W)} - \eta\lambda_1 sgn(W) \tag{20}$$

$$\begin{aligned} \theta \mapsto \theta' &= \theta - \eta \nabla L_{ce(\theta)} - 2\eta\lambda_2 \frac{\theta}{\sqrt{\sum_\theta \theta^2}} \\ &= \left(1 - \frac{2\eta\lambda_2}{\sqrt{\sum_\theta \theta^2}}\right)\theta - \eta \nabla L_{ce(\theta)} \end{aligned} \tag{21}$$

IV. EXPERIMENTS AND RESULTS

A. Datasets

The Institutional Review Board of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine approved this retrospective study. This study was conducted based on the clinical video dataset from the Neurosurgery Department of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. This dataset contains assessment videos of 157 PD patients from 2017 to 2019. Table I shows the demographics and clinical characteristics of these patients, including the age at surgery, disease duration, MDS-UPDRS motor score, and the modified Hoehn and Yahr Scale. Each patient was divided into distinct PD states before (ON and OFF medication) and after surgery (ON and OFF medication and/or ON and OFF deep brain stimulation (DBS)) at different follow-up times. Videos were captured in different PD states independently. All videos were recorded at 30 frames per second (FPS) in the front view with a frame resolution of 1920×1080 or 1280×720. Due to the lack of the LA task in some patients' assessment videos, irregular task execution, or severe camera shake, some videos

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DISTRIBUTION OF THE MDS-UPDRS LA SCORES

LA score	0	1	2	3	4	Total
Number	118	377	256	95	24	870

were deleted. Finally, we studied 148 patients, a total of 483 states and 870 available videos (including right and left LA tasks). The distribution of 870 LA scores is shown in Table II.

The videos of left and right legs were clipped, respectively, and all the video frames of the left leg were flipped horizontally to those of the right leg for simplifying subsequent modeling. For each video, we extracted the first 200 frames as the basic input unit; then, we extracted the human skeleton sequences through the OpenPose [37]. The sequences contained 25 joints of the human body in each frame, and each joint was represented in the form of 2D coordinates (x, y). Subsequently, min-max normalization was performed in each video to eliminate the effect of external factors such as shooting distance.

B. Evaluation Criteria

In this study, the video was considered as the independent sample for the following analysis since patients performed the LA tasks in distinct PD states (such as ON/OFF medication and ON/OFF DBS) at different follow-up times corresponding to different motor fluctuation phases [15]. Unless otherwise specified, the sample-independent five-fold cross-validation (CV) was applied to evaluate the experimental results. However, we also provided the patient-independent five-fold CV in IV.D and IV.H. Specifically, the 870 video samples (or 148 patients) were divided into five fixed folds randomly. Four folds were used for training and one fold for testing. The same sample (i.e., sample-independent) or the samples from the same patient (i.e., patient-independent) did not appear in both training and testing sets simultaneously. In addition, the accuracy (Acc), acceptable accuracy (Acceptable Acc), precision (Prec), recall (Rec), F1 score (F1), and area under the receiver operating characteristic (ROC) curve (AUC) were applied to evaluate the performance. The acceptable accuracy referred to the proportion of samples whose absolute error between the predicted label and true label was no more than 1, and this was considered acceptable in clinical practice [13], [15], [38].

C. Parameter Settings

The proposed network was implemented on an NVIDIA GeForce GTX 1080Ti GPU (11 GB memory) based on PyTorch. The network weights were optimized through the stochastic gradient descent (SGD) method with momentum 0.9, and the batch size was set to 8. The learning rate was initialized to 0.001 and was reduced to 1/10 of the value after the 40th and 60th epochs. The whole training process stopped after the 125th epoch. We set the same random number seed in all experiments to ensure the reproducibility of results. The time spent on each CV was approximately 1.2 h.

TABLE III
CLASSIFICATION RESULTS OF THE PROPOSED MODEL

Sample-independent	Acc (%)	Acceptable Acc(%)	Prec (%)	Rec (%)	F1 (%)	AUC
Score-0	35.59	94.07	60.87	35.59	44.92	0.84
Score-1	81.96	100.00	71.86	81.96	76.58	0.84
Score-2	70.31	100.00	70.87	70.31	70.59	0.88
Score-3	66.32	97.89	66.32	66.32	66.32	0.93
Score-4	75.00	100.00	81.82	75.00	78.26	0.99
Total Acc(%)			70.34 (612/870)			
Total Acceptable Acc(%)			98.97 (861/870)			
Patient-independent	Acc (%)	Acceptable Acc(%)	Prec (%)	Rec (%)	F1 (%)	AUC
Score-0	32.20	97.46	63.33	32.20	42.70	0.80
Score-1	81.96	99.73	68.36	81.96	74.55	0.80
Score-2	65.23	99.61	67.88	65.23	66.53	0.84
Score-3	58.95	94.74	62.22	58.95	60.54	0.91
Score-4	75.00	100.00	81.82	75.00	78.26	1.00
Total Acc(%)			67.59 (588/870)			
Total Acceptable Acc(%)			98.85 (860/870)			

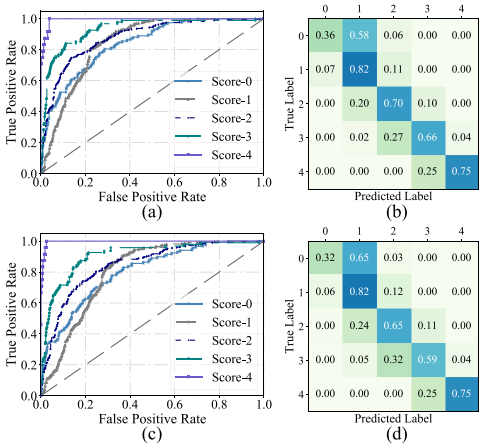

Fig. 5. Sample-based (a), (b) and patient-based (c), (d) experimental results: (a), (c): ROC curves of each score category; (b), (d): confusion matrices.

D. Classification Results

For a more comprehensive evaluation, we provided the sample-independent and patient-independent classification performance of our proposed model by five-fold CV in Table III. It shows a sample-independent overall classification accuracy of 70.34% and an acceptable accuracy of 98.97%, as well as a patient-independent overall classification accuracy of 67.59% and an acceptable accuracy of 98.85%. The AUC values for all categories were above 0.80, and all categories achieved the desirable performance within the acceptable error range, as shown in Fig. 5. Although the classification accuracies of score 0 were less than 36%, the confusion matrices in Fig. 5 (b) and (d) illustrate that most of the score 0 samples were identified as score 1 samples due to the subtle differences between the score 0 and 1 samples; thus, the acceptable accuracies were still greater than 94%. (See discussion for details.)

We also performed five-fold CV experiments by repeating 10 times, as shown in Fig. 6. The fluctuations in the accuracy and AUC value are relatively slight, and the standard deviations are all less than 0.006. In addition, the difference between the sample-independent and patient-independent results is also small, especially the difference in mean AUC values was 0.01. These results confirm the stability of our method.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

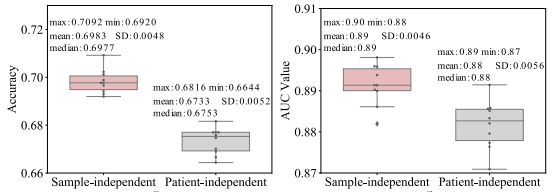

Fig. 6. Sample-independent and patient-independent results of the 10 repeated five-fold CV experiments.

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART MODELS

State-of-the-art Model	Total		Average			
	Acc (%)	Acceptable Acc (%)	AUC	Prec (%)	Rec (%)	F1 (%)
Two-stream CNN[39]	53.91	96.32	0.80	-	34.36	-
ST-GCN (Spatial Configuration)[20]	63.33	97.93	0.86	62.49	61.48	61.63
Js-AGCN[23]	64.94	98.28	0.88	61.75	62.71	60.96
Bs-AGCN[23]	60.23	97.36	0.87	56.48	59.55	57.55
2s-AGCN[23]	66.44	98.85	0.89	65.82	65.79	64.95
Motif-STGCN[25]	64.94	98.16	0.88	63.16	62.59	62.20
Our method	**70.34**	**98.97**	**0.90**	**70.35**	**65.84**	**67.33**

E. Comparison With State-of-the-Art Methods

To demonstrate the superiority of our model in the fine-grained assessment of the LA task, we compared the performance of our model with the state-of-the-art skeleton-based action recognition models [20], [23], [25], [39]. The open-source code of the two-stream CNN [39], ST-GCN[20], joint-stream adaptive graph convolutional network (Js-AGCN) [23], bone-stream adaptive graph convolutional network (Bs-AGCN) [23], two-stream adaptive graph convolutional network (2s-AGCN) [23], and motif-based graph convolutional network with variable temporal dense block architecture (Motif-STGCN) [25] were applied to our LA dataset, and the accuracies were 53.91%, 63.33%, 64.94%, 60.23%, 66.44%, and 64.94%, respectively. As listed in Table IV, by contrast, the accuracy of our proposed model increased to 70.34%, which was significantly higher than that of the state-of-the-art skeleton-based action recognition models.

F. Ablation Study

In this section, we discuss the specific ablation experiments conducted to verify the effectiveness of each component in this model (Table V).

1) Sparse Adaptive Graph Convolutional Unit: We used the ST-GCN [20] as the baseline (Uni-ST-GCN) for our comparisons. Uni-labeling proposed in [20] was used as the partition strategy of the Uni-ST-GCN, which can be regarded as modeling only the physical dependency of joints and having one node type. Compared with the graph convolutional unit based on uni-labeling in the baseline, the SAGCU improved the accuracy and acceptable accuracy by 1.38% and 1.15% (Table V), respectively, and there were significant improvements in other evaluation indicators.

To visually and intuitively demonstrate the effectiveness of the SAGCU, the adaptive masks $\mathcal{M}^1$ and $\mathcal{M}^2$ of the two adjacency matrices in the SAGCU were visualized in Fig. 7.

TABLE V
RESULTS OF ABLATION EXPERIMENTS

Proposed Model	Total		Average			
	Acc (%)	Acceptable Acc(%)	AUC	Prec (%)	Rec (%)	F1 (%)
Uni-ST-GCN (baseline)	63.45	96.90	0.87	62.05	62.38	61.93
Uni-ST-GCN+SAGCU	64.83	98.05	0.87	62.88	63.69	62.91
Uni-ST-GCN+TCM	64.14	97.59	0.86	62.37	64.21	63.13
Uni-ST-GCN+MDALM	64.48	98.16	0.87	65.05	62.76	63.61
Uni-ST-GCN+TCM +MDALM	65.75	97.82	0.88	64.37	63.52	63.50
Uni-ST-GCN+L_{param}	64.02	97.01	0.86	63.24	62.88	62.87
J-stream (Uni-ST-GCN+SAGCU +TCM+MDALM+L_{param})	67.93	98.51	0.89	66.92	65.00	65.47
Jm-stream	64.94	97.47	0.87	63.79	62.95	63.16
Fusion (SA-GCN)	**70.34**	**98.97**	**0.90**	**70.35**	**65.84**	**67.33**

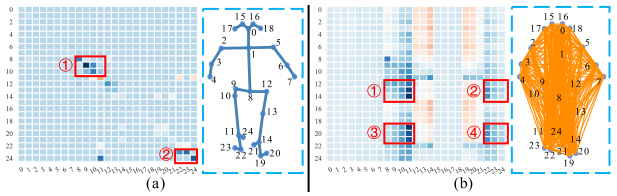

Fig. 7. Visualization of adaptive masks in the SAGCU: (a) physical dependency (blue edge), (b) logical dependency (yellow edge). The darker the blue regions, the more important the adaptive dependency.

The regions with higher response values in $\mathcal{M}^1$ and $\mathcal{M}^2$ represent the stronger importance of the dependency learned by the model, and the regions of strong dependency mainly concentrated in the lower limb regions related to the LA task. All the videos of the left LA task were unified into the videos of the right LA task in data preprocessing; hence, the important lower limb regions are concentrated in the joints 8–11 and 22–24. For example, the dark regions ① and ② in Fig. 7 (a) correspond to the joint regions that the model adaptively enhanced, i.e., the right lower limb joints. The dark regions ① –④ in Fig. 7 (b) correspond to the important dependency between the left and right lower limb joints learned by the model. The dependency between joints 11 and 14 is the strongest, which is the most important in the assessment of the LA task.

Furthermore, to visually reflect that the sparsification strategy selected the discriminative dependency in the graph structure, we visualized the effect of the sparsification strategy on the logical-dependency modeling. $\mathcal{M}^2$ was normalized, and the different threshold values were set to demonstrate the logical dependency. Fig. 8 (a)–(e) and (f)–(j) represent that the threshold values change from small to large, respectively, that is, the most substantial logical dependency is gradually selected. The results indicate that the introduction of the sparsification strategy (Fig. 8 (a)–(e)) makes the model learn the most important logical dependency of the LA task (Fig. 8 (e)). By contrast, the logical dependency determined by this model is relatively redundant without a sparsification strategy (Fig. 8 (f)–(j)).

2) Temporal Context Module: We then verified the effectiveness of the proposed TCM. As shown in Table V, the introduction of the TCM improved many evaluation indicators,

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

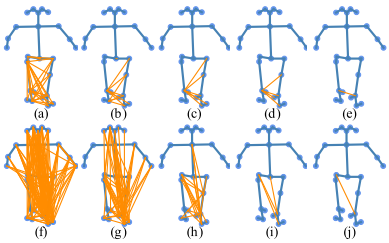

Fig. 8. Impact of the model-driven sparsification strategy on logical dependency. (a)–(e) are the results in the presence of the sparsification strategy and (f)–(j) are the results in the absence of the sparsification strategy. The blue lines represent the inherent physical dependency, and the yellow lines represent the logical dependency.

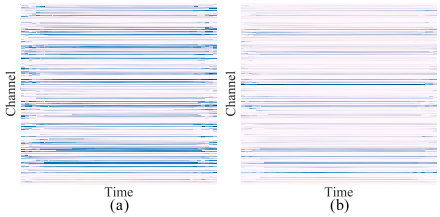

Fig. 9. (a) Input channel-temporal map; (b) channel-temporal map under the guidance of attention. The deeper the color is, the larger the feature value is.

among which the accuracy and acceptable accuracy increased by 0.69% and 0.69%, respectively.

3) Multi-Domain Attention Learning Module: Finally, we verified the performance of the MDALM. First, the MDALM without the TCM was tested, and the accuracy and acceptable accuracy were improved by 1.03% and 1.26%, respectively. After that, the TCM was introduced into the MDALM. The accuracy was 2.3% higher than that of the baseline, and other evaluation indicators were also significantly improved. In addition, we drew the input channel-temporal map and the map after attention enhancement in Fig. 9, intuitively illustrating that the features under the guidance of attention focus on the discriminative regions in the channel domain.

G. Sensitivity Analysis

To further verify the robustness of the model, we conducted parameter sensitivity analysis for the trade-off factors λ_1 and λ_2 of the cost function terms L_{sparse} and L_{param}. We chose $\lambda_1 = 0.03$ and $\lambda_2 = 0.0001$ as the benchmark trade-off parameters through experiments, and we varied these values by approximately 20% to verify the stability of the method. The relative change rate of the assessment indicator was defined as

$$\mathrm{R} = \left| (I_{new} - I_{\lambda_1=0.03,\lambda_2=0.0001}) \Big/ I_{\lambda_1=0.03,\lambda_2=0.0001} \right| \quad (22)$$

The accuracy and acceptable accuracy were regarded as the sensitivity evaluation indicators. Fig. 10 illustrates the similar performance of the model with 25 sets of parameters. The average relative change rate of accuracy is 4.32%, and the average relative change rate of acceptable accuracy is smaller at 0.16%. These results indicate that our model is robust to the changes in the trade-off parameters of the cost function.

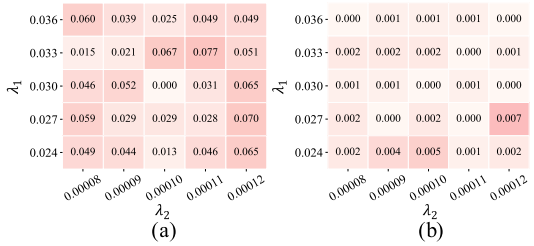

Fig. 10. Sensitivity analysis: (a) accuracy; (b) acceptable accuracy. The values in the grid represent the R values calculated under different combinations of the trade-off factors λ_1 (the horizontal axis) and λ_2 (the vertical axis).

H. Extensive and Objective Performance Evaluation

To demonstrate the rationality and superiority of the proposed method, we presented an extensive and objective performance evaluation with the existing automated quantitative assessment studies of the LA task, as shown in Table VI. Although BSNs and optical flow RGB depth devices can capture relatively accurate motion signals, the existing methods [9], [13]–[15] need to define kinematic features manually; therefore, there are no satisfactory results. Li *et al.* [4] put forward the assessment scheme of the LA task based on videos. However, due to the limitations of the dataset and classifier, a correlation coefficient of only 0.618 was achieved. Our proposed deep learning model can automatically extract significant fine-grained features in an end-to-end manner; thus, the results of our proposed method outperformed those of the existing methods.

These results indicate that the sample-independent accuracy of 70.34% and patient-independent accuracy of 67.59% achieved by our method are reasonable and credible, and our method achieves the best performance among the existing related studies.

V. Discussion

In this study, we proposed a new method for the automated quantitative assessment of the PD LA task. The spatial-temporal graphs were constructed based on the skeleton sequences extracted from videos; then, the fine-grained quantitative assessment was realized through the deep learning model. The SAGCU was proposed to obtain the sparsification features of the graph, followed by the TCM introduced to capture the global dependency of the temporal features, thus capturing fine-grained features in the spatial and temporal dimensions. Subsequently, fine-grained salient features in the channel domain were enhanced by adaptive channel-wise attention learning. These mining strategies of multi-domain discriminative features enable our model to identify subtle differences between different scores in the LA task. Finally, our proposed method achieved accurate and reliable fine-grained assessment results in the clinical dataset, thereby proving its effectiveness and feasibility.

To confirm the effectiveness of our strategies, we analyzed the experimental results comprehensively. The quantitative analysis (Table V) indicates that each strategy resulted in better performance than the baseline. In the qualitative analysis,

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
COMPARISON WITH EXISTING RELATED STUDIES

Author/Year	Resource	Features	Dataset	Models	Performance		
					Correlation	Acc(%)	Acceptable Acc(%)
Giuberti *et al.*[13]/2014	Body sensor network	Kinematic features (time domain)	24 PD; 72 samples	kNN	-	<40	97
Giuberti *et al.*[14]/2015	Body sensor network	Kinematic features (time and frequency domains)	24 PD; 72 samples	kNN	-	~ 50	>90
Parisi *et al.*[15]/2015	Body sensor network	Kinematic features (time and frequency domains)	34 PD; 94 samples	kNN	-	43	97
Ferraris *et al.*[9]/2020	Optical RGB-Depth device	Discriminative kinematic parameters	44 PD	SVM	-	60	-
Li *et al.*[4]/2018	Videos	Kinematic features, spectral features, convex hull	9 PD	Random forest	0.618*	-	-
Ours/2020	Videos	Deep learning features	148 PD; 870 samples	SA-GCN	-	Sample-independent 70.34	98.97
						Patient-independent 67.59	98.85

Note: 0.618* represents the mean correlation between PD severity predictions and ground truth ratings for the LA task.

first, the two adaptive masks in the SAGCU learned the most important dependency of the LA task (Fig. 7). The design of the sparsification strategy rendered the graph features more discriminative. In particular, the model exerted the highest attention weight on the most important dependency in the LA task during the modeling of logical dependency (Fig. 8 (e)). Thereafter, TCM constructed the remote dependency in the temporal dimension, which is crucial in the LA task. Finally, the attention regions in the channel domain were enhanced through the MDALM, and the static spatial information and dynamic temporal information were highlighted through separate branches, thus integrating multi-domain features. Additionally, the features filtered by the attention mask focused on the salient channel locations (Fig. 9). These qualitative analysis results confirmed the effectiveness of the proposed strategies in the classification task.

The limitation of the proposed model is that the accuracies of score 0 were less than 36%. The main reason is that the difference between scores 0 and 1 in the clinical assessment is subtle, such that most of score 0 samples are mistakenly identified as similar score 1 samples by the model. Further, the sample size of the score 0 (i.e., healthy persons) is only 118 in our dataset. However, the acceptable accuracies of score 0 higher than 94% were obtained, which is reasonable and acceptable in clinical practice.

Furthermore, we performed a comparative analysis (Table VI) with the existing studies on the automated assessment of the LA task [4], [9], [13]–[15] from multiple perspectives. The proposed scheme achieved the best performance in accuracy and acceptable accuracy, which confirms its rationality and reliability. The advantages of our study can be summarized as follows: 1) To the best of our knowledge, we have the largest dataset with the largest number of patients and samples; this improves the reliability of the model prediction and makes our study closer to the actual clinical situation. 2) Our method is completely contactless and only needs an ordinary camera to collect data. This avoids the intrusiveness of sensors, the requirement of regular calibration, and the professional calibration process of optical devices, thereby providing convenience for the realization and popularization of the PD automated assessment system. 3) The existing studies adopted the feature extraction method from motion trajectories according to the manually defined rules and ignored the connectivity between joints. However, the deep learning scheme that we adopted automatically extracts spatial connectivity features from skeleton sequences in an end-to-end manner. Moreover, several spatial-temporal modeling strategies were introduced, thus bringing more fine-grained discriminative features.

VI. CONCLUSION

We proposed a novel spatial-temporal network based on the GCN for the automated assessment of the PD LA task in the MDS-UPDRS from videos. Specifically, the SAGCU was proposed to model the spatial structures of physical and logical dependency of skeleton sequences, and the sparse modeling of discriminative features was realized through the sparsification term in the cost function. The TCM was introduced to capture the global temporal dependency. The MDALM was designed by adding the spatial learning branch and the adaptive channel-wise attention learning branch, thereby achieving the multi-domain integration and the enhancement of channel-wise salient features. A comprehensive analysis of the experimental results demonstrated the accuracy and reliability of the proposed scheme, and this scheme outperformed other related methods. Our contactless approach shows great potential for future PD automated assessment and telemedicine.

REFERENCES

[1] L. M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *Lancet Neurol.*, vol. 5, no. 6, pp. 525–535, 2006.
[2] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol., Neurosurg. Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.
[3] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disorders, Off. J. Movement Disorder Soc.*, vol. 23, no. 15, pp. 2129–2170, 2008.
[4] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation," *J. Neuroeng. Rehabil.*, vol. 15, no. 1, p. 97, Dec. 2018.
[5] D. A. Heldman *et al.*, "The modified bradykinesia rating scale for Parkinson's disease: Reliability and comparison with kinematic measures," *Movement Disorders*, vol. 26, no. 10, pp. 1859–1863, Aug. 2011.
[6] S. Patel *et al.*, "Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 864–873, Nov. 2009.
[7] R. Ramsperger *et al.*, "Continuous leg dyskinesia assessment in Parkinson's disease—Clinical validity and ecological effect," *Parkinsonism Rel. Disorders*, vol. 26, pp. 41–46, May 2016.
[8] P. J. M. Bank, J. Marinus, C. G. M. Meskers, J. H. de Groot, and J. J. van Hilten, "Optical hand tracking: A novel technique for the assessment of bradykinesia in Parkinson's disease," *Movement Disorders Clin. Pract.*, vol. 4, no. 6, pp. 875–883, Nov. 2017.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.

[9] C. Ferraris, R. Nerino, A. Chimienti, G. Pettiti, C. Azzaro, and G. Albani, "Automated assessment of motor impairments in Parkinson's disease," *Clin. Neurol. Int.*, vol. 1, no. 1, p. 1010, 2020.

[10] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Automated assessment of levodopa-induced dyskinesia: Evaluating the responsiveness of video-based features," *Parkinsonism Rel. Disorders*, vol. 53, pp. 42–45, Aug. 2018.

[11] Y. Liu *et al.*, "Vision-based method for automatic quantification of parkinsonian bradykinesia," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 1952–1961, Oct. 2019.

[12] A. Sabo, S. Mehdizadeh, K.-D. Ng, A. Iaboni, and B. Taati, "Assessment of parkinsonian gait in older adults with dementia via human pose tracking in video data," *J. Neuroeng. Rehabil.*, vol. 17, no. 1, pp. 1–10, Dec. 2020.

[13] M. Giuberti *et al.*, "Linking UPDRS scores and kinematic variables in the leg agility task of parkinsonians," in *Proc. 11th Int. Conf. Wearable Implant. Body Sensor Netw.*, Jun. 2014, pp. 115–120.

[14] M. Giuberti *et al.*, "Assigning UPDRS scores in the leg agility task of parkinsonians: Can it be done through BSN-based kinematic variables?" *IEEE Internet Things J.*, vol. 2, no. 1, pp. 41–51, Feb. 2015.

[15] F. Parisi *et al.*, "Body-sensor-network-based kinematic characterization and comparative outlook of UPDRS scoring in leg agility, Sit-to-stand, and gait tasks in Parkinson's disease," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1777–1793, Nov. 2015.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[17] 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.

[18] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 816–833.

[19] 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.

[20] 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, pp. 7444–7452.

[21] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5323–5332.

[22] K. Thakkar and P. J Narayanan, "Part-based graph convolutional network for action recognition," 2018, *arXiv:1809.04983*. [Online]. Available: http://arxiv.org/abs/1809.04983

[23] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[24] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7912–7921.

[25] Y.-H. Wen, L. Gao, H. Fu, F.-L. Zhang, and S. Xia, "Graph CNNs with motif and variable temporal block for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8989–8996.

[26] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, Mar. 2018.

[27] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 416–428, Feb. 2019.

[28] T. Yu, C. Guo, L. Wang, H. Gu, S. Xiang, and C. Pan, "Joint spatial-temporal attention for action recognition," *Pattern Recognit. Lett.*, vol. 112, pp. 226–233, Sep. 2018.

[29] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4263–4270.

[30] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1647–1656.

[31] C. Wu, X.-J. Wu, and J. Kittler, "Spatial residual layer and dense connection block enhanced spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.

[32] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction," 2019, *arXiv:1910.02212*. [Online]. Available: http://arxiv.org/abs/1910.02212

[33] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[34] A. Sankar, X. Zhang, and K. C.-C. Chang, "Meta-GNN: Metagraph neural network for semi-supervised learning in attributed heterogeneous information networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2019, pp. 137–144.

[35] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[36] P. N. Druzhkov and V. D. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," *Pattern Recognit. Image Anal.*, vol. 26, no. 1, pp. 9–15, Jan. 2016.

[37] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 17, 2019, doi: 10.1109/TPAMI.2019.2929257.

[38] L. Borzi *et al.*, "Smartphone-based estimation of item 3.8 of the MDS-UPDRS-III for assessing leg agility in people with Parkinson's disease," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 140–147, 2020.

[39] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 597–600.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 28,2023 at 00:34:23 UTC from IEEE Xplore. Restrictions apply.