Multi-scale Sparse Graph Convolutional Network for the Assessment of Parkinsonian Gait

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

***Abstract*—Automated assessment of patients with Parkinson's disease (PD) is urgently required in clinical practice to improve the diagnostic efficiency and objectivity and to remotely monitor the motor disorder symptoms and general health of these patients, especially in view of the travel restrictions due to the recent coronavirus epidemic. Gait motor disorder is one of the critical manifestations of PD, and automated assessment of gait is vital to realize automated assessment of PD patients. To this end, we propose a novel two-stream spatial-temporal attention graph convolutional network (2s-ST-AGCN) for video assessment of PD gait motor disorder. Specifically, the skeleton sequence of human body is extracted from videos to construct spatial-temporal graphs of joints and bones, and a two-stream spatial-temporal graph convolutional network is then built to simultaneously model the static spatial information and dynamic temporal variations. The multi-scale spatial-temporal attention-aware mechanism is also designed to effectively extract the discriminative spatial-temporal features. The deep supervision strategy is then embedded to minimize classification errors, thereby guiding the weight update process of the hidden layer to promote significant discriminative features. Besides, two model-driven terms are integrated into this deep learning framework to strengthen multi-scale similarity in the deep supervision and realize sparsification of discriminative features. Extensive experiments on the clinical video dataset show that the proposed model exhibits good performance with an accuracy of 65.66% and an acceptable accuracy of 98.90%, which is much better than that of the existing sensor- and vision-based methods for Parkinsonian gait assessment. Thus, the proposed method is potentially useful for assessing PD gait motor disorder in clinical practice.**

***Index Terms*—Parkinson's disease, Gait motor disorder, Video-based assessment, Graph convolutional network, Model-driven deep learning**

I. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder, and motor disorder is a typical clinical characteristic of this disease [1]. Presently, the main clinical assessment criterion for PD patients is the Unified Parkinson's Disease Rating Scale (UPDRS) [2], which was updated by the Movement Disorder Society (MDS) in 2008 as MDS-UPDRS [3]. The motor disorder symptoms are rated on a scale of 0-4, where 0 indicates normal and 4 indicates severe condition. The clinical practice with this scale faces two major problems: 1) Full motor assessment by a trained and experienced neurologist is time-consuming, requiring at least half an hour, and the obtained results are subjective with large inter-rater variability; 2) travel limitations, timely follow-up of progressive disease, and sparse distribution of PD specialists also create challenges in providing medical care to PD patients. In addition, due to the ongoing epidemic of coronavirus 2019 (COVID-19) [4], several countries have imposed travel restrictions. Thus, an objective and automated motor function assessment system for remote monitoring of PD patients is urgently required, and this has become a research hotspot for both academic and clinical community.

Gait motor disorder, one of the common motor disorders in PD, is primarily assessed by clinical analysis of patients' stride length, speed, turning, and swinging of arms. Gait assessment is strongly correlated with the severity of the disease [5], and it is an essential part of motor examination based on the MDS-UPDRS [3]. Therefore, automated quantification of gait motor disorder is crucial for realizing automated motor function assessment of PD patients.

Currently, sensor- and vision-based schemes are primarily used for automated quantitative analysis of gait motor disorder in PD patients based on the MDS-UPDRS (or UPDRS). Parisi *et al.* [6], [7] extracted the kinematic features of gait in PD patients in time and frequency domains by applying a body sensor network consisting of three inertial measurement units on the chest and thighs. Then, the k nearest neighbors algorithm was used to achieve an accuracy of 62% and 53%, respectively. Aşuroğlu *et al.* [8] extracted 16 time-domain features and 7 frequency-domain features from ground reaction force signals acquired by gait sensors. These gait features were then associated with the total assessment score of the UPDRS from 0 to 199 using the locally weighted random forest to achieve maximum correlation coefficient of 0.895. Due to the accurate motor signals from the sensors, these sensor-based methods could achieve good results. However, most sensors should be in direct contact with the patient's body, which inevitably affects the assigned movements of PD patients. Thus, the extension of the sensor-based scheme to clinical applications has been limited.

To the best of our knowledge, only Chen *et al.* [9] and Sabo *et al.* [10] have proposed vision-based schemes for the automated assessment of gait motor disorder based on the

R. Guo and X. Qian are with School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200030, China.

X. Qian is the corresponding author. (e-mail: xiaohua.qian@sjtu.edu.cn).

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

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

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes more results of repeated experiments. This material is 409KB in size.

MDS-UPDRS in PD patients. Specifically, Chen *et al.* [9] extracted the contour features of human body from the gait image sequences and calculated the abnormality indexes associated with posture and foot movement from the upper and lower body, respectively. Then, they used a linear regression model to evaluate the overall motor abnormality (MA), and the correlation coefficient between MA and the sum of sub-scores from UPDRS Part III was 0.85. Sabo *et al.* [10] obtained joint coordinates from the videos collected by a Microsoft Kinect sensor and color camera. Then, the univariate regression was utilized to select the features significantly related to the clinical scores. Finally, an accuracy of around 62% was obtained through a multiple regression model. However, the above vision-based methods have two main limitations: 1) It is challenging to use traditional feature engineering for designing mathematical functions to characterize the subtle differences in the gait motor functions. 2) The regression model is based on the linear relationship between gait features and clinical scores and cannot be used to directly predict the scores of gait motor disorder.

With its rapid development, computer vision is widely used in various healthcare fields. For example, Leo *et al.* [11] summarized the applications of computer vision methods in facial information analysis for healthcare. Yeung *et al.* [12] used computer vision algorithms to detect and quantify the care activities of critically ill patients. Liao *et al.* [13] proposed a computer vision algorithm with less interaction and fast convergence to realize 3D medical image segmentation.

In this study, to address the above-mentioned limitations of vision-based methods, we establish a quantification technique for the automated assessment of gait motor disorder in PD patients through front-view videos based on the computer vision technology. Specifically, the joint sequence of human body is extracted from the videos through an advanced human pose estimation model, and the skeleton sequence is constructed by a spatial-temporal graph and then classified by a deep learning model.

However, the quantitative assessment based on front-view videos also faces the following three challenges. Firstly, as shown in Fig. 1, the ill-fitting and loose hospital gown of patients adversely affects the depiction of body contour. Any variation in the distance between the patient and the camera also makes it challenging to characterize the gait movement by traditional feature engineering. Secondly, the assessment of gait based on the MDS-UPDRS can be considered as fine-grained assessment, which is inherently one of the challenges in the field of action recognition. Thirdly, video sequences that play a vital role in the assessment only contain sparse keyframes, although neurologists need to review the entire walking process of PD patients.

To resolve the above-mentioned issues, using the videos from the front view, we propose a two-stream multi-scale sparse spatial-temporal attention-aware graph convolutional network (GCN) under deep supervision to realize automated quantitative assessment of gait motor disorder in PD patients. Specifically, our model has the following characteristics: 1) The joint sequences of patients are extracted from their gait videos by a well-established human pose estimation model; then, these joint sequences and further generated bone sequences are constructed using a two-stream spatial-temporal GCN (2s-ST-GCN) to obtain the features of the entire body rather than those of individual parts of the body. Thus, this model is not affected by the attire and imaging environment of the patient. 2) To fully capture meaningful multi-scale fine-grained features, a multi-scale spatial-temporal attention-aware mechanism (MS-STAM) is introduced. The spatial-temporal attention coefficient matrix in the attention block is utilized to enhance the discriminative spatial-temporal features. Subsequently, a deep supervision strategy is proposed to minimize the cross entropy on each scale and enhance the similarity between different scales. 3) A sparse regularization term is introduced in the cost function of each scale for the sparsification of fine-grained features and selection of discriminative features.

(a) (b) (c)

Fig. 1. An example of recorded videos for gait assessment of PD patients.

The main contributions of this study can be summarized as follows:

- The two-stream spatial-temporal graph convolutional network is proposed to model the topological structure of human joint and bone sequences and to automatically capture the spatial relationship and time dynamics embedded in joints and bones, thereby avoiding the limitation of traditional manual rules for the analysis of spatial mode.
- The multi-scale spatial-temporal attention-aware mechanism under deep supervision is developed to capture the multi-scale fine-grained discriminative spatial-temporal features under a strong correlation among different scales, thereby improving the discriminativeness and robustness of learned features.
- A model-driven sparse strategy is proposed to identify significant discriminative features and effectively eliminate the redundancy of features.

The rest of this paper is organized as follows. In section II, existing related work is reviewed. The proposed method is introduced in section III, and the experimental system and results as well as qualitative and quantitative analysis of results are presented in section IV. The advantages and limitations of the proposed method are discussed in section V. Finally, the study is concluded in section VI.

II. Related Work

A. Video assessment of motor disorder symptom in PD patients based on deep learning

Owing to the development of action recognition [14]–[19]

such as gait recognition [20]–[22], several studies have applied video-based techniques for the automated assessment of motor disorder symptom in PD patients. Li *et al.* [23] extracted human joint sequences from recorded videos of PD patients through a pose estimation method and calculated the movement features. They demonstrated that these features are related to the severity level of motor disorder. Subsequently, they applied random forests for multiclass classification of pathological motion and assessment of clinical ratings corresponding to the UPDRS and Unified Dyskinesia Rating Scale (UDysRS) [24]. Later, they developed a video-based assessment system to quantify the changes in motor disorder severity of levodopa-induced dyskinesia patients and achieved similar or higher performance as compared to the UDysRS [25]. Liu *et al.* [26] developed a lightweight human pose estimation network and applied the support vector machine to generate score ratings corresponding to three bradykinesia-related hand movements (finger tapping, hand clasping, and hand pro/supination) in the MDS-UPDRS. Although all these video-based assessment methods utilized deep learning to achieve pose estimation, they still used traditional engineering scheme to manually design the features for characterizing the motor functions of PD patients.

According to the above survey, although a few studies have evaluated gait videos of PD patients quantitatively, the quantitative assessment of front-view gait videos of PD patients has not been reported yet. Because traditional engineering schemes cannot provide accurate results for motor function assessment based on front-view videos, we developed a novel deep learning framework to extract multi-scale features of the skeleton sequences of PD patients from the front-view videos in an end-to-end way. In addition, we established a graph neural network to characterize the topological properties of human body skeleton.

B. *Fine-grained action recognition based on deep learning*

The major limitation of most of the existing action recognition methods is that they cannot achieve satisfactory performance in fine-grained action recognition with high similarity in appearance and behavior. Earlier studies on fine-grained action recognition [27]–[30] mostly focused on a large number of fine-grained interactions between people and objects as well as on the construction of local context information between people's actions and objects of interest. However, in most practical situations, identifying the subtle differences in the same action is vital to achieve fine-grained classification. Hence, it is critical to explore the most discriminative features in video sequences with high complexity level and overall similarity for fine-grained action classification. To this end, Singh *et al.* [31] located bounding boxes around people using an integrated tracking algorithm and modeled spatial- and temporal information successively through a multi-stream convolutional neural network and a long short-term memory network. Zhu *et al.* [32] proposed a multi-view attention mechanism called channel-spatial-temporal attention block, which consisted of channel-spatial branch, channel-temporal branch, and spatial-temporal branch. These branches were directly embedded into inflated 3D convolution networks based on RGB frames and optical flow to fully utilize the spatial, temporal, and channel information of video sequences.

To avoid the interference of complex environmental factors in the traditional RGB frames and to overcome the limitation of tracking the bounding boxes around people in [31], we obtained the coordinate sequence of human joints from motion videos using a state-of-the-art pose estimation method. In addition, the multi-branch and multi-scale information in deep learning frameworks is often used for the final classification [32]. However, the different features extracted from multi-branch or multi-scale information often lack a strong correlation among various branches or scales. Therefore, we introduced a deep supervision term in the cost function. Specifically, a cross-entropy term is used to minimize the classification errors in each scale. Further, an L2 regularization term is applied to reinforce strong correlation between different scales.

III. METHOD

Fig. 2 shows the architecture of the proposed 2s-ST-AGCN model. The skeleton sequence extracted from the gait videos of PD patients serves as the input of this model, and it is divided into joint and bone streams. Each stream is an independent spatial-temporal GCN (ST-GCN), which is used to naturally simulate the graph structure of the key points of human body and capture the features in both spatial and temporal domains. Then, the MS-STAM under deep supervision is embedded into the middle units of the 2s-ST-GCN for effectively selecting discriminative features in the spatial-temporal dimension. The prediction scores of the joint and bone streams are obtained by fitting fully connected layers on the multi-scale feature vectors, and the final prediction result of the two-stream framework is obtained through a soft-voting score fusion mechanism. The cost function of the network consists of a cross-entropy term that minimizes classification errors in each scale, an L2 regularization term that enhances feature correlation between different scales, and an L1 regularization term that achieves feature sparsification.

A. *2s-ST-GCN structure*

1) *Spatial-temporal graph construction*

The skeleton composed of human joints does not have a regular Euclidean spatial structure, and each joint has its characteristic structural feature and function. Further, structural features between joints are also unique. Thus, the hierarchical representation of human skeleton sequence is constructed through a spatial-temporal graph structure [19]. The original skeleton sequence with N joints and T frames is constructed by the joint coordinates of each frame in spatial and temporal domains. As shown in Fig. 3(a), the complete skeleton sequence can be considered to be composed of the spatial graph of each frame, and it is represented as $\mathcal{J} = \{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_T\}$, where $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$, $t = 1,2, \dots, T$, represents the spatial graph of human skeleton sequence at time t. The node set $\mathcal{V}_t = \{v_i | i = 1, \dots, N\}$ contains N joints of the human body, and the edge set $\mathcal{E}_t = \{v_i v_j | (i,j) \in L\}$ represents the edges formed by the joint set L naturally connected in the human body structure.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068609, IEEE Transactions on Multimedia

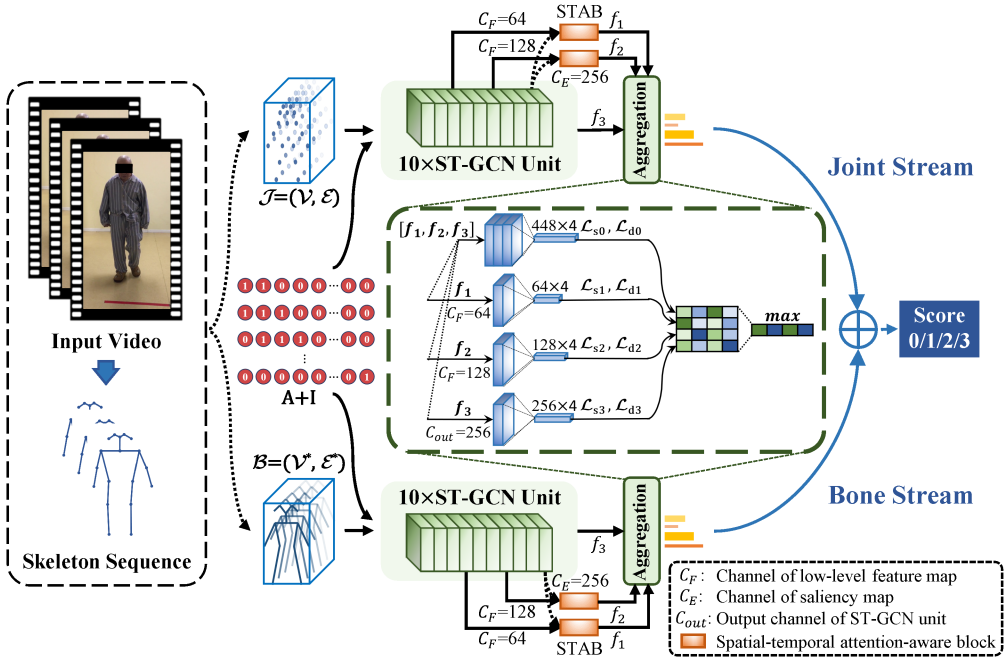


Fig. 2. Architecture of the proposed 2s-ST-AGCN.

Because spatial-domain model cannot represent the useful information of skeleton sequence in the temporal domain, it is necessary to establish a model in both spatial and temporal domains. The resulting undirected joint spatial-temporal graph can be expressed as $\mathcal{J}=\{\mathcal{V},\mathcal{E}\}$, where the node set $\mathcal{V}=\bigcup_{t=1}^{T}\mathcal{V}_t$ contains N joints in each frame, and the edge set $\mathcal{E}=(\bigcup_{t=1}^{T}\mathcal{E}_t)\cup\mathcal{E}_p$ contains the subset composed of naturally connected joints of the human body in each frame and the subset in the temporal domain. The latter term $\mathcal{E}_p=\{v_{ti}v_{(t+1)i}|t=1,\dots,T-1\}$ indicates the edge set formed by connecting the same joints in consecutive frames.

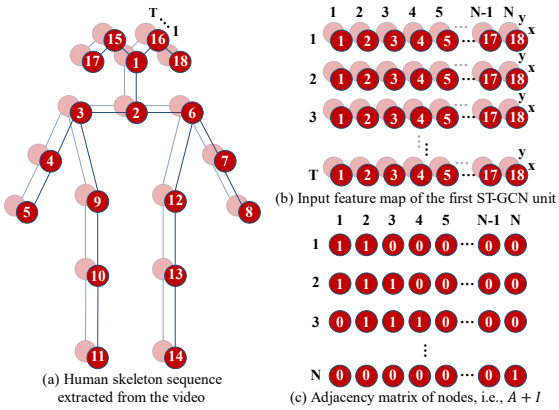


Fig. 3. An example of input data organization in 2s-ST-GCN.

To simulate the direction and distance between adjacent joints of the human body, inspired by [18], we explore the importance of bone information for gait analysis. Specifically, human bones are connected by two joints. According to the centripetal and centrifugal characteristics of the body motion, each bone can be represented as a vector from the source joint to the target joint, which are near and far away from the skeleton gravity center, respectively. If the coordinates of the given source and target joints are $v_1=\{x_1,y_1\}$ and $v_2=\{x_2,y_2\}$, respectively, the bone is represented as $v^*_{v1,v2}=(x_2-x_1,y_2-y_1)$. Each bone tuple is treated as the node information of the target joint. Therefore, the undirected spatial-temporal graph of bone can be expressed as $\mathcal{B}=\{\mathcal{V}^*,\mathcal{E}^*\}$, where the node set $\mathcal{V}^*$ includes the nodes formed by N bone tuples in each frame, and the edge set $\mathcal{E}^*$ includes the subset composed of the nodes naturally connected by the human body in each frame in the spatial domain and the subset formed by connecting the same joints in consecutive frames in the temporal domain.

2) Implementation of the 2s-ST-GCN

Inspired by the ST-GCN proposed in [19], the 2s-ST-GCN is composed of two streams: joint stream and bone stream. Ten ST-GCN units are included in each stream. The residual mechanism [33] is introduced in the ST-GCN units. The network architecture is shown in Fig. 4. The inputs of the joint and bone streams are the coordinates of the joints and bones, respectively. Each ST-GCN unit contains a spatial graph convolutional operation and a temporal convolutional operation.

In the spatial dimension, let the input of each ST-GCN unit be $G_{in}\in\mathbb{R}^{C_{in}\times T\times N}$, where C_{in}, T, and N are the numbers of channels, frames, and joints, respectively. Fig. 3(b) shows the input feature map of the first ST-GCN unit. The spatial graph

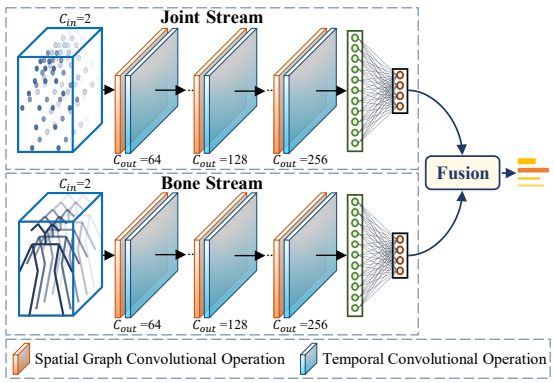


Fig. 4. General architecture of 2s-ST-GCN. Each ST-GCN unit consists of a spatial graph convolutional operation and a temporal convolutional operation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068609, IEEE Transactions on Multimedia

convolutional operation is implemented like the graph convolution proposed in [34]. Specifically, as shown in Fig. 3(c), the adjacency matrix of the internal joints for each frame in the skeleton sequence is $A + I$, where A is constructed by the natural connections of human joints, and the identity matrix I represents the self-connections of all the joints. $A + I$ can simultaneously depict the spatial structural features between joints and the feature information of the joints themselves. Furthermore, because the edges connecting joints in the spatial structure have different significance, a learnable weight matrix M of edges is introduced to form a new adjacency matrix $(A + I) \otimes M$ for scaling the contributions of the same node feature to the other neighboring nodes. M is initialized as an all-ones matrix. Finally, the spatial graph convolutional operation can be realized by

$$G_{out} = M \otimes D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}G_{in}W \quad (1)$$

where $D^{ii} = \sum_j (A^{ij} + I^{ij})$ is the degree matrix, and the renormalization trick (i.e., $D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}$) proposed in [34] is introduced to avoid numerical instabilities. W is the weight function and is realized by 1×1 convolutional operation. After the spatial graph convolutional operation, the output feature map is obtained as $G_{out} \in \mathbb{R}^{C_{out} \times T \times N}$.

To model the temporal information in the skeleton sequence, the classical convolutional operation is directly applied in the temporal dimension. Specifically, since same nodes in all the frames can be naturally organized into two-dimensional sequences, the $\mathcal{T} \times 1$ standard convolutional operation is performed on the output feature map of the spatial graph convolutional operation, where $\mathcal{T}$ is the kernel size of the temporal convolutional operation.

B. *Design of the MS-STAM under deep supervision*

1) *Spatial-temporal attention-aware block (STAB)*

Inspired by [35]–[37], we developed a STAB, whose structure is shown in Fig. 5. Let $F_{in} \in \mathbb{R}^{C_F \times T_F \times N_F}$ be the output feature activation map of a specific layer, where C_F, T_F and N_F are the number of channels, frames, and joints, respectively. The STAB calculates the two-dimensional spatial-temporal attention coefficient matrix $\alpha \in \mathbb{R}^{T_F \times N_F}$, which combines the high-level saliency map and the low-level diversity feature map in the spatial-temporal dimension. Then, α is expanded as $\alpha^* \in \mathbb{R}^{C_F \times T_F \times N_F}$ along the channel dimension, thereby identifying the discriminative spatial-temporal feature regions in each channel and enhancing the feature activation in such regions in the low-level feature map F_{in}. Finally, the output of the STAB is $F_{out} = \alpha^* F_{in}$, which is obtained by multiplying the low-level feature map F_{in} and the corresponding spatial-temporal attention-aware coefficient matrix α^*.

The STAB captures the saliency map at the level of spatial-temporal grid as the feature enhancement signal to drive the perception of spatial-temporal discriminative regions and construct their relationship in the global scope. The output feature map of the last ST-GCN unit is considered as the saliency map $E \in \mathbb{R}^{C_E \times T_E \times N_E}$ to implicitly drive the perception of discriminative regions in the spatial-temporal dimension.

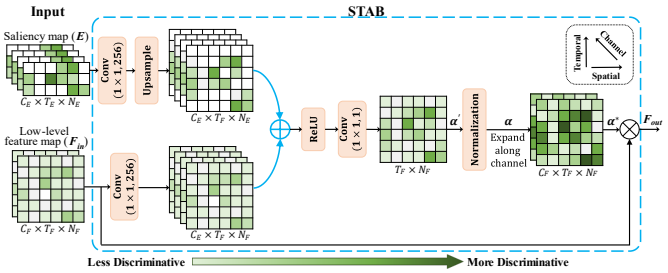


Fig. 5. Implementation of the STAB.

Then, E is upsampled to match the temporal dimension of the low-level feature map F_{in}. Thus, STAB exhibits robust perception ability in the selection of spatial-temporal discriminative feature areas. The specific process of the STAB can be expressed as follows:

$$\alpha = \delta_2(\theta_3(\delta_1(\theta_1(F_{in}) + U(\theta_2(E))))) \quad (2)$$

where θ_1, θ_2, and θ_3 are all realized by 1×1 convolutional operations, and U is the upsampling operation. $\delta_1(\cdot)$ is the rectified linear unit function [38], and $\delta_2(\cdot)$ is the normalization function of the spatial-temporal attention-aware coefficient matrix α, which is obtained as

$$\delta_2(\alpha') = (\alpha' - \alpha'_{min}) / \sum_{i=1}^{T_F \times N_F} (\alpha'_i - \alpha'_{min}) \quad (3)$$

In addition, the STAB can automatically suppress insignificant features in the forward or backward propagation process, and the gradient from insignificant feature areas are gradually reduced in the backward propagation process [35], [39]–[42]. Thus, the parameters of low layers are mainly updated according to the saliency areas perceived by the STAB. The convolutional parameter of the L th layer is updated according to the following relation:

$$\frac{\partial(F_{out}^{L+1})}{\partial(\theta^L)} = \frac{\partial(\alpha^{*L+1} \xi(F_{in}^L, \theta^L))}{\partial(\theta^L)} = \alpha^{*L+1} \frac{\partial(\xi(F_{in}^L, \theta^L))}{\partial(\theta^L)} + \frac{\partial(\alpha^{*L+1})}{\partial(\theta^L)} F_{in}^{L+1} \quad (4)$$

where θ denotes the convolutional parameters. It can be seen that the first gradient term on the right side of (4) is scaled by the attention coefficient to suppress the response of insignificant areas in the spatial-temporal dimension.

2) *Implementation of the MS-STAM under deep supervision*

To construct multi-level spatial-temporal information, we used the MS-STAM to learn different multi-scale spatial-temporal attention-aware coefficients. In this way, each attention block can learn to focus on spatial-temporal discriminative feature areas related to the classification task. In the STAB of each scale, the output of attention block is generated by perceiving spatial-temporal discriminative feature information through the high-level saliency map.

Although the multi-scale feature information generated by the MS-STAM has different feature levels, it ultimately serves the same classification task. Therefore, the combination constraint of deep supervision [43] is used on the output of each fully connected layer in the cost function; thus, the correlation of feature information in different scales is enhanced. This can be expressed as follows:

$$\mathcal{L}_d = -\frac{1}{N} \sum_{i=1}^{N} \sum_{q=1}^{NClass} \sum_{m=1}^{NF} (y_{i,q,m} \cdot \log \hat{y}^*_{i,q,m}) + \lambda_1 \frac{1}{N} \left(\sum_{i=1}^{N} \sum_{q=1}^{NClass} \sum_{m=1}^{NF} \sum_{n=m+1}^{NF} (\hat{y}_{i,q,m} - \hat{y}_{i,q,n})^2 \right)^{\frac{1}{2}} \quad (5)$$

where N is the batch size, and y is the true label. $\hat{y}^*$ and $\hat{y}$ are the predicted outputs after and before the softmax function,

respectively. $NClass$ is the number of sample categories, and NF is the number of fully connected layers fitted under different scales. The first term of $\mathcal{L}_d$ represents the cross-entropy loss, which aims to minimize the classification errors. The second term is the L2 regularization term, which aims to minimize the differences between the classification probability scores obtained by the fully connected layers fitted under different scales and enhance the correlation of discriminative feature's selection among the scales.

C. Model-driven strategy for feature sparsification

In the fine-grained assessment of gait skeleton sequence in PD patients, the discriminative features are often sparse. Therefore, an L1 regularization term is added in the cost function to realize sparsification of discriminative features, i.e.,

$$\mathcal{L}_s = \lambda_2 \sum_{j=1}^{NF} \left\| w_j \right\|_1 \tag{6}$$

where w_j represents the weight of the corresponding fully connected layer, and sparsification is realized by constraining the weight values of all the fully connected layers of different scales.

Furthermore, by combining (5) and (6), the total cost function of the proposed model consists of a deep supervision term that constrains multi-scale branches and the feature sparsification term, which can be expressed as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_d + \mathcal{L}_s \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{q=1}^{NClass}\sum_{m=1}^{NF}\left(y_{i,q,m}\cdot \log \hat{y}^*_{i,q,m}\right) \\ &+\lambda_1 \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{q=1}^{NClass}\sum_{m=1}^{NF}\sum_{n=m+1}^{NF}\left(\hat{y}_{i,q,m}-\hat{y}_{i,q,n}\right)^2\right)^{\frac{1}{2}} \\ &+\lambda_2 \sum_{j=1}^{NF}\left\| w_j \right\|_1 \end{aligned} \tag{7}$$

where λ_1 and λ_2 are trade-off parameters to balance the relative importance of two regularization terms. The specific values of λ_1 and λ_2 are obtained through many experiments. Algorithm 1 shows the calculation process of cost function in the form of a pseudo-code.

Algorithm 1: Function for Total Loss

Input: True label: y, predicted output: $\hat{y}^*, \hat{y}$, batch size: N, parameter: λ_1, λ_2, weights of the four fully connected layers: w_1, w_2, w_3, w_4;
Output: Total loss: $\mathcal{L}$;

$l_1 \leftarrow 0, l_2 \leftarrow 0$;
for $i \leftarrow 1$ *to* N:
  for $q \leftarrow 1$ *to* $NClass$:
    for $m \leftarrow 1$ *to* NF:
      $l_1 \leftarrow l_1 + (-y[i][q][m] \times \log \hat{y}^*[i][q][m])$;
      for $n \leftarrow m+1$ *to* NF:
        $l_2 \leftarrow l_2 + (\hat{y}[i][q][m] - \hat{y}[i][q][n])^2$;
$\mathcal{L}_d \leftarrow (l_1 + \lambda_1 \times l_2^{1/2})/N$;
$\mathcal{L}_s \leftarrow \lambda_2 \times (\|w_1\|_1 + \|w_2\|_1 + \|w_3\|_1 + \|w_4\|_1)$;
$\mathcal{L} \leftarrow \mathcal{L}_d + \mathcal{L}_s$;
Return: $\mathcal{L}$

IV. Experiments and Results

A. Dataset and evaluation metrics

This study was approved by the Institutional Review Board of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. The evaluation dataset was obtained from the Neurosurgery Department of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. It contains motor examination videos of 157 PD patients from January 2017 to July 2019 based on the MDS-UPDRS; the videos were recorded for each patient in different states according to the medication and deep brain stimulation (DBS) states, as well as the postoperative follow-up times. Some videos were excluded due to severe shaking of the camera during video recording and the irregular process of gait assessment. In addition, in clinical practice, some patients with a gait assessment score of 4 could hardly walk, making it impossible to record their gait videos. Consequently, the final number of patients with gait videos was 142, and the total number of states obtained was 441. The videos of these states were rated on a scale of 0-3. Only the first round of the patient's gait was clipped in each video. In practice, there are more patients with scores 1 and 2 and fewer patients with scores 0 and 3. To balance the relative distribution of score categories, we also collected videos of the remaining rounds of multi-round walking in the videos of patients with scores 0 and 3 for data augmentation. However, the expanded video clips in each state were used in the training or validation group (i.e., in the same fold) to ensure the reliability and robustness of the model. Finally, we obtained a total of 725 videos for evaluation.

To comprehensively evaluate the method, we also established an independent test dataset collected by the Neurosurgery Department of Ruijin Hospital from November 2019 to February 2020, including 42 patients with 76 states. The distribution of gait motor disorder scores in our dataset is shown in Table I.

TABLE I
Distribution of The MDS-UPDRS Gait Scores in Our Dataset

Gait score		0	1	2	3	Total
Number	Cross-validation	38	194	154	55	441
	Independent test	8	36	30	2	76

In this study, we focused on the assessment of gait motor disorder based on the skeleton sequence, so joint positions in the pixel coordinate system of the original RGB frames was considered as the input data. The state-of-the-art pose estimation model AlphaPose [44] was applied to extract 2D coordinates of 18 human joints in each frame of videos. Further, each video was converted into a skeleton sequence represented by these coordinates, and the coordinate values were then normalized. Moreover, all coordinates in each video were mapped to a coordinate system with the center point of the human body in the first frame as the origin.

The experimental results were assessed by five-fold cross-validation (CV), i.e., the cross-validation dataset was randomly divided into five fixed independent folds; four folds were used for training, and the remaining one was used for independent testing. The parameter setting of each experiment was the same, and random number seed was set to ensure complete reproducibility of the experimental results. Accuracy (Acc), precision (Prec), recall (Rec), f1-score (F1), and area under the curve (AUC) of receiver operating characteristic curve (ROC) were used as evaluation metrics to evaluate the classification results. These metrics are defined as follows: $\mathrm{Acc} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP})$, $\mathrm{Prec} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, $\mathrm{Rec} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, $\mathrm{F1} = 2 \times (\mathrm{Prec} \times \mathrm{Rec})/(\mathrm{Prec}+\mathrm{Rec})$, where TP, TN, FP, and FN

represent the number of true positive, true negative, false positive, and false negative samples, respectively. Further, in clinical practice, due to the subtle movement difference of assessment scores as well as the subjectivity and inter-rater variability [45]–[47], the assessment score with an error range of one-point fluctuation is considered to be acceptable by neurologists [7], [48], which is consistent with the natural attribute of the MDS-UPDRS [10]. Thus, we also introduced acceptable accuracy (Acceptable Acc) to indicate the accuracy when the error between the model assessment score and the reference score was no more than one point [6], [7], which was deemed to be an acceptable assessment [7], [48].

B. Implementation details

The proposed network was implemented using the deep learning framework of PyTorch. During the experiment, the initial learning rate was set as 1e-4, and the learning rate was reduced to one-tenth of the previous value after the 70th, 80th, and 90th epochs. The batch size was 8. Stochastic gradient descent (SGD) with Nesterov momentum of 0.9 was adopted as the optimization strategy, and the training process ended at the 115th epoch. The average time required for model training using our network on a single standard NVIDIA Geforce GTX 1080Ti graphics card (11 GB memory) was approximately 0.6 h.

C. Qualitative and quantitative analyses

1) Classification results

Quantitative five-fold CV was adopted to evaluate the proposed model on the experimental dataset, and the results are shown in Table II. The total accuracy and acceptable accuracy of the model are 65.66% and 98.90%, respectively. We also calculated Acc, Acceptable Acc, Prec, Rec, F1, and AUC of each category, and the ROC curve of each category is shown in Fig. 6. It is clear that each category achieves reliable results with acceptable accuracies higher than 97% and AUCs greater than 0.79.

Ablation experiments were also conducted to verify the necessity and effectiveness of all strategies, as shown in Table III. The joint-based ST-GCN (Joint-ST-GCN) serves as the baseline network with an accuracy of 57.24%. The 2s-ST-AGCN achieves better performance than the baseline network, and the necessity of the feature sparsification strategy is verified based on the experimental results with and without $\mathcal{L}_s$. Specifically, the total accuracy of the 2s-ST-AGCN is 65.66%, which is 8.42% higher than that of the baseline network. The excellent quantitative analysis results validate the effectiveness of the MS-STAM under deep supervision and the feature sparsification strategy.

TABLE II
CLASSIFICATION RESULTS OF THE PROPOSED MODEL

	Acc (%)	Acceptable Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC
Score-0	71.96	100.00	72.64	71.96	72.30	0.89
Score-1	60.82	99.48	54.13	60.82	57.28	0.79
Score-2	44.16	97.40	59.65	44.16	50.75	0.82
Score-3	83.44	98.16	75.14	83.44	79.07	0.95
Total Acc (%)			65.66			
Total Acceptable Acc (%)			98.90			

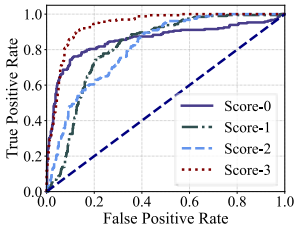


Fig. 6. ROC curves of all score categories.

TABLE III
ABLATION RESULTS OF THE PROPOSED MODEL

Proposed Model	Total		Average			
	Acc (%)	Acceptable Acc (%)	AUC	Prec (%)	Rec (%)	F1 (%)
Joint-ST-GCN (baseline)	57.24	98.21	0.83	57.90	56.87	57.24
2s-ST-GCN	63.03	99.03	0.86	63.52	62.71	62.97
2s-ST-GCN + $\mathcal{L}_s$	63.72	99.31	0.86	63.95	63.24	63.44
2s-ST-GCN + MS-STAM with $\mathcal{L}_d$	64.55	99.45	0.86	65.17	64.20	64.54
2s-ST-AGCN	**65.66**	**98.90**	**0.86**	**65.39**	**65.09**	**64.85**

2) Effectiveness of two-stream framework

To further test the necessity of the two-stream framework, we compared the performance of the proposed two-stream framework with that of three other methods, namely Joint-ST-GCN, 2s-ST-GCN, and joint-based ST-AGCN (Joint-ST-AGCN). The results are shown in Table IV, which confirms that the proposed scheme with two streams, i.e., 2s-ST-AGCN, achieves better performance than the same scheme with only joint stream, i.e., Joint-ST-AGCN, showing a 2.49% improvement in accuracy. Further, it is notable that the 2s-ST-GCN is more reliable than Joint-ST-GCN (Table II and Table III) with an improved accuracy of 63.03%. Thus, these comparison results prove that the two-stream framework is useful for improving the model performance.

TABLE IV
PERFORMANCE COMPARISON OF TWO-STREAM FRAMEWORK

Method	Total		Average AUC
	Acc (%)	Acceptable Acc (%)	
Joint-ST-GCN	57.24	98.21	0.83
2s-ST-GCN	63.03	99.03	0.86
Joint-ST-AGCN	63.17	97.93	0.85
2s-ST-AGCN	65.66	98.90	0.86

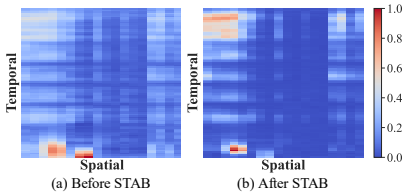
3) Effectiveness of STAB

As shown in Table III, compared to the baseline network, the introduction of the MS-STAM under deep supervision causes a 7.31% improvement in accuracy, and other evaluation metrics are also significantly improved. To prove the necessity of STAB in the MS-STAM, the results of the proposed model without STAB and with STAB† (without saliency map) (Fig. S2) are presented in Table V. The STAB and saliency map improve the accuracy by 1.80% and 1.25%, respectively, which verifies their importance in the MS-STAM (see supplementary material Fig. S3 and Table S3 for the statistically significant analysis). Notably, the saliency map is an indispensable

TABLE V
ABLATION RESULTS OF THE STAB

Model	Total		Average			
	Acc (%)	Acceptable Acc (%)	AUC	Prec (%)	Rec (%)	F1 (%)
Without STAB	63.86	98.07	0.85	62.88	63.07	62.19
With STAB† (without saliency map)	64.41	98.90	0.85	64.41	63.82	63.82
With STAB (with all components)	**65.66**	**98.90**	**0.86**	**65.39**	**65.09**	**64.85**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068609, IEEE Transactions on Multimedia



(a) Before STAB (b) After STAB

Fig. 7. Comparison of the spatial-temporal feature maps before STAB (a) and after STAB (b). The horizontal axis and vertical axis represent the spatial and temporal dimensions, respectively. The color-bar represents the intensity of feature maps.

component of the STAB. The spatial-temporal feature maps before and after STAB (Fig. S2(c)) are shown in Fig. 7 for better visualization. It is clear that saliency areas are greatly strengthened, and the non-significant features are suppressed by the STAB (Fig. 7(b)).

4) Effectiveness of sparsification strategy

We introduced the L1 regularization term for sparsification of features. The results presented in Table III validate the feasibility of our idea. First, the sparsification strategy is introduced in the 2s-ST-GCN, which leads to a 0.69% improvement in accuracy and significant improvement in other evaluation metrics. Subsequently, based on the realization of the MS-STAM under deep supervision, the sparsification strategy is again introduced, which causes a 1.11% improvement in accuracy. As shown in Fig. 8, through the visualization of the attention coefficient matrix, it is evident that the incorporation of sparsification strategy enhances the selectivity of saliency features, and the response values of the attention coefficient in saliency areas become larger and sparser. Furthermore, since sparsification was realized by constraining the weights of the fully connected layers, the distributions of the weight values of four fully connected layers were examined. Fig. 9(a) shows that without the sparsification strategy, the weight values are distributed over a wide range, whereas Fig. 9(b) illustrates that with the sparsification strategy, the weight distribution is sparser, which enhances the selectivity of the discriminative features.

5) Comparison with state-of-the-art skeleton-based action recognition methods

To validate the efficacy of the proposed model in the assessment of Parkinsonian gait, we compared our model with four state-of-the-art action recognition methods [16]–[19], namely the two-stream convolutional neural network (Two-stream CNN) [16], spatial-temporal graph convolutional network (ST-GCN) [19], two-stream adaptive graph convolutional network (2s-AGCN) [18], and motif-based graph convolutional network with variable temporal dense block architecture (Motif-STGCN) [17], using open-source codes. The results presented in Table VI prove that our method

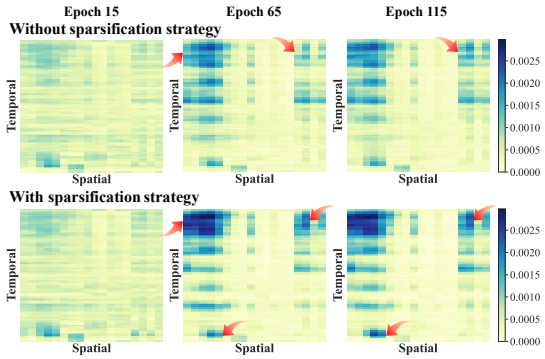
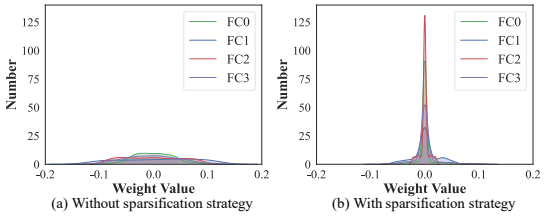


Fig. 8. Visualization of the spatial-temporal attention coefficient matrix in the STAB. Horizontal and vertical axes indicate spatial and temporal dimensions, respectively.



(a) Without sparsification strategy (b) With sparsification strategy

Fig. 9. Visualization of weight distributions of four fully connected layers in the network.

achieves the best performance in various evaluation metrics, especially with a total accuracy greater than 65% and average AUC greater than 0.86 (see supplementary material Fig. S1 for the statistically significant analysis).

We also compared the complexity (including parameters (Params), FLoating point Operations (FLOPs), and memory overhead (Memory)) and running speed (including training phase and inference computational time) of different state-of-the-art models, as shown in Table VI. Params, FLOPs and Memory were all calculated through an open-source tool (https://github.com/Swall0w/torchstat). The training phase time was calculated based on one batch in one stream. Compared to 2s-AGCN [18] with the second-highest accuracy in these state-of-the-art models, our model accelerates the training process and reduces the memory overhead and inference computational time. Overall, our model achieves a balance between classification performance and complexity, and provides a real-time assessment method for Parkinsonian gait.

6) Effectiveness of model-driven method and sensitivity analysis

We incorporated the model-driven terms, i.e., L1 and L2 regularization terms into the cost function and conducted grid-search experiments to determine the optimal parameter values as $\lambda_1 = 0.0001$ and $\lambda_2 = 0.03$. Further, we conducted ablation experiments using these two terms to prove the

TABLE VI
COMPARISON BETWEEN THE PROPOSED MODEL AND EXISTING STATE-OF-THE-ART METHODS

State-of-the-art	Total		Average				Params (M)	FLOPs (G)	Memory (MB)	Training time (s/epoch)	Inference computational time(s)
	Acc(%)	Acceptable Acc(%)	AUC	Prec(%)	Rec(%)	F1(%)					
Two-stream CNN [16]	60.69	98.07	0.84	59.36	60.15	59.29	12.25	0.09	3.15	0.79	0.0012
ST-GCN [19] (Spatial Configuration)	57.24	97.66	0.83	57.48	56.70	56.92	2.63	3.30	70.34	4.81	0.0063
2s-AGCN [18]	62.21	98.34	0.85	62.37	61.96	62.10	3.44	4.48	107.27	12.09	0.0201
Motif-STGCN [17]	57.51	97.66	0.84	58.10	57.08	57.39	1.72	2.54	136.93	16.67	0.0180
2s-ST-AGCN (ours)	**65.66**	**98.90**	**0.86**	**65.39**	**65.09**	**64.85**	**2.81**	**3.53**	**77.39**	**9.32**	**0.0065**

effectiveness of the model-driven strategy. The experimental results are shown in Table VII, which suggests that both regularization terms individually improve the performance of the network to a certain extent and better performance can be achieved by using their combination (see supplementary material Table S1 and Table S2 for the statistically significant analysis).

To further verify the effect of these two parameters on the performance, we performed sensitivity analysis on λ_1 and λ_2. We introduced 10% fluctuation in the positive and negative directions based on the optimal values of λ_1 and λ_2, respectively. Further, to measure the degree of variation in accuracy, the absolute accuracy variation ratio (AAVR) is defined as follows:

$$\mathrm{AAVR} = \left|\frac{Acc_{new} - Acc_{benchmark}}{Acc_{benchmark}}\right| \times 100\% \quad (8)$$

where $Acc_{benchmark}$ is the accuracy obtained by the model under the reference parameter setting ($\lambda_1 = 0.0001$, $\lambda_2 = 0.03$), and Acc_{new} is the accuracy under different parameter setting. The analysis results are shown in Fig. 10. The combination of the two parameters with 10% fluctuation in the positive and negative directions has a minor effect on the final accuracy, and the maximum effect is 2.3%. The sensitivity analysis shows that the proposed model-driven strategy is insensitive to other parameter combinations around the benchmark parameter setting, which further proves the robustness of our model.

TABLE VII
ABLATION COMPARISON RESULTS OF THE MODEL-DRIVEN STRATEGY

Parameter Setting	Total		Average			
	Acc (%)	Acceptable Acc(%)	AUC	Prec (%)	Rec (%)	F1 (%)
$\lambda_1 = 0, \lambda_2 = 0$	64.41	99.59	0.86	64.93	63.96	64.25
$\lambda_1 = 0.0001, \lambda_2 = 0$	64.55	99.45	0.86	65.17	64.20	64.54
$\lambda_1 = 0, \lambda_2 = 0.03$	65.52	98.62	0.86	65.01	64.90	64.68
$\lambda_1 = 0.0001, \lambda_2 = 0.03$	65.66	98.90	0.86	65.39	65.09	64.85

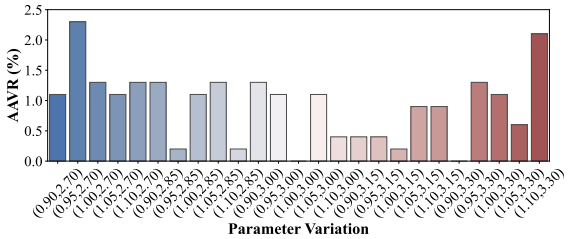


Fig. 10. Sensitivity analysis of parameter λ_1 and λ_2 in our model. The x-axis represents the parameter combination of $\lambda_1(\times 1e-4)$ and $\lambda_2(\times 1e-2)$, and the y-axis represents the absolute accuracy variation ratio.

7) Reliability analysis

Reliability analysis by repetitions: To prove the stability of our method, we performed 5-fold CV experiments for 10 times. In each experiment, the dataset was randomly split into 5-folds. The results are presented in Fig. 11. The accuracies range from 65.10% to 66.48%, and acceptable accuracies range from 98.21% to 99.45%, illustrating that our method is stable in both accuracy and acceptable accuracy. The accuracy at the lower quartile with the scatter point is reported in this paper.

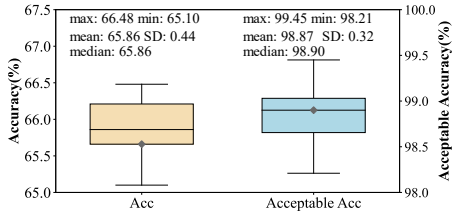


Fig. 11. Results of 10 repeated experiments. The scatter points represent the results reported in this paper.

Results of independent test: To further verify the reliability of the proposed method, we trained the proposed model on the cross-validation dataset and then tested it on the independent test dataset, as shown in Table VIII. The obtained accuracy of 64.47% and acceptable accuracy of 98.68% further confirm the reliability of the proposed method.

TABLE VIII
EVALUATION RESULTS OF THE INDEPENDENT TEST DATASET

Result	Total		Average			
	Acc (%)	Acceptable Acc(%)	AUC	Prec (%)	Rec (%)	F1 (%)
2s-ST-AGCN	64.47	98.68	0.79	54.25	67.71	55.10

8) Comparison with existing related studies

Table IX compares the performance of the proposed model with that of the existing models for quantitative gait assessment of PD patients. Parisi *et al.* [6], [7] proposed a sensor-based assessment of gait with the k-nearest neighbors method and achieved an accuracy of approximately 60% and an acceptance accuracy higher than 90%. Sabo *et al.* [10] tracked gait features from the videos collected through a Microsoft Kinect sensor and color camera, and the performance of their scheme was close to that of the existing sensor-based scheme [7]. Our video-based assessment method with the 2s-ST-AGCN exhibits better performance in terms of both accuracy and acceptable

TABLE IX
COMPARISON WITH EXISTING RELATED STUDIES

Author/Year	Resource	Features	Participants	Methods	Performance	
					Acc (%)	Acceptable Acc(%)
Parisi *et al.*[7]/2015	Body sensor network	Kinematic features in both time and frequency domains	34 PD; 47 samples	K nearest neighbors	62	98
Parisi *et al.*[6]/2016	Body sensor network	Kinematic features in both time and frequency domains	34 PD, 4 HC; 55 samples	K nearest neighbors	53	98
Sabo *et al.* [10]/2020	Microsoft Kinect sensor + camera	16 3D features	14 PD; 398 samples	Multivariate ordinal logistic regression	62.1	97.24
		8 2D features	14 PD; 249 samples	Multivariate ordinal logistic regression	61.4	98.80
Ours/2020	**Videos**	**Discriminative spatial-temporal features**	**142 PD; 441 samples**	**2s-ST-AGCN**	**65.66**	**98.90**

accuracy. This demonstrates that the proposed skeleton-based scheme combined with the GCN outperforms the sensor- and video-based methods with traditional machine learning algorithms, confirming the effectiveness of the proposed assessment scheme. However, these studies used different amounts and types of data and preprocessing methods; hence, a comparison with them may not be justified; nevertheless, it can still provide a valuable reference for the academic community.

V. Discussion

The automated quantitative assessment of gait motor disorder is of great significance for monitoring the health of PD patients. It can help clinicians to perform an accurate and timely follow up on the patients' progress, thereby improving the clinical diagnosis and decision-making process. Currently, the sensor-based automated quantitative scheme is rarely used in clinical applications because the sensors can affect the movements of patients. In addition, the traditional image processing and feature engineering methods used in the vision-based automated quantitative scheme fail to accurately characterize the movements and face considerable challenges in the fine-grained evaluation task. Therefore, in this study, we applied a deep learning framework based on videos to realize automated quantitative assessment of gait motor disorder. The 2s-ST-GCN was used to model the joints of the human body in the spatial-temporal domain. The MS-STAM was designed to effectively extract the fine-grained spatial-temporal features, which was then combined with the constraint of deep supervision to minimize the classification errors and optimize the network performance. Finally, the model-driven items, i.e., L1 and L2 regularization terms, were embedded for the sparsification of discriminative features and the enhancement of strong correlation between different scales, respectively. The proposed model exhibited excellent performance, which validated its immense potential in the quantitative gait assessment of PD patients.

We quantitatively analyzed several experimental results and proved that the proposed model exhibits better performance than the existing state-of-the-art action recognition models, demonstrating its potential in clinical practice. By contrast with these state-of-the-art models, our model has technical advantages. First, the multi-scale spatial-temporal attention-aware mechanism under deep supervision is developed to capture the fine-grained discriminative spatial-temporal features. Specifically, the high-level feature map (i.e., saliency map) and low-level feature map are fused in channel dimension, which is followed by the normalization in spatial-temporal dimension to realize the spatial-temporal attention mechanism, thereby guiding the middle layers to enhance saliency areas and suppress irrelevant features. Afterward, the deep supervision scheme is applied to enhance the similarity between different scales for promoting the network to extract useful salient features. Besides, a sparsification strategy is proposed to identify significant discriminative features. However, our method is designed for a specific task (i.e., gait motor assessment in PD); hence, its generality may be limited. Nevertheless, it can still provide a useful scheme for the motor assessment of other diseases, such as dystonia.

Furthermore, the proposed model was compared with the existing models on PD gait assessment based on the MDS-UPDRS (or UPDRS) in Table IX. Parisi *et al.* [6], [7] used traditional machine learning method to quantitatively assess the gait motor disorder based on the features extracted from the sensor signals. Sabo *et al.* [10] calculated gait features from the joint coordinates of human body and achieved the performance equivalent to that of the sensor-based scheme [7]. Because it is difficult to collect the motion data of patients with the highest severity of gait motor disorder in practice, their experimental dataset did not have any video with an assessment score of 4. However, our accuracy was higher than those of the three studies, demonstrating that the proposed skeleton-based scheme with the novel GCN algorithm can facilitate excellent feature extraction and fine-grain recognition ability. Our model is end-to-end, so it is potentially useful for translational research and clinical practice. Furthermore, compared to these existing studies, the proposed model has following advantages: 1) The video-based scheme is practical in clinical applications, and skeleton feature extraction using deep learning model is automatic and comprehensive. 2) To the best of our knowledge, our video dataset of gait task for PD patients is the largest; thus, our model is more likely to meet the requirements of clinical practice.

Although the proposed model achieved an acceptable accuracy of 98.90%, there is still room for further improvement. We observed that the classification performance of samples with score 2 is weak, and most of them are erroneously assigned score of 1 or 3, which dramatically affects the absolute accuracy of the model. Thus, we aim to improve the fine-grained assessment of score 2 in future work.

VI. Conclusions

To realize automated quantitative assessment of gait motor disorder in PD patients using gait videos, we proposed a new method called 2s-ST-AGCN based on the GCN framework and combined it with the MS-STAM under deep supervision and model-driven scheme. Specifically, the spatial structure and temporal dynamics of the joints and bones were modeled. The introduction of the MS-STAM effectively extracted the discriminative fine-grained spatial-temporal features of the skeleton sequence, and the deep supervision scheme was designed to minimize classification errors and enhance the robustness and discriminativeness of learned features. In addition, model-driven terms included L1 regularization term for feature sparsification and L2 regularization term for strengthening the similarity of different scales.

We conducted experiments on a clinical dataset to prove the efficacy of the proposed model and obtained reasonable results with an acceptable accuracy of 98.90%, which is within an acceptable error range for neurologists. Overall, the proposed method provides a potential tool for boosting the application of automated quantitative assessment of gait videos of PD patients.

REFERENCES

[1] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," *J. Neurol. Neurosurg. psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[2] M. D. S. T. F. on R. S. for P. Disease, "The unified Parkinson's disease rating scale (UPDRS): status and recommendations," *Mov. Disord.*, vol. 18, no. 7, pp. 738–750, 2003.

[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," *Mov. Disord. Off. J. Mov. Disord. Soc.*, vol. 23, no. 15, pp. 2129–2170, 2008.

[4] T. Burki, "Outbreak of coronavirus disease 2019," *Lancet Infect. Dis.*, 2020.

[5] M. E. Morris and R. Iansek, "Characteristics of motor disturbance in Parkinson's disease and strategies for movement rehabilitation," *Hum. Mov. Sci.*, vol. 15, no. 5, pp. 649–669, 1996.

[6] F. Parisi *et al.*, "Inertial BSN-based characterization and automatic UPDRS evaluation of the gait task of Parkinsonians," *IEEE Trans. Affect. Comput.*, vol. 7, no. 3, pp. 258–271, 2016.

[7] 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. Heal. informatics*, vol. 19, no. 6, pp. 1777–1793, 2015.

[8] T. Aşuroğlu, K. Açıcı, Ç. B. Erdaş, M. K. Toprak, H. Erdem, and H. Oğul, "Parkinson's disease monitoring from gait analysis via foot-worn sensors," *Biocybern. Biomed. Eng.*, vol. 38, no. 3, pp. 760–772, 2018.

[9] Y.-Y. Chen *et al.*, "A vision-based regression model to evaluate Parkinsonian gait from monocular image sequences," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 520–526, 2012.

[10] 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.*, 2020.

[11] M. Leo, P. Carcagnì, P. L. Mazzeo, P. Spagnolo, D. Cazzato, and C. Distante, "Analysis of Facial Information for Healthcare Applications: A Survey on Computer Vision-Based Approaches," *Information*, vol. 11, no. 3, p. 128, 2020.

[12] S. Yeung *et al.*, "A computer vision system for deep learning-based detection of patient mobilization activities in the ICU," *npj Digit. Med.*, 2019.

[13] X. Liao *et al.*, "Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[14] S. Zhang *et al.*, "Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks," *IEEE Trans. Multimed.*, vol. 20, no. 9, pp. 2330–2343, 2018.

[15] K. Zhu, R. Wang, Q. Zhao, J. Cheng, and D. Tao, "A Cuboid CNN Model with an Attention Mechanism for Skeleton-based Action Recognition," *IEEE Trans. Multimed.*, 2019.

[16] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2017, pp. 597–600.

[17] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8989–8996.

[18] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12026–12035.

[19] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] J. P. Singh, S. Jain, S. Arora, and U. P. Singh, "Vision-based gait recognition: A survey," *IEEE Access*, vol. 6, pp. 70497–70527, 2018.

[21] S. Li, W. Liu, and H. Ma, "Attentive Spatial–Temporal Summary Networks for Feature Learning in Irregular Gait Recognition," *IEEE Trans. Multimed.*, vol. 21, no. 9, pp. 2361–2375, 2019.

[22] M. Ye, C. Yang, V. Stankovic, L. Stankovic, and S. Cheng, "Distinct feature extraction for video-based gait phase classification," *IEEE Trans. Multimed.*, vol. 22, no. 5, pp. 1113–1125, 2019.

[23] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Automated vision-based analysis of levodopa-induced dyskinesia with deep learning," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 3377–3380.

[24] 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, 2018.

[25] 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 Relat. Disord.*, vol. 53, pp. 42–45, 2018.

[26] 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, 2019.

[27] B. Ni, X. Yang, and S. Gao, "Progressively parsing interactional objects for fine grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1020–1028.

[28] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian, "Interaction part mining: A mid-level approach for fine-grained action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3323–3331.

[29] B. Ni, V. R. Paramathayalan, and P. Moulin, "Multiple granularity analysis for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 756–763.

[30] Y. Zhou, B. Ni, S. Yan, P. Moulin, and Q. Tian, "Pipelining localized semantic features for fine-grained action recognition," in *European conference on computer vision*, 2014, pp. 481–496.

[31] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[32] Y. Zhu and G. Liu, "Fine-grained action recognition using multi-view attentions," *Vis. Comput.*, pp. 1–11, 2019.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[34] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv Prepr. arXiv1609.02907*, 2016.

[35] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.

[36] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Trans. Multimed.*, vol. 21, no. 2, pp. 416–428, 2018.

[37] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Trans. Multimed.*, 2020.

[38] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[39] J. Schlemper *et al.*, "Attention-gated networks for improving ultrasound scan plane detection," *arXiv Prepr. arXiv1804.05338*, 2018.

[40] Y. Lei *et al.*, "CT prostate segmentation based on synthetic MRI-aided deep attention fully convolution network," *Med. Phys.*, 2020.

[41] W. Wang, C. Ye, S. Zhang, Y. Xu, and K. Wang, "Improving Whole-Heart CT Image Segmentation by Attention Mechanism," *IEEE Access*, 2020.

[42] B. Jun Guo *et al.*, "Automated left ventricular myocardium segmentation using 3D deeply supervised attention U-net for coronary computed tomography angiography; CT myocardium segmentation," *Med. Phys.*, 2020.

[43] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*, 2015, pp. 562–570.

[44] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2334–2343.

[45] J. P. Giuffrida, D. E. Riley, B. N. Maddux, and D. A. Heldmann, "Clinically deployable kinesia™ technology for automated tremor assessment," *Mov. Disord.*, 2009.

[46] T. H. Turner and M. L. Dale, "Inconsistent Movement Disorders Society–Unified Parkinson's Disease Rating Scale Part III Ratings in the Parkinson's Progression Marker Initiative," *Movement Disorders*. 2020.

[47] B. Post, M. P. Merkus, R. M. A. de Bie, R. J. de Haan, and J. D. Speelman, "Unified Parkinson's Disease Rating Scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?," *Mov. Disord.*, 2005.

[48] 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.*, 2020.