EGCN++: A New Fusion Strategy for Ensemble Learning in Skeleton-Based Rehabilitation Exercise Assessment

Bruce X. B. Yu, *Member, IEEE*, Yan Liu, *Member, IEEE*, Keith C. C. Chan, *Member, IEEE*, and Chang Wen Chen, *Fellow, IEEE*

***Abstract*—Skeleton-based exercise assessment focuses on evaluating the correctness or quality of an exercise performed by a subject. Skeleton data provide two groups of features (i.e., position and orientation), which existing methods have not fully harnessed. We previously proposed an ensemble-based graph convolutional network (EGCN) that considers both position and orientation features to construct a model-based approach. Integrating these types of features achieved better performance than available methods. However, EGCN lacked a fusion strategy across the data, feature, decision, and model levels. In this paper, we present an advanced framework, EGCN++, for rehabilitation exercise assessment. Based on EGCN, a new fusion strategy called MLE-PO is proposed for EGCN++; this technique considers fusion at the data and model levels. We conduct extensive cross-validation experiments and investigate the consistency between machine and human evaluations on three datasets: UI-PRMD, KIMORE, and EHE. Results demonstrate that MLE-PO outperforms other EGCN ensemble strategies and representative baselines. Furthermore, the MLE-PO's model evaluation scores are more quantitatively consistent with clinical evaluations than other ensemble strategies.**

***Index Terms*—Human action evaluation, model-based fusion, ensemble learning.**

I. Introduction

PHYSICAL therapists often use rehabilitation exercises to aid in the recovery and prevention of various musculoskeletal disorders (e.g., tendonitis, epicondylitis, and mechanical back syndrome). However, patients can rarely afford the expense of routine rehabilitation therapy [1]. People with musculoskeletal disorders are instead encouraged to engage in cost-effective home-based treatment under a rehabilitation therapist's supervision [2]. Some patients struggle to adhere to home-based exercise regimens due to a lack of immediate therapist feedback. This issue can impede therapeutic efficacy and increase overall healthcare expenditure [3]. Scholars have thus researched the use of vision sensors and wearable sensors for home-based physical rehabilitation and health monitoring [4], [5]. When these technologies were introduced, wearable sensors such as inertial measurement units (IMUs) and robot-aided devices with some degrees of freedom were tested in rehabilitation treatments [6], [7]. Wearable sensors have also been combined with vision sensors for this purpose [8]. Although IMU-based methods are promising and have demonstrated good motion data quality [9], they typically must be calibrated for each use and are not as convenient as vision sensors [10].

Since the release of affordable motion sensors such as Kinect, some home-based physical therapy systems [11], [12], [13] using a sensor's human body skeleton data have been developed to assess the accuracy of rehabilitation patients' exercises. These evaluation systems can identify barriers to improving patients' exercise adherence, which motivates patients to complete the exercises as therapists would [14]. The Kinect sensor contains various data-streaming channels such as skeleton, depth, and RGB [15]. In addition to the skeleton channel, the depth channel has been applied for abnormality detection by [16]. The RGB channel can be used as one particular way to get an RGB feed, which can be regarded as video-based action assessment. Prior video-based action assessment works usually tackle Olympic sports based on the RGB video data [17], [18], [19] or both video and 2D pose data [20], [21]. Skeleton data can be more informative than 2D pose data; as such, the current study enriches the literature on exercise evaluation by using the former type.

Skeleton data are streamed as a sequence of skeleton frames, with each containing several skeleton joints. Each joint has two feature groups as shown in Fig. 1: a 3D position and a 3D orientation (i.e., the joint angle). Traditionally, feature engineering methods calculate detailed joint information such as the joint angle [22] and changes in relative joint positions [23] to guide skeleton estimation. Although using features calculated or transformed from raw skeleton data can reduce complexity, doing so may not capture all necessary information [24]. These methods are easily interpretable for abnormality detection but come with

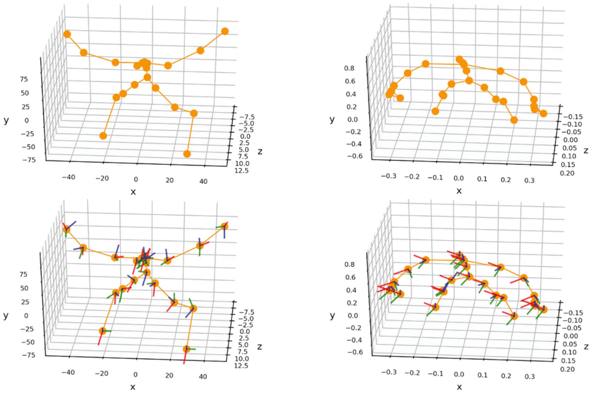
Manuscript received 24 January 2023; revised 16 January 2024; accepted 15 March 2024. Date of publication 20 March 2024; date of current version 6 August 2024. This work was supported in part by RGC Collaborative Research Fund under Grant C7069-19G, in part by the Huawei Technologies under Grant P0038941, and in part by the Grayson-Jockey Club Research Foundation under Grant P0041250. Recommended for acceptance by K.G. Derpanis. *(Corresponding author: Yan Liu.)*

This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.

Bruce X. B. Yu is with the Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, Hangzhou 310027, China (e-mail: xinboyu@intl.zju.edu.cn).

Yan Liu, Keith C. C. Chan, and Chang Wen Chen are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: csyliu@comp.polyu.edu.hk; keithccchan@gmail.com; chencw@comp.polyu.edu.hk).

Digital Object Identifier 10.1109/TPAMI.2024.3378753



UI-PRMD KIMORE

Fig. 1. Visualization of two skeleton features from UI-PRMD and KIMORE datasets. *Upper figures:* 3D position feature of skeleton joints (circles). *Lower figures:* 3D orientation feature of skeleton joints (coordinates on the circles). We aim to use both feature types for exercise assessment.

inter- and intra-subject bias (i.e., each subject might display their own correct performance at different rehabilitation stages). As for the characteristics of skeleton data, the two skeleton features are structurally homogeneous but have heterogeneous physical meanings. The position feature represents the global movement of exercise repetition, which is widely used for action recognition [25]. The orientation feature describes the local attributes of certain skeleton joints; it is relatively more independent. These two feature groups are mutually inconvertible. Exercise abnormalities can be discovered by analyzing global and local skeleton patterns accordingly [26], [27].

However, rehabilitation exercise assessment with skeleton data continues to face difficulties. First, methods such as [28], [29], [30] and [31] mainly rely on one modal setting (i.e., using the orientation feature of skeleton data) but do not capitalize on both position and orientation features. We previously proposed an ensemble-based graph convolutional network (EGCN) [32] framework to fuse these skeleton feature groups at different scales (i.e., early, mid, or late fusion), but this network does not integrate the benefits of each. Second, existing methods are limited in their capability to distinguish incorrect and correct actions: current approaches are either based on fixed geometric features [29], [33] or basic deep learning models such as a convolutional neural network (CNN), long short-term memory [34], or graph convolutional network (GCN) [30]. Several fusion strategies in EGCN have shown encouraging performance. Some could be combined to realize performance improvements. Third, the evaluation metrics (i.e., involving separation degree and distance) used in [28], [29], [30] have not been used with EGCN [32] to determine a model's prediction ability, which is integral to machine assessment. Metrics such as Euclidean distance and correlation, which indicated the consistency between human and machine assessments in [31], were not applied in [32]. An optimal model should be predictive and consistent with human evaluation.

An earlier version of this work appeared in [32]. To further facilitate skeleton-based exercise assessment with position and orientation features, we propose EGCN++, which tackles the aforementioned challenges. Besides providing more technical details in the methodology, this paper offers novel contributions: it presents a new ensemble strategy, more extensive experiments, and investigations of consistency with human evaluation. For the new ensemble strategy, we recognize the advantages of the data-level ensemble and combine this technique with the model-level ensemble (MLE), leading to a fusion strategy called MLE-PO that improves the previous MLE's results. In the more extensive experiments, we expand prior tests of the UI-PRMD [35] and KIMORE [36] datasets to an additional random division protocol and test all proposed methods on a new dataset, EHE [31], for comparison with state-of-the-art baselines. We also revisit the suitability of evaluation criteria in [28], [29] and report experimental results. Based on findings from experiments with the three latest public datasets, UI-PRMD [35], KIMORE [36], and EHE [31], the proposed MLE-PO strategy in our EGCN++ significantly outperforms other ensemble techniques and state-of-the-art GCN-based single-modal methods. Regarding our comparison with human evaluation, we provide a thorough quantitative assessment of the consistency between models' evaluation and human evaluation with datasets for which human evaluation results are available. Evaluation scores are additionally visualized and compared.

The remainder of this paper is organized as follows. Section II introduces related work. In Section III, we detail the proposed EGCN++ framework. Section IV provides experimental results for three benchmark datasets with ablations, comparisons with state-of-the-art methods, and runtime analysis. Section V discusses future directions, followed by a conclusion in Section VI.

II. Related Work

In this section, we review prior work on skeleton-based rehabilitation exercise assessment, graph representation, and ensemble learning.

A. Skeleton-Based Action Assessment

In this subsection, we briefly review related work in terms of datasets and algorithms. Some action evaluation datasets were reviewed in [37], [38], [39]. Datasets such as UI-PRMD [35] and KIMORE [36] are relevant to our study whereas other surveyed datasets either focus on action classification or are not skeleton-based. UI-PRMD used Vicon and Kinect v2 sensors to track repetitions of 10 exercises from 10 healthy subjects to explore good action evaluation models. To build systems to remotely monitor physical rehabilitation, KIMORE data were collected from 78 subjects (44 healthy people and 34 patients with motor dysfunction) via the Kinect v2 sensor. Other representative datasets include SPHERE [33], AHA-3D [40], LAM [41], and EHE [31]. The SPHERE dataset provides information on the body center rather than the entire skeleton, making it somewhat irrelevant for whole-skeleton modeling. The AHA-3D and LAM datasets are not publicly accessible as of this writing. The EHE dataset contains routine morning exercise repetitions from 25 residents of an elderly home, collected in a natural setting using the Kinect v2 sensor. KIMORE and EHE data were obtained from real patients and facilitate clinical evaluation. Such assessment is integral when determining algorithms' effectiveness.

Traditional machine learning models exemplify early algorithms. Multiple hidden Markov models were compared in [33] based on SPHERE. Parmar and Morris [41] used conventional algorithms (e.g., support vector machine, AdaBoosted tree, multi-layer neural network, and dynamic time warping) to evaluate exercise quality. The AdaBoosted tree yielded optimal performance. João et al. [40] devised a per-frame exercise evaluation method but did not consider the whole skeleton sequence, which was tested on the AHA-3D dataset. Elkholy et al. [42] collected a dataset similar to SPHERE [33] and proposed a hidden Markov model-based method with less computational overhead than that developed by [33]. The training process in [42] is supervised by the score of the abnormality degree (on a scale of 1–5) from a professional specialist's evaluation. More recently, a deep learning framework [28] was created to encode skeleton data from the UI-PRMD dataset, supervised by a quality score function. The approach in [28] outperformed certain CNN- and long short-term memory-based models. GCN was featured in [30], [31] and appeared superior to other methods [28], [29].

Approaches to exercise assessment tend to be based on one of two principles: 1) regression, in which case the strategies are supervised by either a score function [28] or clinical scores [42]; or 2) binary classification [30], [31], [32]. Input from clinical experts cannot fully determine assessment validity [43]; discrepancies can arise between experts' evaluations and patients' self-assessment [44]. Training a model using clinical labels renders the model evaluation inadequate for patients. At the same time, supervising the training process with a score function can be meaningless if a predefined score function has already delivered results. Our EGCN++ follows the work of [31] and delivers a numerical evaluation score by using output before the Softmax classifier. This output is then compared with human evaluation.

B. Graph Convolutional Network

Since [45], [46], [47] suggested generalizing CNN to relatively sparse graph data structures, GCN has been adopted to represent skeleton data when classifying human actions [48]. It can capture spatial and temporal attributes. However, when GCN is employed to present various skeleton data features, their physical meanings go overlooked. More advanced GCN models are now available [49], [50], [51] that separately train the skeleton joint and bone streams and aggregate these results. The skeleton bone stream is a transformed version of the skeleton joint position stream. Combining findings from skeleton joints and skeleton bones can enhance action recognition. Aggregating multiple types of skeleton representation data [52], [53] slightly improves action recognition accuracy. [51] works better when integrating the results of skeleton joint and bone modalities; this method is an ideal baseline for skeleton-based action recognition.

We will not elaborate on skeleton bone improvements because exercise assessment is distinct from action recognition: the former concerns the accuracy of a single action, whereas the latter classifies multiple actions. We will instead explore effective ensemble strategies that exploit the position and orientation features of skeleton data. In this work, we design ensemble techniques based on the basic GCN model in [48] for rehabilitation exercise assessment. We also explore the effects of changing the backbone of our EGCN++ using advanced GCN models.

C. Ensemble Learning

Ensemble learning has drawn growing interest due to combining data fusion, data modeling, and data mining into a unified framework [54]. This learning approach applies to classification tasks at the basic data level, feature level, decision level, and model level [54]. A more detailed categorization scheme for data stream classification appears in [55]. Ensemble classification models involve diversity, accuracy, and generalization. Conflicts between these attributes pose obstacles to better model performance. Reducing the model complexity and accelerating training speed constitute additional challenges [54]. Ensemble learning strategies entail forcing submodels' diversity or independence, focusing on local information, and proposing good aggregation mechanisms [56]. Traditional machine learning methods explore ensemble strategies to varying degrees via algorithms such as AdaBoost, bootstrapping, random forest, bagging, voting, and stacking. [57].

Deep learning models can benefit conventional ensemble learning in terms of feature extraction, base learner generation, and ensemble learner formation [58]. However, the costs of training multiple base learners and testing the ensemble learner often increase. Knowledge distillation [59], [60], [61] and ensemble selection-based aggregation criteria [62], [63] have been investigated to lower these costs, but smaller models usually underperform more robust ones. We aim to enhance model performance but do not address knowledge distillation because two feature groups are available.

To realize better performance, typical deep learning-based ensemble methods combine feature-level representations from different data or add decision-level results [64]. Ensemble methods also commonly force feature-level diversity or submodel independence [65]. However, ensuring model diversity remains difficult and does not guarantee good performance [66]. It is possible to improve the performance of ensemble classification models by considering the interconnections and feedback between levels (e.g., the sample level, feature level, and model level); however, more research should be conducted with a strong understanding of data attributes [54].

In this paper, we design an ensemble strategy that integrates data-level and model-level ensembles. This approach outperforms other ensemble strategies proposed in our EGCN++ learning framework.

III. Method

Here, we describe our approach to skeleton-based rehabilitation exercise assessment. We first introduce the GCN model adopted to represent spatial and temporal features within the position and orientation streams of skeleton data. Next, we

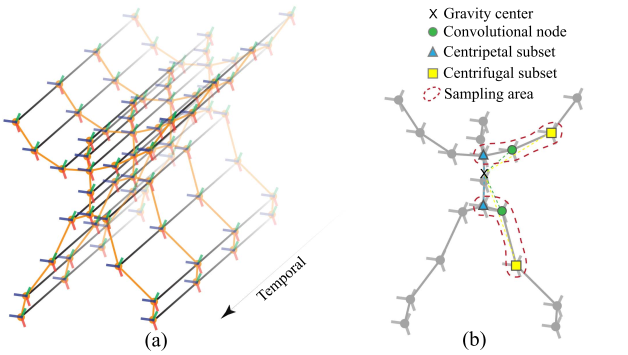


Fig. 2. (a) Illustration of spatiotemporal graph. (b) Illustration of spatial sampling strategy for convolutional operation.

describe ensemble strategies proposed in our EGCN++ learning framework.

Let us denote N samples of exercise repetitions as $S = \{S^{(n)}|n = 1,\ldots,N\}$, where each exercise repetition $S^{(n)}$ can have T skeleton frames collected at regular intervals. Given a specific skeleton structure with J joints, an exercise repetition can be represented as a set of $T \times J$ skeleton joints, which can be written as $S^{(n)} = \{S_{ti}^{(n)} = (P_{ti}^{(n)}, O_{ti}^{(n)})|\ t = 1,\ldots T,\ i = 1,\ldots,J\}$. $P_{ti}^{(n)} = (x, y, z)$ and $O_{ti}^{(n)} = (X, Y, Z)$ denote the position feature and the orientation feature, respectively. (x, y, z) are 3 attributes featuring the 3D cartesian coordinates of the skeleton position. (X, Y, Z) represents 3 attributes that can be transformed into the pitch, roll, and yaw values of the skeleton joint.

Given a sequence of skeleton frames in an exercise repetition $S^{(n)} = (P^{(n)}, O^{(n)})$, which features (x, y, z) and (X, Y, Z), let us use $g(P^{(n)}, \theta_g)$ and $h(O^{(n)}, \theta_h)$ (where θ_g and θ_h are learnable parameters) to denote submodels for learning features from the skeleton position and orientation streams, respectively. The goal is to model the P_{ti} and O_{ti} with proper ensemble strategies that produce higher-quality exercise evaluations through our EGCN++ framework based on different evaluation metrics.

A. Skeleton Data Representation

Taking inspiration from the GCN proposed in [48], we adopt a GCN model to represent skeleton joints' spatiotemporal relationships. Fig. 2(a) displays the constructed spatiotemporal skeleton graph, where joints are represented as vertexes and their natural connections are represented as spatial edges. For the temporal dimension, the black lines connecting corresponding joints between two adjacent skeleton frames are temporal edges. The attributes of each graph vertex are composed of the position and orientation streams of the corresponding skeleton joint. The skeleton graph at time t can be denoted as $\vartheta_t = \{\upsilon_t, \varepsilon_t\}$, where $\upsilon_t = \{\upsilon_{ti}|\upsilon_{ti} = S_{ti}^{(n)}, i = 1,\ldots,J\}$ denotes graph vertexes that can be viewed as corresponding skeleton joints. ε_t denotes spatial edges representing skeleton bones.

Similar to the convolutional operation in the CNN model, the traversal rules of graph convolutional operations rely on the definition of a sampling area. For a graph vertex υ_{ti}, its sampling area is defined by a neighbor set $N(\upsilon_{ti})$. Fig. 2(b) shows this strategy, where the dashed line curve encloses the neighbor set $N(\upsilon_{ti})$. This strategy empirically uses 3 spatial subsets: the vertex denoted by green circles in Fig. 2(b), the centripetal subset (blue triangles) that contains neighboring vertexes closer to the center of gravity, and the centrifugal subset (yellow squares) that contains neighboring vertexes farther from the center of gravity. Suppose $N(\upsilon_{ti})$ has K subsets that can be numerically indexed with a mapping $l_{ti} : N(\upsilon_{ti}) \rightarrow \{0,\ldots,K-1\}$. The convolutional operation of a given graph vertex υ_{ti} on the spatial dimension can then be written as

$$f_{out} = \sum\nolimits_{\upsilon_{tj}\in N(\upsilon_{ti})} \frac{1}{Z_{ti}\left(\upsilon_{tj}\right)} f_{in}\left(\upsilon_{tj}\right) W(l(\upsilon_{tj})), \tag{1}$$

where υ_{tj} represents the graph vertex of a defined neighbor set, $f_{in}(\upsilon_{tj})$ denotes a mapping used to get the attribute vector of υ_{tj}, and $W(l(\upsilon_{tj}))$ is a weight function $W(\upsilon_{ti}, \upsilon_{tj}) : N(\upsilon_{ti}) \rightarrow \mathbb{R}^c$ that can be implemented via a tensor with (c, K) dimensions. Here, c indicates the feature dimensions. $Z_{ti}(\upsilon_{tj}) = |\{\upsilon_{tk}|l_{ti}(\upsilon_{tk}) = l_{ti}(\upsilon_{tj})\}|$ is a normalization term equal to the cardinality of its corresponding subset.

We use an adjacency matrix $\mathbf{A}$ to implement the spatial convolutional layer of a single skeleton frame. The elements of $\mathbf{A}$ show whether a vertex υ_{tj} belongs to one subset of $N(\upsilon_{ti})$. Accordingly, the graph convolution is implemented by performing a 1 × 1 classical 2D convolution and multiplying the output tensor by a normalized adjacency matrix $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{A}\mathbf{\Lambda}^{-\frac{1}{2}}$ on the second dimension, where $\mathbf{\Lambda}^{ii} = \sum_j(\mathbf{A}^{ij}) + \alpha$ is a diagonal matrix with α set to 0.001 to avoid empty rows. Given K sampling strategies $\sum_{k=1}^{K}\mathbf{A}_k$, the graph convolution for a skeleton frame can be expanded from (1) as

$$\mathbf{f}_{out} = \sum\nolimits_{k=1}^{K} \mathbf{\Lambda}_k^{-\frac{1}{2}}\mathbf{A}_k\mathbf{\Lambda}_k^{-\frac{1}{2}}\mathbf{f}_{in}\mathbf{W}_k \odot \mathbf{M}_k, \tag{2}$$

where $\mathbf{M}_k$ is an attention map with the same size of $\mathbf{A}_k$, which indicates the importance of each vertex. $\mathbf{W}_k$ denotes a weight tensor of the 1×1 convolutional operation with $(C_{in},\ C_{out},\ 1,\ 1)$ dimensions, representing the weighting function of (1). $\odot$ reflects the element-wise product operation.

The convolutional operation along the temporal dimension is similar to the implementation of 2T-GCN [31]. Specifically, using the temporal kernel size of Γ, it performs a $1 \times \Gamma$ convolution on the feature map $\mathbf{f}_{out}$. The spatial and temporal graph convolutional layers are each connected with a batch normalization layer and a ReLU layer. To avoid overfitting, a dropout layer is added to a basic GCN block composed of a spatial convolutional layer and a temporal convolutional layer. The residual mechanism is applied to each GCN block to stabilize the training process.

Following the practice of [31], our GCN model is a stack of 9 basic GCN blocks. The first three, middle three, and last three blocks have 64, 128, and 256 output channels, respectively. The strides of the 4th and the 7th blocks are set to 2, while all other blocks use a stride size of 1. We set the temporal kernel size Γ to 9. A global average pooling layer is used to pool the GCN feature map to a 256-dimensional feature vector at the last GCN block. To transform the feature vector to our desired output (i.e.,

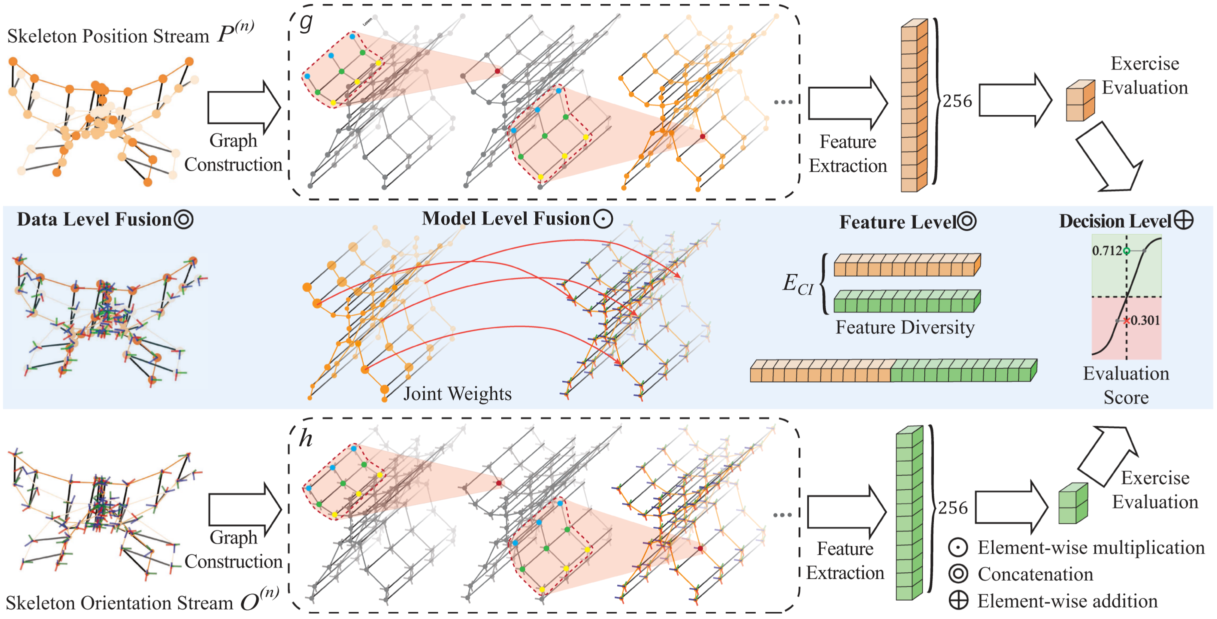


Fig. 3. Illustration of our EGCN++ learning framework. The upper and lower figures show two inputs (i.e., skeleton position and skeleton orientation streams) represented by two submodels (i.e., GCNs). The middle part shows four ensemble strategies at different levels (i.e., data level, model level, feature level, and decision level). $P^{(n)}$ and $O^{(n)}$ can be concatenated at the data level. As for model-level fusion, the feature of the last layer of the pre-trained GCN model on $P^{(n)}$ via the action recognition task can be used as joint weight retrieval to regularize the submodel h for the exercise assessment task. At the feature level, the characteristics of g and h can be concatenated for prediction or to force their diversity. The results can be aggregated for prediction at the decision level.

correct or incorrect), the last layer of the GCN model is a 1×1 2D convolutional layer.

B. EGCN++ Framework

Fig. 3 shows the EGCN++ framework. It includes two submodels, g and h, that respectively represent the skeleton position and orientation features. The submodels g and h are GCN models as defined in Section III-A. Fusion strategies can be applied at different levels in the middle of these two submodels. According to the fusion methods for ensemble learning surveyed in [64], one can use common fusion techniques at the data, feature, and decision levels in ensemble-based methods. For model-level fusion, special fusion strategies are needed based on a clear understanding of the task. In this section, we discuss four groups of ensemble-based methods relevant to our EGCN++ learning framework.

1) Data-Level Ensemble: As shown in Fig. 3, the data-level ensemble method enables fusion for input data $P_{(n)}$ and $O_{(n)}$, which is also known as sample-level ensemble (SLE). Given that $P_{(n)} \in \mathbb{R}^{T\times J\times 3}$ and $O_{(n)} \in \mathbb{R}^{T\times J\times 3}$ share the same graph structure, we concatenate the position and orientation streams along the feature dimension, leading to fused input $S^{(n)} \in \mathbb{R}^{T\times J\times 6}$. We then feed the constructed $S^{(n)}$ to a single GCN model, which can be written as

$$y = \sigma(FC(GAP(g(S^{(n)}, \theta_g)))), \tag{3}$$

where σ is the Softmax classifier, FC is the fully connected convolutional layer, and GAP is the global average pooling layer.

2) Model-Level Ensemble: Unlike other strategies that arbitrarily combine features at specific levels via addition or concatenation, MLEs rely on a comprehensive understanding of the data to be tackled. The position and orientation features can be regarded as the global and local characteristics of an exercise, respectively. The skeleton position feature is frequently used in action classification. For example, based on the GCN model, some model-based fusion methods [67], [68], [69], [70] have proposed attention mechanisms by taking the average of neuron activation values along specific dimensions to improve multimodal action recognition. Inspired by these efforts, we utilize the neuron activation values retrieved from a pre-trained g on the skeleton position stream via an action recognition task as spatiotemporal joint weights to regulate orientation stream training. Existing fusion practices handle heterogeneous multimodal data (i.e., skeleton and RGB modalities) in an attempt to address action recognition related to human–object interaction. Our problem formulation does not involve RGB video data, as none of the investigated exercises include such interaction. Thus, we do not follow the averaging operation on neuron activation values because it tends to smooth out the joint importance along spatial or temporal dimensions of the GCN feature map. Our MLEs fuse joint weights derived from the last GCN block of g, which is a $C_{out} \times T_{out} \times J_{out}$ tensor (see Fig. 3), with the corresponding model representation of the skeleton orientation stream $h(O^{(n)}, \theta_h)$ via element-wise multiplication. We name the first MLE strategy MLE-orientation (MLE-O); it can be written as

$$y = \sigma(FC(GAP(\ g(P^{(n)}, \theta_g) \odot h(O^{(n)}, \theta_h)\ ))), \tag{4}$$

```
Algorithm 1: MLE-PO Optimization.
  Input:
  $P = \{P^{(n)} \mid n = 1, \ldots, N\}$: position stream
  $O = \{O^{(n)} \mid n = 1, \ldots, N\}$: orientation stream
  Procedure:
  1: Train $g$ with position stream $P$ with the action
       recognition task on a whole dataset.
  2: Concatenate $P^{(n)}$ and $O^{(n)}$ along the channel dimension
       to construct skeleton stream $S^{(n)}$.
  3: Extract joint weights $w$ by feeding $P^{(n)}$ to the trained
       submodel $g$.
  4: Feed the constructed skeleton stream $S^{(n)}$ to a submodel
       $h$.
  5: Fuse the features of $h$ and the extracted joint weights $w$
       at the last layer of $h$.
  6: Feed the fused feature in Step 6 to a fully connected
       layer.
  7: Finish a training epoch by Iterating Steps 2–6 with all $N$
       samples.
  Output:
  TrainedMLE-PO including submodels: $g$, $h$
```

where $g(P^{(n)}, \theta_g)$ is the pre-trained model on the action recognition task that classifies different exercises within a dataset. To maintain the mutual independence of submodels g and h, the pre-trained parameters θ_g of $g(P^{(n)}, \theta_g)$ are frozen while training $h(O^{(n)}, \theta_h)$. $\bigodot$ denotes the element-wise product operation.

According to the results from 2T-GCN [31] and EGCN [32], SLE can achieve better performance than single-modal methods. We hence transform our MLE-O model into a multi-level ensemble version called MLE-position orientation (MLE-PO). It takes the SLE feature as input and makes use of the joint weights from MLE-O to regulate the training process. This new fusion strategy (i.e., MLE-PO) can be formulated as follows:

$$y \;=\; \sigma(FC(GAP(g(P^{(n)}, \theta_g) \odot h(S^{(n)}, \theta_h)\ ))). \tag{5}$$

The MLE-PO optimization process is depicted in Algorithm 1.

3) Feature-Level Ensemble: In considering the feature-level ensemble (FLE) method, we investigate two representative strategies: the FLE base (FLE-B), which simply concatenates features; and its extension FLE cosine independence (FLE-CI), which forces feature diversity (see Fig. 3). For FLE-B, we separately extract two 256-dimensional feature vectors from two skeleton feature streams and concatenate their extracted features. This process can be optimized via end-to-end learning for the whole model. FLE-B can be represented as

$$\begin{aligned} y \;=\; & \sigma(FC(GAP(Cat(g(P^{(n)}, \theta_g), \\ & h(O^{(n)}, \theta_h))))), \end{aligned} \tag{6}$$

where Cat is the concatenation operation.

FLE-CI aims to force the diversity of small classifiers, which is a primary motivation of ensemble-based methods. Our FLE-CI is based on FLE-B by forcing the feature-level diversity of FLE-B. Specifically, we follow the local independence training method [65] that penalizes cosine similarity between the features of two submodels to approximate feature-level diversity. The loss objective of cosine independence error (E_{CI}) can be expressed as

$$E_{CI}(f, g) = E[\cos^2(g(P^{(n)}, \theta_g), h(O^{(n)}, \theta_h))], \tag{7}$$

which is optimized together with the cross-entropy (E_{CE}) loss of FLE-B. The overall model objective of FLE-CI is optimized by minimizing losses: $E_{CI} + \lambda E_{CE}$, where λ is a parameter used to balance E_{CI} and cross-entropy loss. λ can be either a learned parameter or a fixed value. We empirically set λ to 0.1.

4) Decision-Level Ensemble: With respect to decision-level ensemble (DLE) approaches, one can employ different training strategies to optimize the entire learning framework. Our EGCN++ accounts for the DLE-full and DLE-dual techniques. For DLE-full, we aggregate the decision-level prediction results and train two submodels h and g together via an end-to-end learning process, which can be written as

$$\begin{aligned} y = & \sigma(FC(GAP(g(P^{(n)}, \theta_g))) \\ & + FC(GAP(h(O^{(n)}, \theta_h)))). \end{aligned} \tag{8}$$

For DLE-dual, the submodels are trained separately, after which their prediction results are aggregated. Compared with DLE-full, the DLE-dual method is more popular when dealing with homogeneous [49], [50], [51] and heterogeneous data fusion [69], [70] owing to its better performance. This method can be represented as

$$\begin{aligned} y \;=\; & \sigma(FC(GAP(g(P^{(n)}, \theta_g)))) \\ & +\; \sigma(FC(GAP(h(O^{(n)}, \theta_h)))). \end{aligned} \tag{9}$$

IV. Experiments

We validate our proposed EGCN++ on three datasets: UI-PRMD [35], KIMORE [36], and EHE [31]. We consider prediction accuracy as well as results' consistency with human evaluation. According to the survey in [71], prior to the development of EHE, UI-PRMD and KIMORE were the two most recent datasets for exercise evaluation. These three datasets are suitable for validating our proposed method. We conduct extensive ablation studies on all datasets to verify the effectiveness of our ensemble scheme.

A. Validation Datasets

UI-PRMD: The UI-PRMD dataset [35] comprises skeleton exercise data from 10 healthy subjects. Each subject performs 10 repetitions of 10 rehabilitation exercises (i.e., E1–10) such as "side lunge", "sit to stand", and "deep squat". For the exercise assessment task, all subjects are asked to perform each exercise in correct and incorrect manners. Subjects simulate the incorrect positioning of patients with musculoskeletal constraints. The 3D motion sensor Kinect v2 and Vicon motion capture provide the position and orientation features of skeleton joints. We use Kinect v2 data for our experiments because these data are preferable to Vicon motion capture data as shown in [31]. The UI-PRMD dataset contains inconsistent samples due to measurement errors and exercise performance with incorrect limbs. Therefore, we follow the consistent version[1] used in [28], which has 1,326 exercise repetitions.

KIMORE: The KIMORE dataset [36] is collected via the Kinect v2 sensor. Data are gathered from 78 subjects across three categories: a control group of experts (CG-E), a control group of non-experts (CG-NE), and a group with pain and postural disorders (GPP). The GPP group includes 34 subjects with motor dysfunctions such as stroke, Parkinson's disease, and back pain. All subjects perform five exercises (i.e., E1–5) such as "lifting of the arms", "trunk rotation", and "squatting". [36] reported no overlap between the clinical total score distributions for CG-E and GPP. This observation implies that we can treat these groups' exercise repetitions as correct and incorrect, respectively. As such, to predict abnormality based on each exercise, we manually segment the dataset based on noticeable features and respectively label the repetitions of 17 experts and 34 patients as correct and incorrect.

EHE: The EHE dataset [31] is collected in a real-world elderly home environment in a natural setting. It contains six morning exercises such as "wave hands", "hands up and down", and "bend waist to left" that aging adults perform in daily life. The exercises are completed by 25 subjects and tracked with the Kinect v2 sensor. In total, 10 of 25 subjects have been diagnosed with Alzheimer's disease of varying severity (ranging from 0 to 10).

B. Evaluation Metrics

This section first introduces the two functions of exercise evaluation score calculation, then it provides metrics for investigating model prediction ability and consistency between human and machine evaluation.

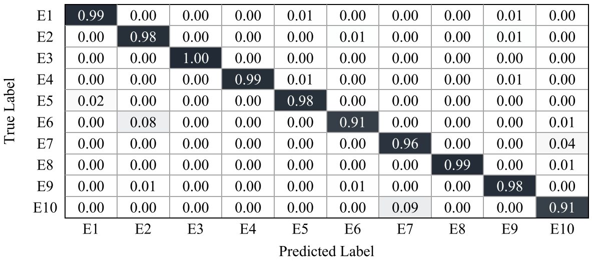
Exercise Evaluation Score Calculation: To conduct a more exhaustive assessment, we quantitatively and qualitatively analyze the result consistency between model evaluation and human evaluation by adopting the evaluation score calculations in [30] and [31]. Specifically, [31] took the probability results of the Softmax layer to infer an exercise repetition's evaluation score. We retrieve the first dimension of the probability distribution from our model's Softmax layer, which can be calculated as

$$f_{score}(a,b) = \frac{e^a}{(e^a + e^b)}, \tag{10}$$

where a and b respectively represent the first and second neuron output values of the fully connected layer, which are then used to calculate the Softmax layer's probability distribution. Alternatively, given the feature before the Softmax layer (i.e., the output of the fully connected layer), the exercise evaluation score can be calculated by using a sigmoid function that transforms the corresponding neuron value into a range of [0,1] [30]. This method can be presented as

$$f_{score}(a) = \frac{1}{(1 + e^{-a})}. \tag{11}$$

Model Prediction Ability: In early work, separation degree (S_D) and distance metrics (D_M) were used to examine a model's


True Label
Predicted Label

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
E1	0.99	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00
E2	0.00	0.98	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00
E3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E4	0.00	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.01	0.00
E5	0.02	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00
E6	0.00	0.08	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.01
E7	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.04
E8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.01
E9	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.98	0.00
E10	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.91

Fig. 4. Confusion matrix of action recognition results on UI-PRMD.

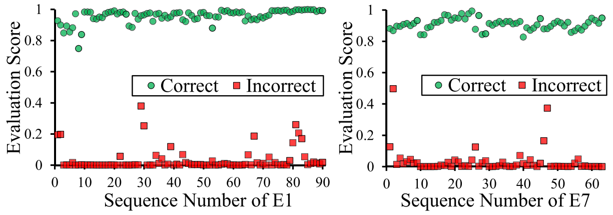


Fig. 5. Evaluation scores (calculated with the sigmoid function) of E1 and E7 in the UI-PRMD dataset based on the training set.

representation ability as defined in [28] and [29], respectively. S_D and D_M each quantify the difference between correct and incorrect evaluation results. For a pair of positive numbers x and y, S_D can be calculated as $S_D(x, y) = \frac{x-y}{x+y} \in [-1, 1]$. Accordingly, S_D between two positive sequences $\boldsymbol{x} = (x_1, \ldots, x_m)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ can be defined as

$$S_D(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} S_D(x_i, y_j). \tag{12}$$

Following the evaluation metric S_D, [30] achieved 0.808 (calculated from UI-PRMD's training accuracy of 99.59%) using the orientation feature. Upon altering the evaluation score calculation without changing the model, [31] achieved an even higher S_D of 0.933 under the same experimental setting. [30] adopted the sigmoid function to calculate the evaluation score; [31] used Softmax. Given the results of [30] and [31], calculating S_D based on the training accuracy cannot properly reflect a model's prediction ability. Instead, we report the S_D based on the results of cross-validation. D_M can be calculated as

$$D_M(x_n, y_n) = \frac{|x_n - y_n|}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_n - y_n)^2}}, \tag{13}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)$ and $\boldsymbol{y} = (y_1, \ldots, y_N)$ are two positive sequences. [27] determined that, based on the training set results, D_M also cannot feasibly evaluate a model's prediction ability: [30], [31] already demonstrated the GCN model's strong representation ability (Fig. 5 visualizes exercise evaluation scores calculated with the sigmoid function, reinforcing the GCN model's capacity to distinguish correct and incorrect exercise repetitions).

Liao et al. [28] divided the UI-PRMD dataset into a training set and a test set to evaluate their model's prediction ability;

[1] https://webpages.uidaho.edu/ui-prmd/

however, results were only reported for E1. Because the KIMORE, UI-PRMD, and EHE datasets are relatively small, we expand on [28] by applying the 5-fold cross-validation criterion used in [27], [31] to evaluate the prediction abilities of different ensemble strategies in our EGCN++. Instead of referring to training accuracy, we investigate evaluation criteria S_D and D_M based on prediction results from the 5-fold cross-validation setting (see Section IV-D).

Consistency between Human and Machine Evaluation: It is worth noting that the numerical score can indicate exercise quality without supervision from subjective human evaluation scores [42] or arbitrary scores calculated by a function as in [28]. To determine whether the newly proposed methods' evaluation scores are consistent with human evaluation, we use two metrics as proposed in 2T-GCN [31]: Euclidean distance and correlation. For an n-dimensional space, the Euclidean distance E_D of two vectors $\boldsymbol{x} = (x_1,\ x_2,\ \ldots, x_n)$ and $\boldsymbol{y} = (y_1,\ y_2,\ \ldots, y_n)$ is calculated as

$$E_D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum\nolimits_{i=1}^{n} (x_i - y_i)^2}. \tag{14}$$

The correlation C_R between $\boldsymbol{x}$ and $\boldsymbol{y}$ can be defined as

$$C_R(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{15}$$

where $\bar{x}$ and $\bar{y}$ are the average values of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Smaller $E_D(\boldsymbol{x}, \boldsymbol{y})$ and larger $C_R(\boldsymbol{x}, \boldsymbol{y})$ indicate that the machine evaluation is more consistent with the normalized severity of Alzheimer's disease observed by a human expert and vice versa. The normalized severity of this condition ranges from 0 to 1 and can be calculated from the clinical evaluation (from 0 to 10, with 0 indicating normal functioning and 10 indicating highly severe Alzheimer's disease) in the EHE dataset [31]. This calculation can also be applied to expert screening scores on exercise repetition in the KIMORE dataset [36].

C. Implementation Details

For cross-validation purposes, we split the UI-PRMD, KIMORE, and EHE datasets based on two division protocols: cross-subject (CS) and random division (RD). The CS protocol can be intuitively more difficult than the RD protocol since the system needs to work for different subjects (i.e., subjects that appear in each cross-validation fold are different). Whereas data from every subject appear in all cross-validation folds of the RD protocol, which can provide a baseline regarding how well the system can learn for an individual and offer practical insight into data collection for real-world scenarios (discussed in Section V). Table I lists the number of exercise repetitions in the CS cross-validation folds of the UI-PRMD and KIMORE datasets. Certain exercises are performed by fewer than five subjects, leading to a lack of exercise repetition in some cross-validation folds.

The proposed MLEs require the submodel $g(P^{(n)}, \theta_g)$ to be pre-trained via the action classification task with the skeleton position feature. The pre-trained model is then used to retrieve joint weights from the position feature to be fused with the orientation feature at the model level. We include each of a dataset's exercise classes in our pre-training implementation. The overall action classification accuracy for the UI-PRMD, KIMORE, and EHE datasets is 96.91%, 98.04%, and 97.81%, respectively. Fig. 4 shows the confusion matrix of action recognition results for the UI-PRMD dataset.

TABLE I
NUMBER OF EXERCISE REPETITIONS FOR TESTING IN CROSS-SUBJECT CROSS-VALIDATION FOLDS (I.E., F1–5) OF DIFFERENT EXERCISES IN THE UI-PRMD AND KIMORE DATASETS

UI-PRMD

Exercise ID	Cross-validation Folds				
	F1	F2	F3	F4	F5
E1	38	38	34	36	34
E2	30	34	14	14	18
E3	30	18	18	18	18
E4	18	36	18	36	32
E5	36	34	36	34	28
E6	20	38	18	32	38
E7	18	18	18	36	36
E8	16	28	16	30	36
E9	20	18	18	36	28
E10	36	18	18	18	18

KIMORE

Exercise ID	Cross-validation Folds				
	F1	F2	F3	F4	F5
Es1	58	55	49	43	50
Es2(L)	53	47	34	31	31
Es2(R)	54	44	34	34	35
Es3(L)	54	48	38	37	33
Es3(R)	51	48	39	38	33
Es4(L)	47	38	49	33	56
Es4(R)	56	61	34	43	26
Es5	58	51	53	48	45

L is left, and R is right.

All ensemble strategies proposed in EGCN++ are optimized via stochastic gradient descent. We set the base learning rate at 0.01. Every models is trained for 50 epochs. The learning rate is decayed by 0.1 at epochs 10 and 30. All experiments are carried out on a workstation with 2 GTX 1080 GPUs.

D. Experiments on the UI-PRMD Dataset

The left and right parts of Table II show the prediction results of exercises on the UI-PRMD dataset with two cross-validation evaluation protocols (CS and RD, respectively). The overall performance of RD is better than that of CS (e.g., 86.14% to 95.00% average performance increase on UI-PRMD), indicating collecting data from a user to tune the system can greatly contribute to the performance. Our newly proposed fusion strategy, MLE-PO, significantly outperforms the state-of-the-art method MLE-O as well as other ensemble strategies such as SLE, FLEs, and DLEs for nearly all exercises under the two protocols. MLE-PO thus combines fusion-related benefits at the data and model levels. The advantage of SLE could be due to the more integrated features of heterogeneous data (i.e., position and orientation features serve as global and local descriptions of an action repetition, respectively). In MLE-O, the joint weights learned from the skeleton position feature can augment the learning process of this feature type (please refer to Section IV-G for ablations).

Table II also indicates slightly improved overall prediction accuracy for FLE-CI and DLE-dual over their corresponding base methods (i.e., FLE-B and DLE-full, respectively) and single-modal methods (i.e., Pos and Ori). FLE-CI and DLE-dual are useful for data fusion at their corresponding scales or positions, which include a mechanism to maintain the submodels' independence. Specifically, FLE-CI uses a cosine similarity loss to force diversity, and DLE-dual trains the submodels separately. These observations demonstrated that maintaining submodels' independence can inform a whole model

TABLE II
COMPARISON OF ENSEMBLE STRATEGIES FOR UI-PRMD DATASET WITH CROSS-SUBJECT AND RANDOM DIVISION PROTOCOLS (ACCURACY IN %)

Exercise ID	Cross-subject									Random Division								
	Pos	Ori	SLE	FLE-B	FLE-CI	DLE-full	DLE-dual	MLE-O	MLE-PO	Pos	Ori	SLE	FLE-B	FLE-CI	DLE-full	DLE-dual	MLE-O	MLE-PO
E1	71.11	73.33	67.22	64.44	72.28	73.33	71.67	83.33	**91.11**	75.44	77.03	76.56	75.31	78.39	75.43	77.19	86.14	**89.95**
E2	82.73	80.91	80.91	81.82	81.82	81.82	82.73	84.55	**90.00**	90.00	87.27	84.55	86.36	87.27	82.73	88.18	92.73	**94.55**
E3	50.98	64.71	72.55	71.57	66.67	58.82	61.76	82.35	**84.31**	54.90	72.55	78.43	77.45	71.57	76.47	74.51	89.22	**91.18**
E4	70.00	75.71	76.43	70.71	73.57	67.86	74.29	79.29	**86.43**	70.71	83.57	77.86	70.71	67.86	77.86	81.43	92.14	**94.29**
E5	86.31	77.98	84.52	73.21	83.33	86.31	80.95	89.88	**93.45**	91.67	88.10	88.10	67.86	75.60	91.07	91.67	92.26	**94.64**
E6	83.56	85.62	82.19	87.67	76.71	81.51	89.04	89.04	**89.73**	87.67	83.56	93.15	82.88	84.93	84.93	89.04	95.21	**97.26**
E7	80.16	87.30	89.68	72.22	80.95	68.25	88.10	92.06	**97.62**	69.05	92.06	90.48	78.57	84.92	79.37	92.06	95.24	**97.62**
E8	65.08	61.90	71.43	79.37	80.95	73.81	57.94	81.75	**84.92**	76.19	80.16	91.27	82.54	78.57	73.02	80.95	89.68	**97.62**
E9	86.67	84.17	78.33	72.50	76.67	86.67	85.83	**95.83**	**95.83**	78.33	85.00	91.67	86.67	82.50	88.33	80.00	92.50	**94.17**
E10	77.78	78.70	78.70	79.63	79.63	75.93	79.63	83.33	**86.11**	84.26	86.11	83.33	80.56	73.15	79.63	89.81	87.96	**92.59**
Average	75.44	77.03	78.57	75.31	77.31	75.43	77.19	86.14	**89.95**	77.39	84.78	86.27	79.08	78.03	81.34	85.54	91.81	**95.00**

Results are based on data collected with the Kinect v2 sensor. Bold and underlined numbers refer to the best and second-best results, respectively. Pos and ori indicate position and orientation features, respectively.

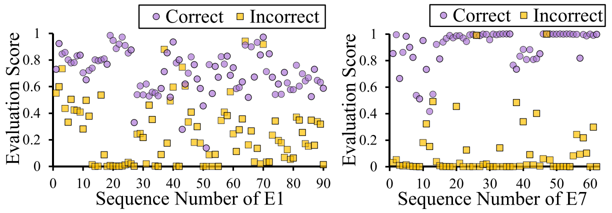


Fig. 6. Evaluation scores (calculated by the sigmoid function) of E1 and E7 using MLE-PO with cross-subject protocol on the UI-PRMD dataset.

TABLE III
COMPARISON OF FUNCTIONS FOR CALCULATING EXERCISE EVALUATION SCORES ON UI-PRMD DATASET USING STATE-OF-THE-ART METHODS

Method	S_D (Std. Dev)	D_M (Std. Dev)
GCN (Sigmoid) [30]	0.3035 (0.1392)	0.7736 (0.6274)
2T-GCN (SoftMax) [31]	0.4336 (0.1741)	0.7830 (0.6117)
EGCN (MLE-O, Sigmoid) [27]	0.4142 (0.0937)	0.8280 (0.5543)
EGCN (MLE-O, SoftMax) [27]	0.5307 (0.1109)	0.8494 (0.5201)
EGCN++ (MLE-PO, Sigmoid)	0.5419 (0.0697)	0.8702 (0.4881)
EGCN++ (MLE-PO, SoftMax)	**0.6485** (0.0703)	**0.8916** (0.4476)

for the focal task. These results generally align with the related data fusion theory (i.e., forcing independence or diversity) [56], [65].

Visualize the Evaluation Score: We visualize and compare the evaluation scores derived from our experimental setting and the setting of prior methods to qualitatively support that the cross-validation approach (introduced in Section IV-C) can better characterize exercise performance. Specifically, Fig. 5 shows the visualized exercise evaluation score retrieved from the training set, which follows the initial investigation of [28] and [29] by keeping the test set the same as the training set. The UI-PRMD dataset groups its exercise repetitions into correct and incorrect categories. Modeling this process as a binary classification task will lead to these types of evaluation scores, indicating the basic model's strong representation ability. However, under this experimental setting, scores do not reflect the exercises' degree of accuracy. Using 5-fold cross-validation generates more meaningful evaluation scores that convey the exercises' likelihood of correctness. This probability can be interpreted as the degree of confidence in the classification learned from the data. Fig. 6 presents the evaluation scores for E1 and E7 in the UI-PRMD dataset using MLE-PO with the CS evaluation protocol. These results reflect quality fluctuations in different exercise repetitions.

Quantitative Analysis: Using the evaluation metrics S_D and D_M (defined in (12) and (13), respectively), we quantify the effects of evaluation score calculation schemes (i.e., Softmax and sigmoid defined in (10) and (11), respectively). As Table III indicates, compared with the sigmoid function, the probability of Softmax leads to a larger disparity between correct and incorrect evaluation scores. The results for EGCN++ in Table III are the average of 10 exercises (see Table 11 in Appendix A, available online). MLE-O and MLE-PO are compared in Table III; as expected, S_D and D_M are generally positively related to the prediction results in Table II. Although quantifying the difference between correct and incorrect evaluation scores demonstrates a model's sensitivity, findings do not reflect the consistency between model evaluation and human evaluation. A model's sensitivity can also be based on the pattern learned from a specific data distribution. More precisely, a model can possess high prediction accuracy but have relatively low S_D and D_M; for example, E7 ($accuracy = 97.62\%, S_D = 0.8426$) and E9 ($accuracy = 95.83\%, S_D = 0.8664$) in the UI-PRMD dataset are negatively related. It thus remains unclear which calculation method produces evaluation scores then align closely with human assessment. We investigate this question using the following two datasets for which human evaluation scores are available.

E. Experiments on the KIMORE Dataset

Table IV shows the results on the KIMORE dataset using the same experimental setting as with the UI-PRMD dataset. Results yield similar implications regarding the impacts of different fusion schemes. By capitalizing on SLE and MLE-O, the MLE-PO fusion strategy performs significantly better than model-level fusion. In particular, for the average prediction accuracies of eight exercises, MLE-PO outperforms MLE-O by 3.42% and 5.97% under the CS and RD evaluation protocols, respectively. Meanwhile, maintaining submodel independence via DLE-dual produces better results than its same-grouped fusion scheme (i.e., DLE-full) and single-modal methods. However, the cosine similarity loss cannot promise improvement in this dataset. This issue may be due to the ad-hoc selection requirement of λ for

TABLE IV
COMPARISON OF ENSEMBLE STRATEGIES FOR KIMORE DATASET WITH CROSS-SUBJECT AND RANDOM DIVISION PROTOCOLS (ACCURACY IN %)

Exercise ID	Cross-subject									Random Division								
	Pos	Ori	SLE	FLE-B	FLE-CI	DLE-full	DLE-dual	MLE-O	MLE-PO	Pos	Ori	SLE	FLE-B	FLE-CI	DLE-full	DLE-dual	MLE-O	MLE-PO
Es1	78.00	77.65	78.04	68.24	69.80	66.27	81.50	79.22	**83.92**	73.33	86.67	85.49	69.02	72.55	72.55	86.67	88.24	**96.47**
Es2(L)	75.51	72.45	78.57	68.88	69.39	78.06	71.43	81.12	**82.14**	76.02	79.59	81.63	74.49	69.90	70.92	79.59	83.16	**92.86**
Es2(R)	71.14	80.10	80.12	70.65	75.62	74.63	77.11	80.60	**85.07**	74.84	78.61	77.65	69.15	67.66	80.10	80.00	83.58	**93.53**
Es3(L)	73.81	84.76	78.05	69.52	68.57	72.38	82.86	77.62	**85.24**	70.95	80.00	78.79	69.52	64.76	73.33	79.05	81.90	**92.86**
Es3(R)	73.68	74.64	74.64	67.46	65.07	66.99	66.51	76.08	**79.90**	76.56	77.99	87.08	75.12	68.42	73.68	79.43	88.04	**89.47**
Es4(L)	80.27	73.99	76.23	76.23	76.23	83.24	81.61	84.75	**85.20**	81.61	78.48	85.65	78.48	74.89	84.75	85.65	87.89	**91.48**
Es4(R)	83.64	79.09	79.55	78.64	79.55	78.64	83.64	84.09	**85.45**	80.91	90.45	87.27	76.36	76.82	81.82	**91.36**	**91.36**	**91.36**
Es5	74.12	77.65	79.22	72.55	64.31	74.12	78.82	77.65	**81.57**	79.61	83.14	86.27	74.12	66.67	77.65	83.92	90.59	**94.51**
Average	76.27	77.54	78.05	71.52	71.07	74.29	77.94	80.14	**83.56**	76.73	81.87	83.73	73.28	70.21	76.85	83.21	86.85	**92.82**

L and R refer to exercises on the left and right sides, respectively. Bold and underlined numbers refer to the best and second-best results, respectively. Pos and ori indicate position and orientation features, respectively.

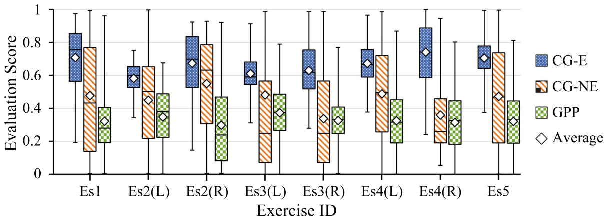


Fig. 7. Visualization of evaluation scores with our model-level ensemble strategy on the KIMORE dataset.

TABLE V
COMPARISON OF FUNCTIONS FOR CALCULATING EXERCISE EVALUATION SCORES ON KIMORE DATASET USING STATE-OF-THE-ART METHODS

Method	C_R (cTS)	C_R (cPO)	C_R (cCF)
GCN (Sigmoid) [30]	0.4447	0.4278	0.4125
2T-GCN (SoftMax) [31]	0.4924	0.4743	0.4544
EGCN (MLE-O, Sigmoid) [27]	0.4908	0.4791	0.4486
EGCN (MLE-O, SoftMax) [27]	0.5240	0.5041	0.4831
EGCN++ (MLE-PO, Sigmoid)	0.5092	0.4960	0.4660
EGCN++ (MLE-PO, SoftMax)	**0.5427**	**0.5223**	**0.5005**

various scenarios [65], which is equal to 0.1 throughout our experimental setting. FLE-CI could potentially achieve better results with proper fixed or learned values of λ.

Visualize the Evaluation Score: The KIMORE dataset [36] provides three clinical evaluation scores based on a clinical questionnaire containing 10 items scored on a scale from 1 to 5 (the higher the better). The first evaluation score is the clinical total score (cTS), which is the sum of the 10 identified scores. The second score is the clinical primary outcome (cPO), which is calculated based on the sum of scores on the first three questions. The third score is the clinical control factor (cCF): the sum of the last seven items concerning postural performance (e.g., postures of the head, right arm, and right leg). We conduct qualitative and quantitative analyses based on these score categories.

Recall that the KIMORE dataset has three subject groups (CG-E, CG-NE, and GPP) that reflect varying levels of exercise capability. Based on the three evaluation scores, the visualized scores for all exercises performed by subjects in CG-E and GPP [36] do not overlap; that is, these groups can be treated as correct and incorrect samples. Hence, we use both samples as training data with the CG-NE sample as test data to infer the latter group's evaluation scores. By doing so, we provide a qualitative view of the match between machine evaluation and clinical evaluation. Fig. 7 shows the box plots of E1–5 for the three subject groups in the KIMORE dataset. Evaluation scores between the first and third quartiles are consistent with the total clinical scores reported in [36]. The average scores also reflect group-based performance differences.

Quantify the Consistency: In addition to providing a qualitative perspective, we further examine the consistency between machine evaluation and human evaluation via quantitative analysis based on the metric C_R. As listed in Table V, compared with the sigmoid calculation, the Softmax probability generates more consistent evaluation scores with human evaluation across a trio of human evaluation metrics (i.e., cTS, cPO, and cCF). The newly proposed MLE-PO from the EGCN++ outperforms state-of-the-art methods [27], [30] in terms of consistency with human evaluation. Expanded results for EGCN++ in Table V appear in Table 12 in Appendix A, available online, with a discussion of the differences between left-side and right-side exercises.

F. Experiments on the EHE Dataset

Table VI displays the results for ensemble strategies on the EHE dataset. Following the results of previous datasets (i.e., UI-PRMD and KIMORE), the MLE-PO fusion strategy outperforms almost all other strategies on all EHE exercises. Compared with MLE-O, the average prediction accuracy respectively improves by 3.97% and 2.59% for the CS and RD evaluation protocols. These results again substantiate the effectiveness of our MLE-PO fusion strategy. MLE-O and SLE respectively achieve the second- and third-best results among other methods, collectively contributing to MLE-PO's excellent performance. Forcing feature-level and decision-level independence does not benefit this dataset. FLE-CI relies on a properly chosen λ to realize good performance. Additionally, compared with our model-level fusion schemes, simple decision-level fusion does not seem to be effective.

Quantify the Consistency: Following the evaluation metrics in [31], we use E_D and C_R to analyze the consistency with human evaluation. Compared with state-of-the-art methods such as [27], [30], [31], the evaluation scores of MLE-PO (calculated

TABLE VI
COMPARISON OF ENSEMBLE STRATEGIES FOR EHE DATASET WITH CROSS-SUBJECT AND RANDOM DIVISION PROTOCOLS (ACCURACY IN %)

Exercise ID	Cross-subject									Random Division								
	Pos	Ori	SLE	FLE-B	FLE-CI	DLE-full	DLE-dual	MLE-O	MLE-PO	Pos	Ori	SLE	FLE-B	FLE-CI	DLE-full	DLE-dual	MLE-O	MLE-PO
E1	86.36	88.89	87.37	86.87	86.36	88.89	88.38	**90.40**	**90.40**	95.96	94.44	94.95	95.45	91.41	95.88	95.45	95.96	**97.98**
E2	75.00	77.08	79.86	79.86	70.83	77.78	76.39	82.64	**84.72**	89.58	95.14	91.67	90.28	80.56	91.35	95.83	97.92	**98.61**
E3	83.67	79.59	81.63	82.14	71.94	77.55	80.10	81.12	**86.22**	85.71	86.22	97.45	97.45	92.35	97.96	91.84	93.88	**97.96**
E4	85.48	77.42	89.25	81.18	80.65	86.01	86.02	87.10	**91.40**	94.77	95.16	97.12	98.39	89.25	97.31	95.42	98.92	**99.46**
E5	64.86	71.62	70.83	70.27	67.57	67.57	71.62	74.32	**82.43**	72.97	68.92	76.00	72.97	66.22	74.32	67.57	81.08	**86.49**
E6	70.42	71.83	73.24	76.06	71.83	76.06	70.42	78.87	**83.10**	69.01	78.87	76.06	71.83	71.83	73.24	74.65	84.51	**87.32**
Average	77.63	77.74	80.36	79.40	74.86	78.98	78.82	82.41	**86.38**	84.67	86.46	88.88	87.73	81.94	88.34	86.79	92.05	**94.64**

Bold and underlined numbers refer to the best and second-best results, respectively. Pos and ori indicate position and orientation features, respectively.

TABLE VII
COMPARISON OF FUNCTIONS FOR CALCULATING EXERCISE EVALUATION SCORES FOR EHE DATASET USING STATE-OF-THE-ART METHODS

Method	$E_D \downarrow$	$C_R \uparrow$
GCN (Sigmoid) [30]	1.7357	0.3458
2T-GCN (SoftMax) [31]	1.5382	0.4878
EGCN (MLE-O, Sigmoid) [27]	1.5893	0.5111
EGCN (MLE-O, SoftMax) [27]	1.5077	0.5531
EGCN++ (MLE-PO, Sigmoid)	1.5456	0.6254
EGCN++ (MLE-PO, SoftMax)	**1.4444**	**0.6713**

via sigmoid or the Softmax probability) can align better with human evaluation (see Table VII). Meanwhile, similar to results for the KIMORE dataset (see Table V), the evaluation score calculated with Softmax is more consistent with human evaluation. Expanded results for the EGCN++ on different exercises are available in Table 13 in Appendix A, available online. Although the results for evaluation metrics E_D and C_R in Table 13 do not perfectly match the prediction accuracy of MLE-PO in Table VI, the trend is that higher prediction accuracy improves the consistency between machine evaluation and human evaluation.

G. Ablation Study

Our MLE methods can be implemented with different training strategies. To validate the superior performance of our MLE-O and MLE-PO shown in Tables II, IV, and VI, we conduct ablation studies with the following MLE-O implementations.

1) *Self-importance:* Use the joint weights derived from h to replace those from g by fixing one θ_h and updating another θ_h.
2) *Swap h and g:* Calculate the joint weights from h and update the θ_g.
3) *No Pre-training:* Optimize h and g together without pre-training g.
4) *Tune θ_g:* Optimize θ_h while updating θ_g of the pre-trained g.
5) *Fix θ_g, Mean Along C_{out}:* Average the joint weights along the C_{out} dimension (see Fig. 8(right)), which is similar to [67], [68].
6) *Implementation with Other Backbones:* Our methods can be implemented with other backbones, such as AGCN [49], MS-G3D [50], and CTR-GCN [51], which were originally designed for action recognition.

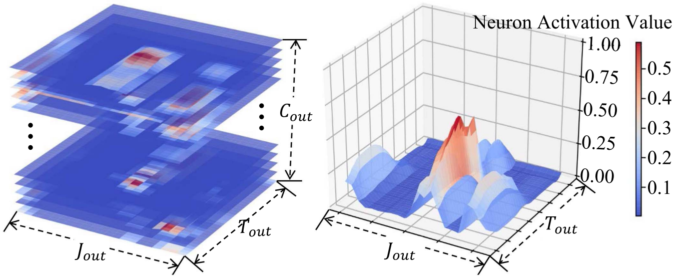


Fig. 8. *Left:* visualization of joint weights derived from neuron activation values of the pre-trained g. *Right:* visualization of mean values along the C_{out} dimension. Larger neuron activation values magnify the joint weights and vice versa.

TABLE VIII
RESULTS OF ABLATION STUDIES FOR MLE ON UI-PRMD, KIMORE, AND EHE DATASETS WITH THE CS PROTOCOL (ACCURACY IN %)

#	Model Implementations	Dataset		
		UI-PRMD	KIMORE	EHE
1	MLE-O (Self Importance)	69.08	76.57	68.28
2	MLE-O (Swap h and g)	72.28	77.02	77.73
3	MLE-O (No pre-training)	77.01	67.78	73.31
4	MLE-O (Tune θ_g)	79.59	72.95	69.59
5	MLE-O (Fix θ_g, Mean Along C_{out})	74.73	77.35	70.27
6	MLE-O (Fix θ_g, AGCN)	70.19	73.33	75.78
7	MLE-O (Fix θ_g, MS-G3D)	85.11	75.24	80.26
8	MLE-O (Fix θ_g, CTR-GCN)	86.34	77.64	81.70
9	MLE-O (Fix θ_g, GCN)	86.14	80.14	82.41
10	MLE-PO (Fix θ_g, GCN)	**89.95**	**83.56**	**86.38**

Best in bold, second best underlined.

Corresponding results for the three datasets (i.e., UI-PRMD, KIMORE, and EHE) appear in Table VIII. Fixing θ_g is consistently more effective than other settings. The adopted backbone GCN model also works better than other tested backbones. In brief, the orientation feature fails to learn joint weight knowledge as Ablations 1 and 2 cannot perform well on these three datasets. Without pre-training Model g via the position feature, Ablation 3 also cannot achieve satisfactory results, indicating that useful information (i.e., joint weights) is learned from pre-training g via the position feature. According to Ablation 4, tuning g

TABLE IX
AVERAGE PREDICTION RESULTS IMPLEMENTED WITH STATE-OF-THE-ART GCN MODELS (SEE POSITION AND ORIENTATION ROWS) AND ENSEMBLE METHODS OF OUR EGCN++ IMPLEMENTED WITH DIFFERENT BACKBONES (I.E., GCN, AGCN, MS-G3D, AND CTR-GCN) (ACCURACY IN %)

Method	UI-PRMD (Kinect v2)				KIMORE				EHE			
	GCN	AGCN	MS-G3D	CTR-GCN	GCN	AGCN	MS-G3D	CTR-GCN	GCN	AGCN	MS-G3D	CTR-GCN
Position	75.44	68.86	77.89	82.17	76.27	72.49	74.76	76.10	77.63	72.21	82.34	84.74
Orientation	77.03	76.96	84.26	86.18	77.54	72.27	75.53	77.38	77.74	68.29	71.15	76.87
SLE	78.57	82.63	85.47	87.53	78.05	76.71	77.19	78.12	80.36	70.78	73.32	85.60
FLE-B	75.31	64.88	71.14	78.86	71.52	70.58	73.37	80.10	79.40	67.51	78.99	82.88
FLE-CI	77.31	78.39	78.39	78.39	71.07	73.3	73.89	78.83	74.86	72.18	77.37	83.64
DLE-full	75.43	71.72	78.74	87.03	74.29	76.79	72.20	78.72	78.98	71.18	80.58	86.06
DLE-dual	77.19	73.54	83.32	84.56	77.94	70.62	76.33	77.18	78.82	70.23	76.54	81.84
MLE-O	86.14	70.19	85.11	86.34	80.14	73.3	75.24	77.64	82.41	75.78	80.26	81.70
MLE-PO	**89.95**	80.22	85.05	87.57	**83.56**	74.78	75.42	77.76	**86.38**	71.75	74.64	85.57

together with the h affects Model g's physical meaning (i.e., joint weights). The results of Ablation 5, "Fix θ_g, Mean Along C_{out}" in Table VIII, show that taking the mean value along the channel dimension can smooth out channel-level importance. Fig. 8(left) illustrates where the visualized joint weights fluctuate along the C_{out} dimension. Consequently, the averaged joint weights (see Fig. 8(right)) can be less informative than the original.

H. Comparison With State-of-the-Art

Sections IV-D, IV-E, and IV-F have compared state-of-the-art exercise assessment methods based on multiple evaluation metrics. Other state-of-the-art methods (designed for action recognition) can also be directly applied to the exercise assessment task as baselines. For this purpose, we first compare state-of-the-art GCN models such as AGCN [49], MS-G3D [50], and CTR-GCN [51] via the single-modal setting (i.e., using either position or orientation). Table IX contains prediction results for these baselines. We next implement our EGCN++ using the GCN baselines as backbones to further explore ensemble strategies proposed in our EGCN++ on the three datasets (see Table IX). Replacing the backbone with other advanced GCN baseline models does not lead to stable improvements. For instance, MS-G3D can improve the single-modal setting of the orientation feature for the UI-PRMD dataset but does enhance the performance of single-modal settings for the other two datasets. This outcome may have arisen because these GCN baselines were designed for action recognition instead of for exercise assessment. With our multi-level fusion approach, the MLE-PO method using the basic GCN outperforms the SLE using CTR-GCN on three datasets. This pattern further confirms that the joint weights learned from the position stream can regularize the orientation stream's training process.

I. Runtime Analysis

It takes about 4 minutes for our model to train one fold of 5-fold cross-validation on two GTX 1080 Ti GPUs with a batch size of 8. Regarding the methods' running time, we provide details including the inference time, the number of model parameters, and floating point operations (FLOPs) in Table X. We use fvcore[2] to calculate FLOPs. We test 20 samples (each sample has 150 skeleton frames and takes around 5 seconds) on a single GTX 1080 Ti with a batch size of 1 and report the average inference time. The runtime analysis indicates that MLE-O and MLE-PO are effective and computationally efficient.

TABLE X
RUNTIME ANALYSIS OF METHODS USED IN THIS RESEARCH

#	Method	Backbone	Inference Time	Parameters	FLOPs	Result
1	Position	GCN	0.094s	3.1M	3.8G	75.44%
2	Position	AGCN	0.109s	3.5M	7.2G	68.86%
3	Position	MS-G3D	0.090s	3.2M	12.2G	77.89%
4	Position	CTR-GCN	0.132s	1.4M	1.9G	82.17%
5	SLE	GCN	0.096s	3.1M	3.8G	78.57%
6	FLE-B	GCN	0.994s	6.2M	7.6G	75.31%
7	FLE-CI	GCN	0.108s	6.2M	7.6G	77.31%
8	DLE-full	GCN	0.100s	6.2M	7.6G	75.43%
9	DLE-dual	GCN	0.094s	6.2M	7.6G	77.19%
10	MLE-O	GCN	0.101s	6.2M	7.6G	86.14%
11	MLE-PO	GCN	0.103s	6.2M	7.6G	89.95%

V. DISCUSSION OF FUTURE DIRECTIONS

We have comprehensively compared several ensemble strategies and assessed the consistency between machine and human evaluations based on all currently available resources. However, avenues remain open for exploration in terms of representation models, domain knowledge, and problem definition.

Among other deep learning methods, recent advances in graph convolution models such as [52], [53], [72], [73], [74] have struggled to further improve the action recognition performance following the introduction of MS-G3D [50] and CTR-GCN [51]. More skeletal representations (e.g., skeleton bone, joint motion, and bone motion) can produce slight improvements in action recognition, which should be explored to potentially enhance the exercise assessment task. Present skeleton-based action recognition mainly ignores the orientation feature. Our work can motivate action recognition studies that account for this characteristic.

Our proposed approach relies solely on data to generate the machine evaluation score. Although this score can be consistent with human evaluation, it does not suggest what is going wrong with an exercise. More domain knowledge should be incorporated into new evaluation standards that can guide trainees by offering timely, or even instantaneous, feedback. [75] collected

[2] fvcore: https://github.com/facebookresearch/fvcore.git

a large real-world dataset that includes many exercises and fine-grained evaluations. Overall, however, large-scale open datasets are lacking in this field.

In addition to obtaining large-scale datasets, data efficiency and domain generalization are important to consider in the future when analyzing results. The RD evaluation protocol outperformed the evaluation CS protocol in our case (see Tables II, IV, and VI). Practitioners might consider gathering new data from subjects to support disease diagnosis or monitor the effects of behavioral therapies. The amount of additional data to be collected from new subjects can be determined using the learning regimes of few-shot learning and efficient transfer learning [76], [77].

Given the need for more helpful machine evaluation with simultaneous feedback, challenges such as segmentation, subject bias, and environmental change stand to be tackled. Along this line, [78] formulated the problem as pose matching. Scholars have yet to identify whether the ground truth is subject-specific (i.e., whether subjects define their ground truth in distinct ways). For example, when people play golf, they might have different wave trajectories. Novel problem definitions involving other sensors such as IMUs [79], biological sensors (e.g., EEG and MRI) [80], or even mixed reality devices [81] can be considered to possibly rectify these issues.

VI. Conclusion

In this paper, we have proposed the EGCN++ framework with various ensemble-based learning strategies for effective skeleton-based exercise assessment. The MLE-PO ensemble strategy fusing data at the data and model levels is superior to other fusion strategies and baselines. We have used several evaluation metrics to validate our strategy's effectiveness. After extensive experiments on the latest UI-PRMD, KIMORE, and EHE datasets, MLE-PO outperforms other ensemble strategies in terms of prediction accuracy. Given the evaluations available in the KIMORE and EHE datasets, MLE-PO can provide machine evaluation scores that are generally more consistent with human evaluation. This result reinforces our strategy's effectiveness. Finally, we have used several training schemes and ablated backbone implementations for our MLE strategy, followed by a runtime analysis that shows MLE-PO to be computationally efficient.

In the future, we aim to develop real-time exercise evaluation methods that can handle segmentation and provide helpful, timely feedback to exercise trainees while capturing more domain knowledge.

Acknowledgment

The authors would like to thank the editors and anonymous reviewers who helped improve this manuscript.

References

[1] S. R. Machlin, J. Chevan, W. W. Yu, and M. W. Zodet, "Determinants of utilization and expenditures for episodes of ambulatory physical therapy among adults," *Phys. Ther.*, vol. 91, no. 7, pp. 1018–1029, 2011.

[2] S. A. Jessep, N. E. Walsh, J. Ratcliffe, and M. V. Hurley, "Long-term clinical benefits and costs of an integrated rehabilitation programme compared with outpatient physiotherapy for chronic knee pain," *Physiotherapy*, vol. 95, no. 2, pp. 94–102, 2009.

[3] S. F. Bassett and H. Prapavessis, "Home-based physical therapy intervention with adherence-enhancing strategies versus clinic-based management for patients with ankle sprains," *Phys. Ther.*, vol. 87, no. 9, pp. 1132–1143, 2007.

[4] V. N. Kumar and C. Kataria, "Efficacy assessment of virtual reality therapy for neuromotor rehabilitation in home environment: A systematic review," *Disabil. Rehabilitation: Assistive Technol.*, vol. 18, pp. 1200–1220, 2023.

[5] V. Antoniou, C. H. Davos, E. Kapreli, L. Batalik, D. B. Panagiotakos, and G. Pepera, "Effectiveness of home-based cardiac rehabilitation, using wearable sensors, as a multicomponent, cutting-edge intervention: A systematic review and meta-analysis," *J. Clin. Med.*, vol. 11, no. 13, 2022.

[6] O. Giggins, D. Kelly, and B. Caulfield, "Evaluating rehabilitation exercise performance using a single inertial measurement unit," in *Proc. 7th Int. Conf. Pervasive Comput. Technol. Healthcare Workshops*, 2013, pp. 49–56.

[7] R. Colombo et al., "Robotic techniques for upper limb evaluation and rehabilitation of stroke patients," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, no. 3, pp. 311–324, Sep. 2005.

[8] Y. Tao, H. Hu, and H. Zhou, "Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation," *Int. J. Robot. Res.*, vol. 26, no. 6, pp. 607–624, 2007.

[9] C. Gu, W. Lin, X. He, L. Zhang, and M. Zhang, "IMU-based mocap system for rehabilitation applications: A systematic review," *Biomimetic Intell. Robot.*, vol. 3, 2023, Art. no. 100097.

[10] B. Yu, Y. Liu, and K. Chan, "A survey of sensor modalities for human activity recognition," in *Proc. 12th Int. Joint Conf. Knowl. Discov.*, Budapest, Hungary, 2020, pp. 2–4.

[11] R. Komatireddy, A. Chokshi, J. Basnett, M. Casale, D. Goble, and T. Shubert, "Quality and quantity of rehabilitation exercises delivered by a 3-D motion controlled camera: A pilot study," *Int. J. Phys. Med. Rehabil.*, vol. 2, no. 4, 2014.

[12] E. Saraee et al., "ExerciseCheck: Remote monitoring and evaluation platform for home based physical therapy," in *Proc. 10th Int. Conf. PErvasive Technol. Related Assistive Environ.*, 2017, pp. 87–90.

[13] H. M. Hondori and M. Khademi, "A review on technical and clinical impact of Microsoft kinect on physical therapy and rehabilitation," *J. Med. Eng.*, vol. 2014, 2014, Art. no. 846514.

[14] K. N. Karmali, P. Davies, F. Taylor, A. Beswick, N. Martin, and S. Ebrahim, "Promoting patient uptake and adherence in cardiac rehabilitation," *Cochrane Database Systematic Rev.*, no. 6, 2014.

[15] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, 2014.

[16] C. Dhiman and D. K. Vishwakarma, "A robust framework for abnormal human action recognition using ∇-transform and zernike moments in depth videos," *IEEE Sensors J.*, vol. 19, no. 13, pp. 5195–5203, Jul. 2019.

[17] P. Parmar and B. T. Morris, "What and how well you performed? A multitask learning approach to action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 304–313.

[18] Y. Tang et al., "Uncertainty-aware score distribution learning for action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9839–9848.

[19] J.-H. Pan, J. Gao, and W.-S. Zheng, "Adaptive action assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8779–8795, Dec. 2022.

[20] J.-H. Pan, J. Gao, and W. S. Zheng, "Action assessment by joint relation graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6331–6340.

[21] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2616–2625.

[22] J.-D. Huang, "Kinerehab: A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," in *Proc. 13th Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2011, pp. 319–320.

[23] R. Altilio, M. Paoloni, and M. Panella, "Selection of clinical features for pattern recognition applied to gait analysis," *Med. Biol. Eng. Comput.*, vol. 55, pp. 685–695, 2017.

[24] S. Sardari et al., "Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review," *Comput. Biol. Med.*, vol. 158, 2023, Art. no. 106835.

[25] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023.

[26] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen, "A survey of vision-based human action evaluation methods," *Sensors*, vol. 19, no. 19, 2019.

[27] B. X. Yu, Y. Liu, X. Zhang, G. Chen, and K. C. Chan, "EGCN: An ensemble-based learning framework for exploring effective skeleton-based rehabilitation exercise assessment," in *Proc. EGCN: An Ensemble-Based Learn. Framework Exploring Effective Skeleton-Based Rehabil. Exercise Assessment*, 2022, pp. 3681–3687.

[28] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 468–477, Feb. 2020.

[29] C. Williams, A. Vakanski, S. Lee, and D. Paul, "Assessment of physical rehabilitation movements through dimensionality reduction and statistical modeling," *Med. Eng. Phys.*, vol. 74, pp. 13–22, 2019.

[30] X. Bruce, Y. Liu, and K. C. Chan, "Skeleton-based detection of abnormalities in human actions using graph convolutional networks," in *Proc. 2nd Int. Conf. Transdisciplinary AI*, 2020, pp. 131–137.

[31] X. Bruce, Y. Liu, K. C. Chan, Q. Yang, and X. Wang, "Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression," *Pattern Recognit.*, vol. 119, 2021, Art. no. 108095.

[32] B. X. Yu, Y. Liu, X. Zhang, G. Chen, and K. C. Chan, "EGCN: An ensemble-based learning framework for exploring effective skeleton-based rehabilitation exercise assessment," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 3681–3687, doi: 10.24963/ijcai.2022/511.

[33] L. Tao et al., "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Comput. Vis. Image Understanding*, vol. 148, pp. 136–152, 2016.

[34] N. Sadawi, A. Miron, W. Ismail, H. Hussain, and C. Grosan, "Gesture correctness estimation with deep neural networks and rough path descriptors," in *Proc. Int. Conf. Data Mining Workshops*, 2019, pp. 595–602.

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

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

[37] M. A. R. Ahad, A. D. Antar, and O. Shahid, "Vision-based action understanding for assistive healthcare: A short review," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1–11.

[38] S. Rahman, S. Sarker, A. N. Haque, M. M. Uttsha, M. F. Islam, and S. Deb, "AI-driven stroke rehabilitation systems and assessment: A systematic review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 192–207, 2022.

[39] F. Frangoudes, M. Matsangidou, E. C. Schiza, K. Neokleous, and C. S. Pattichis, "Assessing human motion during exercise using machine learning: A literature review," *IEEE Access*, vol. 10, pp. 86874–86903, 2022.

[40] J. Antunes, A. Bernardino, A. Smailagic, and D. P. Siewiorek, "AHA-3D: A labelled dataset for senior fitness exercise recognition and segmentation from 3D skeletal data," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[41] P. Parmar and B. T. Morris, "Measuring the quality of exercises," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 2241–2244.

[42] A. Elkholy, M. Hussein, W. Gomaa, D. Damen, and E. Saba, "Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 1, pp. 280–291, Jan. 2020.

[43] J. Sarsfield, D. Brown, C. Langensiepen, N. Sherkat, J. Lewis, and P. Standen, "Reducing clinical subjective discrepancies in evaluation of clinical technology using objective measures," in *Proc. Int. Conf. Disabil., Virtual Reality Assoc. Technol.*, 2018.

[44] J. W. Stokes, J. P. Wanderer, and M. D. McEvoy, "Significant discrepancies exist between clinician assessment and patient self-assessment of functional capacity by validated scoring tools during preoperative evaluation," *Perioper. Med.*, vol. 5, no. 1, pp. 1–8, 2016.

[45] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*.

[46] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," 2015, *arXiv:1506.05163*.

[47] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3844–3852.

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

[49] 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.*, 2019, pp. 12026–12035.

[50] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 143–152.

[51] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13359–13368.

[52] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "InfoGCN: Representation learning for human skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20186–20196.

[53] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," 2022, *arXiv:2208.10741*.

[54] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, pp. 241–258, 2020.

[55] H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet, "A survey on ensemble learning for data stream classification," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–36, 2017.

[56] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscipl. Rev. Data Mining Knowl. Discov.*, vol. 8, no. 4, 2018, Art. no. e1249.

[57] R. Polikar, "Ensemble learning," in *Ensemble Machine Learning*, Berlin, Germany: Springer, 2012, pp. 1–34.

[58] Y. Yang, H. Lv, and N. Chen, "A survey on ensemble learning under the era of deep learning," *Artif. Intell. Rev.*, vol. 56, pp. 5545–5589, 2023.

[59] G. Hinton et al., "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[60] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–118.

[61] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 166–183.

[62] J. Meng, H. Hao, and Y. Luan, "Classifier ensemble selection based on affinity propagation clustering," *J. Biomed. Inform.*, vol. 60, pp. 234–242, 2016.

[63] J. Elmi and M. Eftekhari, "Dynamic ensemble selection based on hesitant fuzzy multiple criteria decision making," *Soft Comput.*, vol. 24, no. 16, pp. 12241–12253, 2020.

[64] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[65] A. S. Ross, W. Pan, L. A. Celi, and F. Doshi-Velez, "Ensembles of locally independent prediction models," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5527–5536.

[66] M. A. Ganaie et al., "Ensemble deep learning: A review," 2021, *arXiv:2104.02395*.

[67] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 469–478.

[68] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1227–1236.

[69] X. Bruce, Y. Liu, and K. C. Chan, "Multimodal fusion via teacher-student network for indoor action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3199–3207.

[70] B. X. Yu, Y. Liu, X. Zhang, S.-H. Zhong, and K. C. Chan, "MMNet: A model-based multimodal network for human action recognition in RGB-D videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3522–3538, Mar. 2023.

[71] Y. Liao, A. Vakanski, M. Xian, D. Paul, and R. Baker, "A review of computational approaches for evaluation of rehabilitation exercises," *Comput. Biol. Med.*, vol. 119, 2020, Art. no. 103687.

[72] N. Trivedi and R. K. Sarvadevabhatla, "PSUMNet: Unified modality part streams are all you need for efficient pose-based action recognition," 2022, *arXiv:2208.05775*.

[73] S. Wang, Y. Zhang, F. Wei, K. Wang, M. Zhao, and Y. Jiang, "Skeleton-based action recognition via temporal-channel aggregation," 2022, *arXiv:2205.15936*.

[74] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, "Language supervised training for skeleton-based action recognition," 2022, *arXiv:2208.05318*.

[75] Z.-Q. Yang, D. Du, X.-Y. Wei, and R. K.-Y. Tong, "Augmented reality for stroke rehabilitation during COVID-19," 2022.

[76] B. X. Yu, J. Chang, L. Liu, Q. Tian, and C. W. Chen, "Towards a unified view on visual parameter-efficient transfer learning," 2022, *arXiv:2210.00788*.

[77] B. X. Yu et al., "Visual tuning," 2023, *arXiv:2305.06061*.

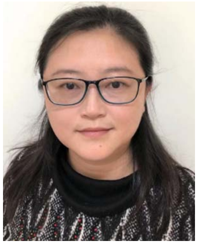
[78] Y. Qiu, J. Wang, Z. Jin, H. Chen, M. Zhang, and L. Guo, "Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training," *Biomed. Signal Process. Control*, vol. 72, 2022, Art. no. 103323.

[79] S. García-de Villa, D. Casillas-Pérez, A. Jiménez-Martín, and J. J. García-Domínguez, "Simultaneous exercise recognition and evaluation in prescribed routines: Approach to virtual coaches," *Expert Syst. Appl.*, vol. 199, 2022, Art. no. 116990.

[80] R. V. Pedroso, A. E. Lima-Silva, P. E. Tarachuque, F. J. Fraga, and A. M. Stein, "Efficacy of physical exercise on cortical activity modulation in mild cognitive impairment: A systematic review," *Arch. Phys. Med. Rehabil.*, vol. 102, no. 12, pp. 2393–2401, 2021.

[81] Y. K. Dwivedi et al., "Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int. J. Inf. Manage.*, vol. 66, 2022, Art. no. 102542.

Bruce X. B. Yu (Member, IEEE) received the PhD degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, where he continued to serve as a post-doc fellow until 2023. He is now a tenure-track assistant professor with the Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, China. His research expertise spans across Big Data analytics, artificial intelligence, and image/video processing. His primary research topic is vision-based human behavior understanding, leading to publications on top venues such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Pattern Recognition*, ICCV, AAAI, and IJCAI. His research outcomes also get international recognition such as best paper and best presentation awards. Besides application-driven research, he also works on fundamental research problems such as 3D human pose reconstruction, multimodal data/sensor fusion, and efficient transfer learning.



Yan Liu (Member, IEEE) received the MSc degree from Business School, Nanjing University, in China, and the PhD degree in computer science from Columbia University, She has been the Principal investigator of more than ten projects and awarded the best paper awards several times. She serves as the organizing committee/technical program committee member for many conferences such as AAAI and ACM Multimedia, and as the reviewer for many journals such as *IEEE Transactions on Neural Networks and Learning Systems* as well as *ACM Transactions on Intelligent Systems and Technology*. She is the director of Cognitive Computing Lab, which focuses on both discovering the secrets of human brain by exploring information technologies, for example, fMRI imaging, and investigating novel computational models and systems by referencing brain structure and mind process, for example, deep learning. In addition to the theoretical study on the leading edge of scientific research, she works on applications with great commercial potentials, such as music therapy for emotional health. Her research interests span a wide range of topics, ranging from brain modeling and cognitive computing, image/video retrieval, computer music to machine learning and pattern recognition.

Keith C. C. Chan (Member, IEEE) received the BMath (Hons.) degree in computer science and statistics and the MASc and PhD degree in systems design engineering from the University of Waterloo, Ontario, Canada. Soon after graduation, he joined the IBM Canada Laboratory in Toronto, Canada, as a software analyst, developing multimedia and software engineering tools, he returned to the academia to join Ryerson University in Toronto and The Hong Kong Polytechnic University where he worked for 25 years. From 2002 to 2008, he served as head of the Department of Computing. His research interests are in artificial intelligence, data science, bioinformatics and software engineering. He has authored and co-authored three books and more than 300 publications in journals and conference proceedings and had also been serving actively as program committee member of more than 20 international conferences annually. He has been active in technology transfer through consulting and contract research and is ranked among the top 2% of scientists in the world in the field of Artificial Intelligence and Image Processing on Stanford University's list.

Chang Wen Chen (Fellow, IEEE) received the BS degree from the University of Science and Technology of China, in 1983, the MSEE degree from the University of Southern California, in 1986, and the PhD degree from the University of Illinois at Urbana-Champaign, in 1992. He is currently chair professor of Visual Computing with The Hong Kong Polytechnic University. Before his current position, he served as dean of the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen from 2017 to 2020. He also served as an Empire Innovation professor with the University at Buffalo, the State University of New York from 2008 to 2021. He was Allen Henry Endow chair professor with the Florida Institute of Technology from 2003 to 2007. He was on the faculty with Electrical and Computer Engineering, University of Rochester from 1992 to 1996 and with the Faculty of Electrical and Computer Engineering, University of Missouri-Columbia from 1996 to 2003. He has served as the editor-in-chief for *IEEE Trans. Multimedia* from 2014 to 2016, and the editor-in-chief for *IEEE Trans. Circuits and Systems for Video Technology* from 2006 to 2009. He has been an editor for several other major IEEE Transactions and Journals, including as senior editor for the *IEEE Journal of Selected Areas in Communications* and the *IEEE Journal of Selected Topics in Signal Processing*. He has served as Conference chair for several major IEEE, ACM, and SPIE conferences related to multimedia communications and signal processing. His research has been funded by both government agencies and industrial corporations. He and his students have received 10 Best Paper Awards or Best Student Paper Awards more than the past two decades. He has also received several research and professional achievement awards. These include the Sigma Xi Excellence in Graduate Research Mentoring Award, in 2003, the Alexander von Humboldt Research Award, in 2009, the University at Buffalo Exceptional Scholar – Sustained Achievement Award, in 2012, the SUNY System Chancellor's Award for Excellence in Scholarship and Creative Activities, in 2016, and the University of Illinois ECE Distinguished Alumni Award, in 2019. He is an SPIE fellow (2007), and an elected member of the Academia Europaea (2021).