



HAL
open science

SSL-Rehab: Assessment of Physical Rehabilitation Exercises Through Self-Supervised Learning of 3D Skeleton Representations

Ikram Kourbane, Panagiotis Papadakis, Mihai Andries

► **To cite this version:**

Ikram Kourbane, Panagiotis Papadakis, Mihai Andries. SSL-Rehab: Assessment of Physical Rehabilitation Exercises Through Self-Supervised Learning of 3D Skeleton Representations. *Computer Vision and Image Understanding*, 2025, 251, pp.104275. 10.1016/j.cviu.2024.104275 . hal-04841807

HAL Id: hal-04841807

<https://imt-atlantique.hal.science/hal-04841807v1>

Submitted on 16 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

SSL-Rehab: Assessment of Physical Rehabilitation Exercises Through Self-Supervised Learning of 3D Skeleton Representations

Ikram Kourbane*, Panagiotis Papadakis and Mihai Andries

IMT Atlantique, Lab-STICC, UMR CNRS 6285, team RAMBO, F-29238 Brest, France,

ARTICLE INFO

Keywords:

Rehabilitation

Quality score assessment

Self-supervised learning

Transfer learning

Graph convolutional networks

Transformer

ABSTRACT

Rehabilitation aims to assist individuals in recovering or enhancing functions that have been lost or impaired due to injury, illness, or disease. The automatic assessment of physical rehabilitation exercises offers a valuable method for patient supervision, complementing or potentially substituting traditional clinical evaluations. However, acquiring large-scale annotated datasets presents challenges, prompting the need for self-supervised learning and transfer learning in the rehabilitation domain. Our proposed approach integrates these two strategies through Low-Rank Adaptation (LoRA) for both pretraining and fine-tuning. Specifically, we train a foundation model to learn robust 3D skeleton features that adapt to varying levels of masked motion complexity through a three-stage process. In the first stage, we apply a high masking ratio to a subset of joints, using a transformer-based architecture with a graph embedding layer to capture fundamental motion features. In the second stage, we reduce the masking ratio and expand the model's capacity to learn more intricate motion patterns and interactions between joints. Finally, in the third stage, we further lower the masking ratio to enable the model to refine its understanding of detailed motion dynamics, optimizing its overall performance. During the second and third stages, LoRA layers are incorporated to extract unique features tailored to each masking level, ensuring efficient adaptation without significantly increasing the model size. Fine-tuning for downstream tasks shows that the model performs better when different masked motion levels are utilized. Through extensive experiments conducted on the publicly available KIMORE and UI-PRMD datasets, we demonstrate the effectiveness of our approach in accurately evaluating the execution quality of rehabilitation exercises, surpassing state-of-the-art performance across all metrics. Our project page is available online.

1. Introduction

Human motion analysis is a highly active research field within computer vision. While the majority of studies in this area focus on action detection and recognition [1, 2, 3, 4], the domain of human movement quality assessment (HMQA) is comparatively understudied. This area involves identifying and quantifying deviations from standard movement patterns and providing feedback on an individual's execution of an action. HMQA is crucial in various domains, including functional capacity evaluation, sports movement optimization, ergonomic risk assessment, and applications in physical therapy and rehabilitation.

The assessment of physical rehabilitation exercises (APRE) is crucial for optimizing patient care and facilitating recovery from injuries or medical conditions. Traditional methods often depend on evaluations by healthcare professionals, which can be time-consuming, subjective, and resource-intensive. The COVID-19 pandemic lockdowns have further highlighted the need for secure, home-based rehabilitation systems that use conventional sensors. These systems provide alternatives to in-person evaluations, ensuring continuity of care during periods of restricted mobility or in isolated areas.


In response to these challenges, there has been a growing interest in leveraging advanced machine learning techniques to

develop objective, data-driven approaches for HMQA. Computer vision techniques provide practical and cost-efficient solutions by utilizing standard RGB cameras or affordable RGB+Depth sensors to capture detailed 3D poses, which serve as crucial inputs for deep learning algorithms [5]. However, training these models requires substantial volumes of diverse data. Furthermore, precise annotation by clinical experts is essential to ensure accurate scoring, adding further complexity to the process. Consequently, the scarcity of large-scale datasets tailored for assessing physical rehabilitation exercises remains a significant obstacle to developing robust and clinically effective models [6].

Recently, the integration of self-supervised learning (SSL) techniques has emerged as a promising approach for acquiring meaningful representations from raw data without the need for explicit labeling in related tasks such as action detection and recognition [7]. However, this approach has not been extensively studied in rehabilitation, where the issue of limited datasets is more significant. Our work introduces a new foundation framework that utilizes SSL pretraining to acquire robust 3D skeletal representations. Through transfer learning, our approach avoids the need for extensive manual annotation. This combination mitigates the risk of overfitting, allowing the model to generalize effectively across diverse datasets and unseen examples.

Similar to recent work in skeleton-based SSL for action recognition Masked Motion Prediction (MAMP) [2], our method adopts a transformer-based architecture to capture the configurations of 3D skeleton motion through masked

*Corresponding author

 ikram.kourbane@imt-atlantique.fr (I. Kourbane)

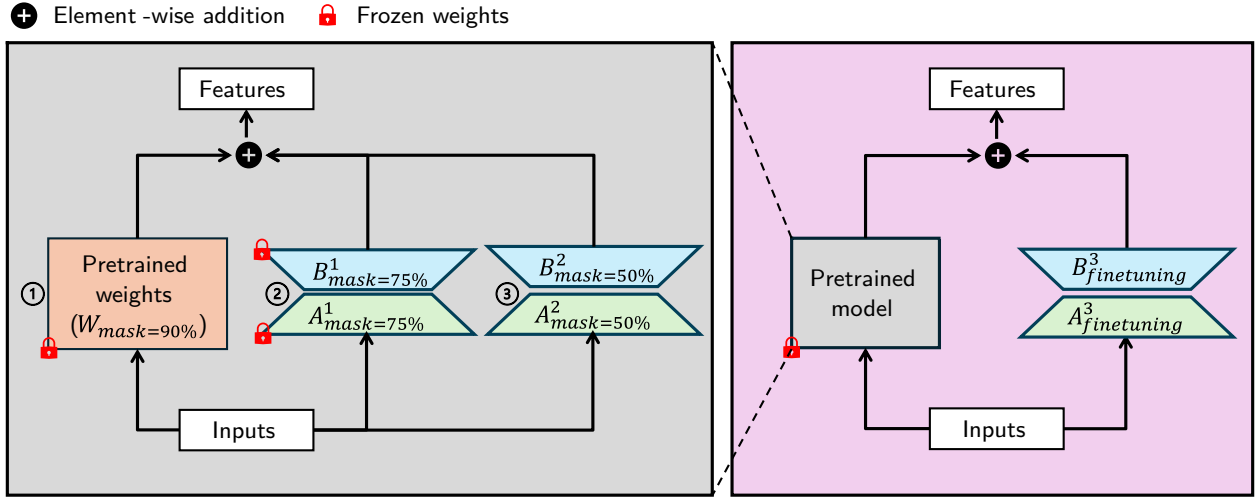


Fig. 1: The proposed training method comprises two phases: pretraining (**left**), aimed at gradually decreasing the masked motion levels, and the fine-tuning phase (**right**). LoRA matrices A and B are incorporated to effectively reduce the model complexity.

motion modeling. MAMP proposes sorting the joints based on their motion intensity, which involves ranking joints based on how much they move throughout an activity sequence. This approach enables MAMP to focus on the most dynamic joints first, as these typically carry more information about the action being performed. For example, in a walking motion, leg joints would exhibit higher motion intensity than arm joints. Unlike [2], which selects a single masking ratio (90%), our method considers different masking ratios to learn 3D skeleton representations (Figure 1). This enables us to provide a more comprehensive understanding of the importance of each joint in the motion sequence, leading to more robust features that capture both subtle and significant motions effectively. Additionally, the model gains adaptability to varying levels of noise or missing information in the input data, enhancing its ability to handle diverse and real-world scenarios. Furthermore, this diversity enriches the learning process during SSL, enabling the model to capture a broader range of motion patterns and nuances in the data.

During pretraining, ensembling models for different masking ratios would lead to a large model size and significant computation time. To address this challenge, we propose using parameter-efficient fine-tuning techniques in the pretraining phase on the NTU-60 dataset [8]. Specifically, we integrate LoRA (Low-Rank Adaptation) layers [9] to learn distinct features tailored to varying degrees of motion complexity while preserving pertinent information integrity (Figure 1). This technique offers the flexibility needed to accommodate diverse motion ranges while maintaining model size. Our approach involves three pretraining stages, each targeting different motion complexity levels. Initially, we utilize MAMP [2] to learn features from 90% masked joints, focusing on capturing high-level motion patterns. Subsequently, we introduce new LoRA layers and gradually reduce the proportion of masked joints to 75% and 50% for the second and third stages, respectively. This allows the model

to pinpoint the joints crucial for capturing intricate motion details.

In the transfer learning phase, we employ LoRA as a fine-tuning technique to leverage the knowledge acquired during pretraining on unlabeled data for the downstream rehabilitation assessment task. This approach harnesses the rich representations learned in pretraining and refines them to capture the specific nuances and complexities of the KIMORE [10] and UI-PRMD [11] datasets.

Instead of utilizing convolution layers [12] as an embedding projector [2], our approach employs a graph convolutional network (GCN) [13]. This adjustment enhances the model's capacity to learn spatial features, resulting in more insightful assessments of exercise quality. We conduct various data augmentations and preprocessing techniques to improve the performance of our model. In summary, the contributions of this work are as follows:

- We propose an SSL-based method for assessing physical rehabilitation exercises. Our approach utilizes decreasing masked motion modeling to learn robust 3D skeleton representations.
- We adopt the LoRA technique for both the pretraining and transfer learning phases of our model. This strategy enables us to effectively preserve pertinent information.
- We conducted extensive experiments on two rehabilitation datasets (KIMORE [10] and UI-PRMD [11]), demonstrating superior performance compared to the state-of-the-art.

The remainder of this paper is organized as follows. Section 2 reviews the related studies of our work. Section 3 explains the proposed method in depth and describes the most important modules of our framework. Section 4 describes the experimental settings. Section 5 analyzes the

obtained results on two public datasets and compares the proposed approach against the state-of-the-art. Finally, Section 6 presents the conclusion of the study and directions of future work.

2. Related work

In this section, we present a thorough review of methodologies related to the APRE. Furthermore, we delve into pertinent literature concerning action recognition utilizing self-supervised learning with skeleton data that fall within the perimeter of the current work.

2.1. Assessment of physical rehabilitation exercises

In recent years, there has been a growing interest in using artificial intelligence-based techniques to improve the accuracy and efficiency of APRE [6, 14]. Early studies focused on probabilistic approaches like Hidden Markov models (HMMs) [15, 16] and mixtures of Gaussian distributions [17] for assessing exercises. However, these approaches require several preprocessing stages, such as feature extraction. This can be time-consuming, and computationally expensive while identifying the optimal parameter values that lead to the best performance can be particularly challenging.

End-to-end deep learning models have demonstrated the capacity to automatically assess a patient's physical abilities based on data collected from wearable [18, 19] or vision sensors [20, 21]. First-generation methods typically classify movements as either correct or incorrect [22], without however providing details on the quality of the movement. More recent methods overcome this limitation by predicting a continuous score for each movement [23, 24, 25, 26] that can be more informative and enable monitoring of subtle progress over time. Also, Du et al. [27] introduced a method to quantify patient performance using a Gaussian Mixture Model (GMM) log-likelihood metric. Their model utilized hierarchical processing of joint displacements across various body parts, incorporating convolutional and recurrent layers to encode correlations in movement data. However, these methods do not explicitly consider the topological structure of the human body. This means that they do not take into account how the different parts of the body are connected to each other and how they move together. To address this limitation, many recent approaches use GCNs to model skeletal constraints among neighboring joints in a non-Euclidean space.

GNNs can extract features from data that are arranged in an irregular graph structure. The spatio-temporal graph convolutional networks (ST-GCN) framework proposed by Yan et al. [28] is the seminal work that captures both spatial and temporal features from skeleton data, achieving remarkable results in classifying actions. In subsequent works, Chowdhury et al. [29] proposed a model that uses a GCN to extract features from skeleton data, followed by an LSTM [30] to predict the output quality score of an exercise. Chen et al. [31] proposed an ensemble-based GCN for movement assessment, which uses a combination of multiple GCNs to

learn more robust features from the movement data. Deb et al. [32] proposed a GCN-based method that can process variable-length inputs using long short-term memory networks (LSTMs) [33] and employs self-attention of body joints indicating their role in predicting assessment scores.

Following this work, [34] merges modified STGCN and transformer [35] architectures to handle spatio-temporal data effectively and identify the most important joints. Also, Réby et al. [20] used a transformer network to learn the long-range dependencies in the input data, using a graph network to learn the spatial and temporal relationships between the different body joints. More recently, [36] proposed a multi-task contrastive learning framework aimed at capturing subtle yet crucial differences within skeleton sequences. This framework is designed to address both the performance metric and assessment quality assurance challenges encountered in physical rehabilitation exercises. However, the aforementioned methods typically involve training models with extensive architectures, such as transformer-based and LSTMs-based models, on relatively small datasets, thereby posing challenges for both training and testing. Specifically, training on small datasets may result in a lack of diversity in the captured motion patterns, limiting the model's ability to accurately assess the quality of rehabilitation exercises across various patient demographics and movement characteristics. To address this issue, we propose employing a self-supervised learning approach to acquire 3D skeleton representations from large-scale datasets [8]. Subsequently, we fine-tune the model using smaller annotated datasets [11, 10] for APRE.

2.2. Self-supervised 3D action recognition

SSL has gained considerable attention in recent years thanks to its capacity to harness large amounts of unlabeled data for training deep neural networks. In the domain of 3D action recognition, numerous methodologies have emerged to exploit the temporal and spatial information inherent in 3D action sequences without necessitating explicit annotations.

Contrastive learning has become increasingly popular in self-supervised 3D action recognition. For instance, Shah et al. [7] leverage contrastive learning to provide valuable guidance across diverse skeleton modalities, while Zhu et al. [37] explore better action data augmentation through this approach. Additionally, Zhou et al. [38] utilize contrastive learning to constrain the distance between confident and ambiguous samples, thereby enhancing the performance of ambiguous action recognition. Moreover, recent works have delved into self-supervised learning combined with GCNs for 3D action recognition [39, 40], enabling models to learn robust representations capturing spatial dependencies between joints.

However, while contrastive learning is effective in capturing spatial relationships, it lacks explicit constraints for exploring the temporal context of motion. To address this

limitation, Mao et al. [2] depart from conventional self-reconstruction objectives by introducing the MAMP framework. This framework explicitly models contextual motion, resulting in significantly improved performance compared to raw skeleton reconstruction methods such as SkeletonMAE [41]. However, this work reconstructs motion using only one ratio of the masked joints, which may limit its ability to capture the full spectrum of motion complexities. In contrast, our approach utilizes various masked motion ratios to achieve a more comprehensive understanding of motion dynamics. In addition, we incorporate Low-Rank Adaptation layers [9] to efficiently reduce the number of trainable parameters while maintaining high performance, ensuring that the model remains both accurate and computationally efficient.

3. Methodology

The necessity for SSL and transfer learning in APRE arises from the challenges associated with acquiring large-scale annotated datasets. Annotating datasets for physical rehabilitation exercises is difficult and resource-intensive due to the need for precise and detailed labeling by domain experts. Additionally, the variability in individual patient movements and the diverse range of exercises further complicate the annotation process.

SSL alleviates this burden by enabling models to learn meaningful representations from unlabeled data, reducing the need for extensive manual annotation. This approach is not only resource efficient but further allows for the use of vast amounts of readily available unlabeled data. Specifically, our approach does not rely on high-level task-specific labels, such as action categories or clinical assessments. Instead, it leverages the inherent structure of motion data. Namely, the joint coordinates, which can be directly inferred from the input. Meanwhile, transfer learning facilitates the adaptation of pre-trained models to APRE. Consequently, by fine-tuning the pre-trained model on a smaller dataset of labeled rehabilitation exercises, we enable effective knowledge transfer and model adaptation.

We propose an SSL method using decreasing masked motion modeling to learn robust 3D skeleton representations (Section 3.2). Our approach, illustrated in Figure 2, comprises two main phases: (i) pretraining on a large-scale dataset, and (ii) fine-tuning on a small-scale rehabilitation dataset. We adopt the LoRA technique for both phases, reducing model size while preserving pertinent information (Section 3.3). Additionally, we employ a GCN-based embedding layer to capture spatial dependencies between joints (Section 3.1). The network architecture is explained in Section 3.4.

3.1. GCN-based embedding layer

Human actions can be viewed as a set of spatio-temporal changes in motion. Inspired by the natural graph representation of the human body, we propose using GCNs as an embedding layer (Figure 2). This approach leverages the

Notation	Definition
G	Graph structure
J	Set of nodes of G
E	Set of edges of G
U	Adjacency graph matrix for G
\bar{U}	Normalized adjacency graph matrix
D	Degree matrix
I	Identity matrix
W	Learnable parameters
k	Number of GCN layers
\oplus	Concatenation operation
\cdot	Dot product
$+$	Element-wise addition
x	Input sequence
n	Number of sequences
T	Number of LoRA stages
S	Number of attention heads
Q	Query matrix in attention
K	Keys matrix in attention
V	Values matrix in attention
M	Masked motion
h	Attention head
MHA	Multi-head attention layer
A	Projection matrix of LoRA
B	Reconstruction matrix of LoRA
$GeLU$	Gaussian Error Linear Units
$ReLU$	Rectified Linear Units
L	Loss function
y	True values
\hat{y}	Predicted values

Table 1
Summary of commonly used notations

inherent graph structure of skeletal data, enabling the capturing of complex relationships between joints and facilitating the extraction of richer and more meaningful features from the skeletal data. In contrast, works such as [2] employ 2D convolution layers that treat each joint independently, which may lead to failure in capturing the inherent structural dependencies within the skeletal data.

A graph, denoted by $G = (J, E)$ is a data structure that consists of a set of nodes J , and a set of edges E , where the edges represent connections between the nodes. GCNs take into account the relationships between the nodes in a graph represented by the adjacency matrix U , which allows them to learn more complex patterns than ordinary neural networks. GCNs learn to represent nodes in a graph by aggregating information from their neighbors. The adjacency matrix of the graph is used to select the neighboring nodes that contribute to each node's representation. The propagation rule typically used in GCNs is defined as:

$$H^{(k+1)} = \sigma \left(D^{-\frac{1}{2}} \bar{U} D^{-\frac{1}{2}} H^{(k)} W^{(k)} \right) \quad (1)$$

where σ is the *ReLU* activation function, $H^{(k)}$ and $W^{(k)}$ are the features and the weights in the k^{th} layer, respectively. The

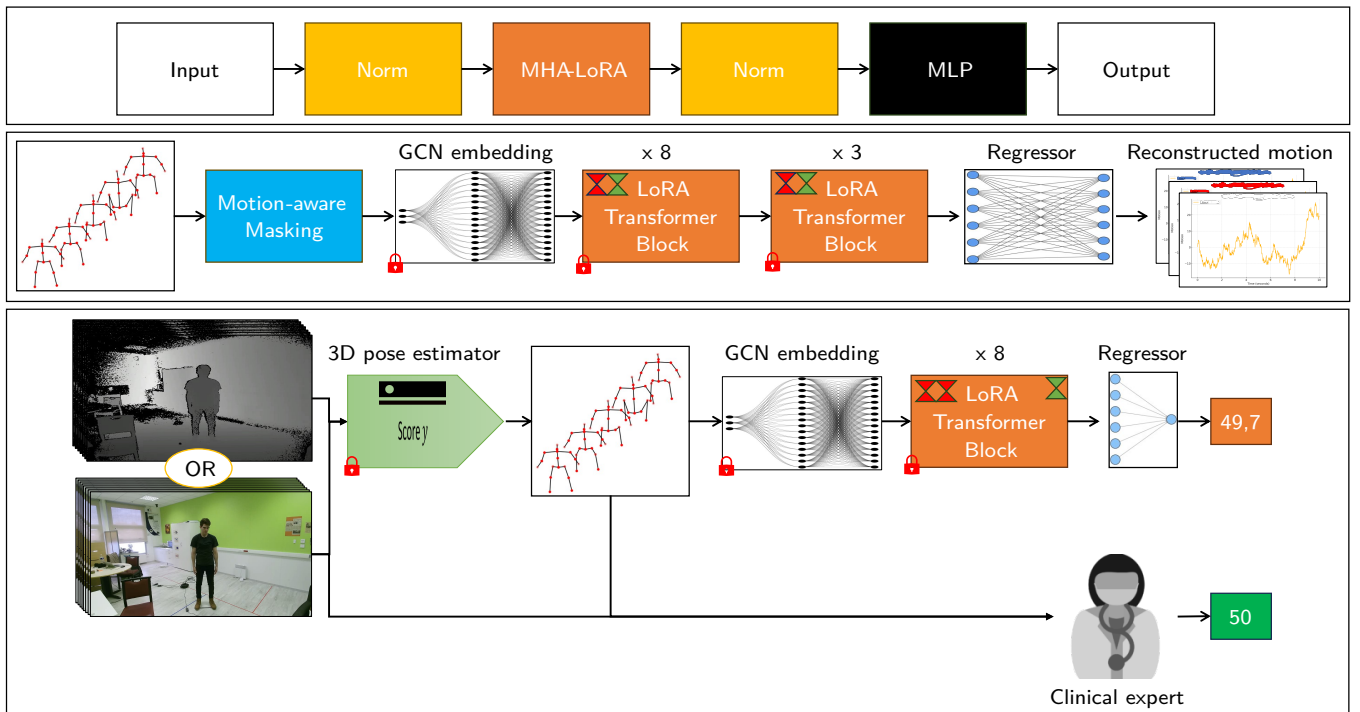


Fig. 2: The overall architecture of the proposed SSL-based assessment of physical rehabilitation exercises is depicted. From top to bottom, the three components represent: the LoRA Transformer Block, the pretraining architecture, and the fine-tuning architecture. Our quality estimation model is trained to align with the score of the clinical expert, who determines the ground truth score.

adjacency matrix $\bar{U} = U + I$ includes self-loops. The term $D^{-\frac{1}{2}} \bar{U} D^{-\frac{1}{2}}$ normalizes the adjacency matrix symmetrically where D is its degree matrix. This normalization helps prevent gradient issues, ensuring stable and efficient training of GCNs. In our architecture, we use two GCN layers that follow the internal structure presented in [13]

3.2. Decreasing masked motion modeling

Recently, Mao et al. [2] proposed that instead of relying on traditional pretext tasks such as masked self-component reconstruction of human joints [41], effective feature representation learning for 3D action recognition depends on explicit contextual motion modeling. Their empirical findings within the MAMP framework demonstrate that using a masking ratio of 90% yields the best results. However, this approach may not be adaptable to varying levels of motion complexity. Additionally, its ability to generalize to different types of movements encountered in rehabilitation exercises may be limited. These exercises encompass a broad spectrum of movements, ranging from fundamental gross motor actions to intricate fine-grained gestures. Each movement type targets distinct muscle groups and contributes uniquely to overall physical recovery.

Our methodology involves a progressive reduction of the proportion of masked joints across different training stages. The numbers in Figure 1 indicate the sequence of training stages. After completing the training of each stage, we freeze the weights before progressing to the next stage. During

fine-tuning, we leverage our pretrained model to adapt it to specific tasks by fine-tuning only a subset of the model parameters. This ensures that the knowledge acquired during pretraining is effectively utilized, while allowing the model to adapt to new data and tasks with minimal adjustments. By systematically adjusting the number of masked joints, we can effectively simulate and capture a diverse array of motion complexities, ultimately leading to more robust and generalized learning. As we will demonstrate in the ablation study of Section 3.2, our approach outperforms MAMP.

In the initial stage, we mask 90% of the joints. This stage presents the model with the most challenging scenarios, forcing it to learn essential features from scratch. By confronting such high levels of occlusion, the model develops a strong foundation of basic motion patterns and relationships between joints. This training technique helps in capturing critical spatio-temporal dependencies within the skeletal data, essential for understanding gross motor movements. Following the high masking stage, we reduce the proportion of masked joints to 75%. At this stage, the task becomes moderately challenging. The model leverages the foundational knowledge acquired in the first stage to refine its understanding and representation of motion patterns. This intermediate stage allows the model to start focusing on more detailed aspects of movement while still dealing with occlusion. It is useful for capturing more intricate movements that involve a combination of gross and fine motor skills. In the final pretraining stage, the proportion of masked

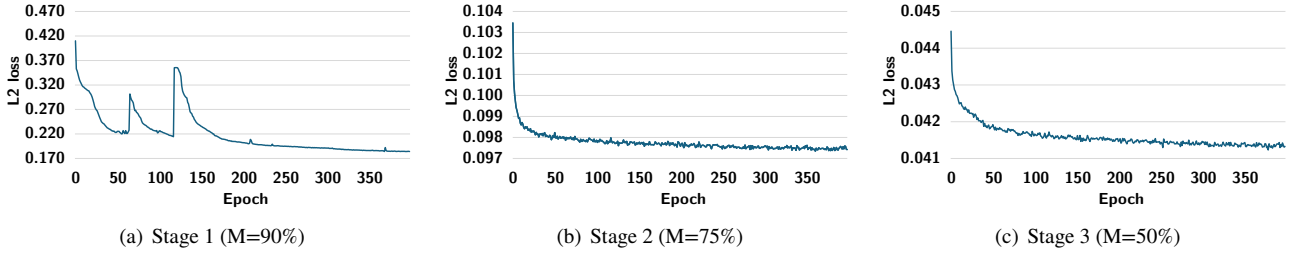


Fig. 3: The pretraining loss curves for the proposed three masked motion modeling stages.

joints is further reduced to 50%. By this stage, the model can effectively utilize the previously learned complex motion patterns to handle simpler, less occluded scenarios. This stage is crucial for capturing fine-grained gestures, which target specific muscle groups and joints. We have adhered to the same masking procedure as the one outlined in MAMP [2] for the initial stage of masking. Initially, the joints are sorted based on their motion intensity, following which the first 90% are masked. Subsequently, in the second and third stages, we adopt a similar strategy by masking 75% and 50% of the joints, respectively. This approach is motivated by the understanding that merely masking 90% of the joints may not yield adequate insights into the significance of each joint for a specific exercise. Instead, by exploring varying levels of masked motion, we can obtain more nuanced and reliable representations.

3.3. Parameter-efficient pretraining and fine-tuning

Parameter efficient finetuning (PEFT) is a strategy to adapt large pre-trained models to specific tasks without retraining the entire model. By minimizing the number of trainable parameters, PEFT reduces computational and storage needs while maintaining performance. LoRA [9] is a PEFT method that reduces parameter space by approximating large weight matrices with lower-rank matrices. This approach captures significant features while discarding redundant information, leading to fewer parameters and often faster convergence [42].

Rather than pretraining a separate model for each stage, our approach advocates for training a single model from scratch in the first stage. The model is trained with 90% masked joints to encourage it to learn generalized representations of movement patterns. As the training progresses, additional LoRA layers are added to the model architecture. These layers are specifically designed to accommodate different levels of masking, allowing the model to learn distinct features for each level without significantly increasing complexity. This strategy not only streamlines the training process by avoiding the need for multiple pretraining models but also ensures that the model remains adaptable to changes in movement complexity.

We adopt the motion reconstruction loss formulation as defined in [2] and illustrate the loss curve for each stage

of our framework in Figure 3. The first stage exhibits a significantly higher loss compared to subsequent stages, which can be attributed to the challenges inherent in masked motion modeling. Consequently, we observe peaks in the loss curve, indicating the model's initial difficulty in reconstructing highly masked motions. This stage requires 400 epochs to stabilize. As training progresses and the masking ratio decreases in the later stages, the model leverages more effectively the learned features, resulting in a smoother loss curve and improved performance.

During the fine-tuning phase, we adapt our pretrained model to the APRE task on KIMORE [10] and UI-PRMD [11] datasets. The regression loss for APRE on the KIMORE and UI-PRMD datasets is defined as:

$$L_{rehab} = \frac{1}{n} \sum_{i=1}^n \|y_i - \hat{y}_i\|^2 \quad (2)$$

where y and \bar{y} are the ground-truth and predicted quality score values.

3.4. Network architecture

The proposed network architecture for the pretraining phase is illustrated in Figure 2 (middle). The input skeleton is first processed by the motion-aware masking module, which masks the joints based on their motion intensity [2]. The output from this module is then fed into a GCN-based embedding layer to learn the spatial dependencies between the joints. Subsequently, the processed data is passed to a transformer-based encoder-decoder, which reconstructs the input motion using a regression layer. The model is trained in an end-to-end manner. The figure represents the final masking stage (50%) where previous LoRA layers are frozen.

In the fine-tuning phase (Figure 2, bottom), the 3D pose is input into the frozen GCN-based embedding and encoder block. We then train the regressor and the LoRA fine-tuning module to estimate the quality score. Note that the red/green triangles represents frozen and learnable LoRa matrices (A and B), respectively.

The Transformer model, as described by [35], is composed of various blocks as shown in Figure 2 (top). We have modified these blocks to incorporate LoRA layers. Specifically, the Multi-Head Attention (MHA) mechanism enables the model to simultaneously attend to information from different

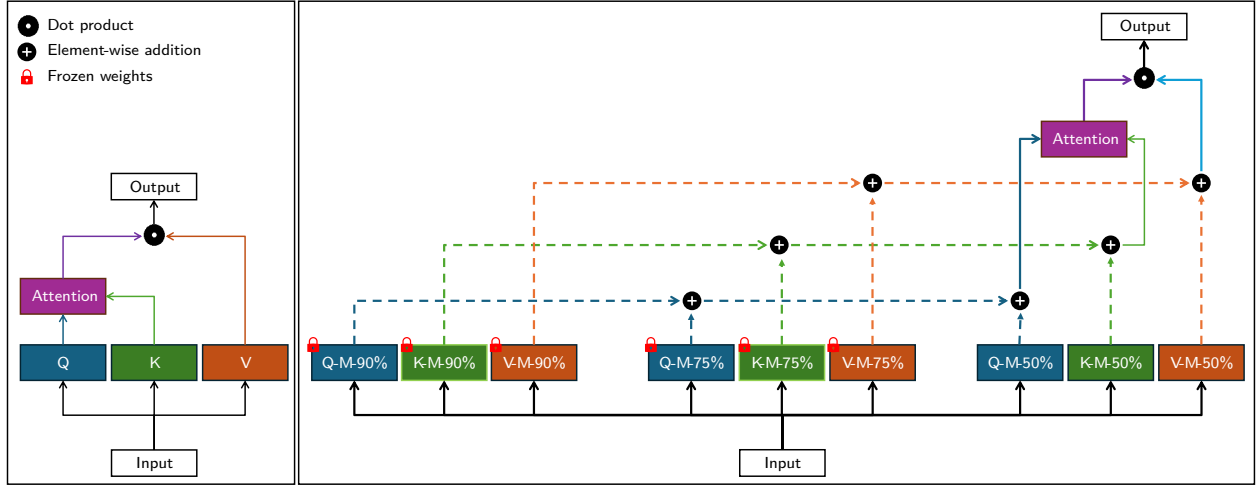


Fig. 4: The left image represents the self-attention layer of the original transformer, while the right image illustrates our proposed multi-stage masked motion modeling module.

representation subspaces at various positions:

$$MHA(x) = (h_1(x) \oplus h_2(x), \dots, \oplus h_S(x)) \cdot W^O \quad (3)$$

Each head is computed as:

$$h_i(x) = Attention(Q_i, K_i, V_i) \quad (4)$$

$$Q_i(x) = x \cdot W_i^Q, K_i(x) = x \cdot W_i^K, V_i(x) = x \cdot W_i^V \quad (5)$$

where x is the input and S is the number of heads, W_i^Q , W_i^K , and W_i^V are projection matrices for the i -th head, and W^O is the output projection matrix.

In our case, we obtain MHA-LoRA by computing the Q_i , K_i and V_i as follows:

$$Q_i(x) = x \cdot W_i^Q + \sum_{j=1}^T x \cdot B_j^Q \cdot A_j^Q \quad (6)$$

$$K_i(x) = x \cdot W_i^K + \sum_{j=1}^T x \cdot B_j^K \cdot A_j^K \quad (7)$$

$$V_i(x) = x \cdot W_i^V + \sum_{j=1}^T x \cdot B_j^V \cdot A_j^V \quad (8)$$

where T is the number of stages, A_j and B_j are LoRA matrices for stage i . Figure 4 illustrates the difference between the original self-attention layer of the transformer and our MHA-LoRA layer. In our approach, a summation operation is performed on each projected query, key, and value from the different LoRA layers before the attention operation. Notably, during stage T training, all layers are frozen except for the corresponding LoRA layer.

4. Experimental settings

In this section, we begin by presenting the datasets used, including their sources and characteristics. In the sequel, we explore data augmentation techniques aimed at enhancing model performance and describe the preprocessing methods employed to prepare the data optimally. Additionally, we provide comprehensive information on the protocols used for training and evaluation in the implementation details. Finally, we discuss the evaluation metrics used to assess model performance.

4.1. Datasets

We conducted our experiments using the NTU-60 dataset [8] for the pretraining phase and two publicly available rehabilitation exercise datasets for the fine-tuning phase, namely KiMORE [10] and UI-PRMD [11].

The **NTU-60** dataset [8] is a prominent and publicly available resource for researchers and developers working in the field of 3D human action recognition. It encompasses over 60,000 video samples categorized into 60 distinct action classes. The dataset captures a broad spectrum of human actions, ranging from commonplace activities like walking and eating to more intricate tasks like playing golf and using tools. Each video sample incorporates both RGB and depth data, providing complementary information about the visual appearance and 3D structure of the human body. This richness allows algorithms to leverage both spatial and temporal cues for improved recognition accuracy. The dataset incorporates variations in lighting, clothing, and camera viewpoints, mimicking real-world scenarios and posing challenges for action recognition algorithms. This fosters the development of robust and generalizable models.

The **KIMORE** dataset [10] is a valuable resource for research on human motion analysis and rehabilitation. It is a well-curated dataset that has been carefully annotated by medical experts. The KIMORE dataset includes a variety

of low-back pain exercises and has three data inputs: RGB, depth videos, and skeleton positions for 25 joints acquired using a Kinect sensor. It was collected from 78 subjects, including 44 healthy subjects and 34 patients with pain and postural disorder (Parkinson, back-pain, stroke). This dataset also provides a set of clinical features, which are invariant among people and selected on the basis of the scope of the exercise.

The **UI-PRMD** dataset [11] contains human motion data collected from healthy individuals performing ten common rehabilitation exercises targeting different body regions. The dataset includes positions and angles of the body joints in the skeletal models provided by the Vicon and Kinect sensors. For each exercise, ten healthy subjects perform ten repetitions in both a correct and incorrect manner. Each sequence is about 20 seconds, and the number of joints is 25 and 39 for Kinect and Vicon, respectively. The performance scores are generated based on a Gaussian mixture model. A scoring function is defined to map the performance metric values into movement quality scores in the range $[0, 1]$. Since this dataset is collected from healthy individuals, the data may be less representative of the movements of patients with injuries or disabilities.

4.2. Data augmentation and pre-processing

Due to the scarcity of annotated data, there is a lack of rehabilitation exercise datasets. The KIMORE [10] and UI-PRMD datasets [11] are small-scale and suffer from a data imbalance problem, especially pronounced in UI-PRMD, where healthy individuals outnumber unhealthy individuals by a significant margin. Training a model with data augmentation could lead to better performances.

In our experiments, we augment the size and diversity of the datasets by generating new motion sequences from the existing data. We introduce variations in speed by randomly adding or removing L frames, ensuring that L falls within the range of $[0 - 25\%]$ of the sequence length. Feedback from clinicians indicates that adjusting the speed of the original sequence does not compromise the quality score, as individuals may perform actions at varying speeds due to factors such as age and physical condition. Experiments validate this data augmentation configuration, ensuring that the sequence's validity is maintained, as the added or removed frames are non-consecutive.

Lastly, to enhance the dataset's diversity, we employed rotation augmentation that introduces controlled variations in skeleton orientation, simulating different poses and viewpoints. By doing so, we strengthen the dataset's ability to generalize across a wider range of real-world scenarios, improving the model's performance. We also use a balanced data loader during training to ensure that each batch of data contains samples from all classes in equal proportions. This is important to avoid overfitting to majority classes.

For the NTU-60 dataset, we apply the same rotation augmentation as used for the rehabilitation datasets to introduce

controlled variations in skeleton orientation. Additionally, we enhance the dataset through random sequence cropping, where sequences are randomly cropped by a proportion between 0 and 0.5. This method helps simulate variations in sequence length and further increases the diversity of the training data, ensuring that the model is exposed to a wide range of scenarios, thus improving its performance on unseen data.

To ensure consistent origin points across all sequences, we employ a sequence-based normalization technique as described in [2]. Specifically, we subtract the spine coordinates of the first frame from each skeleton in the sequence. This standardization enhances model performance by mitigating potential biases introduced by variations in initial skeleton positions. By using a uniformly set reference frame, the model can more accurately learn the underlying patterns and spatial dependencies of the data without being influenced by irrelevant positional discrepancies. This preprocessing step is applied during both the pretraining and fine-tuning phases, maintaining uniformity and consistency in the training process.

4.3. Implementation details

We performed all experiments using the PyTorch framework on a machine with an Intel i7 4.20 GHz processor and four Tesla T4 graphics cards.

During pretraining, the proposed model is trained on the skeletal data of the NTU-60 dataset, where the normalized 3D joint positions of the skeletons are used as input to the GCN-based embedding layer. An extensive grid search was conducted to select the hyperparameters of the GCNs. Specifically, we set the number of layers to $k = 2$ and use the mean function to aggregate information from adjacent joints at each layer in the two GCNs.

We adopt the AdamW optimizer [43] with a weight decay of 0.05 and betas of (0.9, 0.95). We pretrain the network for 400 epochs for each stage with a per-device batch size of 16. The learning rate is decreased from $1e - 3$ to $5e - 4$ following a cosine decay schedule. The LoRA rank is set to 8 for all stages.

During fine-tuning, following [27], the network is trained on a 80%/20% train/validation ratio, following a cross-validation split scheme. We use the Adam optimizer with an initial learning rate of $1e - 4$ and a batch size of 16, with the number of epochs set to 500. In our testing, a rank of 8 strikes a balance between maintaining the integrity of the information and ensuring computational efficiency. Also, we fix the alpha hyperparameter to 1 for all stages.

4.4. Evaluation metrics

In the context of human movement quality assessment, the Mean Absolute Deviation (*MAD*) is ordinarily used to measure the difference between the ground truth movement

quality scores and the predicted ones:

$$MAD = \frac{1}{n} \sum_{i=1}^n \|y - \bar{y}\| \quad (9)$$

where n is the sample size and y refers to the ground truth movement quality scores, which are the actual performance levels observed during the evaluation of movement quality. These scores are derived from expert annotations [10] or standardized assessment methods [11] that accurately reflect the true quality of the movements being evaluated. On the other hand, \bar{y} corresponds to the predicted movement quality scores generated by our model based on the input data.

MAD serves as a straightforward yet effective measure of model performance. It offers a robust evaluation metric that is less sensitive to outliers compared to other measures such as the mean squared error (MSE).

In addition to the MAD metric, we report Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), which are commonly used in the field of APRE to assess the accuracy of predictive models. In our evaluation, lower values across all metrics indicate better performance.

RMSE, measures the average magnitude of the errors between predicted scores and the ground truth. In particular, it is calculated by taking the square root of the average of the squared differences between predicted and actual values. It provides a single measure of the magnitude of prediction errors, with lower values indicating better accuracy. It is sensitive to large errors due to the squaring operation, making it particularly useful for identifying outliers or extreme deviations in predictions. Mathematically, it can be expressed as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y - \bar{y})^2} \quad (10)$$

MAPE, on the other hand, measures the average absolute percentage difference between predicted and actual values. This metric is particularly useful because it expresses the accuracy of the model's predictions as a percentage, making it easier to interpret and compare across different datasets and contexts. Mathematically, it can be expressed as:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left\| \frac{y - \bar{y}}{y} \right\| \times 100 \quad (11)$$

5. Results

In this section, we comprehensively evaluate our approach by comparing it against state-of-the-art methodologies on the two referenced rehabilitation datasets [10, 11]. We also analyze the computational time of our method across the proposed three stages. Additionally, we conduct several ablation studies to validate our contributions

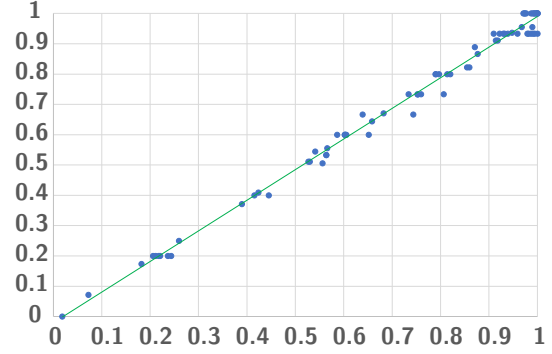


Fig. 5: Comparison of the prediction of the proposed approach \bar{y} against the clinical assessment y for the KIMORE dataset.

5.1. Comparison with state-of-the-art approaches

To ensure a fair assessment of our approach, denoted as *SSL-rehab*, against state-of-the-art deep learning techniques [36, 34, 32, 24, 45, 23, 26, 44, 27, 28], we strictly adhered to the same evaluation criteria and training-test partitioning as specified in [32]. We note that the metric scores provided in our comparison are reported directly from the original papers.

Initially, we present our findings based on the analysis conducted on the ten exercises comprising the UI-PRMD dataset. Subsequently, we provide detailed results for each of the five exercises within the KIMORE dataset. As illustrated in Table 2 and Table 3, our proposed model demonstrates superior performance across multiple evaluation metrics, including MAD , $RMSE$, and $MAPE$. Notably, we achieve the lowest average scores on both the UI-PRMD and KIMORE datasets, with more pronounced improvements observed on the KIMORE dataset.

The superior performance can be attributed to several factors. Firstly, the KIMORE dataset encompasses more complex exercises involving both healthy and unhealthy subjects which calls for more effective training. Additionally, the data collected from the Kinect v2 sensor introduces noise and variability, contrasting with the more precise poses obtained using Vicon in the UI-PRMD dataset. This variability and complexity in the KIMORE dataset pose a greater challenge for developing generalizable models. In contrast, the controlled and straightforward nature of the UI-PRMD dataset allows for various methods to demonstrate relatively comparable performance levels.

To further demonstrate the effectiveness of our method, we visualize the ground truth and predicted movement quality scores for all test sequences in the KIMORE dataset, as shown in Figure 5. As it can be observed, the predictions of our model closely align with the assessments of clinicians, which is a strong indicator that our method can capture the subtle nuances of human movement quality (ideal performance is depicted as the diagonal blue line).

Table 2Results of ten exercises on the UI-PRMD dataset using the evaluation metric *MAD* (**bold** typeface shows best performances)

Exercise	SSL-Rehab	Yao et al [36]	Deb et al [32]	D-STGCN [34]	Song et al [24]	Zhang et al [25]	Liao et al [23]	Li et al [26]	Shahroudi et al [44]	Du et al [27]
1	0.007	0.015	0.009	0.008	0.011	0.022	0.011	0.011	0.018	0.030
2	0.005	0.012	0.006	0.020	0.006	0.008	0.028	0.029	0.044	0.077
3	0.009	0.015	0.013	0.036	0.010	0.016	0.039	0.056	0.081	0.137
4	0.005	0.008	0.006	0.014	0.014	0.016	0.012	0.014	0.024	0.036
5	0.008	0.009	0.008	0.014	0.013	0.008	0.019	0.017	0.032	0.064
6	0.005	0.010	0.006	0.020	0.009	0.008	0.018	0.019	0.034	0.047
7	0.010	0.011	0.011	0.021	0.017	0.021	0.038	0.027	0.049	0.193
8	0.009	0.018	0.016	0.022	0.017	0.025	0.023	0.025	0.051	0.073
9	0.007	0.010	0.008	0.025	0.008	0.027	0.023	0.027	0.043	0.065
10	0.024	0.044	0.031	0.026	0.038	0.066	0.042	0.047	0.077	0.160
Average	0.008	0.015	0.011	0.020	0.014	0.021	0.025	0.027	0.045	0.088

Table 3Results of five exercises on the KIMORE dataset (**bold** typeface shows best performances)

Metric	Exercise	SSL-Rehab	Yao et al [36]	D-STGCN [34]	Deb et al [32]	Song et al [24]	Zhang et al [45]	Liao et al [23]	Yan et al [28]	Li et al [26]	Du et al [27]
MAD	1	0.372	0.444	0.641	0.799	0.977	1.757	1.141	0.889	1.378	1.271
	2	0.291	0.303	0.753	0.774	1.282	3.139	1.528	2.096	1.877	2.199
	3	0.122	0.142	0.210	0.369	1.105	1.737	0.845	0.604	1.452	1.123
	4	0.107	0.121	0.206	0.347	0.415	1.202	0.468	0.842	0.675	0.880
	5	0.280	0.292	0.399	0.621	1.536	1.853	0.847	1.218	1.662	1.864
	Avg	0.234	0.260	0.441	0.582	1.063	1.937	0.965	1.129	1.408	1.467
RMSE	1	0.512	0.569	2.020	2.024	2.165	2.916	2.534	2.017	2.344	2.440
	2	0.354	0.390	1.468	2.120	3.345	4.140	3.738	3.262	2.823	4.297
	3	0.171	0.180	0.487	0.556	1.929	2.615	1.561	0.799	2.004	1.925
	4	0.129	0.148	0.527	0.644	2.018	1.836	0.792	1.331	1.078	1.676
	5	0.363	0.378	0.735	1.181	3.198	2.916	1.914	1.951	2.575	3.158
	Avg	0.305	0.333	1.047	1.305	2.531	2.884	2.108	1.872	2.164	2.699
MAPE	1	1.049	1.105	1.623	1.926	2.605	5.054	2.589	2.339	3.491	3.228
	2	0.742	0.864	0.974	1.272	3.296	10.436	3.976	6.136	5.298	6.001
	3	0.417	0.437	0.613	0.728	2.968	5.774	2.023	1.727	4.188	3.421
	4	0.318	0.341	0.541	0.824	2.152	3.901	2.333	2.325	1.976	2.584
	5	0.752	0.808	1.217	1.591	4.959	6.531	2.312	3.802	5.752	5.620
	Avg	0.655	0.711	0.993	1.268	3.196	6.339	2.647	3.266	4.141	4.170

The results achieved by our method can be attributed to several factors. Firstly, the utilization of decreasing masked motion modeling with SSL enables the model to learn robust 3D skeleton representations from a larger-scale skeletal dataset NTU-60 [8]. Subsequently, transfer learning using LoRA for rehabilitation leverages these features to achieve strong performance. Furthermore, our findings underscore the effectiveness of GCNs and transformers in capturing spatial and temporal features.

5.2. Computational cost

To thoroughly assess performance metrics, computational time, and overall efficiency, we conducted a series of three experiments (Model 1, Model 2 and Model 3). In each, we applied fine-tuning progressively after each stage, enabling us to evaluate the incremental benefits and better understand how each step enhances the model's effectiveness.

We present the training and testing times of the KIMORE dataset, where our model demonstrates efficiency with an average testing time of 22.5 milliseconds per video on a single

Table 4

Computational time of model variants with different pretraining masking strategies.

Model variant	Training type	Training time	Masked motion ratios	MAD	Inference time
Model 1 (MAMP)	Pretraining with Stage 1 masking	1.1 d	90 %	0.472	20.4 ms
	Fine-tuning	1.1 h			
Model 2	Pretraining with Stage 1 and 2 masking	1.8 d	90 %, 75 %	0.341	21.8 ms
	Fine-tuning	1.4 h			
Model 3 (SSL-Rehab)	Pretraining with Stage 1, 2 and 3 masking	2.9 d	90 %, 75 %, 50 %	0.234	22.5 ms
	Fine-tuning	1.6 h			

Tesla T4 GPU. This indicates that our model is capable of providing real-time results, especially considering the availability of real-time skeleton data generation. Additionally, Table 4 shows that incorporating LoRA layers significantly enhances performance without notably increasing inference time.

During pretraining, our strategy of adding LoRA layers only in the second and third stages effectively reduces the model's parameter requirements compared to duplicating the full model size across all stages. In particular, Model 1 in Table 4 has 17.5 million parameters, while Models 2 and 3 increase this by only 2% through the addition of LoRA layers, keeping the model size manageable.

However, training time increases in Models 2 and 3 due to LoRA's need for two forward passes one through the base model and another through the LoRA-specific A and B matrices. Furthermore, training of the first stage is also faster, as the mask size is set to 90%, reducing the input volume to only 10%. In contrast, stages 2 and 3 have mask sizes of 75% and 50%, respectively, leading to a greater input proportion and, consequently, longer training times.

During fine-tuning, our model with 14 million parameters achieves a MAD of 0.234. In contrast, the MAMP approach (Model 1 in Table 4), with 13.2 million parameters, results in a MAD of 0.472. This comparison highlights the significant improvement achieved by our model, despite only a minimal increase in parameter count.

5.3. Ablation studies

We conducted several ablation studies on the challenging KIMORE dataset to examine the specific contributions of individual components within our SSL-Rehab model and identify the optimal configurations for best performance. For these comparisons, we selected the MAD metric to report the performance of the experiments, as it provides a robust measure of accuracy in the context of rehabilitation exercises and is the most commonly used metric in this domain.

5.3.1. Effect of the selected finetuning technique

In the experiment A , we aim to demonstrate the necessity of transfer learning for small-scale rehabilitation datasets. To verify this, we train a transformer architecture [35] from scratch on the KIMORE dataset. Results in Table 5 show

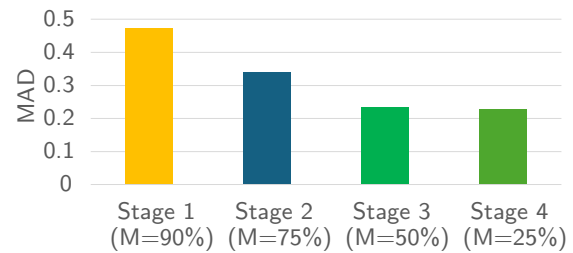


Fig. 6: The proposed method of gradually decreasing the masked motion achieves remarkable MAD improvements in the first three stages on the KIMORE dataset.

that this approach does not yield good performance because starting from scratch lacks the ability to leverage the rich features learned from pre-training on larger datasets (NTU-60). Additionally, training a large model on a small dataset is challenging due to the limited data available for optimization, which leads to suboptimal performance. We did not apply our gradually decreasing motion mask because this approach does not leverage a pre-trained model.

In the second experiment, we conducted a thorough investigation of different fine-tuning strategies through three distinct approaches. Firstly, we employed linear probing (B), which involves freezing all the pre-trained weights and adding a linear layer for regression. This method allows us to evaluate how well the pre-trained features perform when only the final layer is trained to adapt to the specific task. We also

The approach C involved fine-tuning the entire pre-trained model along with the linear layer. We adjusted the parameters of all layers, not just the final one, to allow the model to adapt more fully to the specific task. We systematically tuned the learning rate to optimize the model's performance during the fine-tuning phase, ensuring that the model could leverage the rich features learned during pretraining.

5.3.2. Effect of decreasing masked motion modeling

This experiment is crucial as it validates our proposed approach. We compare Experiment D , which integrates LoRA within the MAMP framework [2] to learn a foundation model exclusively for 90% masked joints, with our proposed

Table 5Ablation studies of the proposed approach on the KIMORE dataset using the *MAD* metric.

Component \ Method	SSL-Rehab	A	B	C	D	E	F	G	H
<i>Training from scratch</i>		✓							
<i>Linear probing</i>			✓			✓			
<i>Full fine-tuning</i>			✓				✓		
<i>MAMP + LoRA fine-tuning</i>					✓				
<i>Gradually increasing the motion mask</i>								✓	
<i>Gradually decreasing the motion mask</i>	✓					✓	✓		✓
<i>GCN-based embedding</i>	✓	✓	✓	✓	✓	✓	✓	✓	
<i>MAD</i>	0.234	0.663	0.534	0.512	0.472	0.380	0.244	0.519	0.241

method, SSL-Rehab. SSL-Rehab involves three stages with varying masking ratios (90%, 75%, and 50%), allowing for a more comprehensive evaluation of the model's adaptability and performance across different levels of motion complexity. Table 5 and Figure 6 illustrate the results, indicating that our approach achieves superior performance compared to the baseline D.

These results can be attributed to the process of gradually decreasing the proportion of masked joints across different stages. This approach allows our model to adapt more effectively to varying levels of motion complexity. In the initial stage, where 90% of joints are masked, the model confronts the most challenging scenarios, enabling it to learn essential features from scratch. Subsequently, as the proportion of masked joints diminishes in later stages, the model leverages the acquired knowledge to tackle less challenging problems. Our method facilitates the capture of a wider range of motion patterns, enriching the learning process. This diversity enhances the model's ability to generalize effectively to unseen data during finetuning.

Furthermore, as shown in Table 5, incorporating the gradually decreasing motion mask into the linear probing *E* and full fine-tuning *F* configurations improved model performance of these approaches (*B* and *C*). However, linear probing did not outperform the LoRA fine-tuning method used in our SSL-Rehab experiment. The LoRA approach preserved adaptability in the intermediate layers, enhancing the model's fine-tuning performance. This contrasts with Linear Probing, where adjustments are limited to the final layer, thereby restricting adaptability.

Moreover, full fine-tuning with a decreasing mask achieved competitive performance, but it requires additional time due to gradient computations across all model parameters. In comparison, the LoRA approach only fine-tunes specific layers, making it significantly more efficient. Additionally, this experiment indicates that fine-tuning the *A* and *B* matrices from the pretraining stages is not very beneficial. Focusing on selective fine-tuning, as done in LoRA, proves more advantageous for efficiently adapting the model to dataset-specific motion patterns.

5.3.3. Gradually increasing vs. decreasing the masked motion

In this experiment, we conduct a comprehensive comparison between two pretraining strategies for our model: one utilizing three stages with increasing masked motion ratios of 50%, 75%, and 90% (*G*), and the other progressively reducing the proportion of masked joints from 90% to 75% to 50% *SSL-Rehab*. The latter approach clearly outperformed the former, showing better results (Table 5).

This superiority can be attributed to several key factors. Firstly, the initial stage, where the model begins training from scratch with 90% of the joints masked, plays a crucial role. Despite its difficulty, this stage enables the model to learn essential features from the ground up, laying a solid foundation for subsequent stages. As the proportion of masked joints decreases in the second and third stages, the task becomes progressively less challenging, allowing the model to refine its representations effectively with the assistance of LoRA layers. In contrast, initializing with a large number of learned weights from the pretrained model to tackle a simpler problem (50% masking) may lead to increased difficulty in subsequent stages. In such cases, the adaptability of LoRA layers may not be sufficient to handle the complexities of the masks.

5.3.4. Effect of choosing the number of stages

In our experiments, we conducted a comprehensive comparative analysis of our model's performance across different training stages and joint masking ratios. Our findings highlight several key insights:

First, we observed diminishing returns beyond the third stage. Specifically, while the initial stages significantly contribute to model improvement, the performance gains taper off after the third stage, suggesting that further stages are not necessary and inefficient (Figure 6). Additionally, we evaluated the impact of various joint masking ratios on model performance. Our results indicate that using a masking ratio of less than 50% does not yield significant performance improvements. These findings underscore the importance of carefully balancing the complexity of the training stages and the degree of joint masking. By optimizing these parameters,

Table 6

Effect of LoRA rank selection on the proposed SSL-Rehab method using the KIMORE dataset.

Rank	MAD	# of parameters
2	0.307	13.2 million
4	0.268	13.3 million
8	0.234	14 million
16	0.252	15.4 million

we can enhance model performance while maintaining efficiency in terms of training and inference time.

5.3.5. Effect of using the GCN-based embedding layer

In the experiment *H*, we compare the proposed GCN-based embedding layer with traditional convolution layers. The results in Table 5 clearly demonstrate that the GCN-based approach enhances performance. This improvement can be attributed to GCNs' superior ability to capture spatial dependencies and intricate patterns within the 3D skeleton data. By leveraging graph-based representations, GCNs excel at modeling relationships between different joints in the skeletal structure, enabling them to extract more nuanced and informative features.

5.3.6. Effect of LoRA rank on performance

To identify the best LoRA rank for our method, we conducted a series of experiments using our SSL-Rehab approach on the Kimore dataset. We systematically evaluated various ranks, specifically 2, 4, 8, and 16, to determine their effects on model performance. The results of these experiments are summarized in Table 6 indicate that a LoRA rank of 8 achieves the highest performance compared to the other configurations. This superior performance at rank 8 can be attributed to its ability to balance expressive power and model complexity. At this rank, the model retains enough capacity to capture complex patterns without the drawbacks of overfitting or excessive computational costs.

6. Conclusions and future work

Our proposed method introduces a comprehensive framework for the robust assessment of physical rehabilitation exercises, leveraging self-supervised learning with 3D skeletal data. Through meticulous experimentation and ablation studies, we have showcased the effectiveness of our approach in gradually decreasing masked motion ratios, thereby facilitating the learning of robust features. This adaptive mechanism enables our model to effectively accommodate varying levels of motion complexity. Furthermore, the incorporation of LoRA layers throughout both stages enhances this process by effectively managing model complexity while maintaining high levels of accuracy and efficiency. Our method offers valuable insights into the importance of modeling motion complexity and leveraging transfer learning for improved generalization of APRE approaches.

In future works, our focus will be on scaling our approach to accommodate larger models and datasets, broadening its applicability and effectiveness. Furthermore, we are committed to exploring the synergy between parameter-efficient fine-tuning and knowledge distillation, with the ultimate goal of optimizing model complexity and reducing inference time.

Acknowledgement

This research was funded by the Région Bretagne and the Conseil départemental du Finistère through the ECFvisuL project, as well as the Institut Carnot Télécom & Société numérique through the FCEval project.

This research benefited from the guidance of Dr. Brice Loddé, Dr. Pierre Balla, and Dr. Thomas Le Rhun, occupational health experts at CHRU Brest, as well as Mrs. Anna Thépaut-Henry, a physiotherapist at Pôle Kiné Plouzané Brest.

License

In accordance with our funding institution's rules regarding open access to results of publicly funded scientific research, the current and all subsequent versions of this article will be published under CC-BY 4.0 license.

References

- [1] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, D. Tao, Tridet: Temporal action detection with relative boundary modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18857–18866.
- [2] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, H. Li, Masked motion predictors are strong 3d action representation learners, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10181–10191.
- [3] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, Q. Miao, Transformer for skeleton-based action recognition: A review of recent advances, *Neurocomputing* (2023).
- [4] M. G. Morshed, T. Sultana, A. Alam, Y.-K. Lee, Human action recognition: A taxonomy-based survey, updates, and opportunities, *Sensors* 23 (2023) 2182.
- [5] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, S. B. i. Badia, Learning to assess the quality of stroke rehabilitation exercises, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 218–228.
- [6] S. Sardari, S. Sharifzadeh, A. Daneshkhan, B. Nakisa, S. W. Loke, V. Palade, M. J. Duncan, Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review, *Computers in Biology and Medicine* (2023) 106835.
- [7] A. Shah, A. Roy, K. Shah, S. Mishra, D. Jacobs, A. Cherian, R. Chellappa, Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18846–18856.
- [8] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.
- [9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022, p. 1.

- [10] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriu, L. Romeo, F. Verdini, The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27 (2019) 1436–1448.
- [11] A. Vakanski, H.-p. Jun, D. Paul, R. Baker, A data set of human body movements for physical rehabilitation exercises, *Data* 3 (2018) 2.
- [12] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems* 25 (2012).
- [13] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: *International Conference on Learning Representations (ICLR)*, 2017, p. 1.
- [14] A. Nogales, M. Rodríguez-Aragón, Á. J. García-Tejedor, A systematic review of the application of deep learning techniques in the physiotherapeutic therapy of musculoskeletal pathologies, *Computers in Biology and Medicine* 172 (2024) 108082.
- [15] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, A. Monteriu, L. Romeo, F. Verdini, A hidden semi-markov model based approach for rehabilitation exercise assessment, *Journal of biomedical informatics* 78 (2018) 1–11.
- [16] J. F.-S. Lin, M. Karg, D. Kulić, Movement primitive segmentation for human motion modeling: A framework for analysis, *IEEE Transactions on Human-Machine Systems* 46 (2016) 325–339.
- [17] A. Vakanski, J. Ferguson, S. Lee, Mathematical modeling and evaluation of human motions in physical therapy using mixture density neural networks, *Journal of physiotherapy & physical rehabilitation* 1 (2016).
- [18] V. Antoniou, C. H. Davos, E. Kapreli, L. Batalik, D. B. Panagiotakos, G. Pepera, Effectiveness of home-based cardiac rehabilitation, using wearable sensors, as a multicomponent, cutting-edge intervention: a systematic review and meta-analysis, *Journal of clinical medicine* 11 (2022) 3772.
- [19] C. E. Lang, J. Barth, C. L. Holleran, J. D. Konrad, M. D. Bland, Implementation of wearable sensing technology for movement: pushing forward into the routine physical rehabilitation care field, *Sensors* 20 (2020) 5744.
- [20] K. Réby, I. Dulau, G. Dubrasquet, M. B. Aimar, Graph transformer for physical rehabilitation evaluation, in: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 2023, pp. 1–8.
- [21] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong, et al., Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review, *Computers in Biology and Medicine* 139 (2021) 104949.
- [22] T. Hamaguchi, T. Saito, M. Suzuki, T. Ishioka, Y. Tomisawa, N. Nakaya, M. Abo, Support vector machine-based classifier for the assessment of finger movement of stroke patients undergoing rehabilitation, *Journal of Medical and Biological Engineering* 40 (2020) 91–100.
- [23] Y. Liao, A. Vakanski, M. Xian, A deep learning framework for assessing physical rehabilitation exercises, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28 (2020) 468–477.
- [24] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Richly activated graph convolutional network for robust skeleton-based action recognition, *IEEE Transactions on Circuits and Systems for Video Technology* 31 (2020) 1915–1925.
- [25] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1112–1121.
- [26] C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, in: *International Joint Conference on Artificial Intelligence*, 2018, p. 1.
- [27] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.
- [28] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: *AAAI Conference on Artificial Intelligence*, 2018, pp. 1–1.
- [29] S. H. Chowdhury, M. Al Amin, A. M. Rahman, M. A. Amin, A. A. Ali, Assessment of rehabilitation exercises from depth sensor data, in: *2021 24th International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2021, pp. 1–7.
- [30] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (1997) 1735–1780.
- [31] M. Chen, Y. Chen, Y. Xu, Q. An, W. Min, Population flow based spatial-temporal eigenvector filtering modeling for exploring effects of health risk factors on covid-19, *Sustainable Cities and Society* 87 (2022) 104256.
- [32] S. Deb, M. F. Islam, S. Rahman, S. Rahman, Graph convolutional networks for assessment of physical rehabilitation exercises, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022) 410–419.
- [33] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, *IEEE transactions on neural networks and learning systems* 28 (2016) 2222–2232.
- [34] Y. Mourchid, R. Slama, D-stgcnt: A dense spatio-temporal graph conv-gru network based on transformer for assessment of patient physical rehabilitation, *Computers in Biology and Medicine* 165 (2023) 107420.
- [35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [36] L. Yao, Q. Lei, H. Zhang, J. Du, S. Gao, A contrastive learning network for performance metric and assessment of physical rehabilitation exercises, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).
- [37] Y. Zhu, H. Han, Z. Yu, G. Liu, Modeling the relative visual tempo for self-supervised skeleton-based action recognition, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13913–13922.
- [38] H. Zhou, Q. Liu, Y. Wang, Learning discriminative representations for skeleton based action recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10608–10617.
- [39] B. Degardin, V. Lopes, H. Proença, Atom: Self-supervised human action recognition using atomic motion representation learning, *Image and Vision Computing* 137 (2023) 104750.
- [40] G. Wang, M. Liu, H. Liu, P. Guo, T. Wang, J. Guo, R. Fan, Augmented skeleton sequences with hypergraph network for self-supervised group activity recognition, *Pattern Recognition* 152 (2024) 110478.
- [41] H. Yan, Y. Liu, Y. Wei, Z. Li, G. Li, L. Lin, Skeletonmae: Graph-based masked autoencoder for skeleton sequence pre-training, *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023) 5583–5595.
- [42] Z. Fu, H. Yang, A. M.-C. So, W. Lam, L. Bing, N. Collier, On the effectiveness of parameter-efficient fine-tuning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 11, 2023, pp. 12799–12807.
- [43] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: *International Conference on Learning Representations*, 2017, p. 1.
- [44] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+d: A large scale dataset for 3d human activity analysis, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.
- [45] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, D.-S. Chen, A comprehensive survey of vision-based human action recognition methods, *Sensors* 19 (2019) 1005.