A Video-Based Augmented Reality System for Human-in-the-Loop Muscle Strength Assessment of Juvenile Dermatomyositis

Kanglei Zhou, Ruizhi Cai, Yue Ma, Qingqing Tan, Xinning Wang, Jianguo Li,
Hubert P. H. Shum, *Senior Member, IEEE*, Frederick W. B. Li, Song Jin, and Xiaohui Liang ✉

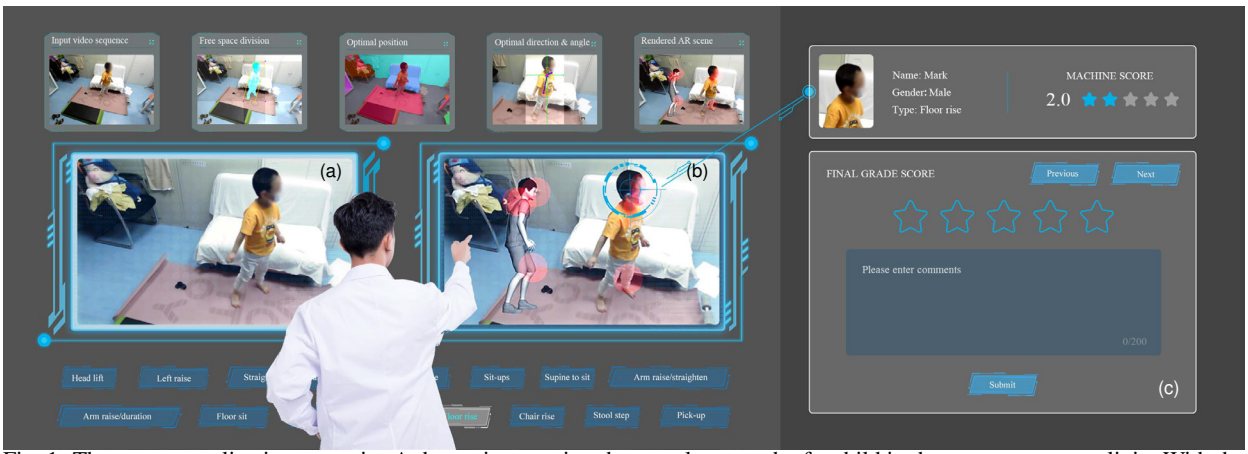

Fig. 1: The system application scenario: A doctor is assessing the muscle strength of a child in the pre-assessment clinic. With the guidance of a demo, the child is asked to complete the specified CMAS action. For a side-by-side comparison, the doctor can see (a) the real-world scene and (b) the augmented scene. Also, the machine score is provided for the doctor to (c) make a final decision.

Abstract— As the most common idiopathic inflammatory myopathy in children, juvenile dermatomyositis (JDM) is characterized by skin rashes and muscle weakness. The childhood myositis assessment scale (CMAS) is commonly used to measure the degree of muscle involvement for diagnosis or rehabilitation monitoring. On the one hand, human diagnosis is not scalable and may be subject to personal bias. On the other hand, automatic action quality assessment (AQA) algorithms cannot guarantee 100% accuracy, making them not suitable for biomedical applications. As a solution, we propose a video-based augmented reality system for human-in-the-loop muscle strength assessment of children with JDM. We first propose an AQA algorithm for muscle strength assessment of JDM using contrastive regression trained by a JDM dataset. Our core insight is to visualize the AQA results as a virtual character facilitated by a 3D animation dataset, so that users can compare the real-world patient and the virtual character to understand and verify the AQA results. To allow effective comparisons, we propose a video-based augmented reality system. Given a feed, we adapt computer vision algorithms for scene understanding, evaluate the optimal way of augmenting the virtual character into the scene, and highlight important parts for effective human verification. The experimental results confirm the effectiveness of our AQA algorithm, and the results of the user study demonstrate that humans can more accurately and quickly assess the muscle strength of children using our system.

Index Terms—Action Quality Assessment, Augmented Reality, Human-in-the-Loop System, Juvenile Dermatomyositis

1 Introduction

Juvenile dermatomyositis (JDM) is the most common idiopathic inflammatory myopathy in children, characterized by skin rashes and muscle weakness [3, 22, 33]. Fig. 2 shows two examples of a typical skin rash and a muscle weakness motion sequence of a child. There are approximately two to four children per million affected by JDM [16, 24, 31]. Due to the acuteness of JDM and the significant harm, early diagnosis and timely treatment are crucial to improving the outcome [13, 25].

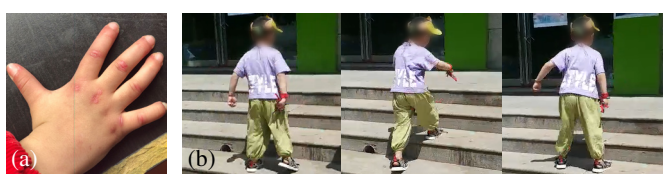

Fig. 2: Typical symptoms of JDM: (a) skin rashes on the hand, (b) difficulty climbing stairs due to muscle weakness.

However, the lack of pediatric specialists with expertise in diagnosing JDM makes missed diagnoses and misdiagnoses extremely common, particularly in developing countries. For example, in 2014, China had 0.43 pediatric specialists per 1,000 children, which was significantly lower than the US at 1.46 [47].

Assessing the degree of muscle involvement in JDM, also known as muscle strength assessment, is a key topic in JDM diagnosis. As shown in Fig. 2(b), a child with JDM staggers while climbing stairs due to muscle weakness. Testing for muscle strength and endurance is one of the primary measures for JDM diagnosis and rehabilitation monitoring. The mainstay of most clinical evaluations in children with JDM is the childhood myositis assessment scale (CMAS), a measure

- *X. Liang is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China, and also with Zhongguancun Laboratory, Beijing, China. E-mail: liang_xiaohui@buaa.edu.cn.*
- *K. Zhou, R. Cai, and Y. Ma are with Beihang University, Beijing, China.*
- *Q. Tan, X. Wang, and J. Li are with the Children's Hospital of Capital Institute of Pediatrics, Beijing, China.*
- *H. Shum and F. Li are with Durham University, Durham, United Kingdom.*
- *J. Song is with Beijing Diannaite Medical Technology Co., Ltd., China.*

Manuscript received 14 October 2022; revised 13 January 2023; accepted 30 January 2023. Date of publication 22 February 2023; date of current version 29 March 2023. Digital Object Identifier no. 10.1109/TVCG.2023.3247092

Fig. 3: Examples of CMAS actions and the corresponding ones from the motion dataset: (a) sit-ups, (b) supine to sit, (c) pick-up.

that incorporates function as well as strength [20]. Usually, children with normal muscle strength achieve a full score of 52 through an independent evaluation of 14 kinds of actions. In contrast, children with a score of less than 52 are considered to have muscle weakness. On the one hand, it is difficult for young children to complete these actions assessed by the CMAS; on the other hand, it is hard for pediatricians to determine accurate scores due to the highly subjective assessment criteria. For example, for the 'supine to sit' action, there is no explicit distinction as to whether the performance is 'very difficult' or 'generally difficult'. Hence, it is necessary to develop a decision support system to provide a reliable and timely economical diagnosis for JDM.

Action quality assessment (AQA) has achieved great success in various fields, such as sports analysis [27] and surgical skill assessment [19]. It aims to develop a system capable of evaluating some specific actions automatically and objectively through a series of input videos. Thus, it can be considered an alternative method of avoiding the influence of personal judgment biases. However, the biomedical community has concerns about fully automated AQA for JDM analysis because it cannot guarantee 100% accuracy, similar to any machine learning algorithm. Also, there is a lack of interface for human experts to understand and be supported by automatic AQA algorithms. This motivates us to research a solution that allows humans to effectively understand and verify AQA results, so as to improve diagnosis effectiveness.

In this work, we propose a human-in-the-loop AQA framework for muscle strength assessment of JDM, which enables effective human verification of AQA-suggested results, thereby allowing automated AQA to support human decisions. Our novel idea is to visualize a virtual character that represents the AQA results, such that users can effectively compare the movements of the real-world patient with the virtual character as a cue for decision-making. Specifically, we first collect a JDM dataset with over 1,000 video clips and construct a 3D animation dataset comprising all the sub-categories of CMAS actions. Next, the muscle strength assessment network trained on the JDM dataset is used to assess a given sample. By using the results of machine scoring, a virtual animation is retrieved from the dataset. Fig. 3 shows three examples of real actions and the corresponding animations.

To allow the most effective comparisons between the real-world patient and the virtual character, we propose a video-based augmented reality system. There is a few virtual/augmented reality (VR/AR) research that shows the potential of usage in an interactive visualization. For example, Robles *et al.* [35] proposed a VR system to support the screening of autism spectrum disorders. Pears *et al.* [30] proposed several innovative AR solutions to deliver medical education. Ours is distinctive in the sense that we focus on the optimal way of visualization for comparisons. In particular, given a feed, we adapt semantic segmentation and pose estimation algorithms to identify the posture of the real-world patient and the objects in the scene. We then identify the optimal position, based on an objective function, to augment the virtual character into the scene. According to the pose of the real-world patient, we also calculate the viewing angle and size of the virtual character, such that the two would be the best aligned for effective comparison. For effective verification, we highlight the important key points of both patients and characters by utilizing the network layer heatmaps.

The experimental results confirm the effectiveness of the proposed AQA algorithm, and the results of the user study demonstrate that our system can assist non-specialists in the more accurate and faster analysis of JDM. The system can be used by the expert to enhance their diagnosis effectiveness. It can also be used by trained non-specialists such as nurses so that they can help prioritize potential serious patients and refer them to experts for diagnosis and treatment.

Our main contributions are:

- We propose a novel framework of human-in-the-loop JDM analysis, which allows humans to effectively verify the result suggested by machine scoring through the visual comparison of real-world patients and virtual characters.
- We propose a video-based augmented reality visualization system that facilitates effective comparisons by adapting computer vision algorithms for scene analysis, and evaluate the optimal way of augmenting the virtual character into the scene.
- We propose a new, large-scale dataset for JDM and an AQA system for JDM analysis. Unlike existing AQA algorithms, ours is designed to facilitate human understanding through the generation of visualization cues.

2 Related Work

VR and AR applications in healthcare, as well as AQA, are closely related to this research. Hence, this section reviews both areas.

2.1 VR/AR in Healthcare

The use of VR and AR technologies in healthcare has been increasing due to the capability they possess to enable a wide range of delivering healthcare applications [9, 15, 23, 32, 34, 36, 40, 43, 44, 49], including the training of physicians and other healthcare professionals, as well as enhancing their ability to provide remote diagnosis services.

In the field of medical training and education [4, 9, 14, 23, 32, 34, 49], VR/AR plays a key role in the interactive visualization. On the one hand, VR/AR technology provides users with vivid visualizations that deepen their understanding; on the other hand, it also provides users with interactive methods, which contribute to the sense of participation. For example, Pears *et al.* [30] proposed several innovative solutions to deliver medical education while maintaining resident and educator safety. Mobile AR allows remote users to view a surgical procedure and interact with the video feed, allowing students to gain a comprehensive understanding of surgery without having to be present in a classroom. According to the results in [9], AR technology can enhance the experiences of medical students by improving knowledge and understanding as well as practical skills, and by facilitating social interaction.

In the field of medical diagnosis [8, 15, 35, 36, 40, 43, 44], VR/AR plays a key role in the useful tool for the human-centered data acquisition. Researchers often build VR/AR scenes to collect human motion data to analyze and diagnose diseases. For example, Wang *et al.* [40] utilized a VR helmet and force feedback gloves to present users with a playful experience for home rehabilitation. This study explores the potential of fully immersing them in a playful experience within a virtual cat bathing simulation. The results demonstrate that playfulness brings a positive impact on the rehabilitation experience in VR. Therefore, designing VR/AR games [8, 44] for disease diagnosis is appealing. However, long-term motion interactions in a VR/AR environment can bring users cybersickness. Some studies [36, 43] to adopt mitigation measures to avoid discomfort by predicting cybersickness.

Different from these medical systems, our system enables non-specialists to make easy-to-judge assessments by comparing the patient's movement with the virtual animation produced by AQA, thereby allowing AQA to support human decisions. By side-by-side comparison, they can easily observe the difference between them.

2.2 Action Quality Assessment

AQA aims to quantify how well actions are performed from the same class, which can be used as an alternative to avoid personal judgment bias [17]. Recently, AQA has gained widespread attention due to its wide applications such as rehabilitation medicine [1, 26], athletic competition [37, 48], and specific skills assessment [10, 11].

Different classification criteria can be used to categorize these AQA methods. As for the input format, these frameworks include exemplar-based [2, 46, 48] and exemplar-free [37, 42] methods depending on whether exemplars are used or not. The former usually involves selecting a set of exemplars along with the target sample as input, while the latter does not. For example, Bai *et al.* [2] learned the relative score of the input action relative to the exemplar, and used the exemplar's

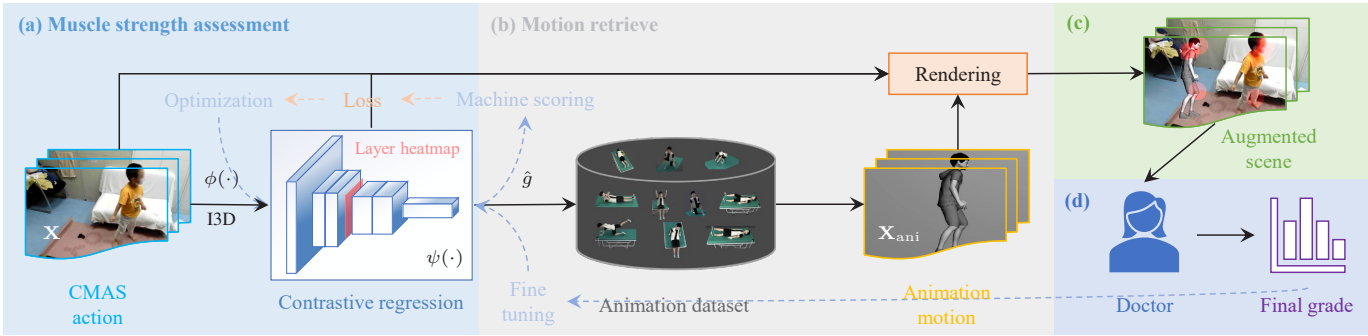

Fig. 4: Pipeline of the proposed system: (a) muscle strength assessment using contrastive regression (Sect. 3.2), (b) motion retrieve from the animation dataset, (c) AR effects rendering (Sect. 3.3), and (d) assisted assessment.

score to calculate the final quality score. Compared to exemplar-free methods, this method is easier to regress relative scores. As for data output, existing AQA methods can be divided into methods based on quality scores, methods based on grades, and methods based on rank orders. For example, Parmar *et al.* [27] classify the levels of cerebral palsy rehabilitation exercises as 'good' or 'bad'. Since the scores of individual actions in CMAS are determined by different execution levels, our method falls into the grade-based method. As for processing flow, AQA generally entails three steps: feature extraction, feature aggregation, and score regression. To avoid over-fitting caused by few samples, most existing methods adopt powerful backbones such as C3D [38] and I3D [5] as the feature extractor, which are usually pre-trained on large action recognition datasets. Common feature aggregation and regression methods [28, 29] include LSTM, TCN, MLP, *etc.*

Different from prior works, we propose a human-in-the-loop AQA framework for muscle strength assessment based on contrastive regression. The AQA-suggested results are finally used to generate augmented visualization cues for effective human verification.

3 Methodology

Inspired by the fact that direct assessment requires more expertise than spotting differences, we design an AR system that can assist non-specialists such as clinical nurses in assessing the muscle strength of children with JDM. For example, it is possible for clinical nurses to use our system to perform pre-assessment of patients. So they can help prioritize potential serious patients and refer them to experts for further diagnosis and treatment, which will reduce the workload of experts from spending their time examining negative or non-urgent cases. As shown in Fig. 1, a user can easily compare the difference between the real patient and the virtual character through our system. Similar to any machine learning algorithm, it cannot guarantee 100% accuracy for existing AQA methods while we use it in biomedical applications that require a lot of responsibility. To achieve this, a novel human-in-the-loop AQA framework is embedded in our system, enabling effective human verification of the AQA-suggested results.

In this section, we first provide an overview of the whole system. Then, we detail the proposed AQA algorithm using contrastive regression. Finally, we describe the setting of the user study.

3.1 System Design

As can be seen in Fig. 4, there are four components in our pipeline: (a) muscle strength assessment using contrastive regression (Sect. 3.2), (b) motion retrieve from the animation dataset, (c) AR effects rendering (Sect. 3.3), and (d) assisted assessment. The muscle strength assessment algorithm produces results for searching virtual motions from the animation dataset. The use of video-based AR effects to augment real-world scenes can provide users with valuable cues for assessment.

Given an unobserved sample, our AQA algorithm outputs a predicted grade $\hat{g}$, which is used to search a corresponding virtual motion from the constructed 3D animation dataset. To superimpose the virtual animation in the real-world scene, we adapt computer vision algorithms for scene understanding and evaluate the optimal way of augmenting the virtual character in the scene. Thus, users can compare the real-world patient and the character as a means to understand and verify the AQA results. When there is a large difference between the real movement and the virtual animation, it is thus easy for them to detect the difference, and then re-evaluate this action by carefully reviewing the assessment criteria. Notably, even people without prior experience in muscle strength assessment can make an accurate diagnosis by comparing the machine results facilitated by the virtual character to the real patient.

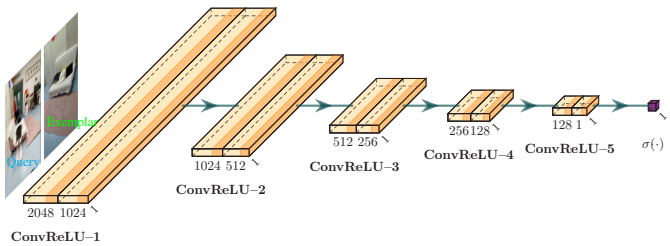

Fig. 5: Network architecture of the contrastive regression module.

3.2 Muscle Strength Assessment

Our AQA algorithm aims to automatically assess the muscle strength of the given action. Given an action sequence $\mathbf{X} \in \mathbb{R}^{T\times W\times H\times 3}$ composed of T frames of size $W \times H$, the output is the predicted grade $\hat{g}$, which is supervised by its ground-truth g. Thus, we need to first obtain the video-level representation and then use it for grading.

3.2.1 Video-Level Representation Extraction

Since JDM is a rare disease, the collected JDM dataset (Sect. 4.1) is with fewer than 100 samples for each action. It is easily over-fitting when training large-sized models. Previous works [2, 45, 48] have hypothesized that the action recognition features can also be applied to the AQA task. So we choose I3D [5] pre-trained on the Kinetcis-400 dataset [18] as the feature extractor.

Considering that 3D CNNs [5, 38] are always memory- and computation-intensive, we first divide the whole video sequence $\mathbf{X}$ into M small clips $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M \in \mathbb{R}^{T'\times W\times H\times 3}$ with equal length T'. Then, these clips are fed to the feature extractor $\phi(\cdot)$ to obtain the corresponding clip-level features $\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_M \in \mathbb{R}^C$. Next, we can obtain the video-level representation $\boldsymbol{f} \in \mathbb{R}^C$ by average pooling, where C is the dimension size.

$$\boldsymbol{f} = \mathrm{AvgPool}([\phi(\mathbf{X}_1), \phi(\mathbf{X}_2), \cdots, \phi(\mathbf{X}_M)]). \tag{1}$$

3.2.2 Contrastive Regression

Motivated by the fact that directly assessing action quality is more difficult than comparing a sample with exemplars, we propose the contrastive regression module for muscle strength assessment. Fig. 5 illustrates the network architecture of the proposed contrastive regression module. A better representation space is developed by ensuring that the distance between two similar samples is small, while the distance between two dissimilar samples is large. The distance between two samples in the representation space can, therefore, already reflect their semantic relationship, if they belong to the same category.

Our goal is to regress the difference/similarity between the input action and exemplars. A set of representative samples from the training

set is selected by a pediatric specialist for inference. The exemplars are randomly selected during training, which enhances the robustness of our system against any mistake by the pediatric specialist in selecting the representative samples. For an action with G types of grades, given an input sample $\mathbf{X}$ and the corresponding exemplars $\mathbf{X}^i_{\text{emp}}$ ($i = 1, 2, \cdots, G$), we can obtain the video-level representations $\boldsymbol{f}$ and $\boldsymbol{f}^i_{\text{emp}}$ through Eq. (1). Next, $\boldsymbol{f}$ and $\boldsymbol{f}^i_{\text{emp}}$ are concatenated together and fed into our contrastive regression module $\psi(\cdot)$ to obtain the similarity:

$$\hat{s}_i = \sigma(\psi(\text{concat}(\boldsymbol{f}, \boldsymbol{f}^i_{\text{emp}}))), \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. A similarity score indicates how likely the sample is to belong to the corresponding grade. The final grade is determined by the highest similarity score:

$$\hat{g} = \arg\max_i(\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_G). \tag{3}$$

If the training set contains more than one set of exemplars, we can use an ensemble strategy to determine the final grade by selecting multiple sets of samples. The pseudo-code for inference of our algorithm is elaborated on Algorithm 1, where E sets of samples are selected.

Algorithm 1: Inference procedure of our AQA algorithm for JDM analysis.

Input: A given action $\mathbf{X}$, the training set $\mathcal{T}$, the ensembles E and grades G.
Output: The corresponding predicted grade $\hat{g}$.

```
Obtain the video-level representation f by Eq. (1);
Select E sets of representative samples {X^{1,1}_emp, X^{1,2}_emp, ..., X^{1,E}_emp, ...,
  X^{E,G}_emp} from the training set T;
Initialize the vote variable ŝ ∈ R^G to the zero vector;
for i ← 1, 2, ..., E do
    Initialize the temporary variable s_tmp ∈ R^G to zeros;
    for j ← 1, 2, ..., G do
        Obtain the video-level representation f^{i,j}_emp by Eq. (1);
        Calculate s^i_tmp based on f^{i,j}_emp by Eq. (2);
    Calculate the temporary variable ĝ_tmp based on s_tmp by Eq. (3);
    Add a vote for the ĝ_tmp-the element of the vote variable ŝ;
Calculate the final grade ĝ based on ŝ by Eq. (3);
```

3.2.3 Optimization

Existing AQA methods [2,48] ignore the intrinsic heterogeneity among the feature spaces of action recognition and AQA, *i.e.*, the model pre-trained on action recognition datasets can be sub-optimal for AQA. To compensate for this, we first design feature distance loss to regularize the heterogeneous feature space:

$$\mathcal{L}_{\text{dis}} = -\left(d_{ij} \ln \hat{d}_{ij} + (1 - d_{ij}) \ln(1 - \hat{d}_{ij})\right). \tag{4}$$

where we use the cosine similarity to measure the distance $\hat{d}_{ij}$ between $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$. When $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ belong to the same grade, the ground-truth distance d_{ij} is set to 1; otherwise, the ground-truth is set to 0.

We train the network by imposing the classification task and use the cross-entropy function to define the score regression loss:

$$\mathcal{L}_{\text{sco}} = -\sum_{i=1}^{G} (g_i \ln \hat{g}_i + (1 - g_i) \ln(1 - \hat{g}_i)). \tag{5}$$

Finally, the overall loss $\mathcal{L}$ can be obtained by:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{dis}} + \lambda_2 \mathcal{L}_{\text{sco}}, \tag{6}$$

where λ_1 and λ_2 are factors that balance the scale of each loss term, distinguishing the importance of two objectives and ensuring that the feature extractor and the score regression converge synchronously.

3.3 AR Effects Rendering

We augment the character in the optimal position and orientation so that the real-world patient and the virtual character are compared as a means of understanding and verifying the AQA results.

3.3.1 Position

As can be seen in Fig. 1, there are different objects and the patient in the real-world scene. Due to the visual comparison of human movements, we are much concerned with the location and orientation of the patient. In this way, it is necessary to find where the patient is located so that we can put the character in a suitable free place.

To achieve this, we utilize a video instance tracking and segmentation algorithm [41] to track the patient during the assessment. Thus, we can obtain a mask sequence of the patient. Fig. 6(a) shows the result of tracking and segmentation where a mask is in cyan for a frame of the patient and the bounding box is in yellow. Let $P_i = (x_i, y_i)$ for $i \in [1, 4]$ denote the four vertices of the rectangle bounding box, we can divide the whole space into four free sub-spaces.

We first calculate four outer vertices of the bounding box by:

$$\begin{aligned} P'_1 &= (x_{\min}, y_{\min}), P'_2 = (x_{\min}, y_{\max}), \\ P'_3 &= (x_{\min}, y_{\min}), P'_4 = (x_{\max}, y_{\min}), \end{aligned} \tag{7}$$

where $x_{\max}$, $x_{\min}$, $y_{\max}$, $y_{\min}$ denote the maximum as well as minimum x- and y-axis coordinate values of P_i. Then, there are four adjacent combinations of P'_i to form four corresponding lines $P'_1P'_2$, $P'_2P'_3$, $P'_3P'_4$, $P'_4P'_1$. Naturally, they split the free space into four parts: left, bottom, right, and top, denoted as $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$, and $\mathcal{A}_4$. Fig. 6(a) shows that there are four corners that cover two adjacent sub-spaces. Particularly, all the corner areas will not be considered prior to the placement of the character for the best visualization effect.

Generally, the larger the area of the free sub-spaces, the more likely it is to be selected for positioning the character. At the same time, the expected sub-space should have the capacity to accommodate the outer rectangle $P'_1P'_2P'_3P'_4$ of the bounding box, so that it satisfies a proportional insertion. The problem can be modeled as:

$$\begin{aligned} \min_i \quad & -h_i \times w_i, \ \forall i \in [1, 4] \\ \text{s.t.} \quad & h_i > y_{\max} - y_{\min}, \\ & w_i > x_{\max} - x_{\min}, \end{aligned} \tag{8}$$

where h_i and w_i denote the vertical length and the horizontal length of the i-th sub-space respectively. In case there is no solution to Eq. (8) (*i.e.*, none of the four sub-spaces can positively contain the rectangle $P'_1P'_2P'_3P'_4$), we remove the constraints, find the optimal solution again, and scale the virtual character accordingly.

Generally, the character will be placed at the center of the optimal sub-space. However, the character in the area center may sometimes be suspended in the air, which is unfriendly for visual comparison. As shown in Fig. 6(b), we thus use a semantic segmentation algorithm for scene understanding so that the virtual character can be placed on the ground or the sleeper sofa, thus enhancing the visualization experience. Therefore, if we can find a suitable plane in the optimal sub-space, we can place the character there. If not, we will place it in the center.

3.3.2 Direction and Angle

Notably, our animation dataset is 3D in nature so we can render it from all possible angles. To ensure the same orientation as the real-world patient, it is necessary to identify the face angle and direction.

We determine the front-to-back and side-to-side orientation of the patient's face separately. Firstly, the required coordinates as shown in Fig. 6(c), *i.e.*, the hip P_{H}, the neck P_{N}, and the left shoulder P_{L}, can be easily obtained by a pose estimation method [21]. Note that the neck is obtained by the median interpolation between two shoulders. The two directed lines $\overrightarrow{P_{\text{H}}P_{\text{N}}}$ and $\overrightarrow{P_{\text{N}}P_{\text{L}}}$ form two vectors, denoted as $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Thus, we can then determine the front-to-back orientation by calculating the normal direction $\boldsymbol{n}$ between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

$$\boldsymbol{n} = \boldsymbol{v}_1 \times \boldsymbol{v}_2. \tag{9}$$

Since the three points are located at the core of the human body and are relatively stable, we can utilize the right-hand rule to determine the real patient's direction. If the direction of $\boldsymbol{n}$ is perpendicular to the

Fig. 6: Illustrations of (a) sub-space division, (b) scene segmentation, (c) key points and local coordinate system, and (d) the rendered AR effect.

image plane inwards, the patient faces forward; otherwise, the patient faces backward. Using the depth information provided by the pose estimation, a rough 3D angle can be calculated.

As shown in Fig. 6(c), we establish a local coordinate system along the x and y axes with the neck P_{N} as the origin. Intuitively, if $\boldsymbol{v}_2$ falls into the first and second quadrants, the patient faces to the right of the camera; otherwise, the patient faces to the left. Furthermore, if $\boldsymbol{v}_2$ falls into the first and fourth quadrants, the patient faces backward of the camera; otherwise, the patient faces forward.

3.3.3 Attention Effects

Since different body parts contribute differently to AQA, identifying their importance is crucial. However, it is impossible to pre-define key parts for each action because key parts vary by patients' symptoms. Motivated by the success of attention mechanisms in deep learning, we can obtain adaptive key parts for different patients and characters by using intermediate layer heatmaps of our network, as shown in Fig. 4.

Firstly, we need to determine which parts are most important. It is easy to obtain a feature map of the intermediate network layer, and resize it to match the input size, denoted as $\mathbf{F} \in \mathbb{R}^{T\times W\times H}$. Also, we can obtain a mask sequence as same as Sect. 3.3.1, denoted as $\mathbf{M} \in \mathbb{R}^{T\times W\times H}$. The attention map is calculated by:

$$\mathbf{A} = \mathrm{softmax}(\mathbf{F} \odot \mathbf{M}), \tag{10}$$

where the function $\mathrm{softmax}(\cdot)$ normalizes the response of the human body to $[0, 1]$. Responses are greater for important parts.

Secondly, we need to determine which of these important parts correspond to the key points. The same as Sect. 3.3.2, we can obtain a sequence of poses of the patient containing J key points. For each frame, the importance α_i of i-th key point is obtained by summing the responses in its neighborhood $\mathcal{N}_i$ with the radius r: $\alpha_i = \sum_{j\in\mathcal{N}_i} A_{ij}$. Next, we calculate the i-th joint angle difference ω_i between the patient (ω_1^i) and the character (ω_2^i) by $|\omega_1^i - \omega_2^i|$. Then, a highlighted key point is determined if the response and the joint angle difference values exceed the average values: $\alpha_i > \frac{1}{J}\sum_j \alpha_j$ and $\omega_i > \frac{1}{J}\sum_j \omega_j$.

Fig. 6(d) shows the final rendered scene where the knees of the patient and the character are highlighted in red circle shading. It indicates that the knee joint is important during floor rising, and that the knee joint angle of the character and the patient differs significantly.

3.4 User Study Design

The purpose of this study is twofold: the first is to assess the gap between machine scoring and expert scoring; the second is to determine whether the proposed system can assist non-specialists in assessing muscle strength more accurately for children with JDM. The former can be verified by comparing the ground truth with the machine scoring, as demonstrated in Sect. 3.2. The user study aims to explore how the latter can be verified. We need to investigate two questions:

Q1. How *fast* are the participants when assessing a given sample?
Q2. How *accurate* are the assessment results of participants?

To answer these questions, we have designed two tasks (Sect. 3.4.1). Participants are asked to solve the two tasks (Sect. 3.4.2). For both tasks, we have recorded results, time usage, and some personal information. In addition, we describe the detailed study procedure (Sect. 3.4.3).

3.4.1 Tasks

Two tasks are designed to evaluate the effectiveness of the system:

T1. CMAS Assessment Participants are required to assess the muscle strength of patients according to CMAS rules.

T2. Assisted Assessment Participants are required to make a final decision based on the machine score and the augmented scene.

3.4.2 Apparatus and Participants

The user study has been conducted at the Children's Hospital of Capital Institute of Pediatrics in Beijing, China.

The 3D animations are made by a Unity3D program. The system includes the patient, server, and doctor ends. The patient end is primarily responsible for collecting data, which is then sent to the server for processing and calculation. Based on the results calculated on the server, the doctor end renders the AR scene to assist doctors in assessment. Using our system, it is possible to complete the assessment at home via remote diagnosis. Our system is currently in a prototype stage and is primarily used for pre-assessment in clinics. After simple training for the patient, the doctor opens the patient end to record actions. At the doctor end, the rendered AR scenes can be seen for assessment within a few seconds of delay. Data and results are automatically saved.

Instead of pediatric specialists, our system is primarily designed to assist clinic doctors in pre-assessment. A total of 30 non-paid general practitioners participate in the study to ensure its repeatability, and each participant needs to assess all actions for each patient. Half of the participants are males and half are females. Ages range from 25 to 50, and the mean, median, and standard deviation are 34.1, 32, and 7.8, respectively. No vision issues are reported by participants. The assessment has been conducted on five representative patients selected by a pediatric specialist. One girl and four boys are aged 8, 5, 5, 5, and 15, with one mild case, two moderate cases, and two severe cases. In all actions they perform, all scores are covered, indicating that the symptom of representative patients has a high degree of diversity.

3.4.3 Procedure

Due to the COVID-19 epidemic, we recorded AR effects on videos and created an online form. The action can therefore be prevented from being repeated by the child. During the user study, participants are invited for a remote assessment via the Internet. Each participant must sign a consent form before the study procedure can begin.

The study procedure begins with an explanation of the study design. Questions could be posed at any time. After the explanation, participants can assess the training examples for each task. Then, the actual study formerly starts. Participants can assess the muscle strength of patients by watching their actions. For task T2, participants can compare the real action and the virtual animation. They should make a final decision. After finishing each sub-task, the answer and the time usage are automatically recorded. There are 30 (6×5) sub-tasks, comprising 6 actions per child (5 children). Among the two tasks, the action order is different to eliminate possible negative influences. After completing each sub-task, participants can click the 'pause' button to take a break. The break time cannot be recorded until they click the 'start' button to the next. Participants fill out a questionnaire after solving all tasks regarding their age and experience with the application. The study takes approximately half an hour.

4 Experiments

We first introduce the experimental setup, including datasets, evaluation metrics, and implementation details. And then we show lots of

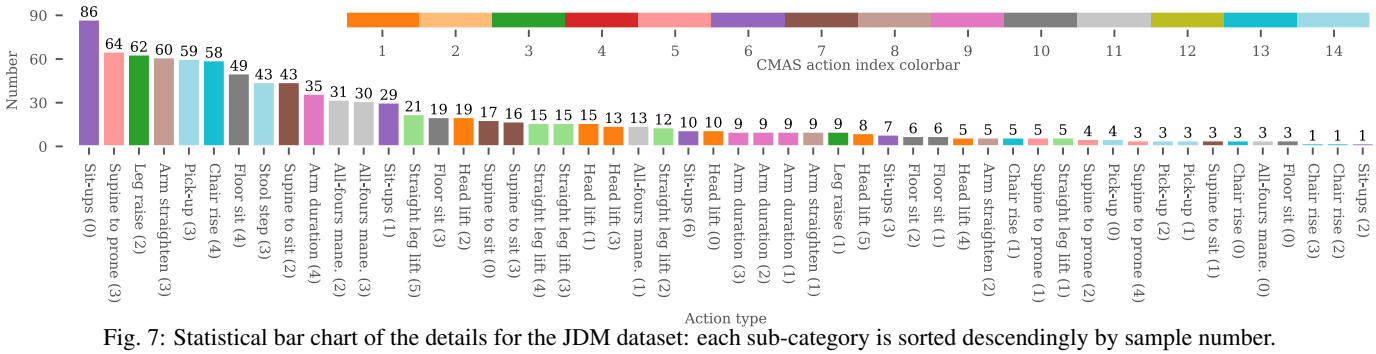

Fig. 7: Statistical bar chart of the details for the JDM dataset: each sub-category is sorted descendingly by sample number.

qualitative and quantitative experiments and analyze the results.

4.1 Datasets

The details of the collected dataset are shown in Fig. 7. Different colors represent different actions, and each sub-category is sorted in descending order by the sample number.

In total, 14 types of actions are included in the dataset. Based on different evaluation criteria, these actions can be divided into three categories: time-related actions, position-related actions, and uncertain actions. A time-related action is evaluated primarily on the basis of how long it takes to complete the action. A position-related action is primarily evaluated in terms of the magnitude and position of movement of a body part. Uncertain actions require subjective judgments about how difficult they are to perform. The uncertain action is more difficult to evaluate, while the others can be easily measured by simple distance metrics. Among them, we do not consider actions that miss some sub-categories and we prefer to select actions that require subjective judgments. As a result, six challenging actions have been selected to conduct this study. To boost data collection efficiency and increase the number of samples, we use an iPad, an iPhone, and two different USB cameras to collect the data. The video resolution contains (1080, 1920), (1080, 1912), (1920, 1080), and (1080, 1914), and the frame rate contains 30 fps and 29 fps. Furthermore, the duration of different actions varies, even for the same type of action, leading to differences in the total number of frames. To ensure a unified input, all samples have been normalized to the same resolution and frame number.

4.2 Evaluation Metrics

In addition to accuracy, we also use two evaluation metrics to validate the performance of the proposed AQA algorithm.

Spearman's Rank Correlation Coefficient Similar to previous works [11,17,37], we adopt the Spearman's rank correlation coefficient ρ to evaluate the algorithm performance. The coefficient ρ measures the correlation between two statistical variables, which is defined as:

$$\rho = \frac{\sum_n (g_n - g_{\text{avg}})(\hat{g}_n - \hat{g}_{\text{avg}})}{\sqrt{\sum_n (g_n - g_{\text{avg}})^2 \sum_n (\hat{g}_n - \hat{g}_{\text{avg}})^2}}, \tag{11}$$

where g_{avg} and $\hat{g}_{\text{avg}}$ denote the average values of the ground truth and predicted score vectors, respectively. The rank correlation between ground truth and predicted scores increases as ρ increases.

Relative ℓ_2 Distance We also adopt the relative ℓ_2 distance (R-ℓ_2) [48] to measure the algorithm performance. It eliminates the negative impact of different actions spanning different scoring intervals. Given the highest and lowest scores $g_{\max}$ and $g_{\min}$, R-ℓ_2 is defined as:

$$\text{R-}\ell_2 = \frac{1}{N} \sum_n^N \left(\frac{|g_n - \hat{g}_n|}{g_{\max} - g_{\min}} \right)^2 \times 100, \tag{12}$$

where N is the sample number. Fisher's z-value is used to measure the average performance across actions.

Table 1: Comparison results on the MTL-AQA dataset.

Methods	Spe. Coe.	Rel. R-ℓ_2 Dis.
I3D + MLP [37]	0.9231	0.468
USDL [37]	0.9066	0.654
I3D + MLP [48]	0.9381	0.394
CoRe [48]	0.9512	0.260
Ours	**0.9537**	**0.245**

4.3 Implementation Details

We have implemented all experiments using the Pytorch framework with two Nvidia RTX 3090 GPUs.

The initial learning rate is set to 10^{-3} for regression and 10^{-4} for I3D. We use Adam to optimize the network, and the weight decay is set to 10^{-5}. The balance factors λ_1 and λ_2 are set to 1. Both the training and testing batch size is set to 1. The neighborhood radius is set to 20. In the prediction process, ten sets of exemplars are selected for inference, and the final score is determined via ensemble voting. For all experiments, 103 frames are extracted for each video clip and split into ten clips, each containing 16 frames with overlap between clips. For non-square frames, we use the padding strategy to convert them into squared ones. We use the frame-shifting strategy to augment the samples in the training phase: all frames are shifted left or right by 0 to 5 frames and the rest is filled with empty images. With a test ratio of 0.3, we divide the dataset into the training set and the test set. In the training phase, a hybrid training strategy is proposed to boost the assessment performance: the pre-trained model is first fine-tuned on all 14 types of actions in an unsupervised manner [7], followed by fine-tuning on each type of action separately. To re-balance imbalance samples, we adopt a re-sampling strategy [12] to train the network.

4.4 Results and Analysis

We first compare our model with the state-of-the-art. Then, the ablation results are given. Next, we show large qualitative and quantitative results. Finally, we present and analyze the results of the user study.

4.4.1 Comparisons with State-of-the-Art

Our method is compared with the state-of-the-art on a public AQA dataset and our dataset for muscle strength assessment.

We first use our proposed method to perform the AQA task on the MTL-AQA dataset [28]. It is a quite large-scale AQA dataset and is commonly used as the benchmark dataset to evaluate the effectiveness of AQA methods. The results are reported in Table 1. We compare two state-of-the-art AQA methods, where [37] is based on direct regression and [48] on learning a relative quality score. Table 1 shows that our method achieves better performance than the others [37,48]. The fact that our method still improves on this dataset is very encouraging since the state-of-the-art has reached close to the limit. Thus, our method is quantitatively demonstrated to be effective.

Then, two methods have been implemented for muscle strength assessment. The whole pipeline of our method can be divided into feature extractor (as well as feature aggregation) and score regression. To ensure a fair comparison, all methods utilize the I3D network as a feature extractor and the average pooling as a feature aggregator, whereas the score regression differs. Most AQA methods [17,37,48]

Table 2: Accuracy (%) and Spearman's rank coefficient on six CMAS actions with different methods of muscle strength assessment.

Methods	Head lift		Supine to sit		Arm straighten		Floor sit		Floor rise		Chair rise	
	Acc.	Spe. Coe.	Acc.	Spe. Coe.	Acc.	Spe. Coe.	Acc.	Spe. Coe.	Acc.	Spe. Coe.	Acc.	Spe. Coe.
SVM	43.48	0.3139	58.33	0.5351	52.17	0.3246	68.00	0.7761	73.91	0.6718	90.00	0.4716
MLP	66.67	0.7626	83.83	0.8361	73.33	0.3731	55.56	0.1284	78.26	0.3969	81.25	0.5878
Ours	**86.96**	**0.8393**	**91.67**	**0.9256**	**80.00**	**0.5357**	**76.00**	**0.7840**	**87.50**	**0.8710**	**93.75**	**0.9326**

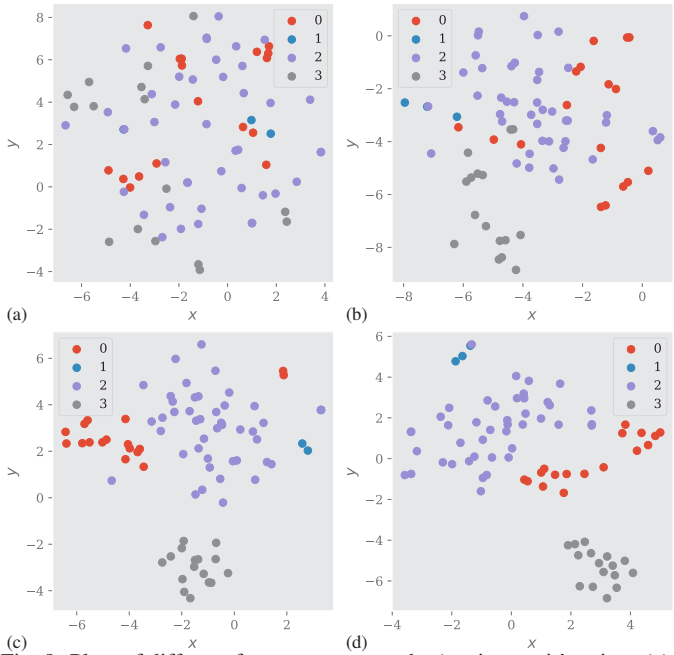

Fig. 8: Plots of different feature spaces on the 'supine to sit' action: (a) the original space, (b) MLP, (c) ours w/o $\mathcal{L}_{\text{dis}}$, and (d) ours.

adopt the MLP layers to regress the quality score. Similarly, we also use several MLP layers to regress the grade of muscle strength. Different from these methods, a simple yet effective contrastive regression module is proposed. Therefore, we have implemented an MLP-based method for comparison. In addition, we have also implemented a linear support vector machine (SVM) [6, 50] based method for muscle strength assessment. The comparison results on six CMAS actions are shown in Table 2. Among these actions, our method achieves the best performance, indicating the effectiveness of our proposed method. The other methods perform poorly, showing that directly transferring the current AQA methods to muscle strength assessment is ineffective. For example, our method on the 'supine to sit' action achieves an accuracy of 92.67%, whereas the MLP only achieves that of 58.33%.

We use the t-distributed stochastic neighbor embedding (t-SNE) [39] to visualize the different feature spaces in Fig. 8. The original feature space in Fig. 8(a) makes it almost impossible to distinguish any subcategory. In contrast, the feature space in Fig. 8(b) processed by MLP has a certain degree of identification, while the feature space in Fig. 8(d) processed by our method is almost linearly separable. In this regard, our proposed method is qualitatively demonstrated to be effective.

4.4.2 Ablation Study

This ablation study aims to verify the effectiveness of the core designs. All the experiments of the ablation study are performed on the 'supine to sit' action. Table 3 shows the corresponding results.

Effectiveness of Fine-Tuning Existing AQA methods [2, 48] ignore the large transfer gap from the action recognition to action quality assessment. To address this issue, we first fine-tune the pre-trained backbone model on all 14 types of actions in an unsupervised manner and then fine-tune it on each action separately. According to the second line (w/o FT) in Table 3, the removal of the first fine-tuning step results in a 25.00% reduction in accuracy and a 0.6895 reduction in correlation. Thus, fine-tuning the model is essential for improving assessment performance, especially for the JDM dataset, where each action only spans several dozen. This also demonstrates the effectiveness of the hybrid training strategy.

Table 3: Ablation results of the 'supine to sit' action.

Methods	Acc. (%)	Spe. Coe.	Rel. R-ℓ_2 Dis.
Ours	91.67	0.9256	0.0231
w/o FT	66.67 $^{\downarrow 25.00}$	0.2361 $^{\downarrow 0.6895}$	0.1296 $^{\uparrow 0.0190}$
w/o RS	75.00 $^{\downarrow 16.67}$	0.7166 $^{\downarrow 0.2090}$	0.0972 $^{\uparrow 0.0741}$
w/o CR	83.83 $^{\downarrow 7.84}$	0.8361 $^{\downarrow 0.0895}$	0.0471 $^{\uparrow 0.0190}$
w/o $\mathcal{L}_{\text{dis}}$	88.24 $^{\downarrow 3.43}$	0.8975 $^{\downarrow 0.0281}$	0.0523 $^{\uparrow 0.0292}$

Effectiveness of Re-Sampling As shown in Fig. 7, all action data exhibit a typical long-tailed distribution, and even for each type of action, the distribution of its sub-categories is highly imbalanced. In the case of the 'head lift' action, there are five grades, with the numbers (10, 15, 19, 13, 5, 8). According to the third row (w/o RS) of Table 3, the removal of the re-sampling strategy results in a 16.67% reduction in accuracy and a 0.2090 reduction in correlation. In addition, the network always produces the same grades in this case. Thus, it is evident that re-sampling is efficient in fixing the class imbalance problem, which is crucial for improving the accuracy of muscle strength measurement.

Effectiveness of Contrastive Regression Inspired by the contrastive AQA methods [17, 48], we propose a contrastive regression module for muscle strength assessment. According to the fourth line (w/o CR) of Table 3, the removal of the proposed contrastive regression results in a 7.84% reduction in accuracy and a 0.0895 reduction in correlation. In this way, the contrastive regression module contributes to improving the performance of muscle strength assessment.

Effectiveness of the Distance Loss For each action, we also fine-tune the backbone during the training phase. To constrain the feature space, we propose a distance loss, where the optimal model requires low distance and high similarity between the same subcategories. Similarly, we eliminate this loss function and maintain all other modules unchanged to determine whether it is effective. Based on the results in the fifth row (w/o $\mathcal{L}_{\text{dis}}$) of Table 3, the elimination of the distance loss results in a 3.43% reduction in accuracy and a 0.0281 reduction in correlation. After removing this loss, the feature space in Fig. 8(c) has more indistinguishable samples than the feature space in Fig. 8(d). Hence, the distance loss is demonstrated to be effective in both the qualitative and quantitative results.

Effectiveness of the Number of Ensembles For the regression module, the number of ensembles is a significant hyper-parameter affecting the results. We conduct several experiments on the 'supine to sit' action with different values of layers, ranging from 0 to 4. The corresponding results are shown in Fig. 10. When the ensemble is 0, the model is equivalent to removing contrastive regression, directly predicting muscle strength. There is a gradual increase in accuracy rate and correlation coefficient as the number of ensembles increases. At the same time, the relative R-ℓ_2 distance decreases, suggesting that the number of ensembles is an important factor in improving assessment performance. By using a greater number of ensembles, the model is able to refer to more exemplars while incurring a greater computational cost. Thus, we only use one ensemble to conduct the other experiments.

4.4.3 Visualization

We visualize the accuracy as well as loss during the training phase, the confusion matrix, and the inference results of several examples.

Fig. 9: Three samples and their predicted grade distribution using our proposed method: the first column to the fifth column denote the five different frames, and the final column represents the plot of the corresponding predicted grade similarity distribution.

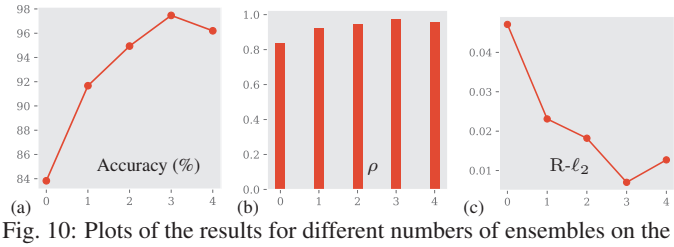

Fig. 10: Plots of the results for different numbers of ensembles on the 'supine to sit' action: (a) accuracy, (b) ρ, and (c) R-ℓ_2.

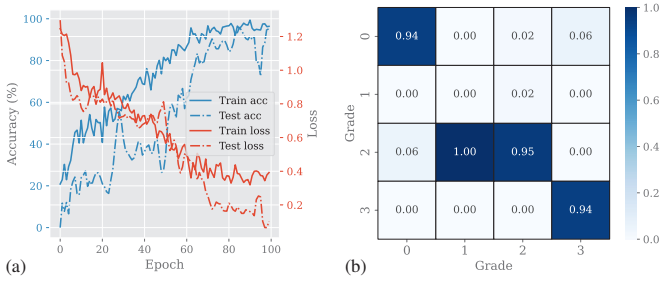

Fig. 11: The plot of (a) accuracy and loss curves during the training phase and the heatmap of (b) the confusion matrix.

Case Study In Fig. 9, three action samples have all been successfully assessed. The samples are #003 for the 'supine to lift' action, #007 for the 'arm straighten' action, and #011 for the 'floor rise' action, respectively. Five frames have been selected to illustrate the performing process of each action, which are shown in the first to fifth columns of Fig. 9. Please refer to the supplementary video for the complete rendered effect. The rendered scene can be viewed in conjunction with the virtual animation and the real action at the same time, where the key joints have been highlighted with transparent circles. The last column of Fig. 9 is the predicted grade distribution, where Figs. 9(a) and 9(b) are two successful cases and Fig. 9(c) shows a failed case.

In Fig. 9(b), the young girl can easily raise her arms, which is within grade 3. According to the grade distribution, the model output is the most similar to grade 3, indicating that this sample is the most similar to the exemplar of grade 3 and thus the predicted grade is 3. This demonstrates that our proposed muscle strength assessment algorithm is effective. We can also see that the arm joints are highlighted, indicating that the arm joints are important and that the virtual character differs from the real patient. Even though there are few differences between the two, the clinic doctor will still choose to believe in the machine score because he is confident that the two have the same level of action. In Fig. 9(c), the little boy sways when he gets up from the floor, and judging subjectively with different degrees of difficulty is quite difficult. Unfortunately, the model incorrectly identifies mild difficulty (grade 3) as severe (grade 1). When observing that there is a significant difference between the real patient and the virtual character, the doctor makes a final assessment using CMAS rules. Finally, the action in Fig. 9(c) is still successfully assessed as grade 3. By using a simple visual comparison rather than a direct assessment based on CMAS rules, our system can assist doctors in their assessment and reduce their workload. Once the machine makes a mistake, doctors can make a final decision to improve the assessment performance.

Our muscle strength assessment algorithm is video-based, so pose estimation is not required. During the AR effects rendering phase, determining the direction for inserting the character and visualizing key parts requires pose estimation. Hence, once the pose estimation fails, the character may not be accurately placed in the real environment. In this way, the assisted assessment functionality may be affected so doctors cannot use our system to verify the machine results. In this scenario, the accuracy is equal to that of the algorithm at least. It should be noted that the results of our algorithm are still significantly better than clinical doctors. As a result, our system is still capable of assisting clinical doctors in the assessment process and improving the performance of the assessment, even if the pose estimation fails.

Accuracy and Loss Fig. 11(a) shows the accuracy and loss curves during the training phase with 100 epochs for an experiment on the 'supine to sit' action. During the training process, the training error gradually decreases, the accuracy gradually increases, and the test results are also consistent. The test accuracy fluctuates more than the training accuracy as a result of fewer samples in the test set. After 50 epochs, the average loss in the test set is smaller than the loss in the training set, as the accuracy gradually approaches the accuracy of the training set. Thus, it demonstrates that our algorithm is so effective that few samples are needed to converge without over-fitting.

Confusion Matrix Fig. 11(b) shows the confusion matrix of one inference on the 'supine to sit' action. Each element of the confusion matrix means that the action with the i-th grade is assessed as the j-th

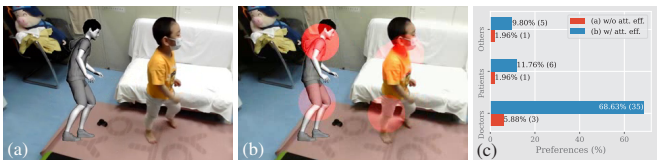

Fig. 12: Comparative study of attention effects: (a) augmented scene, (b) augmented scene with attention effects, (c) user preferences.

grade. Accordingly, the greater the number of elements in the main diagonal, the more accurate the model is at assessment. As can be seen in Fig. 11(b), all samples of grade 1 are incorrectly assessed as grade 2, while the others are more accurate. Among four types of sub-categories, the number of samples with grade 1 is only three, accounting for 3.8% of the total. During training, samples with grade 1 are over-sampled, but the similarity between the augmented samples is high, resulting in incorrect assessments. Additionally, the actions of grades 1 and 2 are similar, so grade 1 is incorrectly assessed as grade 2.

4.4.4 User Study

A comparative study is first presented to evaluate the effectiveness of attention effects. Next, we evaluate the effectiveness of our machine scoring algorithm by comparing its results with those of the CMAS assessment (T1). Finally, we evaluate the effectiveness of the human verification functionality in our system by comparing the results of the assisted assessment (T2) and those of the CMAS assessment (T1).

Effectiveness of Attention Effects In Sect. 3.3.3, we highlight the key parts of both the patient and the character, allowing users to complete the task the most effectively. Notice that a better user experience may not directly impact accuracy, but it reflects in other aspects such as ease of operation and effectiveness of visualisations. Thus, a comparative study has been conducted to investigate the user preference for two different rendered effects in Figs. 12(a) and 12(b). The result is shown in Fig. 12(c) where 90% of respondents, including doctors, parents, and others, prefer to highlight key parts for visual comparison. This demonstrates that it is easier to identify their movements' differences using Fig. 12(b) than Fig. 12(a). In this way, we next verify the virtual character and the attention effects as a whole.

Effectiveness of Machine Scoring Due to personal bias and a lack of professional diagnostic experience, it is difficult for general practitioners to accurately assess muscle strength. However, the machine scoring algorithm can circumvent these difficulties. Combining Fig. 13(a) and Table 2, it is encouraging that our machine scoring algorithm has achieved greater accuracy than the CMAS assessment (T1) on all actions. For example, participants achieve 88.89% accuracy on the 'supine to sit' action when performing task T1, which is lower than the machine accuracy of 91.67%. The participants are mainly general practitioners, so it is evident that our machine scoring algorithm is more effective. This provides the potential to improve the assessment performance of clinical doctors with the aid of machine scoring.

Effectiveness of Visual Comparison By combining AR technology and visual comparison, our system aims to provide effective verification for clinical doctors. If our algorithm makes a mistake, there is a discrepancy between the actions of the virtual character and the real patient, and then clinical doctors can make a final decision based on CMAS rules. Using our system (T2), participants can easily identify and correct errors for machine scoring. In Fig. 13(a), participants achieve 96.66% accuracy on the 'supine to sit' action when performing task T2, in which the accuracy is increased by 7.77% than machine scoring and 4.99% than the CMAS assessment (T1). This indicates that our system using visual comparison can assist clinical doctors in improving their assessment accuracy. In addition, we have examined how long participants spent performing the two tasks. In Fig. 13(b), participants spent less time performing task T2 than task T1. Therefore, it demonstrates that our system using visual comparison can also assist clinical doctors in improving their assessment efficiency.

Discussion Because our algorithm is relatively objective, it can achieve a more accurate result than participants who are limited by personal bias. However, it cannot achieve 100% accuracy, similar to any machine learning algorithm. Hence, the proposed algorithm cannot be directly applied to biomedical applications. To this end, we propose an AR system that incorporates visual comparison for effective human verification. On the one hand, the human-in-the-loop nature is helpful to avoid possible errors/biases introduced by the machine scoring algorithm. On the other hand, the visual comparison facilitated by AR rendering effects improves the efficiency of assessment, since the visual comparison is more effective than the direct assessment based on CMAS rules. As compared with the CMAS assessment used by clinical doctors, the results of the user study prove that our system can assist them in assessing muscle strength more accurately and rapidly.

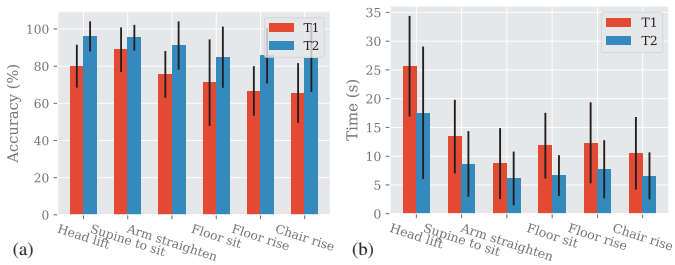

Fig. 13: Plots of average accuracy and time usage for the user study.

5 Conclusions

A method for assisting clinical doctors in assessing the muscle strength of children with JDM is explored in this work. Through side-by-side comparison, we aim to construct augmented scenes to assist clinical doctors in assessing muscle strength and monitoring the rehabilitation progress of patients. On the one hand, the proposed algorithm can avoid personal bias when grading actions performed by patients. On the other hand, the use of visual comparison supported by AR technology allows effective human verification, thereby reducing the negative consequences of misjudgment on the part of the machine scoring algorithm. The experimental results show that clinical doctors without expertise can make faster and more accurate assessments with our system.

Limitations and Future Work Since different types of actions are assessed according to different rules, it is difficult to establish a general algorithm that can be used to adaptively assess any action in CMAS. Furthermore, taking into account the distribution of the sample, this work has studied six challenging CMAS actions for the assessment of muscle strength. The principle of muscle strength assessment is to infer the patients' muscle strength score from their movements. Therefore, we assess muscle strength by classifying actions into different levels. Different from the known action recognition, since actions of different grads also tend to be less variable, our muscle strength assessment task is more fine-grained and thus more challenging. Additionally, the imbalanced class size and small sample size complicate our task. Therefore, we propose a new pre-training method to boost assessment performance using the re-sampling technique. To assist doctors in effective verification, we incorporate visual comparison to propose an AR system. As our system is still in the prototype stage, its ease of functionality needs to be further verified before clinical use.

The following future work needs to be explored further: (1) Based on different CMAS rules, we can divide the 14 types of actions into three evaluation types. We will establish automatic assessment algorithms for each type of action. This allows us to assess muscle strength across all actions in a comprehensive manner. (2) The corresponding character for a given action is selected from the previously generated animation dataset, which does not contain any personalized design. Therefore, it is possible to generate more realistic motions by combining the motion-style transfer technique. (3) We will conduct further user studies to assess the repeatability and usability of our system. Further optimization and validation of our system will be also necessary before clinical use. In conclusion, it is necessary to solve these problems for accurate muscle strength assessment and comfortable clinical usage.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Project Number: 62272019).

REFERENCES

[1] M. Antunes, R. Baptista, G. Demisse, D. Aouada, and B. Ottersten. Visual and human-interpretable feedback for assisting physical activity. In *European Conference on Computer Vision*, pp. 115–129. Springer, 2016. 2

[2] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang. Action quality assessment with temporal parsing transformer. *arXiv preprint arXiv:2207.09270*, 2022. 2, 3, 4, 7

[3] M. Batthish and B. M. Feldman. Juvenile dermatomyositis. *Current rheumatology reports*, 13(3):216–224, 2011. 1

[4] E. Buchanan, G. Loporcaro, and S. Lukosch. On the effectiveness of conveying bim metadata in vr design reviews for healthcare architecture. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 806–807. IEEE, 2022. 2

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4724–4733, 2017. 3

[6] L. Chen, K. Zhou, J. Jing, H. Fan, and J. Li. Solution path algorithm for twin multi-class support vector machine. *Expert Systems with Applications*, 210:118361, 2022. 7

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020. 6

[8] M. A. Cidota, P. J. Bank, P. Ouwehand, and S. G. Lukosch. Assessing upper extremity motor dysfunction using an augmented reality game. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 144–154. IEEE, 2017. 2

[9] P. Dhar, T. Rocks, R. M. Samarasinghe, G. Stephenson, and C. Smith. Augmented reality in medical education: students' experiences and learning outcomes. *Medical Education Online*, 26(1):1953953, 2021. 2

[10] H. Doughty, D. Damen, and W. Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6057–6066, 2018. 2

[11] H. Doughty, W. Mayol-Cuevas, and D. Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7862–7871, 2019. 2, 6

[12] Y. Fu, L. Xiang, Y. Zahid, G. Ding, T. Mei, Q. Shen, and J. Han. Long-tailed visual recognition with deep models: A methodological survey and evaluation. *Neurocomputing*, 2022. 6

[13] D. Gorgos. Early treatment is best for long-term outcome of juvenile dermatomyositis. *Dermatology Nursing*, 16(5):461–463, 2004. 1

[14] J. Hochreiter, S. Daher, A. Nagendran, L. Gonzalez, and G. Welch. Touch sensing on non-parametric rear-projection surfaces: A physical-virtual head for hands-on healthcare training. In *2015 IEEE Virtual Reality (VR)*, pp. 69–74. IEEE, 2015. 2

[15] J. Hombeck, M. Meuschke, L. Zyla, A.-J. Heuser, J. Toader, F. Popp, C. J. Bruns, C. Hansen, R. R. Datta, and K. Lawonn. Evaluating perceptional tasks for medicine: A comparative user study between a virtual reality and a desktop application. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 514–523. IEEE, 2022. 2

[16] C. Hutchinson, B. Feldman, S. Li, M. Patterson, and E. TePas. Juvenile dermatomyositis and polymyositis: epidemiology, pathogenesis, and clinical manifestations. *World Journal of Pediatrics*, 16, 2020. 1

[17] H. Jain, G. Harit, and A. Sharma. Action quality assessment using siamese network-based deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2260–2273, 2020. 2, 6, 7

[18] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[19] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9522–9531, 2021. 2

[20] D. J. Lovell, C. B. Lindsley, R. M. Rennebohm, S. H. Ballinger, S. L. Bowyer, E. H. Giannini, J. E. Hicks, J. E. Levinson, R. Mier, L. M. Pachman, et al. Development of validated disease activity and damage indices for the juvenile idiopathic inflammatory myopathies: Ii. the childhood myositis assessment scale (cmas): a quantitative tool for the evaluation of muscle function. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 42(10):2213–2219, 1999. 2

[21] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 4

[22] L. McCann, A. Juggins, S. Maillard, L. Wedderburn, J. Davidson, K. Murray, and C. Pilkington. The juvenile dermatomyositis national registry and repository (uk and ireland)—clinical characteristics of children recruited within the first 5 yr. *Rheumatology*, 45(10):1255–1260, 2006. 1

[23] H. M. Mentis, I. Avellino, and J. Seo. Ar hmd for remote instruction in healthcare. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 437–440. IEEE, 2022. 2

[24] C. Oddis, C. Conte, V. Steen, and T. Medsger Jr. Incidence of polymyositis-dermatomyositis: a 20-year study of hospital diagnosed cases in allegheny county, pa 1963-1982. *The Journal of rheumatology*, 17(10):1329–1334, 1990. 1

[25] L. M. Pachman, B. E. Nolan, D. DeRanieri, and A. M. Khojah. Juvenile dermatomyositis: new clues to diagnosis and therapy. *Current treatment options in rheumatology*, 7(1):39–62, 2021. 1

[26] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, and M. Mirmehdi. Online quality assessment of human movement from skeleton data. In *BMVC*, pp. 153–166, 2014. 2

[27] P. Parmar and B. T. Morris. Measuring the quality of exercises. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2241–2244. IEEE, 2016. 2, 3

[28] P. Parmar and B. T. Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 304–313, 2019. 3, 6

[29] P. Parmar and B. Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 20–28, 2017. 3

[30] M. Pears, M. Yiasemidou, M. A. Ismail, D. Veneziano, and C. S. Biyani. Role of immersive technologies in healthcare education during the covid-19 epidemic. *Scottish Medical Journal*, 65(4):112–119, 2020. 2

[31] P. Pelkonen, H. Jalanko, R. Lantto, A.-L. Mäkelä, M. Pietikäinen, H. Savolainen, and P. Verronen. Incidence of systemic connective tissue diseases in children: a nationwide prospective study in finland. *The Journal of rheumatology*, 21(11):2143–2146, 1994. 1

[32] I. Phelan, M. Arden, C. Garcia, and C. Roast. Exploring virtual reality and prosthetic training. In *2015 IEEE Virtual Reality (VR)*, pp. 353–354. IEEE, 2015. 2

[33] A. Ramanan and B. M. Feldman. Clinical features and outcomes of juvenile dermatomyositis and other childhood onset myositis syndromes. *Rheumatic Disease Clinics*, 28(4):833–857, 2002. 1

[34] N. Renu. Applications of ar and vr technologies in healthcare marketing. *Journal of Marketing Management*, 9(2):35–39, 2021. 2

[35] M. Robles, N. Namdarian, J. Otto, E. Wassiljew, N. Navab, C. M. Falter-Wagner, and D. Roth. A virtual reality based system for the screening and classification of autism. *IEEE Trans. Vis. Comput. Graph.*, 28(5):2168–2178, 2022. 2

[36] A. Singla, S. Göring, D. Keller, R. R. R. Rao, S. Fremerey, and A. Raake. Assessment of the simulator sickness questionnaire for omnidirectional videos. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 198–206. IEEE, 2021. 2

[37] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020. 2, 6

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015. 3

[39] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[40] Q. Wang, B. Kang, and P. O. Kristensson. Supporting playful rehabilitation in the home using virtual reality headsets and force feedback gloves. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 504–513. IEEE, 2022. 2

[41] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 1328–1338, 2019. 4

[42] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang. Tsa-net: Tube self-

attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4902–4910, 2021. 2

[43] Y. Wang, J.-R. Chardonnet, F. Merienne, and J. Ovtcharova. Using fuzzy logic to involve individual differences for predicting cybersickness during vr navigation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 373–381. IEEE, 2021. 2

[44] R. Wetzel and T. Kreienbühl. Breathe to dive: Exploring a virtual reality game for treatment of cystic fibrosis. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 412–416. IEEE, 2019. 2

[45] A. Xu, L.-A. Zeng, and W.-S. Zheng. Likert scoring with grade decoupling for long-term action assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3232–3241, 2022. 3

[46] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. *arXiv preprint arXiv:2204.03646*, 2022. 2

[47] W. Xu and S.-C. Zhang. Chinese pediatricians face a crisis: should they stay or leave? *Pediatrics*, 134(6):1045–1047, 2014. 1

[48] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pp. 7919–7928, 2021. 2, 3, 4, 6, 7

[49] A. Zafar and M. S. Farooq. Augmented reality in healthcare education for human anatomy. 2021. 2

[50] K. Zhou, Q. Zhang, and J. Li. Tsvmpath: Fast regularization parameter tuning algorithm for twin support vector machine. *Neural Processing Letters*, pp. 1–26, 2022. 7