Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events

Hong-Bo Zhang[1,2] · Li-Jia Dong[1,3] · Qing Lei[2] · Li-Jie Yang[3] · Ji-Xiang Du[3] ⓘD

Abstract

Most existing action quality assessment (AQA) methods provide only an overall quality score for the input video and lack an evaluation of each substage of the movement process; thus, these methods cannot provide detailed feedback for users. Moreover, the existing datasets do not provide labels for substage quality assessment. To address these problems, in this work, a new label-reconstruction-based pseudo-subscore learning (PSL) method is proposed for AQA in sporting events. In the proposed method, the overall score of an action is not only regarded as a quality label but also used as a feature of the training set. A label-reconstruction-based learning algorithm is built to generate pseudo-subscore labels for the training set. Moreover, based on the pseudo-subscore labels and overall score labels, a multi-substage AQA model is fine-tuned from the PSL model to predict the action quality score of each substage and the overall score for an athlete. Several ablation experiments are performed to verify the effectiveness of each module. The experimental results show that our approach achieves state-of-the-art performance.

Keywords Action quality assessment · Substage quality assessment · Label reconstruction · Pseudo-subscore learning · Multi-substage AQA model

1 Introduction

Action quality assessment (AQA), which aims at assessing the action quality of a performer and providing detailed feedback to help the executor improve his/her action quality, has received increasing attention in the field of computer vision. AQA is widely used in sports quality scoring [1–9], exercise quality evaluation during the rehabilitation training of stroke patients [10–14] and completion quality evaluation of other exercises in daily life [15, 16].

The purpose of AQA in sporting events is to provide an automatic objective assessment mechanism that simulates subjective human perception in assessing the quality of actions completed by athletes. On the basis of related field experience and knowledge of sports evaluation, a referee will give an overall score based on the performance of an athlete in each of multiple action substages. For example,

the diving process can be divided into five substages: start, takeoff, drop, entry, and end. The overall score for diving performance is computed using the execution score for each substage and the difficulty level of the action, as shown in Fig. 1. Accordingly, this paper argues that stagewise quality evaluation is a reasonable and beneficial approach for building an accurate sports evaluation system for AQA in sporting events.

More importantly, substage quality assessment is an effective way to obtain detailed feedback on the performance of athletes in each substage of sports actions, which is more meaningful than simply providing an overall score. However, most existing methods treat AQA as a score regression problem, in which the overall scores are taken as the training labels to train an AQA regression model to directly map input videos to quality scores. Such methods ignore the details of the individual substages and cannot provide meaningful instruction for users. To fill these gaps, this paper discusses how to evaluate the action quality of various substages in sports when only an overall score is provided for each performance, as is the case in almost all published datasets, and aims to find an effective AQA model that can generate a quality score for each substage and thus

✉ Ji-Xiang Du
  jxdu@hqu.edu.cn

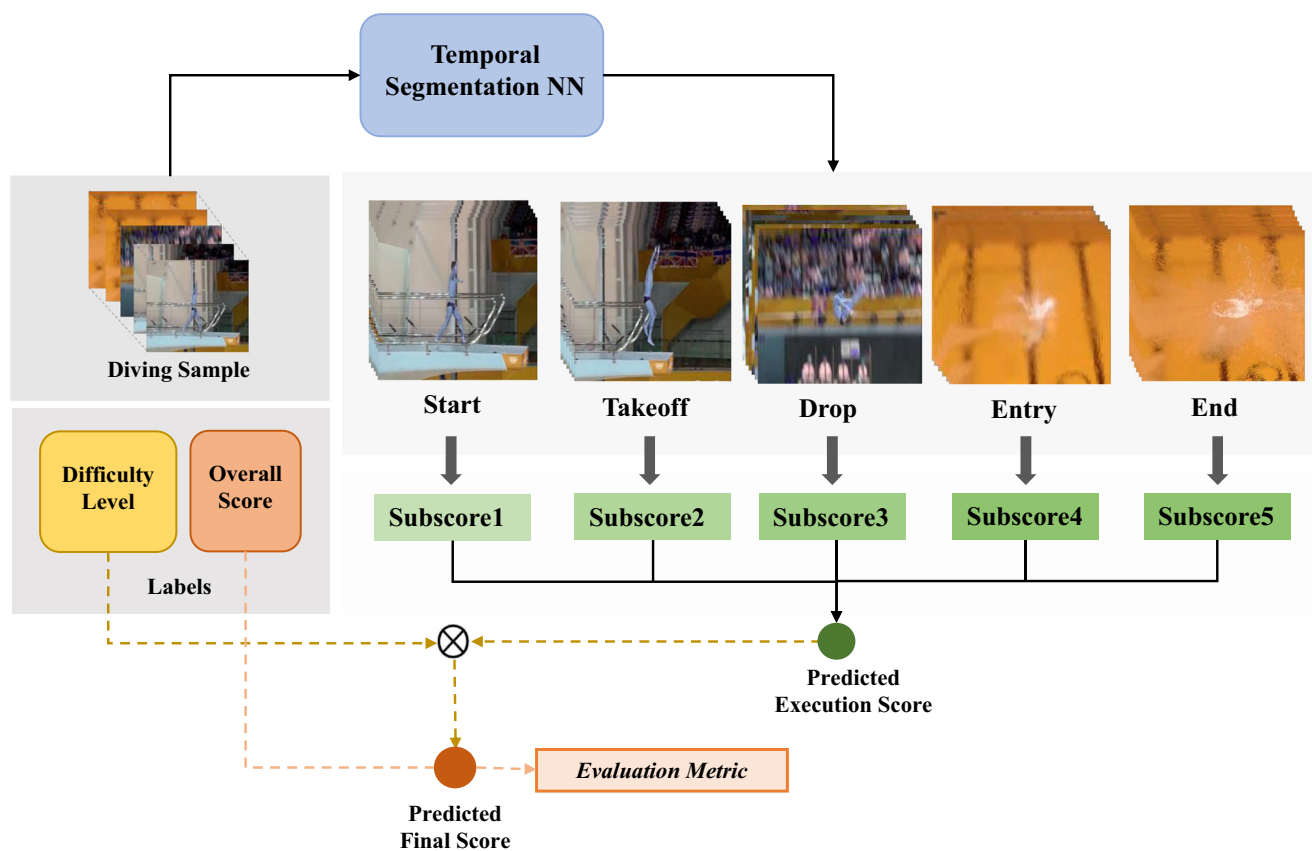Extended author information available on the last page of the article.

Fig. 1 Example of multi-substage AQA in diving

obtain a more accurate overall quality score based on the performance of an athlete in all substages.

However, due to the lack of quality labels for each substage, accurate performance evaluation in each substage is a challenging problem. In the existing work on AQA, all of the available datasets provide only overall score labels as indicators of action quality rather than individual substage labels. Only the UNLV-Diving dataset [1], which was annotated by [2], provides a segmentation label for each substage, and it still lacks a quality score label for each substage. In addition, the manual annotation of substage scores requires professional knowledge and experience, which is difficult to obtain.

To solve these problems, this work aims to automatically generate pseudo substage score labels for the training data and then train a multi-substage AQA model based on the generated pseudo-subscores in combination with the overall score labels. To this end, we propose a label-reconstruction-based pseudo-subscore learning (PSL) method, in which the overall score is regarded not only as a label but also as an input feature. We decompose the regression process from visual features to the overall score into two stages: mapping from visual features to subscores and mapping from subscores

to the overall score. Accordingly, the network structure of the proposed PSL model is mainly composed of three parts: a backbone network for the feature extractor, a label decomposition network for the subscore generator, and a label construction network for the overall score generator. In particular, in the label decomposition network, the overall score label is used as a feature and is fused with the visual features of the substage segments to generate subscores. The label decomposition network is followed by the label construction network; together, they form the label reconstruction network (LRN). After pseudo-subscore generation, the multi-substage AQA model is fine-tuned from the PSL model. The calculation process in this work is summarized in Fig. 2 and includes two modules: the PSL module and the AQA model learning module.

In the PSL module, the label reconstruction loss is computed to train the model. After training, each training video and its overall score are fed into the PSL model again, and the output of the label decomposition network is taken as the pseudo-subscore labels of the training video. In the multi-substage AQA module, the overall score feature in the label decomposition network is removed to form the subscore prediction network (SPN), which replaces the label
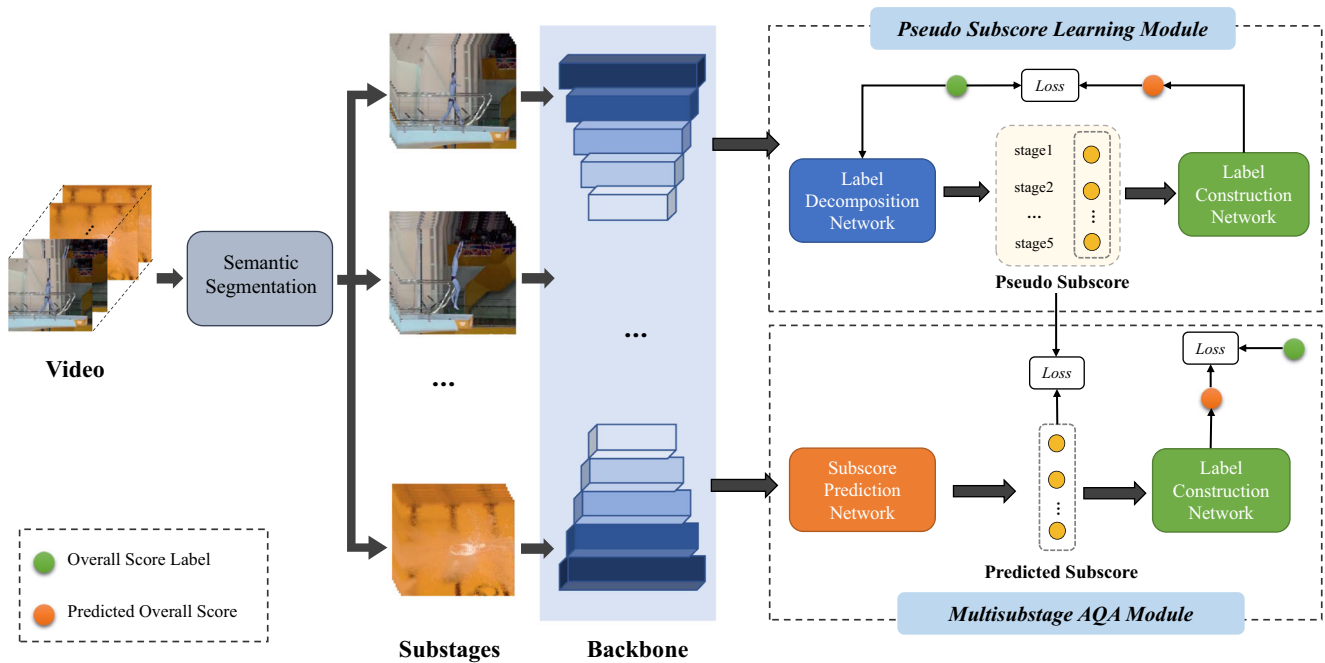
Fig. 2 Calculation process applied in this work. It involves temporal semantic segmentation, feature representation, a PSL model, and a multi-substage AQA model

decomposition network in the PSL model to form the AQA model. Finally, the AQA model is retrained based on the generated pseudo-subscore and overall score labels.

The contributions of this work are summarized as follows:

(1) We propose a novel multi-substage AQA method that can not only evaluate the overall quality of sports actions but also give a score for each action substage as detailed feedback.

(2) We propose a label-reconstruction-based PSL method to generate pseudo-subscore labels for the training data. The generated pseudo-subscore labels and original overall score labels are then applied together to train the proposed AQA model.

(3) Ablation experiments are performed to verify the effectiveness of each module. In addition, a new score prediction task for the Olympic diving finals is designed in an illustrative experiment. The experimental results show that the proposed method achieves state-of-the-art performance.

The remainder of this paper is organized as follows. Section 2 introduces related works, and Section 3 describes the algorithms used to implement the proposed method. Section 4 presents and discusses the experimental results, and Section 5 concludes the paper.

2 Related work

Action understanding is a classic research topic in computer vision. Previous works have focused on the problem of action recognition, which is regarded as a classification problem. In recent years, an increasing number of AQA tasks have been deeply studied by scholars, usually by using a regression model to predict action quality scores. In both cases, the basic approach used is action feature learning. In this section, a brief summary of the methods of action feature learning and recognition is presented first, and then, a detailed review of AQA methods is provided.

2.1 Action feature learning and recognition

Action recognition in video is widely applied in many fields. How to encode the action information of a video into high-level features has become a key issue for such tasks.

In addition to the traditional methods of video action recognition, the extraction of spatiotemporal information from video based on deep learning networks has become a popular focus of research in recent years. To process video data with temporal and spatial information, 3D convolutional networks have been widely used [17]. 3D convolutional networks extend 2D convolutions to 3D space to encompass the time dimension. Similarly, since recurrent neural network (RNN) and long short-term memory

(LSTM) structures can process temporal signals, they have also been used to extract features of video frames.

On the basis of the 3D convolutional neural network (CNN) model named C3D, several action feature learning methods have been proposed, such as the 3D ResNet network [18], the 3D ResNeXt network [19], and an inflated 3D convolutional neural network (I3D) [20]. Qiu et al. [21] proposed a pseudo-3D convolutional network based on ResNet, in which the 3D convolution kernel is decoupled into two 2D convolution kernels. To reduce the number of parameters, a pretrained 2D CNN network was used to improve the performance of the 3D convolutional network. He et al. [22] divided C3D into two paths, a slow path and a fast path, and designed an end-to-end network for action recognition. The slow path was used to capture spatial semantics, and the fast path was used to capture motion information with temporal resolution.

2.2 AQA methods

Consistent with the annotation style of the available datasets, most AQA studies have applied a regression network to map input videos to performance scores and have used the ranking accuracy or score error as the final evaluation metric. A general method was designed by [3] to evaluate the quality of seven different sports, including diving and figure skating. Similar to other methods, the training of the network in that method was also based on the overall score of the movement. Through graph-based joint relationship modeling, the authors of [23] designed a method that could be used to assess action quality in gymnastics competitions and surgical procedures.

AQA research has also focused on expanding the scope of application of the developed methods beyond that of the original research with the aim of solving more practical problems. For example, Parmar et al. [24] designed a method for the automatic evaluation of levels of piano playing. In the remote environment motivated by COVID-19, this method, as a pioneering work on the automatic evaluation of teaching skills, is expected to play an important role in future online teaching. In addition to the assessment of the level or quality of specific skills in professional fields or in daily life [10–12, 15, 16, 24], AQA tasks have also been applied in rehabilitation [13, 14] and sporting events [1–3, 5–9, 25–28]. Due to the differences in the domain rules and action characteristics in different applications, special evaluation methods need to be developed for different tasks. In this work, we focus on the application of AQA in sporting events, which is discussed in detail below.

In sports, the overall AQA score is composed of an execution score and a difficulty level. The difficulty level is predefined in accordance with the action routines, whereas the execution score is assessed by referees. Movement in sports is highly variable, and AQA in sporting events usually involves features of multiple action substages. [1, 4, 29] considered that the overall score for diving is obtained by multiplying the execution score by the difficulty level, and the difficulty level was used as a variable in the experiment. Similarly, [8, 25] combined the execution score and the difficulty level for diving evaluation. Nekoui et al. [8] combined postures and scoring rules, while [25] combined different key fragments to generate the final feature for AQA. Xiang et al. [2] divided the process of diving into different stages, discussed the contributions of these different stages to the overall score, and evaluated diving performance on this basis. Zeng et al. [27] combined the dynamic and static information in a video and considered the periodic contributions of different stages to the final score over a long period of movement. A dynamic and static context-aware hybrid attention network was designed for quality assessment over long periods in sports. Xu et al. [5] combined the generation of two kinds of scores in figure skating—the total element score (TES) and the total program component score (PCS)—to design an automatic evaluation method for figure skating based on two kinds of attention mechanisms.

Other AQA methods for use in sporting events have also been reported. Parmar et al. [7] designed a multitask learning framework to complete AQA tasks by combining the pose, category and overall score of diving movements. Tang et al. [28] proposed a score distribution learning method for perceiving uncertainty to account for the subjectivity of the scores of different judges. Considering that most existing AQA methods are limited to individual behaviors and lack the ability to model asymmetric relationships between subjects, Gao et al. [9] designed an AQA method for synchronized diving and surgical operation. Nekoui et al. [26] proposed a network structure named EAGLE-Eye, based on visual and posture cues and consisting of two modules: a pose heatmap extractor and an appearance feature extractor. This method quantifies the quality of execution of an action based on fine- and coarse-grained time dependencies.

3 Method

In this section, we describe the proposed method in detail. As shown in Fig. 2, we present the pipeline of our work in three parts: temporal semantic segmentation and feature representation, the PSL model, and the multi-substage AQA model.

3.1 Temporal semantic segmentation and feature extraction

To obtain detailed action quality feedback, the input video V is first segmented into several substage clips, $V = \{p_1, ..., p_n\}$, $p_i \in \mathbb{R}^{W*H*T}$, where W and H are the image width and height, respectively; T is the length of each clip; and n is the number of clips. In this work, we use an off-the-shelf encoder–decoder temporal convolutional network (ED-TCN) [30] as the model for temporal semantic segmentation. After temporal segmentation, each clip corresponds to one substage of the action to be assessed. Temporal segmentation is an offline process in this work. The core component of the proposed method lies in the further processing of the segmented clips, which are provided as input to the feature backbone network for feature extraction.

In this paper, a pseudo-3D CNN model (P3D) [21] is selected as the backbone network for feature extraction. P3D uses a $1 \times 3 \times 3$ spatial convolution and a $3 \times 1 \times 1$ temporal convolution to simulate the commonly used $3 \times 3 \times 3$ 3D convolution. These parameters represent the time, width, and height dimensions, respectively. Through simplification compared with a traditional 3D CNN of the same depth, P3D can obtain accurate action recognition results with fewer parameters and an improved operation speed. To ensure fair comparisons with state-of-the-art methods, the network structure of P3D used in the proposed method is set to be the same as that in the literature [2].

3.2 PSL model

In the first module of the proposed method, the PSL model is built to obtain the pseudo-score values for each substage. This model consists of three parts: a feature backbone network, a subscore generator (label decomposition) and an overall score generator (label construction). The network structure of the PSL model is shown in Fig. 3. The feature backbone network is followed by the subscore generator. In accordance with the purpose of the LRN, the subscore generator generates subscores for the input substage clips, whereas the overall score generator calculates the corresponding overall score from the subscores produced by the subscore generator. Thus, the subscore and overall score generators in the PSL model together constitute the LRN structure.

Let the features extracted from the feature backbone network be denoted by $f_i \in \mathbb{R}^m$, where m is the number of feature dimensions and i is the substage index. After feature extraction, the input substage clips are transformed into a feature set $V = \{f_1, ..., f_n\}$. In PSL, the overall score label is also regarded as a feature and is embedded into the features of each substage to obtain new features $\{f_{l1}, ..., f_{ln}\}$, where $f_{li} \in \mathbb{R}^{m+1}$. These new features are used as the input to the subscore generator.

The subscore generator (label decomposition) is constructed as a fully connected network. In this work, this network includes five fully connected layers. The numbers of nodes in these layers are $m + 1$, $m/4$, $m/32$, $m/128$, and 1. The first layer takes the label-embedded features as input. The structure of these fully connected layers will be further discussed in the ablation study section. The $Sigmoid$ function is applied as the activation function in the last fully connected layer, and the output is the predicted subscore. The calculation process of the subscore generator can be expressed as follows:

$$
\begin{aligned}
f_{li} &= L \oplus f_i \\
f_{li}^t &= FC(W^t, f_{li}^{t-1}) \\
sbs_i &= Sigmoid(f_{li}')
\end{aligned}
\tag{1}
$$

where $\oplus$ represents the feature concatenation operation; $FC(\cdot)$ represents the fully connected operation; L represents the score label; f_{li}' represents the output of the last fully connected layer; $W^t = \{w_i^t \| i = 1, 2, ..., h^t\}$ represents the parameters of the t-th layer, with h^t representing the number of parameters of the fully connected operation in this layer; and sbs_i represents the subscore for the i-th substage. A subscore generator with the same structure is built for each substage in order to predict the subscores independently.

The overall score generator (label construction) takes the subscores for each substage as its input and predicts the overall score of the video, which is achieved by means of a fully connected layer, as expressed in (2):

$$
S = Sigmoid \left(\sum_{i=1}^{n} (w_i \times sbs_i) \right)
\tag{2}
$$

where S denotes the predicted overall score of the input video and the w_i are the weight parameters in this fully connected layer.

Finally, in this work, an end-to-end PSL model is built by concatenating the feature backbone network, the subscore generator and the overall score generator. To train the PSL model, considering that the overall score in the sporting events of interest is the product of an execution score and a difficulty level, we use two scores as the labels of each input video in this work: the overall score and the execution score. The mean square error (MSE) function is applied as the label reconstruction loss $loss_{psl}$ to measure the error between the overall score label and the predicted score. In the case in which the execution score is used as the label, the output of the PSL model is multiplied by the difficulty level

to generate the predicted overall score. The loss function $loss_{psl}$ is defined as follows:

$$loss_{psl} = \frac{1}{N} \sum_{i=1}^{N} (S_i - L_i)^2$$
$$= \frac{1}{N} \sum_{i=1}^{N} ((S_i^e - L_i^e)^2 \times d_i^2) \qquad (3)$$

where N is the number of training videos. S_i denotes the predicted overall score for the i-th training sample, and L_i denotes the overall score label. Similarly, S_i^e denotes the predicted execution score for the i-th training sample, and L_i^e denotes the execution score label. d_i denotes the difficulty level of the i-th sample. According to the score computation rules for diving competitions, $S_i = S_i^e \times d_i$ and $L_i = L_i^e \times d_i$.

After PSL model training, the pseudo-subscores of the training set are obtained. The training videos are fed into the trained PSL model, and the pseudo-scores for the various substages of each video, $SBS = \{sbs_1, ..., sbs_n\}$, are extracted by the subscore generator. It should be emphasized that for the PSL model, all operations involve only the training data, and none of the test data are used. Finally, the LRN-based PSL algorithm is summarized in Algorithm 1.

Algorithm 1 LRN-based PSL algorithm.

Input : Training set: video V and overall score label L
Output: Subscores for each video: $SBS = \{sbs_1, ..., sbs_n\}$

1: Segment the input video V into a clip set $\{p_1, ..., p_n\}$ by means of ED-TCN;
2: **while** $loss_{psl}$ is not converged **do**
3:      Select a batch of videos from the training data;
4:      **for** Each video **do**
5:          **for** $i = 1 : n$ **do**
6:              Compute the features of each clip: $f_i = P3D(p_i)$;
7:              Embed the label L into the features of each clip: $f_{li} = L \oplus f_i$;
8:              Compute the subscore for each clip: $sbs_i = Sigmoid(f'_{li})$;
9:          **end for**
10:          Compute the overall score: $S = Sigmoid(\sum_{i=1}^{n}(w_i \times sbs_i))$;
11:      **end for**
12:      Compute the loss: $loss_{psl} = \frac{1}{N} \sum_{i=1}^{N} ((S_i^e - L_i^e)^2 \times d_i^2)$;
13: **end while**
14: Obtain the PSL model in the last round;
15: Compute pseudo-subscores SBS for each video in training set via (1);

3.3 Multi-substage AQA model

In the second module, the multi-substage AQA model is implemented. In the PSL model, the labels of the training data are applied as the features; however, for AQA, the labels are unknown and need to be predicted. Therefore, we need to transfer the PSL model to the AQA model. In detail, the label features are removed from the label decomposition network of the PSL model, and a new SPN is built. This SPN is also constructed as a fully connected network consisting of five fully connected layers, with m, $m/4$, $m/32$, $m/128$, and 1 nodes. The output of the feature backbone network is fed into the SPN, which is followed by the label construction network to form the multi-substage AQA model, as shown in Fig. 4. The calculation process of the multi-substage AQA model is expressed as shown in (4):

$$f_i^t = FC(W^t, f_i^{t-1})$$
$$S_i^{sub} = Sigmoid(f_i')$$
$$S = Sigmoid\left(\sum_{i=1}^{n} w_i \times S_i^{sub}\right) \qquad (4)$$

where S_i^{sub} and S are the quality scores for each substage and the overall video, respectively.

After PSL, for each training video, we obtain pseudo-subscore labels SBS and an overall score label L. The proposed multi-substage AQA model is trained based on these two sets of labels. To conduct this training, the MSE loss function is again applied to compute the subscore and overall prediction errors. The loss function for the i-th substage can be formulated as shown in (5):

$$loss_i^{sub} = (S_i^{sub} - sbs_i)^2 \qquad (5)$$

Then, the total loss function used to train the multi-substage AQA model is defined as the sum of the substage and overall losses, as shown in (6):

$$loss_o = (S - L)^2$$
$$loss_{total} = loss_o + \sum_{i=1}^{n} loss_i^{sub} \qquad (6)$$

where $loss_o$ is the loss in terms of the predicted overall score.

4 Experiment

4.1 Experimental setting

Dataset In this paper, the UNLV-Diving dataset is used to verify the effectiveness of the proposed method. This
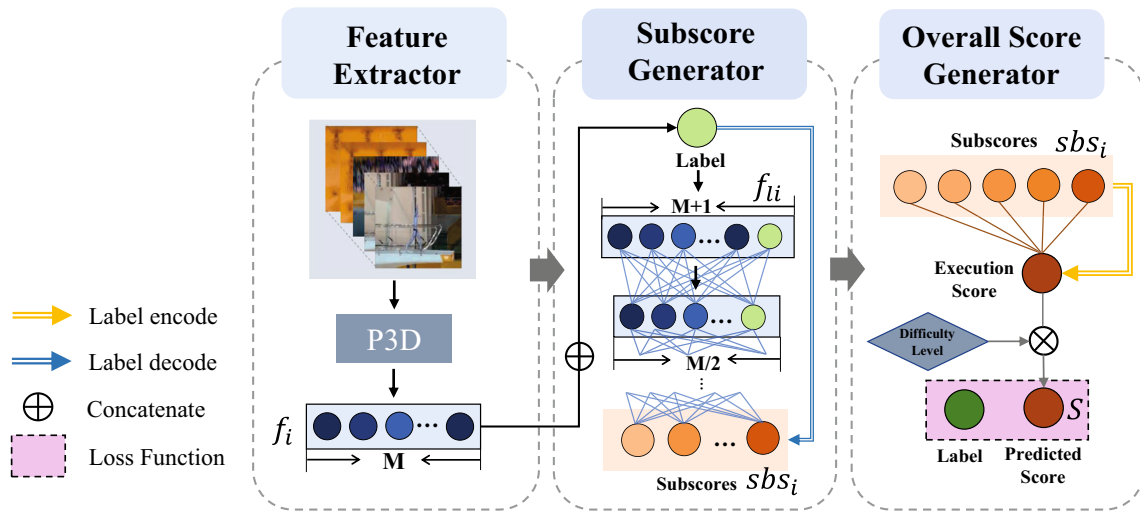
Fig. 3 Illustration of the PSL model. It consists of a feature extractor, a subscore generator and an overall score generator

dataset includes 370 videos from the semifinals and finals of the men's 10-m platform at the 2012 London Olympics. Each video sample includes overall score, difficulty level and substage segmentation labels. The difficulty level values range from 2.7 to 4.1, and the overall score values range from 21.6 to 102.6. In the original split, 300 of the videos in the dataset are used as the training set, and the remaining 70 videos are used as the test set.

In accordance with the characteristics of the UNLV-Diving dataset, we design a new AQA task called finals ranking prediction, in which the semifinal videos are used as the training data to train the model and predict the scores and ranking of the athletes in the finals. Specifically, in the finals

of the men's 10-m platform at the 2012 London Olympics, there were 12 athletes who took turns performing six rounds of diving actions, corresponding to a total of 72 videos. These videos are used as the test videos for the AQA model, and the remaining 298 video samples from the semifinals are used as the training set. After the sixth round, the total score for each athlete is calculated and ranked.

Evaluation metrics To evaluate the performance of the proposed method, Spearman's correlation coefficient (SRC) ρ, as defined in (7), is used to measure the accuracy of score ranking. In some works, the MSE [6] and the mean Euclidean distance (MED) [4, 25] have also been applied to
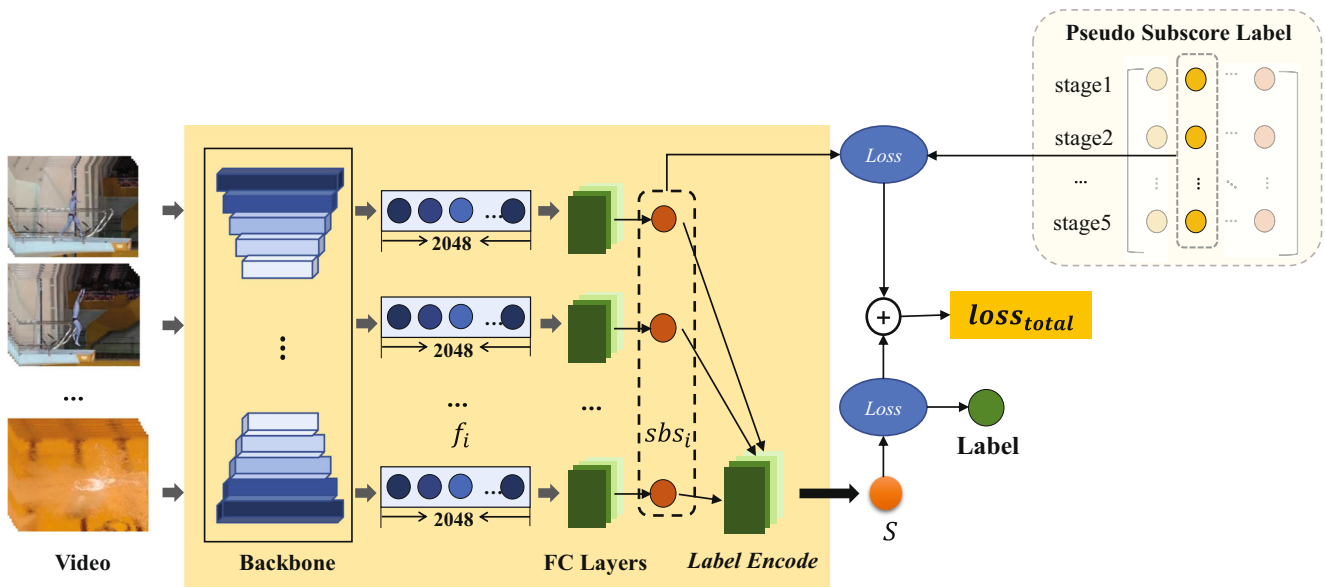

Fig. 4 Illustration of the multi-substage AQA regression model

measure the accuracy of the predicted scores. In this work, all of these evaluation metrics are adopted.

$$\rho = \frac{cov(R_S, R_G)}{\sigma_{R_S}\sigma_{R_G}} \tag{7}$$

$$MED = \frac{1}{N}\sum_{i=1}^{N}|s_i - g_i| \tag{8}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(s_i - g_i)^2 \tag{9}$$

where R_S and R_G represent the ranking sequences of the test videos in terms of the predicted scores and the ground truth, respectively; $cov(\cdot, \cdot)$ is the covariance of these two ranking sequences; and σ_{R_S} and σ_{R_G} are the standard deviations of R_S and R_G, respectively. The larger the value of ρ is, the more relevant the predicted value is to the ground truth, and the better the effect of the method. For the MED and MSE, s_i denotes the predicted score, and g_i denotes the ground truth; the smaller either of these metric values is, the smaller the distance between the predicted score and the ground truth, and thus, the better the performance.

For the finals ranking prediction task, in addition to calculating the SRC, MED and MSE of the test video ranking sequence, we can also calculate these evaluation metrics for the 12 athletes in the finals. The total scores of the 12 athletes are used to calculate these SRC, MED, and MSE values, which are denoted by A-SRC, A-MED and A-MSE, respectively. All of these indicators are used to evaluate the accuracy of the proposed method in predicting the outcome of the finals.

$$A - SRC = \frac{cov(R_S^A, R_G^A)}{\sigma_{R_S^A}\sigma_{R_G^A}} \tag{10}$$

$$A - MED = \frac{1}{12}\sum_{i=1}^{12}\left|s_i^A - g_i^A\right| \tag{11}$$

$$A - MSE = \frac{1}{12}\sum_{i=1}^{12}\left(s_i^A - g_i^A\right)^2 \tag{12}$$

where R_S^A and R_G^A represent the ranking sequences of the total scores of the 12 athletes in the finals in terms of the predicted and ground-truth values, respectively, and $\sigma_{R_S^A}$ and $\sigma_{R_G^A}$ are the standard deviations of R_S^A and R_G^A, respectively. For A-MED and A-MSE, s_i^A denotes the predicted total score of the i-th athlete, and g_i^A denotes the ground-truth total score of the i-th athlete.

Implementation details In the experiment, the substage segmentation model was trained based on ED-TCN [30]. Each diving video was segmented into five substages. A P3D model pretrained on a kinetics dataset was used as the feature backbone network. The input size of the feature backbone network was set to $16 \times 160 \times 160$, which means that each substage was sampled into 16 frames and that the images were resized to 160×160. The dimensionality m of the features extracted by the P3D model was 2,048. Due to the randomness of frame sampling, during the testing process, we repeated the sampling process 35 times and took the average of the corresponding evaluation metrics as the final experimental result.

In the training of the PSL and AQA models, the dropout probability was set to 0.5 after average pooling was performed. The learning rate was initially set to 0.0001 and decayed every 30 training steps with a decay rate of 0.1. The Adam optimizer was applied to train these models. During network optimization, L2 regularization with a weight decay value of 0.0005 was used to mitigate overfitting.

As mentioned in Section 3.2, there are two values that can be used as labels in the LRN: the overall score and the execution score for each action. The overall scores were normalized by means of max–min normalization, whereas since the execution scores ranged from 0 to 30, they were simply divided by 30 for normalization.

4.2 Ablation study

To verify the effectiveness of PSL and the pseudo-subscores for AQA, an ablation study is presented in this section. In this experiment, we used the AQA model as the baseline model, without using the PSL model or the pseudo-sub scores. To train the baseline model, only the overall score loss $loss_o$ in (6) was applied. We implemented the proposed multi-substage AQA model in two ways: training from scratch (TS) and fine-tuning from the PSL model (FTPSL). For FTPSL, the feature backbone network and overall score generation network of the PSL model were loaded to initialize the corresponding modules in the AQA model, and the SPN was trained from scratch. The experimental results are shown in Table 1, where "OSL" and "ESL" denote baseline models trained based on the overall score and execution score labels, respectively.

These experimental results show that the use of the pseudo-subscores can improve the SRC value and obviously decrease the MSE and MED values. In the original task for the UNLV-Diving dataset, the MSE values of the OSL models with PSL trained using the TS and FTPSL strategies are reduced by 70.12 and 44.34, respectively, compared to the MSE value of the OSL model without PSL, and the MED values of these two models are similarly reduced by 2.36 and 1, respectively. The SRC value of the model with PSL trained with the TS strategy is the same as that of the OSL baseline, but the proposed model trained with the TFPSL strategy shows an improvement of 0.02. We also

Table 1 Ablation experiments of the proposed method

Method	Original task			Final ranking prediction task		
	SRC	MSE	MED	SRC	MSE	MED
OSL	0.78	130.71	8.59	0.34	165.99	11.24
OSL+PSL (TS)	0.78	60.59	6.23	0.55	84.81	7.85
OSL+PSL (FTPSL)	0.80	86.37	7.59	0.61	70.76	7.06
ESL	0.87	85.24	5.66	0.74	52.68	6.60
ESL+PSL (TS)	0.85	38.99	4.98	0.81	42.53	3.59
ESL+PSL (FTPSL)	0.87	38.68	4.80	0.75	51.68	5.15

find similar results for the ESL models, for which the MSE values of the models with PSL are reduced by 46.25 and 46.56 with respect to that for the model without PSL, and the MED values are decreased by 0.68 and 0.86, respectively.

Similar results are also found for the finals ranking prediction task. For the OSL models, the MSE and MED values of the models with PSL are decreased by 81.18, 95.23, 3.39 and 4.18 relative to the baseline, and the SRC values are improved by 0.21 and 0.27. For the ESL models, the MSE and MED values of the models with PSL are decreased by 10.15, 1, 3.01 and 1.45, and the SRC values are improved by 0.07 and 0.01. All of these comparative results show that PSL and the generated pseudo-subscore labels can improve the accuracy of the ranking and score predictions of the AQA model.

In the subscore generator (label decomposition) of the PSL model, a fully connected network with layers of decreasing size is used to regress the scores. We also discuss the performance of different fully connected structures, including "2049-1024-512-64-16-1", "2049-512-64-16-1", "2049-64-16-1", "2049-16-1", and "2049-1", where each numerical value represents the number of nodes in the corresponding layer and "-" represents the fully connected operation. The experimental results are shown in Table 2. For this experiment, we again use the SRC, MSE and MED as the evaluation indicators. The second structure achieves the best performance in terms of all of these indicators and is used in the proposed method.

Table 2 Comparison of the results of different fully connected (FC) network structures

FC Structure	SRC	MSE	MED
2049-1024-512-64-16-1	0.81	96.84	6.86
2049-512-64-16-1	**0.85**	**38.99**	**4.98**
2049-64-16-1	0.81	84.64	6.53
2049-16-1	0.82	99.43	7.43
2049-1	0.84	61.64	5.81

The structure used in our paper is indicated in **bold**

4.3 Comparison with state-of-the-art methods

To verify the effectiveness of the proposed method, we compare it with state-of-the-art methods.

Comparison on the original task for the UNLV-Diving dataset We applied the proposed method to the original task for the UNLV-Diving dataset. The results are compared in Table 3.

For clarity, the existing methods can be divided into two categories: methods that use difficulty level labels and methods that do not use difficulty labels. As seen from Table 3, the best SRC, MED and MSE values of the proposed method are 0.87, 4.80, and 38.68, respectively. Compared with the existing AQA methods, the proposed

Table 3 Comparison of the results of the proposed method with the results of state-of-the-art methods

Method	DL	SRC	MED	MSE
S3D [2]	×	0.86	6.90	97.46
C3D-AVG-STL [7]	×	0.83	-	-
JRG [23]	×	0.76	-	-
AIM [9]	×	0.74	-	-
Metric Learning [6]	×	0.76	-	105.62
EAGLE-Eye [26]	×	0.83	-	-
SCN+ATCN [31]	×	0.85	-	-
USDL [32]	×	0.81	-	-
Adaptive [33]	×	0.83	-	-
C3D+SVR [1]	✓	0.74	-	-
C3D+CNN [4]	✓	0.80	7.78	-
ScoringNet [25]	✓	0.84	5.36	-
FALCONS [8]	✓	0.85	-	-
MRSM (ESL) [29]	✓	0.88	-	73.92
Ours (OSL+TS)	×	0.78	6.23	60.59
Ours (OSL+FTPSL)	×	0.80	7.59	86.37
Ours (ESL+TS)	✓	0.85	4.98	38.99
Ours (ESL+FTPSL)	✓	**0.87**	**4.80**	**38.68**

DL: difficulty level

The bold entries represents the best result

Table 4 Comparison of the proposed method with S3D on the finals ranking prediction task

Method	SRC	MSE	MED	A-SRC	A-MSE	A-MED
S3D (full video)	0.59	101.55	8.45	0.63	4179.87	59.12
S3D (entry substage)	0.61	78.00	8.12	0.78	4381.17	62.56
Ours (OSL+TS)	0.55	84.81	7.85	**0.84**	1612.82	36.60
Ours (OSL+FTPSL)	0.61	70.76	7.06	0.75	1352.71	33.06
Ours (ESL+TS)	**0.81**	**42.53**	**3.59**	0.81	**676.80**	**17.78**
Ours (ESL+FTPSL)	0.75	51.68	5.18	**0.84**	748.62	23.57

The bold entries represents the best result

method achieves state-of-the-art performance in terms of the SRC, MED and MSE results except that it is slightly inferior to [29] in SRC. Compared with [29], the SRC value of the proposed method is 0.01 lower, but the MSE of the proposed method is reduced by 35.24. This means that the scores estimated by the proposed method are closer to the ground truth, while the predicted score ranking results are almost equal to those of [29]. Compared with the stacking segmental P3D (S3D) method [2], the SRC value of the proposed method is improved by 0.1, and the MED and MSE values are reduced by 2.1 and 61.78, respectively, indicating that our method is more accurate in predicting both the ranking and the absolute scores.

Comparison on the finals ranking prediction task The proposed multi-substage AQA model was applied to predict the performance scores and ranking of the athletes in the finals of the men's 10-m platform at the 2012 London Olympics. To verify the effectiveness of the proposed method for this task, we compare it with S3D [2], which achieves state-of-the-art performance on the original task for the UNLV-Diving dataset. The experimental results are shown in Table 4.

In [2], S3D was implemented via two strategies: with the input sampled from the full video and with the input sampled only from the entry substage. As seen in Table 4, for the 70 test videos, the SRC value of the proposed method is 0.81, an improvement of 0.2 compared with that of S3D. The MSE value of the proposed method is 42.53, corresponding to a decrease of 35.47. Similarly, the MED value of the proposed method is reduced by 4.53. We also find similar results for the total scores and rankings of the athletes. The A-SRC value of the proposed method is 0.81, which is improved by 0.03 compared with that of S3D. Similarly, the A-MSE and A-MED values are decreased by 3,704.37 and 44.78, respectively.

To further evaluate the performance of the proposed method for the finals ranking prediction task, the predicted total score for each athlete is shown in Fig. 5. The top three athletes are also listed in this figure. In Fig. 5, the only identifiers used for the 12 athletes are their nationalities and initials. The blue line represents the predicted score, and the green line represents the true score. We find that for the top three predictions, the recall and accuracy of the proposed method are both 100%. However, the proposed method fails to predict the champion among the 12 athletes.

Fig. 5 Predicted and ground-truth scores of the 12 athletes in the finals. The abscissa presents the athlete identifiers, and the ordinate indicates the total scores
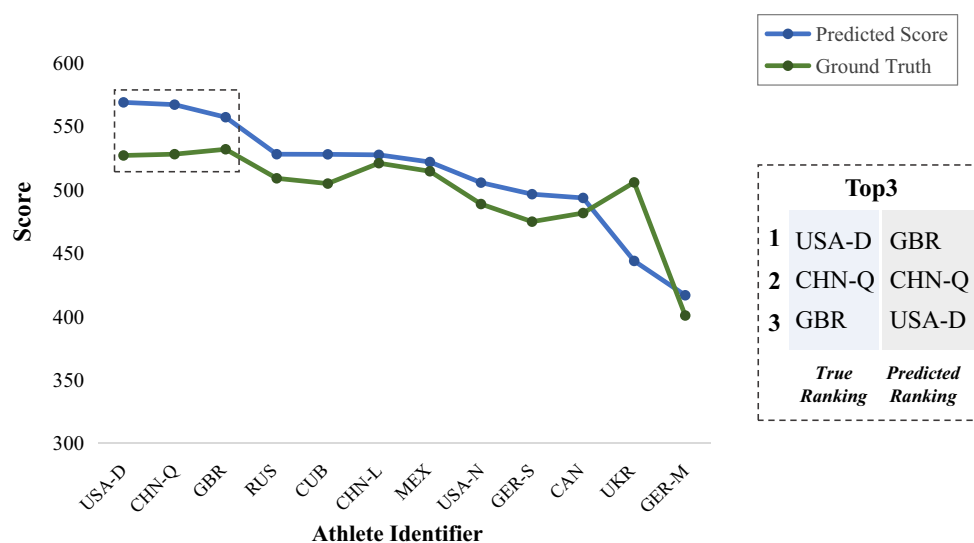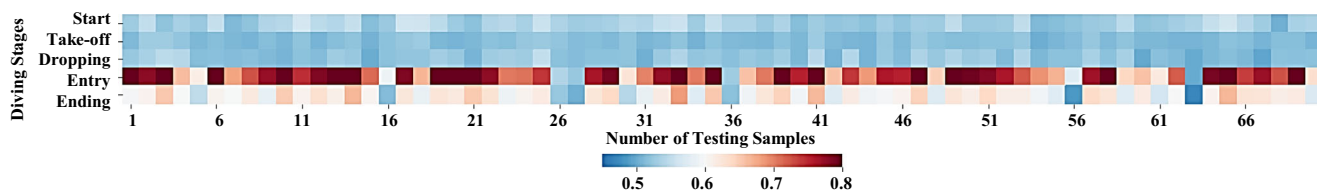
Fig. 6 Visualization of the execution scores for the individual substages of the test videos

Specifically, in the prediction results, the champion and third-place athletes are swapped. In fact, the performance of the top three athletes in the finals was very close. The accurate evaluation of these athletes requires a more fine-grained method, which will be addressed in our future work.

4.4 Substage quality assessment feedback

Another advantage of the proposed method is that it provides detailed substage motion quality feedback for each action. To analyze the performance of the proposed method for substage quality assessment, we visualize the normalized execution scores for each substage of the diving actions in the test set in Fig. 6. For this experiment, the ESL+FTPSL model was selected to output the subscores.

From Fig. 6, we find that the subscores for the first three substages of the test samples seem to be close, ranging from 0.45 to 0.6. For the last two substages, however, the distributions of the quality scores are uneven and distinguishable. These observations indicate that in diving, the performance of athletes in the entry and end stages is a key factor for assessing their overall execution scores. Accordingly, three samples with high, medium and low overall scores are selected; the ground truths for the overall scores of these samples are 92.75, 70.2 and 42.9, and the values predicted by the proposed method are 94.86, 89.25 and 50.68, respectively.

For the last two substages, representative frames and the predicted execution scores for these three samples are shown in Fig. 7. From this figure, we can see that for Sample 1, which shows a good entry posture and the smallest splash, our method gives the highest execution scores for these two substages. For Sample 3, which shows the largest splash, the proposed method produces the lowest scores. These results further verify that the proposed method can provide an accurate evaluation of the last two substages of the diving action.

In contrast, the proposed method does not seem to be able to distinguish the quality of the first three actions; this inability is manifested in the small range of scores. The reason for this may be that in diving, the human posture in the first three substages changes rapidly, and it is difficult to capture fine-grained discriminative features in these substages. Thus, we will also investigate this problem in future work.

Finally, we discuss the computation time of the proposed method. We conducted all of experiments on a device with one NVIDIA RTX 3090 graphics card. The average computation time for each video in the testing process was about 0.34 s, excluding the preliminary video segmentation stage. Although the proposed method can quickly predict the action quality scores from segmented videos, unfortunately, most temporal segmentation networks (TSNs), including the video segmentation model used in this paper, are offline tools. Therefore, this method cannot yet be used for real-time sporting events.
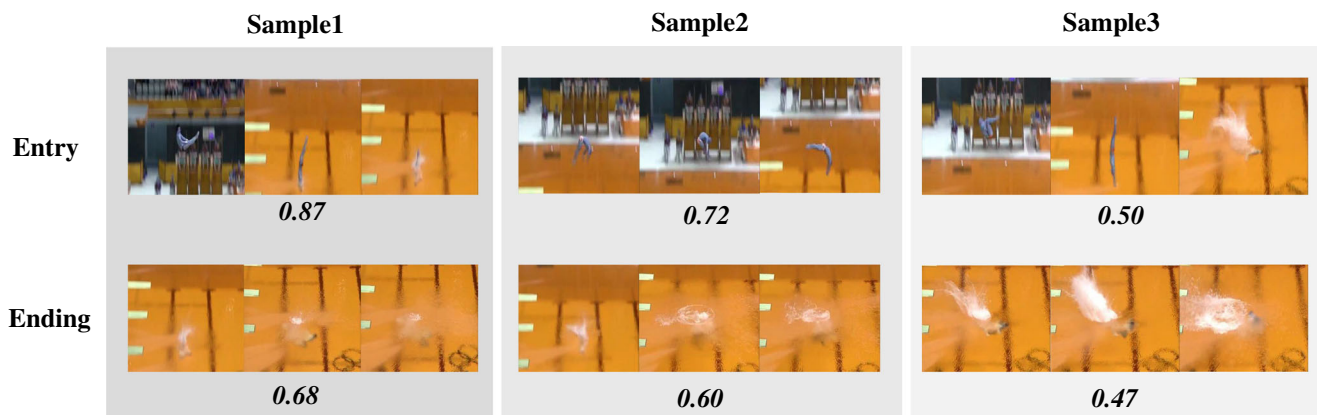

Fig. 7 Execution scores for the last two substages of selected samples

5 Conclusion

This paper has proposed a label-reconstruction-based pseudo-subscore learning (PSL) method for AQA in sporting events. In the proposed method, pseudo-subscore labels are generated for the training data. Then, a multi-substage AQA model is trained based on these pseudo-subscores in combination with the original overall score labels. Experimental results obtained on a public dataset verify the effectiveness of the proposed method.

However, there are some limitations to our work. First, in this work, we use the same TSN and substage labels as in the previous work on the AQA method known as S3D [2] to achieve video segmentation. This approach was verified to have a very high segmentation accuracy (96.6%) in [2]. However, the accuracy of the TSN will affect the performance of the proposed method. Regarding this problem, we argue that the TSN belongs to the scope of action temporal segmentation (ATS). In contrast, our work focuses on the AQA problem, and the TSN is an open and basic module in the proposed method. It can easily be replaced by the latest and most effective ATS model. In future work, we will also consider an AQA method without substage segmentation or build a multitask model to solve the TSN and AQA problems simultaneously.

Second, the frame sampling strategy leads to unstable video features. For the feature learning network, the input video should be sampled to form an image sequence of a fixed length. In most existing methods, the random sampling method is adopted. However, for the same video, the scores predicted based on different sampling results will be somewhat different. This is a common problem in 3D video feature learning. As mentioned in the experimental section, in this work, we averaged the evaluation metrics over 35 repeated experiments to obtain the final results to mitigate this problem. In future work, we will also pursue a more stable and robust video feature learning method.

Declarations

- Conflict of interest: The authors declare that they have no conflict of interest.
- Availability of data and materials: The data generated during and/or analysed during the current study is available from the first author or corresponding author on reasonable request.

References

1. Parmar P, Morris BT (2017) Learning to score olympic events. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 76–84. https://doi.org/10.1109/CVPRW.2017.16

2. Xiang X, Tian Y, Reiter A, Hager GD, Tran TD (2018) S3d: stacking segmental p3d for action quality assessment. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE, pp 928–932

3. Parmar P, Morris B (2019) Action quality assessment across multiple actions. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1468–1476

4. Li Y, Chai X, Chen X (2018) End-to-end learning for action quality assessment. In: Pacific rim conference on multimedia. Springer, pp 125–134

5. Xu C, Fu Y, Zhang B, Chen Z, Jiang Y-G, Xue X (2019) Learning to score figure skating sport videos. IEEE Trans Circuits Syst Video Technol

6. Jain H, Harit G, Sharma A (2021) Action quality assessment using siamese network-based deep metric learning. IEEE Trans Circuits Syst Video Technol 31(6):2260–2273. https://doi.org/10.1109/TCSVT.2020.3017727

7. Parmar P, Morris BT (2019) What and how well you performed? a multitask learning approach to action quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 304–313

8. Nekoui M, Omar Tito Cruz F, Cheng L (2020) Falcons: fast learner-grader for contorted poses in sports. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 900–901

9. Gao J, Zheng W-S, Pan J-H, Gao C, Wang Y, Zeng W, Lai J (2020) An asymmetric modeling for action assessment. In: European conference on computer vision. Springer, pp 222–238

10. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD (2018) Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot Comput Assist Surgery 14(1):1850

11. Wang Z, Fey AM (2018) Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int J Comput Assist Radiology Surgery 13(12):1959–1970

12. Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller P-A (2018) Evaluating surgical skills from kinematic data using convolutional neural networks. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 214–221

13. Liao Y, Vakanski A, Xian M (2020) A deep learning framework for assessing physical rehabilitation exercises. IEEE Trans Neural Syst Rehabilitation Eng 28(2):468–477

14. Blanchard N, Skinner K, Kemp A, Scheirer W, Flynn P (2019) Keep me in, coach!: a computer vision perspective on assessing acl injury risk in female athletes. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1366–1374

15. Doughty H, Damen D, Mayol-Cuevas W (2018) 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

16. Doughty H, Mayol-Cuevas W, Damen D (2019) 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

17. Shi Q, Zhang H-B, Li Z, Du J-X, Lei Q, Liu J-H (2022) Shuffle-invariant network for action recognition in videos. ACM Trans Multimed Comput Commun Appl, vol 18(3). https://doi.org/10.1145/3485665
18. Hara K, Kataoka H, Satoh Y (2017) Learning spatio-temporal features with 3d residual networks for action recognition. In: 2017 IEEE international conference on computer vision workshops (ICCVW), pp 3154–3160. https://doi.org/10.1109/ICCVW.2017.373
19. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 6546–6555. https://doi.org/10.1109/CVPR.2018.00685
20. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 4724–4733. https://doi.org/10.1109/CVPR.2017.502
21. Qiu Z, Yao T, Mei T (2017) Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE international conference on computer vision, pp 5533–5541
22. Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6202–6211
23. Pan J-H, Gao J, Zheng W-S (2019) Action assessment by joint relation graphs. In: Proceedings of the IEEE international conference on computer vision, pp 6331–6340
24. Parmar P, Reddy J, Morris B (2021)
25. Li Y, Chai X, Chen X (2018) Scoringnet: learning key fragment for action quality assessment with ranking loss in skilled sports. In: Computer vision – ACCV 2018, pp 149–164
26. Nekoui M, Cruz FOT, Cheng L (2021) Eagle-eye: extreme-pose action grader using detail bird's-eye view. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 394–402
27. Zeng L-A, Hong F-T, Zheng W-S, Yu Q-Z, Zeng W, Wang Y-W, Lai J-H (2020) Hybrid dynamic-static context-aware attention network for action assessment in long videos. In: Proceedings of the 28th ACM international conference on multimedia, pp 2526–2534
28. Tang Y, Ni Z, Zhou J, Zhang D, Lu J, Wu Y, Zhou J (2020) Uncertainty-aware score distribution learning for action quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9839–9848
29. Dong L-J, Zhang H-B, Shi Q, Lei Q, Du J-X, Gao S (2021) Learning and fusing multiple hidden substages for action quality assessment. Knowl-Based Syst 229:107388. https://doi.org/10.1016/j.knosys.2021.107388
30. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD (2017) Temporal convolutional networks for action segmentation and detection. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1003–1012. https://doi.org/10.1109/CVPR.2017.113
31. Wang J, Du Z, Li A, Wang Y (2020) Assessing action quality via attentive spatio-temporal convolutional networks. In: Peng Y, Liu Q, Lu H, Sun Z, Liu C, Chen X, Zha H, Yang J (eds) Pattern recognition and computer vision. Springer, Cham, pp 3–16
32. Tang Y, Ni Z, Zhou J, Zhang D, Lu J, Wu Y, Zhou J (2020) Uncertainty-aware score distribution learning for action quality assessment. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9836–9845. https://doi.org/10.1109/CVPR42600.2020.00986
33. Pan J-H, Gao J, Zheng W-S (2021) Adaptive action assessment. IEEE Trans Pattern Anal Mach Intell:1–1. https://doi.org/10.1109/TPAMI.2021.3126534

Hong-Bo Zhang received a Ph.D. in Computer Science from Xiamen University in 2013. Currently, he is an associate professor with the School of Computer Science and Technology of Huaqiao University. He is the member of Fujian key laboratory of big data intelligence and security. His research interests include computer vision and pattern recognition.

Li-Jia Dong received the B.S. degree from Dalian Minzu University, China, in 2019. Now, she is currently pursuing the M.S. degree at Huaqiao University. Her research interests include image processing and computer vision.

Qing Lei received a Ph.D. from the Cognitive Science Department of Xiamen University, China. She joined the faculty of Huaqiao University in 2005. Her research interests include human motion analysis and object detection/recognition.

Li-Jie Yang received a Ph.D. in software engineering from the University of Macau. Currently, she is an associate professor with the school of Computer Science and Technology, Huaqiao University, Xiamen, China. Her research interests include computer graphics and computer animation.

Ji-Xiang Du received a Ph.D. in Pattern Recognition and Intelligent System from the University of Science and Technology of China (USTC), Hefei, China in 2005. He is currently a professor at the School of Computer Science and Technology at Huaqiao University. He is the director of Fujian key laboratory of big data intelligence and security. His current research interests mainly include pattern recognition and machine learning.

Affiliations

Hong-Bo Zhang[1,2] · **Li-Jia Dong**[1,3] · **Qing Lei**[2] · **Li-Jie Yang**[3] · **Ji-Xiang Du**[3] 🆔

Hong-Bo Zhang
zhanghongbo@hqu.edu.cn

Li-Jia Dong
donglijia@stu.hqu.edu.cn

Qing Lei
leiqing@hqu.edu.cn

Li-Jie Yang
yanglijie@hqu.edu.cn

[1] Department of Computer Science and Technology,
Huaqiao University, Jimei, Xiamen, 361000, Fujian, China

[2] Xiamen Key Laboratory of Computer Vision and Pattern
Recognition, Huaqiao University, Jimei, Xiamen, 361000,
Fujian, China

[3] Fujian Key Laboratory of Big Data Intelligence and Security,
Huaqiao University, Jimei, Xiamen, 361000, Fujian, China