Pairwise Contrastive Learning Network for Action Quality Assessment

Mingzhe Li[1], Hong-Bo Zhang[1(✉)], Qing Lei[2], Zongwen Fan[2], Jinghua Liu[3], and Ji-Xiang Du[3]

[1] College of Computer Science and Technology, Huaqiao University, Xiamen, China
limingzhe@stu.hqu.edu.cn, zhanghongbo@hqu.edu.cn
[2] Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Xiamen, China
{leiqing, zongwen.fan}@hqu.edu.cn
[3] Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen, China
{liujinghua, jxdu}@hqu.edu.cn

Abstract. Considering the complexity of modeling diverse actions of athletes, action quality assessment (AQA) in sports is a challenging task. A common solution is to tackle this problem as a regression task that map the input video to the final score provided by referees. However, it ignores the subtle and critical difference between videos. To address this problem, a new pairwise contrastive learning network (PCLN) is proposed to concern these differences and form an end-to-end AQA model with basic regression network. Specifically, the PCLN encodes video pairs to learn relative scores between videos to improve the performance of basic regression network. Furthermore, a new consistency constraint is defined to guide the training of the proposed AQA model. In the testing phase, only the basic regression network is employed, which makes the proposed method simple but high accuracy. The proposed method is verified on the AQA-7 and MTL-AQA datasets. Several ablation studies are built to verify the effectiveness of each component in the proposed method. The experimental results show that the proposed method achieves the state-of-the-art performance.

Keywords: Action Quality Assessment, Pairwise Contrastive Learning Network, Consistency Constraint, Video Pair, Relative Score

1 Introduction

Action quality assessment (AQA) is the task of evaluating how well an action is performed. The potential value of AQA has been gradually explored in many real-world scenarios. For instance, in sports scoring [32,35,7,14,30,40,18,21], daily skill evaluation [8,31,9,25,4], and medical rehabilitation [15,36,29,16,34]. In addition, with the rapid development of the computer vision community, AQA methods are constantly emerging and improving, and the scenarios of AQA application are gradually enriched. In this work, we focus on the problem of AQA in sports events.

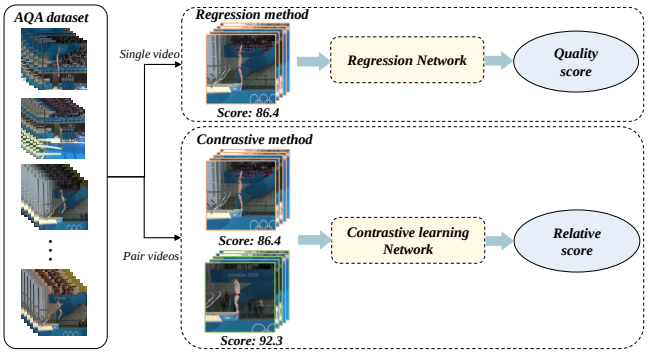

Fig. 1. An overview comparison of the contrastive learning with regression learning. In contrastive learning, the regression model is transformed into comparing the performance of a given video with another one. And the output is the relative score between video pair

As a common solution for AQA in sports events, the score regression methods are utilized to map the input videos to quality scores [7,27,35,26]. However, this strategy ignored the subtle difference between various videos, which is the key reminder to predict action quality score. For example, in diving competition, due to the same scene and similar appearance of the athletes, the difference of an action performed by athletes is hard to discern. In addition, even the same score will also appear on actions with different difficulty degrees. It is very difficult to represent such complex changes through single regression learning. Thus, how to achieve better performance for quality score prediction in sports remains a challenge worth exploring.

An intuitive idea to solve this problem is to train a specific model to learn the difference between videos. To achieve this goal, we rethink the problem of AQA and observe that the performance ranking of different athletes plays a crucial role in AQA task. Inspired by the idea of pairwise learning to rank (LTR) task that compares each pair of data to obtain the ranking, we extend the AQA problem to the contrastive learning problem of video pairs, as shown in Fig. 1. In the proposed contrastive learning strategy, the video pair is applied as input, the relative score between these two videos is applied as the label. And a new pairwise contrastive learning network (PCLN) is designed to learn the mapping from the video pair to the relative score.

Furthermore, PCLN is fused with a basic regression network to form an end-to-end AQA model. In this work, the basic regression network is built by feature extraction network followed by multilayer perceptron (MLP). In the training stage, video pair is randomly composited from the training set. The features of these two videos are obtained by the feature extraction network. Then these extracted features are fed into MLP to predict the quality score. In PCLN, a regression network including the feature fusion module, convolution module and fully connected layers, is applied to predict the relative score. In order to train

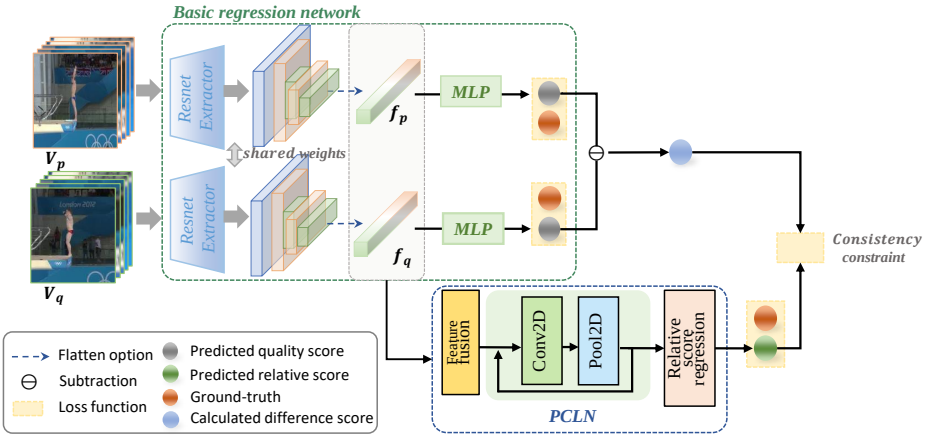

Fig. 2. Pipeline of our proposed model. The video pair is fed into the feature extractor as the input, then the score regressor is used to predict the score of these two videos, and PCLN is designed to learn the mapping from the video pair to the relative score

the proposed model, three constraints are designed in this work: minimization of the error between predicted score and ground truth, minimization of the error between predicted relative score and ground truth, a new consistency constraint between basic regression network and PCLN. The overall framework of the proposed method is shown in Fig. 2.

More importantly, although the amount of computation of the proposed method in the training phase is larger than that of the basic regression network, in the testing phase, only basic regression network is employed to predict the quality score of the input video, and the PCLN module is not necessary. The computational complexity of the testing phase does not increase. Another important advantage behind the proposed method is that the combination of any two videos can expand the number and diversity of samples in the training process, resulting in better accuracy and generalization. In summary, the contributions of this work are listed as follows:

(1) We extend the quality score regression to relative score prediction and propose a new end-to-end AQA model to enhance the performance of the basic regression network. The basic regression network and PCLN are combined during the training, but in the testing, only basic regression network is employed, which makes the proposed method simple but high accuracy.

(2) A novel pairwise LTR-based model PCLN is proposed to concern the subtle difference between videos. A new consistency constraint between PCLN and basic regression network is defined.

(3) The experimental results based on the public datasets show that the proposed method achieves the better performance compared with existing methods. Ablation experiments are also conducted to verify the effectiveness of the each component of proposed method.

2 Related Work

The purpose of AQA task is to automatically evaluate the quality of an action. In recent years, many AQA methods have been proposed and it made rapid progress in the computer vision community. Most of the previous studies used the mainstream regression method to solve AQA problem, and some works has begun to use contrastive learning methods.

In the regression-based methods, there are two kinds of methods according to the form of input data: skeleton-based methods and appearance-based methods. In skeleton-based methods, Pirsiavash et al. [27] proposed a framework for learning spatio-temporal features from human skeleton joint sequences, which extracted action features using discrete cosine transform (DCT) and predicted scores using linear support vector regression (SVR). Pan et al. [21] processed specific joint action information according to the relationship between joint locations, combined joint common module and joint difference module for human joint action learning. More recently, they continued to propose an adaptive method [20], which adaptively designed different assessment architectures for different types of actions.

Moreover, many attempts devote to acquire more detailed appearance based information to improve the assessment performance. For example, Parmar and Tran Morris [26] utilized C3D network to acquire spatio-temporal features and performed score regression using the SVR and LSTM. Later, they built an AQA dataset named MTL-AQA [24], and devoted to exploit the representation of the action and its quality to improve the performance of AQA model. Xiang et al. [37] proposed to apply P3D on each stage of diving video and then fuse the stage-wise features into P3D to regress the subscores. Tang et al. [32] proposed an uncertainty-aware score distribution learning approach, which described score as probability distribution to predict quality score. Dong et al. [3] proposed a learning and fusion network of multiple hidden substages to assess athlete performance by segmenting videos into five substages.

Furthermore, several methods defined the AQA problem as a pairwise ranking formulation. Doughty et al. [4] formulated the problem as pairwise and overall ranking of video collections, and proposed a supervised deep ranking method. They also trained [5] a rank-specific temporal attention modules, which processed higher and lower skills parts separately. Yu et al. [39] developed a group-aware regression tree to replace the traditional score regression. In addition, different from the methods which only used the vision features to assess action quality, in [32,39], the extra referee information is added to improve AQA results. Jain et al. [11] proposed a binary classification network to learn the similarities between videos. After that, it was transferred to the score regression task. In the score regression network, each input video and expert video, which has the highest score in the dataset, were combined as input for model training and score prediction. However, only selecting single expert sample as the reference is difficult to model the diversity of action scores and video differences in AQA task. This strategy adopted two-stage approach, which cannot ensure that the parameters learned from the binary classification task were applicable to the score regres-

sion network. Different from the above methods, in this work, we propose a new approach to learn the difference between video pairs, and build an end-to-end model to improve the performance of the basic regression network.

3 Approach

In this section, we introduce the proposed AQA method for sports events in details, including the feature extractor, score regressor and PCLN module.

3.1 Problem Formulation

Given a sport video $V = \{v_i\}_{i=1}^{L}$, where v_i represents the i-th frame in a video and L is the length of input video, the goal of AQA is to automatically generate the score based on the performance of athlete. It can be defined as:

$$S = \Theta(V) \tag{1}$$

where $\Theta(\cdot)$ represents the score prediction function, and S represents the predicted action quality score.

The goal of the proposed method is to find a more effective regression function $\Theta(\cdot)$. As shown in Fig. 2, in the training process of the proposed method, a pair of videos are applied as input and a multitask framework is proposed. In addition to learning the quality score of each video, we also learn the relative score between the input two videos. The AQA problem can be reformulated as:

$$[S_p, S_q, \Delta S] = \Upsilon(V_p, V_q) \tag{2}$$

where V_p and V_q represent the pair of input videos, $\Upsilon(\cdot, \cdot)$ represents the proposed algorithm, S_p and S_q represent the predicted quality score of the corresponding video respectively, and ΔS represents the predicted relative score between two videos.

In the proposed method, PCLN is built to learn the difference between videos and provide a more accurate evaluation result. To train PCLN, a video pair $< V_p, V_q >$ is generated from the original video dataset $V = \{V_1, V_2, ..., V_n\}$ by a combinations way. As shown in Eq. 3, the total number of video pairs can be reached $C(n, 2)$:

$$C(n, 2) = \frac{n!}{2! * (n-2)!} = \frac{n * (n-1)}{2} \tag{3}$$

where n represents the number of videos in the training set.

3.2 Feature Extraction and Score Regression

Given a pair of input videos $< V_p, V_q >$, the effective vision features should be extracted first. There are generally two types of methods to extract the features: 3D convolution-based methods and temporal encoder-based methods. Limited

by the computational scale, 3D convolution-based methods [1,33] usually require to sample a short clip with fixed-length from the video. However, due to the randomness of the sampling strategy, the features extracted from the sampled short clip are unstable, so that the score prediction will fluctuate a lot. Therefore, we use the latter strategy to compute video features in this work.

In temporal encoder-based methods, the image feature of each video frame is extracted by the feature backbone network. In this work, we use ResNet [10] model as the feature extractor. After that, a temporal encoder network [13] is applied to encode the temporal information of the feature sequence. By doing so, the higher-level and stable video feature are obtained. The encoder network is comprised of two stacked encoding blocks, and each encoding block is composed of the 1×1 temporal convolution, specific activation function and maxpooling layer across temporal series. This feature extraction process can be defined as Eq. 4.

$$f_i = \mathbb{E}(\mathcal{F}(V_i)), i = p, q \tag{4}$$

where f_p and f_q represent the features of the input videos $< V_p, V_q >$ respectively, $\mathcal{F}(\cdot)$ represents the ResNet model and $\mathbb{E}(\cdot)$ represents the temporal encoder network.

Finally, a fully connected (FC) network is designed to regress the action quality score and form a basic regression network. Referring to the previous works [22,24,3], the FC network contains three fully connected layers: $D \times 4096$, 4096×2048 and 2048×1, where D is the dimension of the video feature. Based on the above definition, the basic regression network can be defined as:

$$S_i = \Theta(V_i) = \mathcal{R}(f_i) = \mathcal{R}(\mathbb{E}_c(\mathcal{F}(V_i)), i = p, q \tag{5}$$

where $\mathcal{R}(\cdot)$ represents the score regressor.

3.3 PCLN Model

As most of the same sport events are competed in similar environment, the differences between the same competition videos are often very subtle, and there are slight differences in how the athletes perform on the same actions. For example, in the diving competition, the referees primarily pay attention to the size of the splash, the degree of the athlete's leg bending, the execution standard of the action and so on. Although these factors are difficult to observe, they greatly affect the accuracy of scoring. In order to learn the differences between videos to assist the final scoring task, we build a separate branch for the pairwise video based LTR network named PCLN to learn the relative scores. The detailed structure of PCLN model is shown in Fig. 3.

First, 1D convolutional layer is carried out for each encoded video feature in temporal. It can further encode features to capture higher level action information. Then the feature matrices of the two videos are connected by matrix

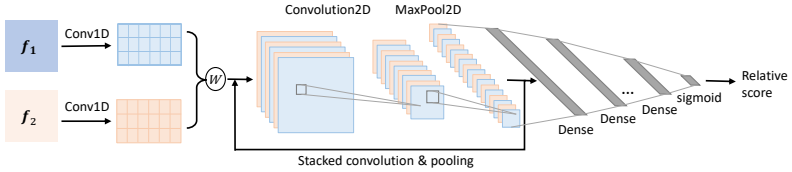

Fig. 3. Structure of PCLN model. The temporal encoded features f_1 and f_2 are applied as inputs

multiplication. To be detailed, we form the fusion process as follows:

$$\begin{aligned} f'_p &= \sigma(w_{(0)} \otimes f_p + b_{(0)}), \\ f'_q &= \sigma(w_{(0)} \otimes f_q + b_{(0)}), \\ f_{(0)} &= f'_p \circ f'_q \end{aligned} \tag{6}$$

where f'_p and f'_q represent the output feature map of 1D convolutional layer, $w_{(0)}$ is the parameters in this layer, $\otimes$ means convolution operator and b is the corresponding bias vector, $\sigma(\cdot)$ represents the ReLU activation function, $f_{(0)}$ represents the connected matrix, $\circ$ means matrix multiplication operator.

Secondly, stacked 2D convolutional and pooling layers are performed on the connected matrix, then a high-level representation of the interaction between the two videos can be obtained. Finally, we use a general MLP module to predict the relative score of the pair videos. There are four layers in this MLP module, and the number of nodes in each layer is 64, 32, 8 and 1 subsequently. The calculation process of PCLN can be defined as follows:

$$\begin{aligned} f'_{(i)} &= \sigma(w_{(i)} \otimes f_{(i-1)} + b_{(i)}), \\ f_{(i)} &= Mp(f'_{(i)}), \\ \Delta S &= \mathcal{R}_d(f_{(c)}) \end{aligned} \tag{7}$$

where $f'_{(i)}$ represent the output of 2D convolutional operation in i-th layer, $w_{(i)}$ and $b_{(i)}$ are the parameters and bias vector in i-th layer, and $i = 1, ..., c$, c is the number of layers in the PCLN, it is set to 2 in the experiment. $Mp(\cdot)$ represents the maxpooling operation, and $\mathcal{R}_d(\cdot)$ represents the MLP module for relative score.

3.4 Module Training and Inference

To train the proposed AQA model, we formulate three constraints to learn effective relative scores between different videos and accurate athlete quality scores simultaneously. First, the fundamental requirement of the AQA task is to obtain accurate quality scores, which requires minimizing the error of the predicted score of the input video pair. Therefore, for each video pair $< V_p, V_q >$,

$< \tilde{S}_p, \tilde{S}_q >$ represents the ground truth of action quality score, the loss function of the basic regression network is defined as:

$$\mathcal{L}_{bs} = \frac{1}{2} \sum_{i=p,q}^{N} (\tilde{S}_i - S_i)^2 \tag{8}$$

Similarly, it also needs to minimize the error between the predicted relative score and the corresponding ground truth. In this task, the absolute value of two score labels between the input video pair is applied as ground truth. Therefore, the loss function of PCLN model can be defined as:

$$\mathcal{L}_{ds} = (\Delta S - |\tilde{S}_p - \tilde{S}_q|)^2 \tag{9}$$

Furthermore, a consistency constraint is defined for basic regression network and PCLN to improve the performance of the proposed AQA model. This consistency constraint confines the PCLN predicted relative score is equal to the calculated difference score from the two quality scores predicted by basic regression network. Therefore, a consistency loss function is defined as:

$$\mathcal{L}_{rs} = (\Delta S - |S_p - S_q|)^2 \tag{10}$$

Finally, the overall loss function of the proposed AQA model can be summarized as:

$$\mathcal{L} = \mathcal{L}_{bs} + \mathcal{L}_{ds} + \mathcal{L}_{rs} \tag{11}$$

In the testing phase, we only utilize the basic regression network, which includes the feature extractor, temporal encoder and regression network, to predict the quality score. No matter how many branches and constraints we add to the model during training phase, the proposed framework can still guarantee lower complexity during testing, which is different from the previous AQA studies.

4 Experimental Results and Discussion

The proposed method is evaluated on the AQA-7 [22] and MTL-AQA [24] datasets. Ablation study is applied to verify the effectiveness of each component of the method.

4.1 Datasets and Evaluation Metric

AQA-7 Dataset. The AQA-7 dataset includes 1189 videos from 7 sports captured in Summer and Winter Olympics: 370 videos from single 10m diving, 176 videos from gymnastic vault, 175 videos from big air skiing, 206 videos from big air snowboarding, 88 videos from synchronous 3m diving, 91 videos from synchronous 10m diving, and 83 videos from trampoline. All of the videos in this dataset have a fixed length of 103 frames except that trampoline videos are much longer than other sports. Thus, in this experiment, the trampoline videos are excluded according to the setting in [22]. In addition, AQA-7 dataset only

provides the final score of each video as the label. The split of training set and testing set follows the official setting.

MTL-AQA Dataset. The MTL-AQA dataset is the largest diving dataset released in 2019. It contains 1412 diving videos collected from 16 different events. In addition, the dataset includes the 10m platform and 3m springboard, both male and female athletes. Different kinds of labels are provided for each video in the dataset, such as final score, difficulty degree and execution score given by the referees. We follow the split setting as [24] suggested: 1059 videos are used for training, while 353 videos are used for testing.

Evaluation Metric. To be comparable with the existing AQA methods [26,27], the Spearman's rank correlation (SRC) is adopted to measure the rank correlation between ground-truth and predicted results. The higher the SRC, the better. The calculation can be expressed as:

$$\rho = \frac{\sum_i (h_i - h)(k_i - k)}{\sqrt{\sum_i (h_i - \bar{h})^2 \sum_i (k_i - \bar{k})^2}} \tag{12}$$

where h and k denote the rankings of the two sequences respectively. We use Fisher's z-value [6] to measure the average correlation coefficient across actions as previous work [32,39,20].

4.2 Implementation Details

We implement the proposed model using PyTorch, and it is trained on single Nvidia RTX 3090 GPU. ResNet-50 pretrained on ImageNet [2] is used as image feature extractor. The proposed model is trained for 200 epochs. Adam [12] optimizer with initial learning rate of 0.0001 is applied and the decay rate is set as 0.5. In the experiments on AQA-7 and MTL-AQA, all the video frames are resized to 224×224, and each video contains 103 frames as [24,22] suggested.

In diving sports, the final score is generated by multiplying the execution score and the difficulty degree, and the execution score is the average score provided by judges [32,3]. In MTL-AQA dataset, since there are difficulty degree and execution score labels for each video, we implement the proposed method in two scenarios: execution score prediction (ESP) and final score prediction (FSP). In ESP scenario, execution score is used as the training label, and in the inference stage, the predicted execution score is multiplied by the difficulty degree to obtain the final score. In the FSP scenario, the final score is predicted directly. In all experiments, the SRC of the final score is used as the evaluation metric. In addition, the final scores in both datasets are normalized with min-max normalization operation, and the execution score in MTL-AQA is divided by 30 for normalization since the range of execution score is from 0 to 30. Code is available at `https://github.com/hqu-cst-mmc/PCLN`.

4.3 Results on AQA-7 Dataset

Comparison with the state-of-the-art methods. In order to evaluate the effectiveness of the proposed method, it is compared with the existing AQA ap-

Table 1. Comparison results of the proposed method with the state-of-the-art methods. "-" means that the result did not provide in the literature

Network	Year	Diving	Gym Vault	Skiing	Snow board	Sync. 3m	Sync. 10m	Avg. SRC
Pose+DCT [27]	2014	0.5300	-	-	-	-	-	-
ST-GCN [38]	2018	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-LSTM [22]	2019	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
C3D-SVR [22]	2019	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
JRG [21]	2019	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL [32]	2020	0.8099	0.7570	0.6538	**0.7109**	0.9166	0.8878	0.8102
DML [11]	2020	0.6900	0.4400	-	-	-	-	-
FALCONS [19]	2020	0.8453	-	-	-	-	-	-
HalluciNet [23]	2021	0.8351	-	-	-	-	-	-
EAGLE-Eye [18]	2021	0.8331	0.7411	0.6635	0.6447	0.9143	0.9158	0.8140
Adaptive [20]	2021	0.8306	0.7593	0.7208	0.6940	0.9588	0.9298	0.8500
Ours	2022	**0.8697**	**0.8759**	**0.7754**	0.5778	**0.9629**	**0.9541**	**0.8795**

proaches. Unlike the works [27,11,3,19] that only performed experiments on part of sports to verify the robustness of the proposed method, we conduct experiments based on all the sports in the AQA-7 dataset. The experimental results are shown in Table 1. Obviously, the proposed method achieves the state-of-the-art performance in the terms of average SRC. In addition, the proposed method obtains significant improvement for all action classes except Snowboard compared to the recent USDL approach [32], which utilizes label distribution learning to replace original regression task. Another recent work Adaptive [20] reached a SRC of 0.85 for the average performance but the approach relied on human skeleton data and required different assessment architectures for different categories of sports in AQA. Compared with these two recent works, the average SRC of the proposed approach gains improvement of 0.0693 and 0.029 respectively, which clearly verifies the effectiveness of the proposed method in AQA problem.

Influence of feature extractor. As mentioned in Section 3, there are two methods to extract video features: 3D convolution-based (3DCNN) method and temporal encoder-based method. In order to discuss their performance for AQA task, we apply P3D [28] as the feature extraction network to replace the temporal encoder of basic regression network to build a 3DCNN regression network. And the comparison results can be found in Fig. 4. In the experiment, both of these methods are trained from the same training set, and one sample is randomly selected for testing. This sample is evaluated 10 times by two models respectively.

In Fig. 4, the green line is the predicted results using 3DCNN regression network, the yellow line is the results of basic regression network and the gray line is the ground truth, which is 83.25 for the sample "diving_007". From the comparison results, it can be observed that the predicted score of basic regression network is fixed 83.07 for 10 times testing, while the results of 3DCNN regression network is unstable. The main reason is that a random sampling process is

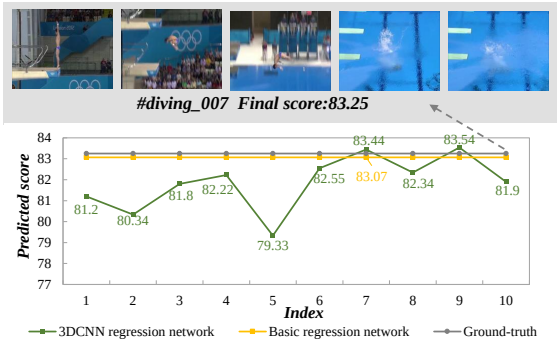

Fig. 4. Comparison results of temporal feature encoder and 3D convolution network

Table 2. Comparison results of the proposed method with different video pair number

Batch Size	Video Pairs	Diving	Gym Vault	Skiing	Snow board	Sync. 3m	Sync. 10m	Avg. SRC
8	28	0.8458	0.7723	0.7311	0.5285	0.9255	0.9317	0.8276
16	120	0.8597	0.8185	0.7455	0.5750	0.9639	0.9412	0.8621
32	496	**0.8697**	**0.8759**	**0.7754**	**0.5778**	**0.9629**	**0.9541**	**0.8795**
64	2016	0.8189	0.8518	0.7295	0.4927	-	-	-

required before feature calculation in 3DCNN, and the sampling results will affect the predicted results of the model. Therefore, in this work, we employ temporal encoder-based method to extract video features.

Influence of the number of video pairs. We apply the 2-combinations method to expand the dataset. Suppose that n is the number of the selected video in each iteration, which is the batch size in the training process. According to Eq. 3, when n is set to 8, 16, 32 and 64, the number of video pairs in each iteration is 28, 120, 496 and 2016 correspondingly. In order to verify the influence of the number of video pairs, the predicted score with different batch size is reported in Table 2. The experimental results show that the number of video pairs has a positive impact on the accuracy of assessment. It means the more video pairs used, the more effectively difference information between videos can be obtained. When n is set to 32, our proposed method achieves the highest average SRC of 0.8795, and the performance in each category is the best. Since the number of training set in "sync. 3m" and "sync. 10m" are less than 64, the experiment cannot be carried out when n is set to 64.

Ablation study for exploring the effectiveness of each module. To verify the effectiveness of each proposed component in this work, an ablation study is performed on AQA-7 dataset. The proposed method is composed of basic regression network, PCLN and consistency constraint. We discuss the contribu-

Table 3. Ablation study of different component in the proposed method

Basic Regression Network	PCLN	Consistency Constraint	Diving	Gym Vault	Skiing	Snow board	Sync. 3m	Sync. 10m	Avg. SRC
✓	×	×	0.8604	0.8156	0.7314	0.5755	0.9432	0.9417	0.8504
✓	✓	×	0.8656	0.8701	0.7624	0.5759	0.9547	0.9530	0.8721
✓	✓	✓	**0.8697**	**0.8759**	**0.7754**	**0.5778**	**0.9629**	**0.9541**	**0.8795**

Table 4. Comparison of our approach with existing methods on the MTL-AQA dataset

Methods	Year	Sp. Corr.
Pose+DCT* [27]	2014	0.2682
C3D-SVR* [26]	2017	0.7716
C3D-LSTM* [26]	2017	0.8489
MSCADC-STL [24]	2019	0.8472
MSCADC-MTL [24]	2019	0.8612
C3D-AVG-STL [24]	2019	0.8960
C3D-AVG-MTL [24]	2019	0.9044
USDL [32]	2020	0.9066
C3D-AVG-SA&HMreg [17]	2021	0.8970
Ours(FSP)	2022	0.8798
Ours(ESP)	2022	0.9230

* These results were taken from [24].

tions of PCLN and consistency constraint under basic regression network. In this experiment, the batch size is set 32. The experimental results are shown in Table 3. Without PCLN and consistency constraint, the average SRC of basic regression network is 0.8504. PCLN can improve the average SRC by 0.0217, while consistency constraint brings a 0.0174 improvement in average SRC based on PCLN. From these results, it can be observed that PCLN and consistency constraint are effective to obviously improve the performance of basic regression network.

4.4 Results on MTL-AQA Dataset

Comparison with state-of-the-art methods. In order to further verify the robustness and effectiveness of the proposed method, we extend the same experiment on MTL-AQA dataset. Table 4 shows the comparison results of the proposed method with the existing methods. Since the difficulty degree and execution score given by referees are available in this dataset, we further conduct experiments under two different scenarios (ESP and FSP) to verify the effectiveness of the proposed approach. The experimental results show that the proposed model in ESP scenario achieves the best SRC 0.923. In addition, it can be clearly observed that compared with FSP, using execution score as the label is more conducive to the learning of the proposed method.

Table 5. Comparison results of different video pair number on the MTL-AQA dataset

Batch Size	Video Pairs	Score Label	
		FSP	ESP
8	28	0.8729	0.9094
16	120	0.8750	0.9118
32	496	0.8777	0.9188
64	2016	**0.8798**	**0.9230**

Table 6. Ablation study on different assessment structures on the MTL-AQA dataset, all of the models use 2016 pairs of video

Basic regression network	PCLN	Consistency constraint	Score Label	
			FSP	ESP
✓	×	×	0.8745	0.9095
✓	✓	×	0.8788	0.9196
✓	✓	✓	**0.8798**	**0.9230**

Influence of the number of video pairs. Similarly, we further verified the influence of the number of video pairs on MTL-AQA dataset, and the results can be found in Table 5. The experimental results show that when n is set to 64 (i.e., the number of video pairs in each batch is 2016), the proposed method achieved the highest SRC of 0.8798 and 0.9230 in FSP and ESP scenarios respectively. With the increase number of video pairs, the SRC value increases gradually in both scenarios. Therefore, in all of the ablation experiments on MTL-AQA dataset, the batch size is set to 64.

Ablation study for exploring the effectiveness of each module. We also conduct the ablation study on the MTL-AQA dataset to further verify the effectiveness of each component in the proposed method. The experimental results are shown in Table 6. From these results, we get similar conclusions with the previous experiment based on AQA-7 dataset. When adding PCLN module with the basic regression network, the SRC value is improved by 0.0043 and 0.0101 in FSP and ESP scenarios, respectively. When consistency constraint is employed, the SRC value achieves 0.8798 and 0.923. Especially, compared with the basic regression network in ESP scenario, it is improved by 0.0135. These experimental results show that our proposed PCLN and consistency constraint can also improve the performance of basic regression network. This can further verify the effectiveness and robustness of the proposed model for AQA task.

4.5 Qualitative Evaluation.

In Fig. 5, we show some exemplars of predicted scores by different methods in details to quantitatively analyze the effectiveness of the proposed method. The first video pair shows two actions that both use “Tuck Position” but the

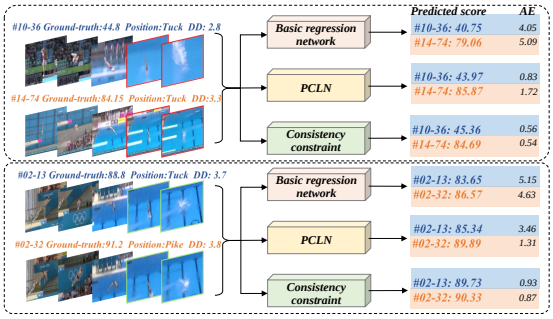

Fig. 5. Predicted results of samples. DD represents the difficulty degree. AE represents the absolute error between the predicted score and ground truth

difficulty degree is different. There is a large gap between the final scores of the two videos because of the different performance of the athletes, especially the splash size in the red border frames. In the other case, these two athletes perform different actions "Tuck Position" and "Pike Position" with very close difficulty degree. But both execution scores of these two actions are equal to 24. From the comparison results in Fig. 5, it can be clearly observed that the absolute error of each module is gradually decrease, and the proposed method gives the predicted score that is closer to the ground truth. These results can further verify the effectiveness of the proposed method.

5 Conclusions

In this paper, we propose a new contrastive learning model for AQA, which is capable of exploring the subtle difference in sports videos. In the proposed method, we adopted a more stable feature extraction strategy, a basic regression network and PCLN module were applied to predict the quality score and relative score simultaneously. Moreover, a consistency constraint was defined to train the proposed method. The experimental results showed that the proposed method has achieved the state-of-the-arts performance. However, in this work, we only use very simple score error to calculate the loss of predicted results. In the future, more assessment information, such as the scoring pattern of each referee, can be applied to improve the accuracy of the proposed method in AQA tasks.

Acknowledgement

This work was supported by the Natural Science Foundation of China (No. 61871196, 62001176); Natural Science Foundation of Fujian Province of China (No. 2019J01082, 2020J01085, 2022J01317); Scientific Research Funds of Huaqiao University (No. 21BS122) and the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (No. ZQN-YX601).

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Dong, L.J., Zhang, H.B., Shi, Q., Lei, Q., Du, J.X., Gao, S.: Learning and fusing multiple hidden substages for action quality assessment. Knowledge-Based Systems p. 107388 (2021). https://doi.org/https://doi.org/10.1016/j.knosys.2021.107388, https://www.sciencedirect.com/science/article/pii/S095070512100650X
4. Doughty, H., Damen, D., Mayol-Cuevas, W.: 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)
5. Doughty, H., Mayol-Cuevas, W., Damen, D.: The pros and cons: Rank-aware temporal attention for skill determination in long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7862–7871 (2019)
6. Faller, A.J.: An average correlation coefficient. Journal of Applied Meteorology **20**(2), 203–205 (1981)
7. Farabi, S., Himel, H.H., Gazzali, F., Hasan, B., Kabir, M., Farazi, M., et al.: Improving action quality assessment using resnets and weighted aggregation. arXiv preprint arXiv:2102.10555 (2021)
8. Fard, M.J., Ameri, S., Darin Ellis, R., Chinnam, R.B., Pandya, A.K., Klein, M.D.: Automated robot-assisted surgical skill evaluation: Predictive analytics approach. The International Journal of Medical Robotics and Computer Assisted Surgery **14**(1), e1850 (2018)
9. Gao, J., Zheng, W.S., Pan, J.H., Gao, C., Wang, Y., Zeng, W., Lai, J.: An asymmetric modeling for action assessment. In: European Conference on Computer Vision. pp. 222–238. Springer (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
11. Jain, H., Harit, G., Sharma, A.: Action quality assessment using siamese network-based deep metric learning. IEEE Transactions on Circuits and Systems for Video Technology **31**(6), 2260–2273 (2020)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
14. Lei, Q., Du, J.X., Zhang, H.B., Ye, S., Chen, D.S.: A survey of vision-based human action evaluation methods. Sensors **19**(19), 4129 (2019)
15. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9522–9531 (2021)
16. Malpani, A., Vedula, S.S., Chen, C.C.G., Hager, G.D.: Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In: International Conference on Information Processing in Computer-Assisted Interventions. pp. 138–147. Springer (2014)

17. Nagai, T., Takeda, S., Matsumura, M., Shimizu, S., Yamamoto, S.: Action quality assessment with ignoring scene context. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 1189–1193. IEEE (2021)
18. Nekoui, M., Cruz, F.O.T., Cheng, L.: 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 (2021)
19. Nekoui, M., Tito Cruz, F.O., Cheng, L.: Falcons: Fast learner-grader for contorted poses in sports. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3941–3949 (2020). https://doi.org/10.1109/CVPRW50498.2020.00458
20. Pan, J., Gao, J., Zheng, W.: Adaptive action assessment. IEEE Transactions on Pattern Analysis & Machine Intelligence (01), 1–1 (nov 5555). https://doi.org/10.1109/TPAMI.2021.3126534
21. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6331–6340 (2019)
22. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1468–1476. IEEE (2019)
23. Parmar, P., Morris, B.: Hallucinet-ing spatiotemporal representations using a 2d-cnn. Signals **2**, 604–618 (09 2021). https://doi.org/10.3390/signals2030037
24. Parmar, P., Morris, B.T.: 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)
25. Parmar, P., Reddy, J., Morris, B.: Piano skills assessment. arXiv preprint arXiv:2101.04884 (2021)
26. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–28 (2017)
27. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: European Conference on Computer Vision. pp. 556–571. Springer (2014)
28. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541 (2017)
29. Reiley, C.E., Hager, G.D.: Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: International conference on medical image computing and computer-assisted intervention. pp. 435–442. Springer (2009)
30. Roditakis, K., Makris, A., Argyros, A.: Towards improved and interpretable action quality assessment with self-supervised alignment. In: The 14th PErvasive Technologies Related to Assistive Environments Conference. p. 507–513. PETRA 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3453892.3461624, `https://doi.org/10.1145/3453892.3461624`
31. Sardari, F., Paiement, A., Hannuna, S., Mirmehdi, M.: Vi-net—view-invariant quality of human movement assessment. Sensors **20**(18), 5258 (2020)
32. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: 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 (2020)

33. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
34. Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.: Data-derived models for segmentation with application to surgical assessment and training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 426–434. Springer (2009)
35. Wang, J., Du, Z., Li, A., Wang, Y.: Assessing action quality via attentive spatiotemporal convolutional networks. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 3–16. Springer (2020)
36. Wang, T., Wang, Y., Li, M.: Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 668–678. Springer (2020)
37. Xiang, X., Tian, Y., Reiter, A., Hager, G.D., Tran, T.D.: S3d: Stacking segmental p3d for action quality assessment. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 928–932. IEEE (2018)
38. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
39. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7919–7928 (2021)
40. Zeng, L.A., Hong, F.T., Zheng, W.S., Yu, Q.Z., Zeng, W., Wang, Y.W., Lai, J.H.: 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 (2020)