

# Group-aware Contrastive Regression for Action Quality Assessment

Xumin Yu\*, Yongming Rao\*, Wenliang Zhao, Jiwen Lu<sup>†</sup>, Jie Zhou

Department of Automation, Tsinghua University, China

State Key Lab of Intelligent Technologies and Systems, China

Beijing National Research Center for Information Science and Technology, China

yuxm20@mails.tsinghua.edu.cn; raoyongming95@gmail.com;

zhaowl20@mails.tsinghua.edu.cn; {lujiwen, jzhou}@tsinghua.edu.cn

## Abstract

Assessing action quality is challenging due to the subtle differences between videos and large variations in scores. Most existing approaches tackle this problem by regressing a quality score from a single video, suffering a lot from the large inter-video score variations. In this paper, we show that the relations among videos can provide important clues for more accurate action quality assessment during both training and inference. Specifically, we reformulate the problem of action quality assessment as regressing the relative scores with reference to another video that has shared attributes (e.g., category and difficulty), instead of learning unreferenced scores. Following this formulation, we propose a new **Contrastive Regression (CoRe)** framework to learn the relative scores by pair-wise comparison, which highlights the differences between videos and guides the models to learn the key hints for assessment. In order to further exploit the relative information between two videos, we devise a group-aware regression tree to convert the conventional score regression into two easier sub-problems: coarse-to-fine classification and regression in small intervals. To demonstrate the effectiveness of CoRe, we conduct extensive experiments on three mainstream AQA datasets including AQA-7, MTL-AQA and JIGSAWS. Our approach outperforms previous methods by a large margin and establishes new state-of-the-art on all three benchmarks.

## 1. Introduction

Action quality assessment (AQA), which aims to evaluate how well a specific action is performed, has attracted growing attention in recent years since it plays a crucial role in many real world applications including sports [9, 20, 13, 23, 24, 22, 31, 21], healthcare [17, 39, 26, 38, 41, 42] and others [5, 6]. Unlike conventional action recognition tasks

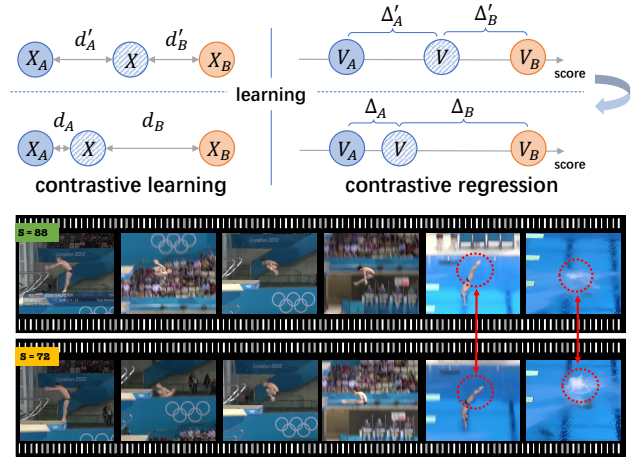


Figure 1: Our Contrastive Regression (CoRe) framework for action quality assessment. Inspired by contrastive learning that learns representation by encouraging the distances of samples (e.g.,  $d_A$  and  $d_B$ ) to reflect their semantic relationship, we learn an AQA model to regress the relative scores (e.g.,  $\Delta_A$  and  $\Delta_B$ ) to reflect the differences of action quality among videos. By comparing two videos with different scores, CoRe encourage the model to learn from differences between videos for assessment.

that focus on action classification [12, 33, 32, 27, 15, 7, 34] and detection [40, 16, 28, 37, 18], AQA is more challenging as it requires the model to predict fine-grained scores from videos that describe the same action. Considering the differences between videos and large variations in scores, we argue that a key to addressing this problem is to discover the differences among the videos and predict scores based on the differences.

Many efforts have been made to tackle this problem over the past few years [19, 22, 6, 35, 20]. Most of them formulate the AQA as a regression problem, where the scores are directly predicted from a single video. While some promis-

\*Equal contribution. <sup>†</sup>Corresponding author.

ing results have been achieved, AQA still faces three challenges. First, since the score labels are usually annotated by human judges (*e.g.*, the score of the diving game is calculated by summarizing scores from different judges, then multiplied by the degree of difficulty), the subjective appraisal of judges makes accurate score prediction quite difficult. Second, the difference between videos for AQA is very subtle, since actors are usually performing the same action in a similar environment. Last, most current models are evaluated based on Spearman’s Rank, which may not faithfully reflect the prediction performance (see our discussions in Section 4.1).

Towards a better AQA framework that can utilize the differences among the videos to predict the final rating, we borrow the merits from the concept of contrastive learning [10, 4]. Contrastive learning (Figure 1, top-left) aims to learn a better representation space where the distance  $d_A$  between two similar samples  $X, X_A$  is enforced to be small while the distance  $d_B$  between the dissimilar ones  $X, X_B$  is encouraged to be large. Therefore, the distance in the representation space can already reflect the semantic relationship between two samples (*i.e.*, if they are from the same category). Analogically, in the context of AQA, we aim to learn a model that can map the input video into the score space where the differences between the action qualities can be measured by the relative scores ( $\Delta_A, \Delta_B$ ). Motivated by this, we propose a **Contrastive Regression (CoRe)** framework for the AQA task. Unlike previous works which aim to predict the scores directly, we propose to regress the relative scores between an input video and several exemplar videos as references.

Moreover, as a step towards more accurate score prediction, we devise a group-aware regression tree (GART) to convert the relative score regression into two easier sub-problems: (1) coarse-to-fine classification. We first divide the range of the relative score into several non-overlapping intervals (*i.e.*, groups) and then use a binary tree to allocate the relative score to a certain group by performing classification progressively; (2) regression in a small interval. We perform regression inside the group where the relative score lies and predict the final score. As another contribution, we design a new metric, called relative L2-distance ( $R-\ell_2$ ) to more precisely measure the performance of action quality assessment by considering the intra-class variance.

To verify the effectiveness of our method, we conduct extensive experiments on three mainstream AQA datasets containing both Olympic and surgical actions, namely AQA-7 [20], MTL-AQA [22] and JIGSAWS [8]. Experiments results demonstrate our method largely outperforms the state-of-the-art on the three benchmarks under the Spearman’s Rank Correlation (81.0% to 84.0% on AQA-7, 92.7% to 95.1% on MTL-AQA and 70% to 85% on JIGSAWS) and new proposed  $R-\ell_2$  metric, which clearly

shows the advantages of our proposed contrastive regression framework.

## 2. Related Work

The past few years have witnessed the rapid development of AQA. The mainstreams of AQA methods formulate AQA as a regression task based on reliable scores labels given by expert judges. For example, Gordan *et al.* [9] propose to use the trajectory of the skeleton to solve the problem of gymnastic vault action quality assessment in their pioneer work. Pirsiavash *et al.* [24] use DCT to encode body pose as input features. SVR [1] is also used to build the mapping from the features to the final score. Thanks to the great success of deep learning in action recognition tasks, Parmar *et al.* [21] show that the spatio-temporal features from C3D [30] can better encode the video data and significantly improve the performance. They also propose a large-scale AQA dataset and explore all-action models to further enhance the scoring performance. Following [21], Xu *et al.* [35] propose a model containing two LSTM to learn the multi-scale features of videos. Pan *et al.* [19] propose to use spatial and temporal relation graphs to model the interaction among the joints. In addition, they also propose to use I3D [3] as a stronger backbone network to extract spatio-temporal features. Parmar *et al.* [22] propose a larger AQA dataset with more annotations for various tasks. The idea of multi-task learning is also introduced to improve the model capacity for AQA. Recently, Tang *et al.* [29] propose a new uncertainty-aware score distribution learning (USDL) to reduce the underlying ambiguity of the action score labels from human judges. Different from this line of works, several methods [39, 5, 6, 2] formulate AQA as a pair-wise ranking task. However, they mainly focus on longer and more ambiguous tasks and only predict an overall ranking, which might limit the application of AQA where some quantitative comparisons are required. In this work, we present a new contrastive regression framework to simultaneously rank videos and predict accurate scores, which makes our method distinguished from previous works.

## 3. Approach

The overall framework of our method is illustrated in Figure 2. We will describe our method in detail as follows.

### 3.1. Contrastive Regression

**Problem Formulation.** Most existing works [19, 22, 6, 35, 20, 29] formulate AQA as a regression task, where the input is a video containing the target action and the output is the predicted quality score of the action. Note that in some AQA tasks (*e.g.*, diving), each video is associated with a degree of difficulty for each video (which is a known constant). The final score is the multiplication of the action quality score (*i.e.*, raw score) and the degree of diffi-

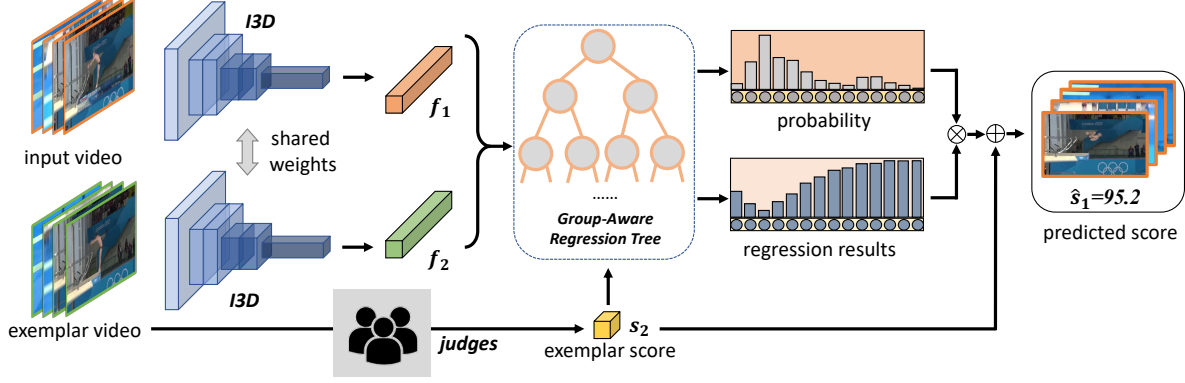


Figure 2: The pipeline of our proposed group-aware contrastive regression method. We first sample an exemplar video for each input video according to the category and degree of difficulty of the action. We then feed the video pair into a shared I3D backbone to extract spatio-temporal features and combine these two features with the reference score of the exemplar video. Finally, we pass the combined feature to the group-aware regression tree and obtain the score difference between the two videos. During inference, the final score can be computed by averaging the results from multiple different exemplars.

culty. Since the degree of difficulty is already known, we only need to predict the action quality score following [29]. Formally, given the input video  $v = \{F^i\}_{i=1}^L$  with action quality label  $s$ , the regression problem is to predict the action quality  $\hat{s}$  based on the input video:

$$\hat{s} = \mathcal{R}_\Theta(\mathcal{F}_\mathcal{W}(v)), \quad (1)$$

where  $\mathcal{R}_\Theta$  and  $\mathcal{F}_\mathcal{W}$  are the regressor model and the feature extractor parameterized by  $\Theta$  and  $\mathcal{W}$ , respectively. The regression problem is usually solved by minimize the mean-square error between the predicted score and the ground-truth score:

$$\mathcal{L}_{\text{AQA}}(\Theta, \mathcal{W}|v) = \text{MSE}(\hat{s}, s), \quad (2)$$

where  $\Theta$  and  $\mathcal{W}$  are the parameters of regression model and feature extractor.

However, since the action videos are usually captured in similar environments (e.g., diving competitions often take place in aquatics centers), it is difficult for the model to learn the diverse scores based on videos with subtle differences. To this end, we propose to reformulate the problem as regressing relative score between the input and an exemplar. Let  $v_m = \{F_m^i\}_{i=1}^{L_m}$  denotes the input video, and  $v_n = \{F_n^i\}_{i=1}^{L_n}$  denotes the exemplar video with score label  $s_n$ . The regression problem can be re-written as:

$$\hat{s}_m = \mathcal{R}_\Theta(\mathcal{F}_\mathcal{W}(v_m), \mathcal{F}_\mathcal{W}(v_n)) + s_n. \quad (3)$$

This formulation can be also viewed as a form of residual learning [11], where we aim to regress the difference of the scores between the input video and a reference video.

**Exemplar-Based Score Regression.** We now describe how to implement the CoRe framework for the AQA problem. Since we aim to regress the relative score between

the input and the exemplar, how to select the exemplar becomes critical. To make the input and the exemplar comparable, we tend to select the video that shares some certain attributes (e.g., category and degree of difficulty) with the input video as the exemplar. Formally, given an input video  $v_m$  and the corresponding exemplar  $v_n$ , we first use an I3D [3] to extract the features  $\{f_n, f_m\}$  following [29, 22], and then aggregate them with the score of the exemplar  $s_n$ :

$$f_{(n,m)} = \text{concat}([f_n, f_m, s_n/\epsilon]), \quad (4)$$

where  $\epsilon$  is a normalizing constant to make sure  $s_n/\epsilon \in [0, 1]$ . We then predict the score difference of the pair through a regressor  $\mathcal{R}_\Theta$  as  $\Delta s = \mathcal{R}_\Theta(f_{(n,m)})$ .

### 3.2. Group-Aware Regression Tree

Although the contrastive regression framework can predict the relative score  $\Delta s$ ,  $\Delta s$  usually takes values from a wide range (e.g., for diving,  $\Delta s \in [-30, 30]$ ). Therefore, predicting  $\Delta s$  directly is still of great difficulty. To this end, we devise a group-aware regression tree (GART) to solve the problem in a divide-and-conquer manner. Specifically, we first divide the range of  $\Delta s$  into  $2^d$  non-overlapping intervals (namely ‘‘groups’’). We then construct a binary regression tree with  $d-1$  layers, of which the leaves represent the  $2^d$  groups, as is illustrated in Figure 3. The decision process of group-aware regression tree follows a coarse-to-fine manner: in the first layer, we determine whether the input video is better or worse than the exemplar video; in the following layers, we gradually make a more accurate prediction about how much the input video is better/worse than the exemplar. Once we have reached the leaf nodes, we can know which group the input video should be classified and we can then perform regression in the corresponding small interval.

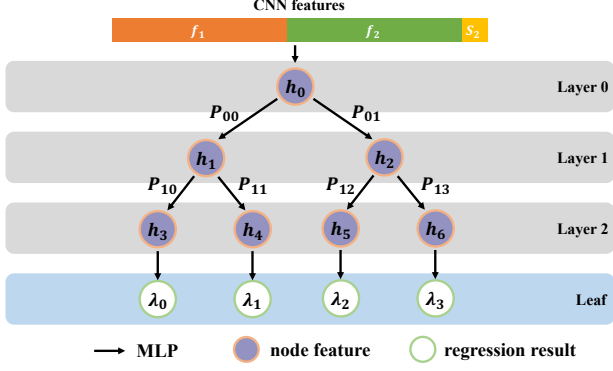


Figure 3: The architecture of the proposed group-aware regression tree. Given the video features and the reference score, the regression tree determines the score difference in a coarse-to-fine manner, where a sequence of binary classification tasks is performed at first (purple nodes) and the regression modules in the leaf layer then give the final prediction (white nodes).

**Tree Architecture.** We adopt the binary tree architecture to perform the regression task. To begin with, we perform an MLP to  $f_{(n,m)}$  and use the output as an initialization of the root node feature. We then perform the regression in a top-down manner. Each node takes the output feature from its parent node as input and produces the binary probability together with the updated feature. The probability of each leaf node can be computed by multiplying all the probabilities along the path to the root. We use the Sigmoid to map the output of each leaf node to  $[0, 1]$ , which is the predicted score difference w.r.t. the corresponding group.

We then describe our partition strategy to define the boundary of each group. First, we collect the list of score differences of all possible training video pairs  $\delta = [\delta_1, \dots, \delta_T]$ . Then, we sort the list in an ascending order to obtain  $\delta^* = [\delta_1^*, \dots, \delta_T^*]$ . Given the group number  $R$ , the partitioning algorithm gives the bounds of each interval  $\mathcal{I}^r = (\zeta_{\text{left}}^r, \zeta_{\text{right}}^r)$  as:

$$\begin{aligned} \zeta_{\text{left}}^r &= \delta^* \left( \left\lfloor (T-1) \times \frac{(r-1)}{R} \right\rfloor \right), \\ \zeta_{\text{right}}^r &= \delta^* \left( \left\lfloor (T-1) \times \frac{r}{R} \right\rfloor \right), \forall i = 1, 2, \dots, R, \end{aligned} \quad (5)$$

where we use  $\delta^*(i)$  to represent the  $i$ -th element of  $\delta$ . It is worth noting that the partition strategy is non-trivial. If we simply uniformly divide the whole range into multiple groups, the pairs of videos in the training set of which the differences of scores lie in some certain group may be unbalanced (see Figure 4 for details).

**Optimization.** We train the regression tree by imposing a classification task on the leaf probabilities and a regression task on the ground-truth interval. Specifically, when

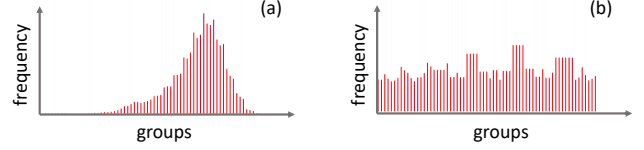


Figure 4: The distribution of the differences of scores in the training set under different partition strategy. (a) Uniform partition. We can observe a large variation of frequency among different groups. (b) The proposed grouping strategy in Equation (5). The training pairs belonging to each group are balanced.

the ground-truth score difference of the input pairs  $\delta$  is in  $i$ -th group, i.e.  $\delta \in (\zeta_{\text{left}}^i, \zeta_{\text{right}}^i)$ , the one-hot label classification  $L = \{l_r\}$  is defined by assigning 1 to the  $i$ -th node and the regression label is set as  $\sigma_i = \frac{\delta - \zeta_{\text{left}}^i}{\zeta_{\text{right}}^i - \zeta_{\text{left}}^i}$ .

For each video pair in the training data with classification label  $\{l_r\}_{r=1}^R$  and regression label  $\{\sigma_r\}_{r=1}^R$ , the objective function for the classification task and regression task can be written as:

$$\begin{aligned} J_{\text{cls}} &= - \sum_{r=1}^R (l_r \log(P_r) + (1 - l_r) \log(1 - P_r)) \\ J_{\text{reg}} &= \sum_{r=1}^R \mathbb{I}(l_r = 1) (\hat{\sigma}_r - \sigma_r)^2, \end{aligned}$$

where  $\{P_r\}_{r=1}^R$  and  $\{\hat{\sigma}_r\}_{r=1}^R$  are the predicted leaf probabilities and regression results. The final objective function for the video-pair is:

$$J = J_{\text{cls}} + J_{\text{reg}}. \quad (6)$$

**Inference.** The overall regression process of the proposed group-aware regression tree can be written as:

$$\mathcal{R}_{\Theta}(f_{(n,m)}) = \hat{\sigma}_{r^*} (\zeta_{\text{right}}^{r^*} - \zeta_{\text{left}}^{r^*}) + \zeta_{\text{left}}^{r^*}, \quad (7)$$

where  $r^*$  is the group with the highest probability. In our implementation, we also adopt a multi-exemplar voting strategy. Given an input video  $v_{\text{test}}$ , we select  $M$  exemplars from training data to construct  $M$  pairs using these  $M$  different exemplars  $\{v_{\text{train}}^m\}_{m=1}^M$  whose scores are  $\{s_{\text{train}}\}_{m=1}^M$ . The process of multi-exemplar voting can be summarized as:

$$\hat{s}_{\text{test}}^m = \mathcal{R}_{\Theta}(\mathcal{F}_{\mathcal{W}}(v_{\text{test}}, v_{\text{train}}^m)) + s_{\text{train}}^m, \quad (8)$$

$$\hat{s}_{\text{test}} = \frac{1}{M} \sum_{m=1}^M \hat{s}_{\text{test}}^m, m = 1, 2, \dots, M. \quad (9)$$



Table 1: Comparisons of Spearman’s Correlation and R- $\ell_2$  Distance on the AQA-7 dataset. ‡ indicates our implementation.

Sp. Corr	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg. Corr.	Year
Pose+DCT [24]	0.5300	0.1000	–	–	–	–	–	2014
ST-GCN [36]	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433	2018
C3D-LSTM [21]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165	2017
C3D-SVR [21]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937	2017
JRG [19]	0.7630	0.7358	0.6006	0.5405	0.9013	<b>0.9254</b>	0.7849	2019
I3D+MLP* [29]	0.7438	0.7342	0.5190	0.5103	0.8915	0.8703	0.7472	2020
USDL [29]	0.8099	0.7570	0.6538	<b>0.7109</b>	0.9166	0.8878	0.8102	2020
I3D + MLP*‡	0.8685	0.6939	0.5391	0.5180	0.8782	0.8486	0.7601	
CoRe + GART*	<b>0.8824</b>	<b>0.7746</b>	<b>0.7115</b>	0.6624	<b>0.9442</b>	0.9078	<b>0.8401</b>	
R- $\ell_2(\times 100)$	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg. R- $\ell_2$	Year
C3D-SVR [21]	1.53	3.12	6.79	7.03	17.84	4.83	6.86	2017
USDL [29]	0.79	2.09	4.82	4.94	0.65	<b>2.14</b>	2.57	2020
I3D + MLP*‡	0.81	2.54	6.06	5.31	1.41	3.08	3.20	
CoRe + GART*	<b>0.64</b>	<b>1.78</b>	<b>3.67</b>	<b>3.87</b>	<b>0.41</b>	2.35	<b>2.12</b>	

## 4. Experiments

### 4.1. Datasets and Experiment Settings

**Datasets.** We perform experiments on three widely used AQA benchmarks including AQA-7 [20], MTL-AQA [22] and JIGSAWS [8]. For more details about the datasets, please refer to the Supplementary.

**Evaluation Protocols.** In order to compare with the previous work [19, 20, 22, 29] in AQA, we adopt Spearman’s rank correlation as an evaluation metric. Spearman’s correlation is defined as:

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}}, \quad (10)$$

where  $p$  and  $q$  represent the ranking for each sample of two series respectively. We also follow the previous work to use Fisher’s z-value [20] when measure the average performance across actions.

We also propose a stricter metric to measure the performance of AQA models more precisely, called relative L2-distance (R- $\ell_2$ ). Given the highest and lowest scores for an action  $s_{\max}$  and  $s_{\min}$ , R- $\ell_2$  is defined as:

$$\text{R-}\ell_2(\theta) = \frac{1}{K} \sum_{k=1}^K \left( \frac{|s_k - \hat{s}_k|}{s_{\max} - s_{\min}} \right)^2, \quad (11)$$

where  $s_k$  and  $\hat{s}_k$  represent ground-truth score and prediction for  $k$ -th sample, respectively. We use R- $\ell_2$  instead of traditional L2-distance because different actions have different scoring intervals. Comparing and averaging  $\ell_2$  distance among different classes of actions is meaningless and confusing. Our proposed R- $\ell_2$  is different from Spearman’s correlation: Spearman’s correlation focuses more on the

ranks of the predict scores while our R- $\ell_2$  focuses on the numerical values.

**Implementation Details:** We adopt the I3D model pre-trained on Kinetics [3] dataset as the feature extractor  $\mathcal{F}_{\mathcal{V}}$ . For all the experiments, we set the depth of GART to  $d = 5$  and the node feature dimension as 256. The initial learning rate is  $1e-3$  for the regression tree and  $1e-4$  for the I3D backbone. We use Adam [14] optimizer, and weight decay is set to zero. we select 10 exemplars for an input test video during inference and vote for the final score using the multi-exemplar voting strategy. In experiments on AQA-7 and MTL-AQA, we follow [29, 19, 20, 22] to extract 103 frames for each video clip, and segment them into 10 overlapping snippets, each containing 16 continuous frames. In JIGSAWS, we follow [29] to evenly sampled out 160 frames to form 10 non-overlapping 16-frame snippets. In AQA-7 and JIGSAWS, we select the exemplar video only according to the coarse category of the video. For example, if the input video is from *single diving-10m platform* in AQA-7, we randomly select an exemplar video from the training set of *single diving-10m platform* in AQA-7. In MTL-AQA dataset, since there are annotations about the degree of difficulty (DD) for diving sports, we select the exemplar based on both the category and the degree of difficulty. Note that this implementation is consistent with the real-world scenario since DD is known to all judges before the action is completed.

We report the performance of the following methods in experiments including the baseline method and different versions of our methods<sup>1</sup>:

- **I3D + MLP** and **I3D + MLP\*** (Baseline) : Most existing works adopt this strategy. We use I3D [3] to en-

<sup>1</sup>We use \* to indicate that we did not use DD in both training and test.

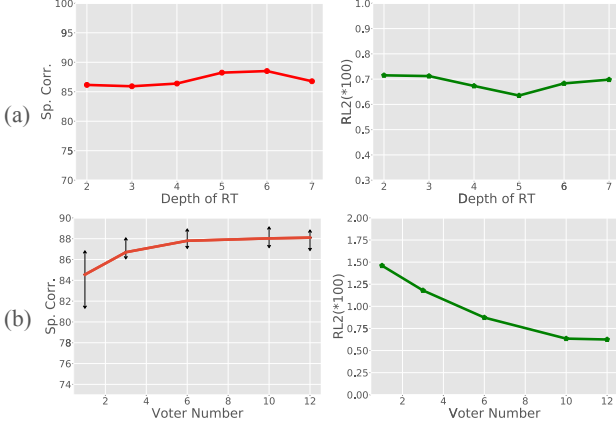


Figure 5: Effects of the depth of regression tree (a) and the number of exemplars for voting (b).

code a single input video, and predict the score based on the feature with a 3-layer MLP. MSE loss between the prediction and the ground-truth is used to optimize the model.

- **CoRe + MLP** and **CoRe + MLP\***: We reformulate the regression problem as mentioned in Sec. 3.1. We choose exemplar videos from the training set to construct the video pairs and also use MSE loss for optimization.
- **I3D + GART** and **I3D + GART\***: We replace the regression sub-network (MLP) with our group-aware regression tree in the baseline method. We use the loss defined in Equation (6)
- **CoRe + GART** and **CoRe + GART\***: The proposed method in Section. 3.

Note that we did not evaluate some of them on the AQA-7 and JIGSAWS datasets due to the absence of degree of difficulty annotations.

## 4.2. Results on AQA-7 dataset

The experiment results of our method and other AQA approaches on AQA-7 are shown in Table 1. The state-of-the-art method USDL [29] uses Gaussian distribution to create a soft distribution label for each video, which can reduce the subjective factor from human judges on original labels. We achieve the same goal with contrastive regression. We also provide the results of the baseline I3D + MLP\* on this dataset, which clearly show the performance improvement obtained by our method. We reach the best results on almost all classes in AQA-7 under both Spearman's correlation and R- $\ell_2$ . Our method achieves 8.95%, 2.32%, 8.83%, -6.82%, 3.01% and 2.25% performance improvement for each sports class compared with USDL under Spearman's

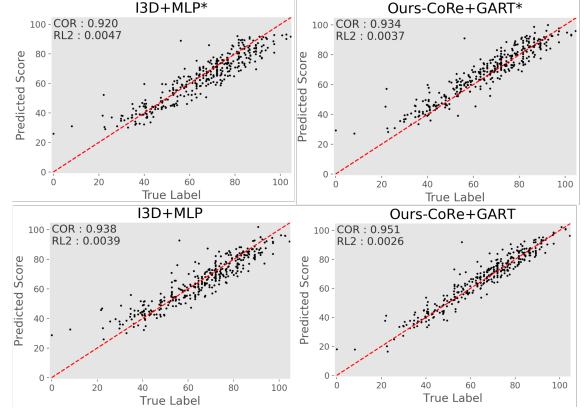


Figure 6: A comparison of different methods in scatter plot. Each point in the figure represents a video in the test set. The red line indicates the prefect predictions.

rank. Meanwhile, we achieve 0.15, 0.31, 1.15, 1.07, 0.24, -0.21 performance improvement under R- $\ell_2$ . For the average correlation and average R- $\ell_2$  performance, we have nearly 3.7% and 0.45 improvements compared to USDL model, clearly showing the effectiveness of our model.

We also conduct several analysis experiments to study the effects of the *depth* of the regression tree and the vote number  $M$  in the multi-exemplar voting on *Diving* class of AQA-7 dataset.

**Effects of the depth of regression tree.** In the regression tree module, the *depth* of the tree is a significant hyper-parameter determining the architecture of the regression tree. We conduct several experiments on *Diving* class of AQA-7 dataset with different values of *depth*, ranging from 2 to 7, and set the  $M$  to 10. As shown in Figure 5, our model performs better when *depth* is 5 and 6, where the total number of groups is 32 and 64. However, there is a little drop in performance when *depth* is smaller than 4 or bigger than 7. In general, our model is robust to different depths.

**Effects of the number of exemplars for voting.** The number of exemplars used in the inference phase is another important hyper-parameter. A larger number for  $M$  means the model can refer to more exemplars while leading to larger computational cost. We conduct experiments on *Diving* class to study the impact of  $M$ . Figure 5 shows the result when the *depth* of regression tree is set to 5. We observe that with  $M$  increasing, the performance becomes better and the variance is lower. The improvement on Sp. Corr. becomes less significant when  $M$  exceeds 10. We can also find the same trend for R- $\ell_2$ .

## 4.3. Results on MTL-AQA dataset

Table 2 shows the performance of existing methods and our method on MTL-AQA dataset. Since the degree of dif-

Table 2: Comparisons of performance with existing methods on the MTL-AQA dataset. ‡ indicates our implementation.

Method (w/o DD)	Sp. Corr.	R- $\ell_2(\times 100)$	Year
Pose+DCT [24]	0.2682	—	2014
C3D-SVR [21]	0.7716	—	2017
C3D-LSTM [21]	0.8489	—	2017
MSCADC-STL [22]	0.8472	—	2019
C3D-AVG-STL [22]	0.8960	—	2019
MSCADC-MTL [22]	0.8612	—	2019
C3D-AVG-MTL [22]	0.9044	—	2019
I3D + MLP* [29]	0.8921	0.707	2020
USDL [29]	0.9066	0.654	2020
MUSDL* [29]	0.9158	0.609	2020
I3D + MLP*‡	0.9196	0.465	
CoRe + GART*	<b>0.9341</b>	<b>0.365</b>	

Method (w/ DD)	Sp. Corr.	R- $\ell_2(\times 100)$	Year
USDL <sub>DD</sub> [29]	0.9231	0.468	2020
MUSDL [29]	0.9273	0.451	2020
I3D + MLP	0.9381	0.394	
CoRe + GART	<b>0.9512</b>	<b>0.260</b>	

Table 3: Ablation study on MTL-AQA dataset

Method	Ablation	Sp. Corr.	R- $\ell_2(\times 100)$
I3D + MLP	Baseline	0.9381	0.394
I3D + GART	+ GART	0.9403	0.366
CoRe + GART	+ CoRe	<b>0.9512</b>	<b>0.260</b>

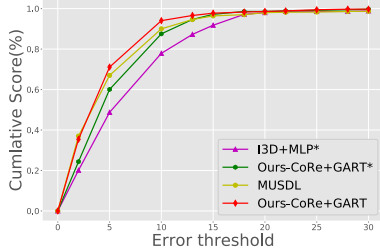


Figure 7: Cumulative score curve on MTL-AQA dataset. The larger the area under the curve indicates the better performance.

difficulty (DD) annotations are available for diving actions in MTL-AQA, we also verify the effects of DD on this dataset. We divide all methods into two types: some use the DD labels in the training phase (bottom part of the table) and the others (upper part of the table) do not. We see CoRe + GART\* achieves respectively 2.0% and 0.244 improvement compared to MUSDL\* [29] under Spearman’s rank and R- $\ell_2$  metric without DD labels. By training with the degree of

Table 4: Comparisons of performance with existing methods on the JIGSAWS dataset.

Sp. Corr.	S	NP	KT	Avg. Corr.
ST-GCN [36]	0.31	0.39	0.58	0.43
TSN [21]	0.34	0.23	0.72	0.46
JRG [19]	0.36	0.54	0.75	0.57
USDL [29]	0.64	0.63	0.61	0.63
MUSDL [29]	0.71	0.69	0.71	0.70
I3D + MLP*	0.61	0.68	0.66	0.65
CoRe + GART*	<b>0.84</b>	<b>0.86</b>	<b>0.86</b>	<b>0.85</b>

R- $\ell_2(\times 100)$	S	NP	KT	Avg.
I3D + MLP*	4.795	11.225	6.120	7.373
CoRe + GART*	<b>5.055</b>	<b>5.688</b>	<b>2.927</b>	<b>4.556</b>

difficulty, our method becomes even better, achieving 2.6% and 0.191 improvements compared to MUSDL under the two metrics. We conjecture that there are two reasons: one is that we can select more suitable exemplars, the other reason is that our method can exploit more information about the action from the degree of difficulty. To have an intuitive understanding of the differences between our method and baseline methods, we visualize the prediction results in form of a scatter plot in Figure 6. We see our method is much more accurate compared to the baseline. By using the degree of difficulty information, the performance of our method can be further improved, where almost all the points are near the red line in the middle of the picture. In Figure 7, we show the cumulative score curves of our methods and SOTA method MUSDL [29]. Given the error threshold  $\epsilon$ , the samples whose absolute differences between their prediction and ground-truth are less than  $\epsilon$  will be regarded as positive samples. It can be observed that under any error threshold, CoRe + GART (red line) shows a stronger ability to predict accurate scores.

**Ablation Study.** We further conduct an ablation study for our method. The results are shown in Table 3. Comparing I3D + MLP and I3D + GART, we see when replacing MLP with our group-aware regression tree, the performance is improved by 0.0022 and 0.028 under Spearman’s rank metric and R- $\ell_2$  metric, which demonstrates the effectiveness of the designs of GART. The performance is further improved when replace the I3D baseline with our proposed CoRe framework. The above results demonstrate the effectiveness of the two components of our method.

**Case Study.** In order to have a deeper understanding of the behavior of our model, we present a case study in Figure 11. Based on the comparison between input and exemplar, the regression tree determines the relative score from coarse to fine. The first layer of the regression tree tries to determine which video is better, and the following layers try to make

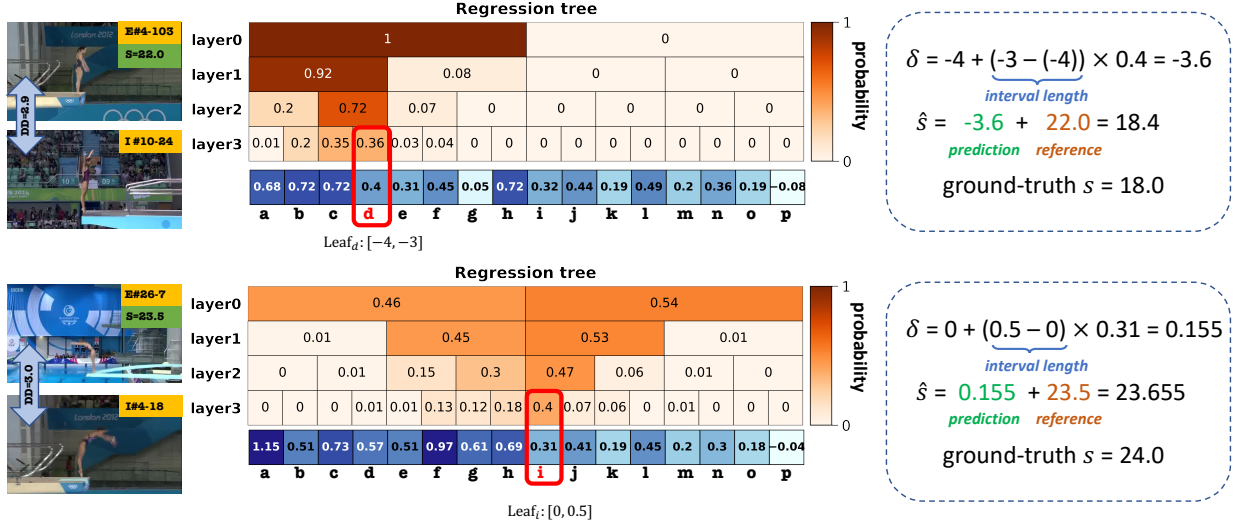


Figure 8: Case study. The videos marked with  $E$  and  $I$  in the upper left corner are the exemplar and the input video, respectively. Each pair of exemplar and input videos have the same degree of difficulty (DD). We show the probability output for each layer of the regression tree and the regression value for each leaf on the right. We take the regression value of the leaf node with the highest probability as the final regression result. The very small errors between our prediction results with ground-truths demonstrate the effectiveness of our method.

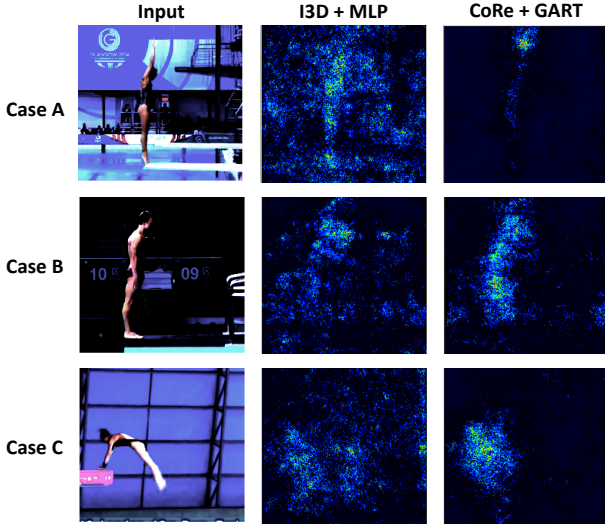


Figure 9: Visualization. We show the visualization result on MTL-AQA using Grad-CAM [25]. Our method can focus on the regions that are critical to assess the action quality.

the prediction more accurate. The first case in the figure shows the behavior when the difference between input and exemplar is large, and the second case shows the behavior when the difference is small. In both situations, our model can give satisfactory predictions.

#### 4.4. Results on JIGSAWS

We also conduct experiments on this surgical action dataset JIGSAWS. Four-fold cross-validation is used following previous works [29, 19]. Table 4 shows the experiment results. CoRe + GART\* largely improves the previous state-of-the-arts. Our method also obtains a more balanced performance in different action classes.

#### 4.5. Visualization

To further prove the effectiveness of our method, we visualize the baseline model (I3D + MLP) and our best model (CoRe + GART) using Grad-CAM [25] on MTL-AQA, as is shown in Figure 9. We observe that our method can focus on certain regions (hands, body, etc.), which indicates our contrastive regression framework can alleviate the influence caused by the background and pay more attention to the discriminative parts.

### 5. Conclusions

In this paper, we have proposed the *CoRe* framework for action quality assessment, which learns the relative scores based on the exemplars. We have also devised a group-aware regression tree to convert the conventional score regression into a coarse-to-fine classification task and a regression task in small intervals. The experiments on three AQA datasets have demonstrated the effectiveness of our approach. We expect the introduction of *CoRe* provides a new and generic solution for various AQA tasks.



## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant U1813218, Grant U1713214, and Grant 61822603, in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI), and in part by a grant from the Institute for Guo Qiang, Tsinghua University.

## A. Datasets

We now describe the datasets we used in our experiments in detail.

**AQA-7** [20]: AQA-7 contains 1,189 samples from seven different actions collected from winter and summer Olympic Games. It contains two dataset released before: UNLV-Dive [21] is named *single diving-10m platform* in AQA-7, contains 370 samples. UNLV-Vault [21] is named *gymnastic vault* in AQA-7, contains 176 samples; The other action classes are newly collected in this dataset: *synchronous diving-3m springboard* contains 88 samples and *synchronous diving-10m platform* contains 91 samples. *big air skiing* contains 175 samples and *big air snowboarding* contains 206 samples.

**MTL-AQA** [22]: The MTL-AQA dataset contains all kinds of *diving* actions, which is the largest AQA dataset up to date. There are 1,412 samples collected from 16 difference world events. The annotations in this dataset are various, including the degree of difficulty (DD), scores from each judge (totally 7 judges), type of diver’s action, and the final score. We adopt the evaluation protocol suggested in [22] in our experiments.

**JIGSAWS** [8]: JIGSAWS is a surgical actions dataset containing 3 type of surgical task: “*Suture(S)*”, “*NeedlePassing(NP)*” and “*Knotted(KT)*”. For each task, each video sample is annotated with multiple annotation scores assessing different aspects of surgical actions, and the final score is the sum of those sub-scores. We adopt a similar four-fold cross validation strategy as [19, 29].

## B. More Discussions

**More analysis on the regression tree.** To better understand the prediction process of the regression tree, we also investigate the prediction accuracy of each layer in the regression tree on the MTL-AQA dataset, as shown in Figure 10. We also compare the results with two baseline methods. Comparing *CoRe + GART* and *GART*, we can see *CoRe + GART* performs better in each layer under all values of  $K$ , which indicates measuring relative score between input and exemplar is more effective than predicting the final score directly. Comparing two *CoRe*-based methods, we see the group-aware regression tree measures relative score more accurately.

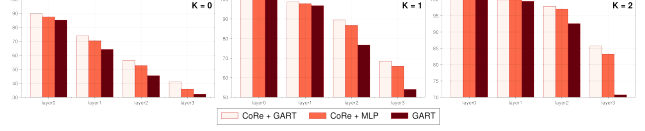


Figure 10: Classification accuracy for each layer of the group-aware regression tree. *CoRe + GART* is our final method, combining contrastive regression and group-aware regression tree together. *CoRe + MLP* uses an MLP to replace the regression tree and the *GART* method only keeps the regression tree without using the contrastive regression framework.  $K$  is a tolerance threshold, which indicates classifying a pair into the nearest- $K$  groups is still regarded as a correct classification.

**More analysis on CoRe.** Another advantage of our proposed *CoRe* is that *CoRe* could alleviate the subjectiveness from human judges by predicting the difference, despite the fact that the exemplar video is also annotated by human judges. Formally, we can assume a score  $\mathbf{x} = x + n$  can be decompose as the actual value  $x$  and a subjectiveness term  $n$  that subjects to normal distribution  $\mathcal{N}(0, \sigma^2)$ . If we directly predict  $\mathbf{x}$ , the variance of subjectiveness term is  $\sigma^2$ . By introducing  $M$  exemplar videos with scores  $\{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ , our goal is to predict the difference

$$\delta = \frac{1}{M} \sum_i (\mathbf{x} - \mathbf{x}_i), \quad (12)$$

which also subjects to a normal distribution:

$$\delta \sim \mathcal{N}\left(\frac{1}{M} \sum_i (x - x_i), \frac{2}{M} \sigma^2\right) \quad (13)$$

We see the prediction becomes closer to the actual value when  $M > 2$ . The empirical results in Figure 5(b) in the original paper also support our assumption.

**More analysis on  $R\text{-}\ell_2$ .** To more precisely measure the AQA performance, we propose a stricter metric, called relative L2-distance ( $R\text{-}\ell_2$ ), to measure the performance of the score prediction model. We use  $R\text{-}\ell_2$  instead of traditional L2-distance because different actions may have different scoring intervals. Comparing and averaging  $\ell_2$  distance among different classes of actions is may be confusing in some cases. Given the highest and lowest scores for an action  $s_{max}$  and  $s_{min}$ ,  $R\text{-}\ell_2$  is defined as:

$$R\text{-}\ell_2(\theta) = \frac{1}{K} \sum_{k=1}^K \left( \frac{\max(|s_k - \hat{s}_k| - \theta, 0)}{s_{max} - s_{min}} \right)^2. \quad (14)$$

$s_k$  and  $\hat{s}_k$  represent ground-truth score and prediction for  $k^{th}$  sample.  $\theta$  is a tolerance threshold. If error between

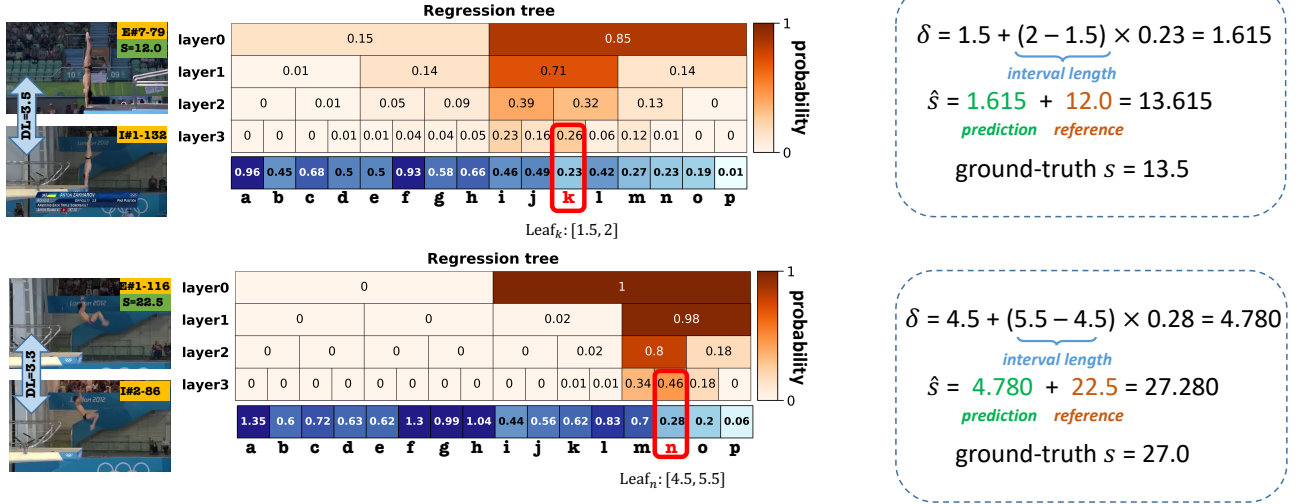


Figure 11: Case study. The videos marked with  $E$  and  $I$  in the upper left corner are the exemplar and the input video, respectively. Each pair of exemplar and input videos have the same degree of difficulty (DD). We show the probability output for each layer of the regression tree and the regression value for each leaf on the right. We take the regression value of the leaf node with the highest probability as the final regression result.

prediction and ground-truth is less than the threshold, the error will be ignored.  $K$  is the size of dataset.

Compared to previous metrics like Spearman’s correlation, the proposed  $R\text{-}\ell_2$  metric has two key advantages: 1) our metric can judge a single prediction while Spearman’s correlation requires the whole test set, which makes our metric more flexible; 2) our metric is stricter and more reasonable especially when the test set is relatively small. For example, diver A and diver B get score of 95 and 65 respectively by human professional judges. If the predictions of these two actions are 80 and 30, it is a prefect prediction under the Spearman’s correlation metric, while our metric can clearly reflect the prediction performance.

### C. Case study

We conduct two more case studies here, as shown in Figure 11. Based on the comparison between the input and the exemplar, the regression tree determines the relative score from coarse to fine. The first layer of the regression tree tries to determine which video is better, and the following layers try to make this determination more accurate. The first case in the figure shows the behavior when the difference between the pair is small, while the second case shows the behavior when this difference is large. When the difference between the two videos is large, it is easy to make the prediction. While the difference is small, the classification task is more difficult, but our method can still give a relatively accurate judgment. We see the proposed contrastive regression framework and the regression tree are two key techniques to achieve accurate score prediction.

### References

- [1] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. 2007. 2
- [2] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. Am I a baller? basketball performance assessment from first-person videos. In *ICCV*, pages 2196–2204, 2017. 2
- [3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 2, 3, 5
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 2
- [5] Hazel Doughty, Dima Damen, and Walterio W. Mayol-Cuevas. Who’s better, who’s best: Skill determination in video using deep ranking. *CoRR*, abs/1703.09913, 2017. 1, 2
- [6] Hazel Doughty, Walterio W. Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *CVPR*, pages 7862–7871, 2019. 1, 2
- [7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210, 2019. 1
- [8] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmadi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamin Bejar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MIC-CAIW*, page 3, 2014. 2, 5, 9
- [9] Andrew Gordon. Automated video assessment of human performance. 11 2001. 1, 2

- [10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3
- [12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. In Johannes Fürnkranz and Thorsten Joachims, editors, *ICML*, pages 495–502, 2010. 1
- [13] Marko Jug, Janez Pers, Branko Dezman, and Stanislav Kovacic. Trajectory based assessment of coordinated human activity. In James L. Crowley, Justus H. Piater, Markus Vincze, and Lucas Paletta, editors, *JCVS*, pages 534–543, 2003. 1
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014. 5
- [15] Hongyang Li, Jun Chen, Ruimin Hu, Mei Yu, Huafeng Chen, and Zengmin Xu. Action recognition using visual attention with reinforcement learning. In Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis, editors, *ACM MM*, pages 365–376, 2019. 1
- [16] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. *CoRR*, abs/1907.09702, 2019. 1
- [17] Anand Malpani, S. Swaroop Vedula, Chi Chiung Grace Chen, and Gregory D. Hager. Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In Danail Stoyanov, D. Louis Collins, Ichiro Sakuma, Purang Abolmaesumi, and Pierre Jannin, editors, *IPCAI*, pages 138–147, 2014. 1
- [18] Alberto Montes, Amaia Salvador, and Xavier Giró-i-Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *CoRR*, abs/1608.08128, 2016. 1
- [19] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *ICCV*, 2019. 1, 2, 5, 7, 8, 9
- [20] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, pages 1468–1476, 2019. 1, 2, 5, 9
- [21] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPRW*, pages 76–84, 2017. 1, 2, 5, 7, 9
- [22] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? A multitask learning approach to action quality assessment. In *CVPR*, 2019. 1, 2, 3, 5, 7, 9
- [23] Matej Perše, Matej Kristan, Janez Pers, and Stanislav Kovačič. Automatic evaluation of organized basketball activity using bayesian networks. 03 2007. 1
- [24] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *ECCV*, pages 556–571, 2014. 1, 2, 5, 7
- [25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 8
- [26] Yachna Sharma, Vinay Bettadapura, Thomas Plötz, Nils Y. Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Video based assessment of osats using sequential motion textures. 2014. 1
- [27] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014. 1
- [28] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, pages 1207–1216, 2019. 1
- [29] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, 2020. 2, 3, 5, 6, 7, 8, 9
- [30] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 2
- [31] Vinay Venkataraman, Ioannis Vlachos, and Pavan K. Turaga. Dynamical regularity for action analysis. In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, *BMVC*, pages 67.1–67.12, 2015. 1
- [32] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015. 1
- [33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 1
- [34] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 1
- [35] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y. Jiang, and X. Xue. Learning to score figure skating sport videos. *TCSVT*, pages 1–1, 2019. 1, 2
- [36] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018. 5, 7
- [37] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, pages 2678–2687, 2016. 1
- [38] Qiang Zhang and Baixin Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In *MMAR*, page 19–24, 2011. 1
- [39] Qiang Zhang and Baixin Li. Relative hidden markov models for video-based evaluation of motion skills in surgical training. *TPAMI*, (6):1206–1218, 2015. 1, 2
- [40] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2933–2942, 2017. 1
- [41] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, Mark A. Clements, and Irfan A. Essa. Automated assessment of surgical skills using frequency analysis. In Nasir Navab, Joachim Hornegger, William M. Wells III, and

Alejandro F. Frangi, editors, *MICCAI*, pages 430–438, 2015. [1](#)

- [42] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L. Sarin, and Irfan A. Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *IJ-CARS*, (3):443–455, 2018. [1](#)