Group-aware Contrastive Regression for Action Quality Assessment
Supplementary Material

A. Datasets

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

AQA-7 [3]: 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 [4] is named *single diving-10m platform* in AQA-7, contains 370 samples. UNLV-Vault [4] 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* cantains 175 samples and *big air snowboarding* contains 206 samples.

MTL-AQA [5]: 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 [5] in our experiments.

JIGSAWS [1]: 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 [2, 6].

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 1. 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.

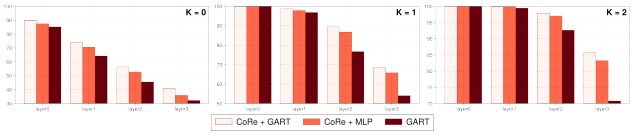

Figure 1: Classification accuracy for each layer of the group-aware regression tree. *CoRe + GART* is our final method, combining contrastive regression and group-aware regression tree together. *CoRe + MLP* uses an MLP to replace the regression tree and the *GART* method only keeps the regression tree 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 σ^2. By introducing M exemplar videos with scores $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$, our goal is to predict the difference

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

which also subjects to a normal distribution:

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

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.

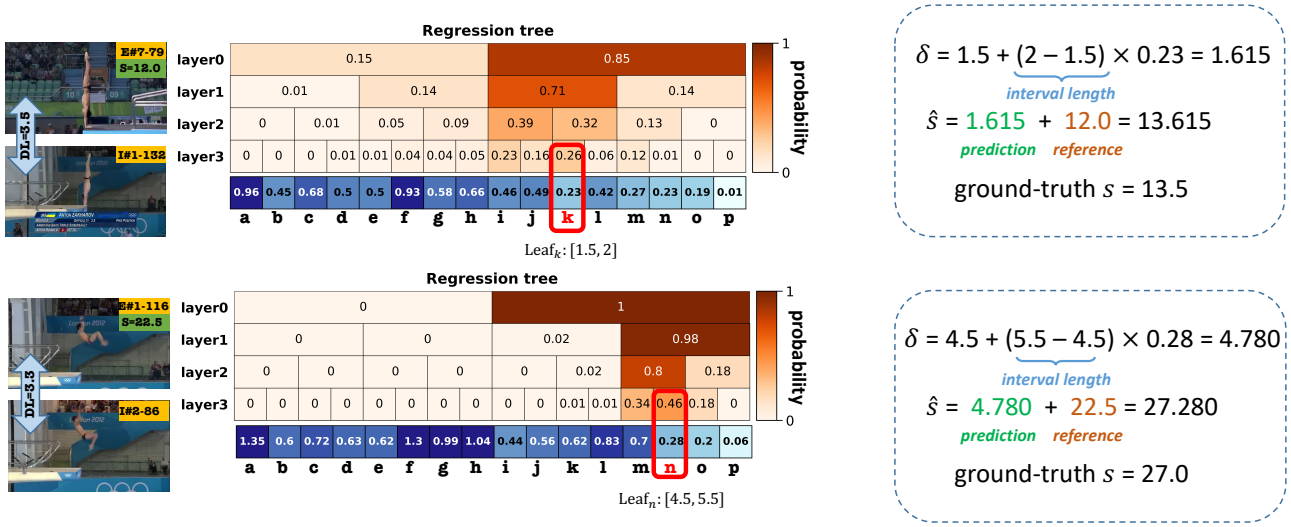

Figure 2: Case study. The videos marked with E and I in the upper left corner are the exemplar and the input video, respectively. Each pair of exemplar and input videos have the same degree of difficulty (DD). We show the probability output for each layer of the regression tree and the regression value for each leaf on the right. We take the regression value of the leaf node with the highest probability as the final regression result.

More analysis on R-ℓ_2. To more precisely measure the AQA performance, we propose a stricter metric, called relative L2-distance (R-ℓ_2), to measure the performance of the score prediction model. We use R-ℓ_2 instead of traditional L2-distance because different actions may have different scoring intervals. Comparing and averaging ℓ_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-ℓ_2 is defined as:

$$\text{R-}\ell_2(\theta) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\max(|s_k - \hat{s}_k| - \theta, 0)}{s_{max} - s_{min}}\right)^2. \quad (3)$$

s_k and $\hat{s}_k$ represent ground-truth score and prediction for k^{th} sample. θ is a tolerance threshold. If error between 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-ℓ_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 2. 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] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Bejar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAIW*, page 3, 2014. 1

[2] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *ICCV*, 2019. 1

[3] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, pages 1468–1476, 2019. 1

[4] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPRW*, pages 76–84, 2017. 1

[5] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? A multitask learning approach to action quality assessment. In *CVPR*, 2019. 1

[6] 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. 1