FineDiving: A Fine-grained Dataset for Procedure-aware Action Quality Assessment

Jinglin Xu*, Yongming Rao*, Xumin Yu, Guangyi Chen, Jie Zhou, Jiwen Lu†
Department of Automation, Tsinghua University, China
Beijing National Research Center for Information Science and Technology, China
{xujinglinlove, raoyongming95}@gmail.com; yuxm20@mails.tsinghua.edu.cn;
guangyichen1994@gmail.com; {jzhou, lujiwen}@tsinghua.edu.cn

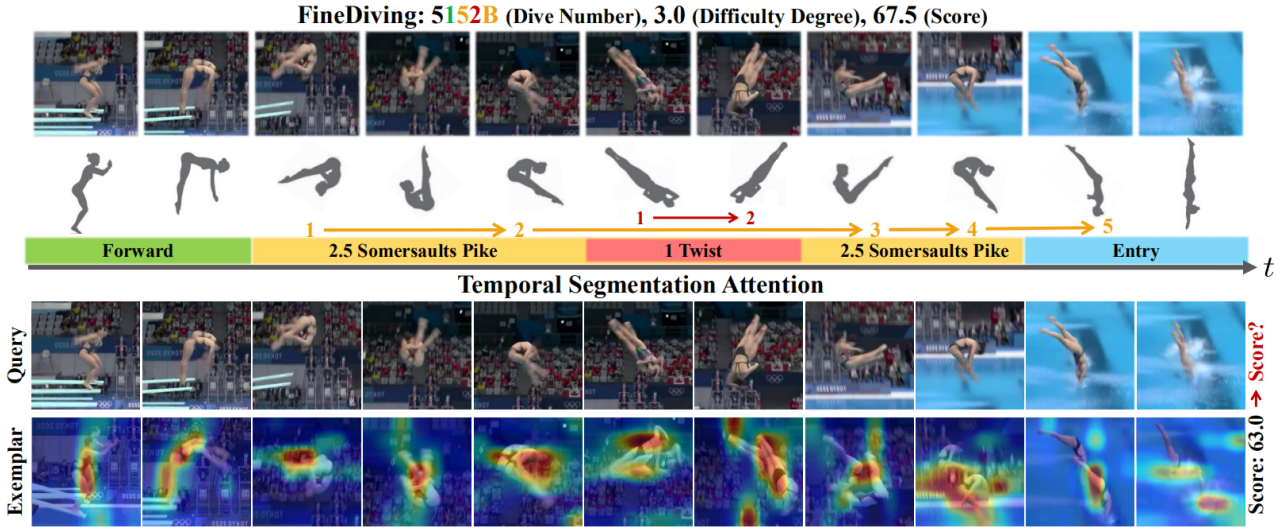

Figure 1. An overview of the **FineDiving** dataset and procedure-aware action quality assessment approach. FineDiving is a fine-grained sports video dataset with detailed annotations on action procedures. It provides a potential for proposing an action quality assessment approach with better interpretability via constructing a new Temporal Segmentation Attention module between query and exemplar instances.

Abstract

Most existing action quality assessment methods rely on the deep features of an entire video to predict the score, which is less reliable due to the non-transparent inference process and poor interpretability. We argue that understanding both high-level semantics and internal temporal structures of actions in competitive sports videos is the key to making predictions accurate and interpretable. Towards this goal, we construct a new fine-grained dataset, called FineDiving, developed on diverse diving events with detailed annotations on action procedures. We also propose a procedure-aware approach for action quality assessment, learned by a new Temporal Segmentation Attention module. Specifically, we propose to parse pairwise query and exemplar action instances into consecutive steps with diverse semantic and temporal correspondences. The procedure-aware cross-attention is proposed to learn embeddings between query and exemplar steps to discover their semantic, spatial, and temporal correspondences, and further serve for fine-grained contrastive regression to derive a reliable scoring mechanism. Extensive experiments demonstrate that our approach achieves substantial improvements over the state-of-the-art methods with better interpretability. The dataset and code are available at https://github.com/xujinglin/FineDiving.

1. Introduction

Competitive sports video understanding has become a hot research topic in the computer vision community. As one of the key techniques of understanding sports action, Action Quality Assessment (AQA) has attracted growing attention in recent years. In the 2020 Tokyo Olympic Games, the AI scoring system in gymnastics acted as a judge for assessing the athlete's score performance and providing feedback for improving the athlete's competitive skill, which reduces the controversies in many subjective scoring events,

*Equal contribution. †Corresponding author.

e.g., diving and gymnastics.

AQA is a task to assess how well an action is performed by estimating a score after analyzing the performance. Unlike conventional action recognition [2, 8, 13, 17, 26, 33, 34, 36, 40–43, 45] and detection [21, 24, 46, 50], AQA is more challenging since an action can be recognized from just one or a few images while the judges need to go through the entire action sequence to assess the action performance. Most existing AQA methods [1,5,6,9,18,27–31,39,44,48] regress on the deep features of videos to learn the diverse scores, which is difficult for actions with a small discrepancy happening in similar backgrounds. Since the diving events are usually filmed in a similar environment (i.e., aquatics centers) and all the videos contain the same action routine, that is "take-off", "flight", and "entry", while the subtle differences mainly appear in the numbers of both somersault and twist, flight positions as well as their executed qualities. Capturing these subtle differences requires the AQA method not only to parse the steps of diving action but also to explicitly quantify the executed qualities of these steps. If we judge the action quality only via regressing a score on the deep features of the whole video, it would be a confusing and non-transparent assessment of the action quality, since we cannot explain the final score via analyzing the performances of action steps.

Cognitive science [22, 32] shows that humans learn to assess the action quality by introducing fine-grained annotations and reliable comparisons. Inspired by this, we introduce these two concepts into AQA, which is challenging since existing AQA datasets lack fine-grained annotations of action procedures and cannot make reliable comparisons. If we judge the action quality using coarse-grained labels, we cannot date back to a convincing reason from the final action quality score. It is urgent to construct a fine-grained sports video dataset for encouraging a more transparent and reliable scoring approach for AQA.

To address these challenges, we construct a new competitive sports video dataset, "FineDiving" (short for Fine-grained Diving), focusing on various diving events, which is the first fine-grained sports video dataset for assessing action quality. FineDiving has several characteristics (Figure 1, the top half): (1) Two-level semantic structure. All videos are annotated with semantic labels at two levels, namely, action type and sub-action type, where a combination of the presented sub-action types produces an action type. (2) Two-level temporal structure. The temporal boundaries of actions in each video are annotated, where each action is manually decomposed into consecutive steps according to a well-defined lexicon. (3) Official action scores, judges' scores, and difficulty degrees are collected from FINA.

We further propose a procedure-aware approach for assessing action quality on FineDiving (Figure 1, the bottom half), inspired by the recently proposed CoRe [47]. The proposed framework learns procedure-aware embeddings with a new Temporal Segmentation Attention module (referred to as TSA) to predict accurate scores with better interpretability. Specifically, TSA first parses action into consecutive steps with semantic and temporal correspondences, serving for procedure-aware cross-attention learning. The consecutive steps of query action are served as queries and the steps of exemplar action are served as keys and values. Then TSA inputs pairwise query and exemplar steps into the transformer and obtains procedure-aware embeddings via cross-attention learning. Finally, TSA performs fine-grained contrastive regression on the procedure-aware embeddings to quantify step-wise quality differences between query and exemplar, and predict the action score.

The contributions of this work are summarized as: (1) We construct the first fine-grained sports video dataset for action quality assessment, which contains rich semantics and diverse temporal structures. (2) We propose a procedure-aware approach for action quality assessment, which is learned by a new temporal segmentation attention module and quantifies quality differences between query and exemplar in a fine-grained way. (3) Extensive experiments illustrate that our procedure-aware approach obtains substantial improvements and achieves the state-of-the-art.

2. Related Work

Sports Video Datasets. Action understanding in sports videos is a hot research topic in the computer vision community, which is more challenging than understanding actions in the general video datasets, e.g., HMDB [16], UCF-101 [37], Kinetics [2], AVA [11], ActivityNet [7], THUMOS [10], Moments in Time [23] or HACS [49], due to the low inter-class variance in motions and environments. Competitive sports video action understanding relies heavily on available sports datasets. Early, Niebles *et al.* [25] introduced the Olympic sports dataset into modeling the motions. Karpathy *et al.* [14] provided a large-scale dataset Sports1M and gained significant performance over strong feature-based baselines. Pirsiavash *et al.* [31] released the first Olympic judging dataset comprising of Diving and Figure Skating. Parmar *et al.* [29] released a new dataset including Diving, Gymnastic Vault, Figure Skating for better working on AQA. Bertasius *et al.* [1] proposed a first-person basketball dataset for estimating performance assessment of basketball players. Li *et al.* [20] built the Diving48 dataset annotated by the combinations of 4 attributes (i.e., back, somersault, twist, and free). Xu *et al.* [44] extended the existing MIT-Figure Skating dataset to 500 samples. Parmar *et al.* [28, 30] presented the MTL-AQA dataset that exploits multi-task networks to assess the motion. Shao *et al.* [34] proposed the FineGym dataset that provides coarse-to-fine annotations both temporally and semantically for facilitating action recognition. Recently, Li *et al.* [19] developed

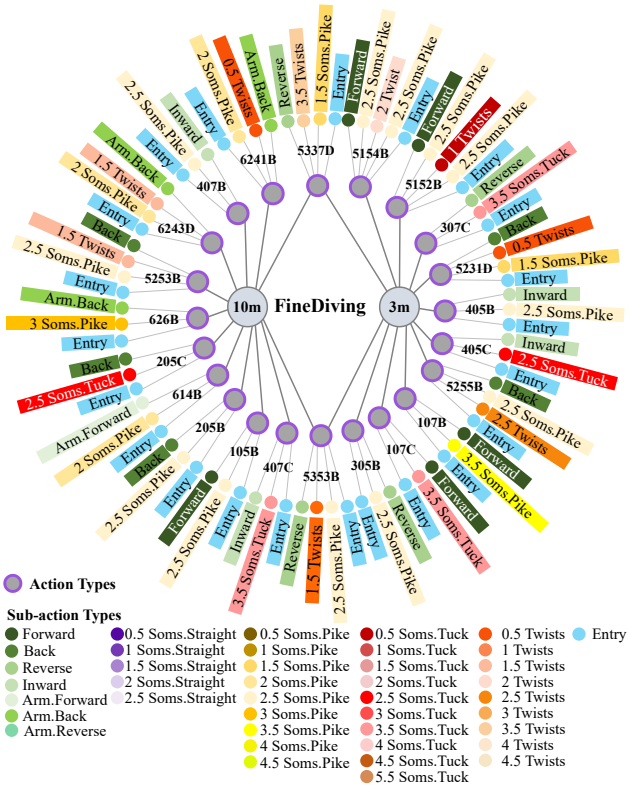

Figure 2. Two-level semantic structure. Action type indicates an action routine described by a dive number. Sub-action type is a component of action type, where each combination of the sub-action types can produce an action type and different action types can share the same sub-action type. The green branch denotes different kinds of take-offs. The purple, yellow, and red branches respectively represent the somersaults with three positions (i.e., straight, pike, and tuck) in the flight, where each branch contains different somersault turns. The orange branch indicates different twist turns interspersed in the process of somersaults. The light blue denotes entering the water. (Best viewed in color.)

a large-scale dataset MultiSports with fine-grained action categories with dense annotations in both spatial and temporal domains for spatio-temporal action detection. Hong *et al.* [12] provided a figure skating dataset VPD for facilitating fine-grained sports action understanding. Chen *et al.* [3] proposed the SMART dataset with fine-grained semantic labels, 2D and 3D annotated poses, and assessment information. Unlike the above datasets, FineDiving is the first fine-grained sports video dataset for AQA, which guides the model to understand action procedures via detailed annotations towards more reliable action quality assessment.

Action Quality Assessment. Most existing methods formulate AQA as a regression on various video representations supervised by the scores. In early pioneering work, Pirsiavash *et al.* [31] first formulated AQA and proposed to map the pose-based features, spatio-temporal interest points, and hierarchical convolutional features to the

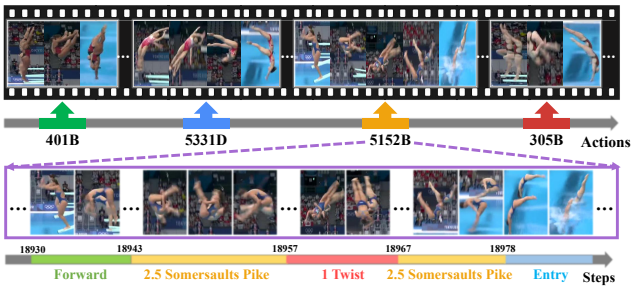

Figure 3. Two-level temporal structure. The action-level labels describe temporal boundaries of valid action routines, while the step-level labels provide the starting frames of consecutive steps in the procedure. (Best viewed in color.)

scores by using SVR. Parmar *et al.* [29] utilized the spatio-temporal features to estimate scores and demonstrated its effectiveness on actions like Diving, Gymnastic Vault, and Figure Skating. Bertasius *et al.* [1] proposed a learning-based approach to estimate motion, behaviors, and performance assessment of basketball players. Li *et al.* [18] proposed to combine some network modification with ranking loss to improve the AQA performance. Doughty *et al.* [6] assessed the relative overall level of skill in a long video based on the video-level pairwise annotation via the high-skill and low-skill attention modules. Parmar *et al.* [28] introduced the shared concepts of action quality among actions. Parmar *et al.* [30] further reformulated the definition of AQA as a multitask learning in an end-to-end fashion. Besides, Xu *et al.* [44] proposed to use self-attentive and multiscale skip convolutional LSTM to aggregate information from individual clips, which achieved the best performance on the assessment of Figure Skating samples. Pan *et al.* [27] assessed the performance of actions visually from videos by graph-based joint relation modeling. Recently, Tang *et al.* [39] proposed to reduce the underlying ambiguity of the action score labels from human judges via an uncertainty-aware score distribution learning (USDL). Yu *et al.* [47] constructed a contrastive regression framework (CoRe) based on the video-level features to rank videos and predict accurate scores. Different from previous methods, our approach understands action procedures and mines procedure-aware attention between query and exemplar to achieve a more transparent action assessment.

3. The FineDiving Dataset

In this section, we propose a new fine-grained competitive sports video dataset FineDiving. We will introduce FineDiving from dataset construction and statistics.

3.1. Dataset Construction

Collection. We search for diving events in Olympics, World Cup, World Championships, and European Aquatics Cham-

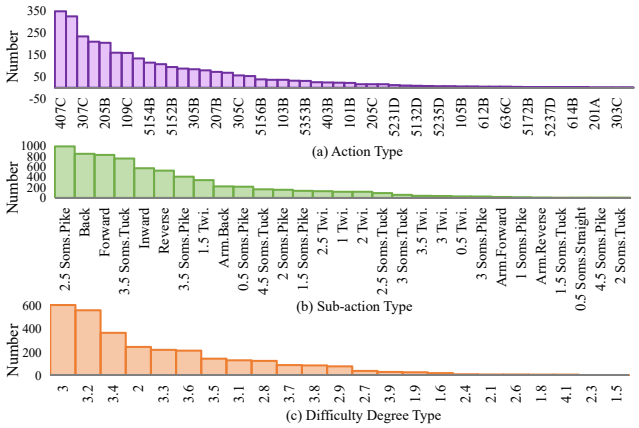

Figure 4. Statistics of FineDiving. (a) The action-type distribution of action instances. (b) The sub-action type distribution of action instances. (c) The difficulty degree distribution of action instances.

pionships on YouTube, and download competition videos with high-resolution. Each official video provides rich content, including diving records of all athletes and slow playbacks from different viewpoints.

Lexicon. We construct a fine-grained video dataset organized by both semantic and temporal structures, where each structure contains two-level annotations, shown in Figures 2 and 3. Herein, we employ three professional athletes of the diving association, who have prior knowledge in diving and help to construct a lexicon for subsequent annotation.

For semantic structure in Figure 2, the action-level labels describe the action types of athletes and the step-level labels depict the sub-action types of consecutive steps in the procedure, where adjacent steps in each action procedure belong to different sub-action types. A combination of sub-action types produces an action type. For instance, for an action type "5255B", the steps belonging to the sub-action types "Back", "2.5 Somersaults Pike", and "2.5 Twists" are executed sequentially.

In temporal structure, the action-level labels locate the temporal boundary of a complete action instance performed by an athlete. During this annotation process, we discard all the incomplete action instances and filter out the slow playbacks. The step-level labels are the starting frames of consecutive steps in the action procedure. For example, for an action belonging to the type "5152B", the starting frames of consecutive steps are 18930, 18943, 18957, 18967, and 18978, respectively, shown in Figure 3.

Annotation. Given a raw diving video, the annotator utilizes our defined lexicon to label each action and its procedure. We need to accomplish two annotation stages from coarse- to fine-grained. The coarse-grained stage is to label the action type for each action instance and its temporal boundary accompanied with the official score. The fine-grained stage is to label the sub-action type for each step in the action procedure and record the starting frame of each step. Both coarse- and fine-grained annotation stages adopt a cross-validating labeling method. Specifically, we employ six workers who have prior knowledge in the diving domain and divide data into six parts without overlap. The annotation results of one worker are checked and adjusted by another, which ensures annotation results are double-checked. To improve the annotation efficiency, we utilize an effective toolbox [38] in the fine-grained annotation stage. Under this pipeline, the total time of the whole annotation process is about 120 hours.

Table 1. Comparison of existing sports video datasets and FineDiving. *Score* indicates the score annotations; *Step* is fine-grained class and temporal boundary; *Action* is coarse-grained class and temporal boundary; *Tube* contains fine-grained class, temporal boundary, and spatial localization.

Localization	#Samples	#Events	#Act. Clas.	Avg.Dur.	Anno.Type
TAPOS [35]	16294	/	21	9.4s	Step
FineGym [34]	32697	10	530	1.7s	Step
MultiSports [19]	37701	247	66	1.0s	Tube
Assessment	#Samples	#Events	#Sub-act.Typ.	Avg.Dur.	Anno.Type
MIT Dive [31]	159	1	/	6.0s	Score
UNLV Dive [29]	370	1	/	3.8s	Score
AQA-7-Dive [28]	549	6	/	4.1s	Score
MTL-AQA [30]	1412	16	/	4.1s	Action,Score
FineDiving	3000	30	29	4.2s	Step,Score

3.2. Dataset Statistics

The FineDiving dataset consists of 3000 video samples, covering 52 action types, 29 sub-action types, and 23 difficulty degree types, which are shown in Figure 4. These statistics will be helpful to design competition strategy and better bring athletes' superiority into full play. Table 1 reports more detailed information on our dataset and compares it with existing AQA datasets as well as other fine-grained sports datasets. Our dataset is different from existing AQA datasets in the annotation type and dataset scale. For instance, MIT-Dive, UNLV, and AQA-7-Dive only provide action scores, while our dataset provides fine-grained annotations including action types, sub-action types, coarse- and fine-grained temporal boundaries as well as action scores. MTL-AQA provides coarse-grained annotations, i.e., action types and temporal boundaries. Other fine-grained sports datasets cannot be used for assessing action quality due to a lack of action scores. We see that FineDiving is the first fine-grained sports video dataset for the AQA task, filling the fine-grained annotations void in AQA.

4. Approach

In this section, we will systematically introduce our approach, whose main idea is to construct a new temporal segmentation attention module to propose a reliable and transparent action quality assessment approach. The overall architecture of our approach is illustrated in Figure 5.

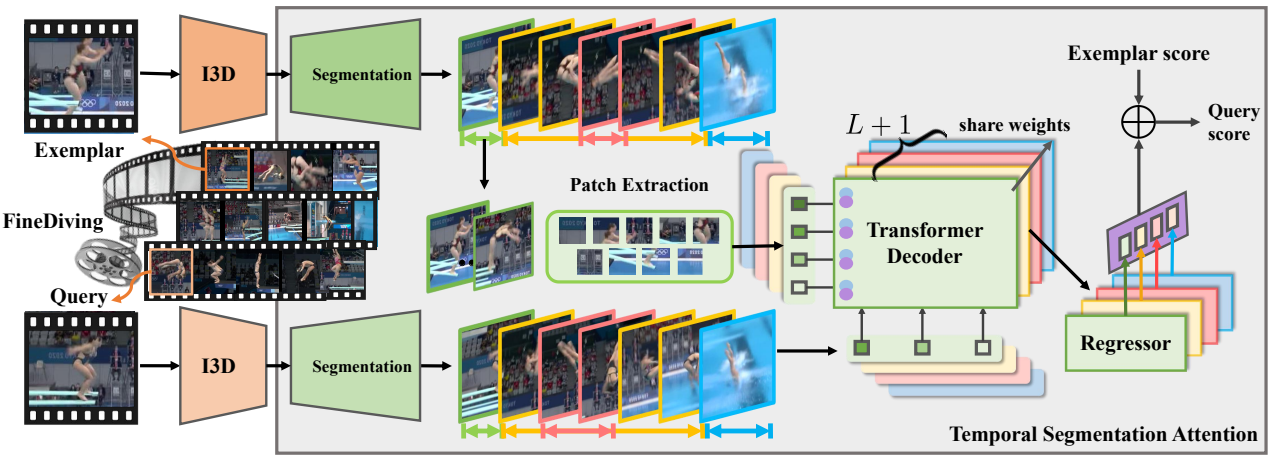

Figure 5. The architecture of the proposed procedure-aware action quality assessment. Given a pairwise query and exemplar instances, we extract spatial-temporal visual features with I3D and propose a Temporal Segmentation Attention module to assess action quality via successively accomplishing procedure segmentation, procedure-aware cross-attention learning, and fine-grained contrastive regression. The temporal segmentation attention module is supervised by step transition labels and action score labels, which guides the model to focus on exemplar regions that are consistent with the query step and quantify their differences to predict reliable action scores.

4.1. Problem Formulation

Given pairwise query X and exemplar Z instances, our procedure-aware approach is formulated as a regression problem that predicts the action quality score of the query video via learning a new Temporal Segmentation Attention module (abbreviated as TSA). It can be represented as:

$$\hat{y}_X = \mathcal{P}(X, Z|\Theta) + y_Z \tag{1}$$

where $\mathcal{P} = \{\mathcal{F}, \mathcal{T}\}$ denotes the overall framework containing I3D [2] backbone $\mathcal{F}$ and TSA module $\mathcal{T}$; Θ indicates the learnable parameters of $\mathcal{P}$; $\hat{y}_X$ is the predicted score of X and y_Z is the ground-truth score of Z.

4.2. Temporal Segmentation Attention

There are three components in TSA, that is, procedure segmentation, procedure-aware cross-attention learning, and fine-grained contrastive regression.

Procedure Segmentation. To parse pairwise query and exemplar actions into consecutive steps with semantic and temporal correspondences, we first propose to segment the action procedure by identifying the transition in time that the step switches from one sub-action type to another.

Suppose that L step transitions are needed to be identified, the procedure segmentation component $\mathcal{S}$ predicts the probability of the step transition occurring at the t-th frame by computing:

$$[\hat{p}_1, \cdots, \hat{p}_L] = \mathcal{S}(\mathcal{F}(X)), \tag{2}$$

$$\hat{t}_k = \underset{\frac{T}{L}(k-1)<t\leq\frac{T}{L}k}{\arg\max} \hat{p}_k(t) \tag{3}$$

where $\hat{p}_k \in \mathbb{R}^T$ is the predicted probability distribution of the k-th step transition; $\hat{p}_k(t)$ denotes the predicted probability of the k-th step transiting at the t-th frame; $\hat{t}_k$ is the prediction of the k-th step transition.

In Equation (2), the component $\mathcal{S}$ is composed of two blocks, namely "down-up" (b_1) and "linear" (b_2). Specifically, the b_1 block consists of four "down-m-up-n" sub-blocks, where m and n denote specified dimensions of output along the spatial and temporal axes, respectively. Each sub-block contains two consecutive convolution layers and one max pooling layer. The b_1 block increases the length of I3D features $\mathcal{F}(X)$ using convolution layers along the temporal axis and reduces the dimension of $\mathcal{F}(X)$ using max-pooling layers along the spatial axis. The "down-up" block advances the visual features $\mathcal{F}(X)$ in deeper layers to contain both deeper spatial and longer temporal views for procedure segmentation. The "linear" block further encodes the output of the b_1 block to generate L probability distributions $\{\hat{p}_k\}_{k=1}^{L}$ of L step transitions in an action procedure. Besides, the constrain in Equation (3) ensures the predicted transitions being ordered, i.e., $\hat{t}_1 \leq \cdots \leq \hat{t}_L$.

Given the ground-truth of the k-th step transition, i.e., t_k, it can be encoded as a binary distribution p_k, where $p_k(t_k) = 1$ and $p_k(t_s)|_{s\neq k} = 0$. With the prediction $\hat{p}_k$ and ground-truth p_k, the procedure segmentation problem can be converted to a dense classification problem, which predicts the probability of whether each frame is the k-th step transition. We calculate the binary cross-entropy loss between $\hat{p}_k$ and p_k to optimize $\mathcal{S}$ and find the frame with the greatest probability of being the k-th step transition. The objective function can be written as:

$$\begin{aligned}\mathcal{L}_{\text{BCE}}(\hat{p}_k, p_k) = &-\textstyle\sum_t (p_k(t)\log\hat{p}_k(t) \\ &+ (1-p_k(t))\log(1-\hat{p}_k(t))).\end{aligned} \tag{4}$$

Minimizing $\mathcal{L}_{\text{BCE}}$ makes distributions $\hat{p}_k$ and p_k closer.

Procedure-aware Cross-Attention. Through procedure segmentation, we obtain $L+1$ consecutive steps with semantic and temporal correspondences in each action procedure based on L step transition predictions. We leverage the sequence-to-sequence representation ability of the transformer for learning procedure-aware embedding of pairwise query and exemplar steps via cross-attention.

Based on $\mathcal{S}$, the query and exemplar action instances are divided into $L+1$ consecutive steps, denoted as $\{S_l^X, S_l^Z\}_{l=1}^{L+1}$. Considering that the lengths of S_l^X and S_l^Z may be different, we fix them into the given size via downsampling or upsampling to meet the requirement that the dimensions of "query" and "key" are the same in the attention model. Then we propose procedure-aware cross-attention learning to discover the spatial and temporal correspondences between pairwise steps S_l^X and S_l^Z, and generate new features in both of them. The pairwise steps complement each other and guide the model to focus on the consistent region in S_l^Z with S_l^X. Here, S_l^Z preserves some spatial information from the feature map (intuitively represented by patch extraction in Figure 5). The above procedure-aware cross-attention learning can be represented as:

$$S_l^{r'} = \text{MCA}(\text{LN}(S_l^{r-1}, S_l^Z)) + S_l^{r-1}, \quad r = 1 \cdots R \tag{5}$$

$$S_l^r = \text{MLP}(\text{LN}(S_l^{r'})) + S_l^{r'}, \quad r = 1 \cdots R \tag{6}$$

where $S_l^0 = S_l^X$, $S_l = S_l^R$. The transformer decoder [4] consisted of alternating layers of Multi-head Cross-Attention (MCA) and MLP blocks, where the LayerNorm (LN) and residual connections are applied before and after every block, respectively, and the MLP block contains two layers with a GELU non-linearity.

Fine-grained Contrastive Regression. Based on the learned procedure-aware embedding S_l, we quantify the step deviations between query and exemplar by learning the relative scores of pairwise steps, which guides the TSA module to assess action quality via learning fine-grained contrastive regression component $\mathcal{R}$. It is formulated as:

$$\hat{y}_X = \frac{1}{L+1} \sum_{l=1}^{L+1} \mathcal{R}(S_l) + y_Z \tag{7}$$

where y_Z is the exemplar score label from the training set. We optimize $\mathcal{R}$ by computing the mean squared error between the ground truth y_X and prediction $\hat{y}_X$, that is:

$$\mathcal{L}_{\text{MSE}} = \|\hat{y}_X - y_X\|^2. \tag{8}$$

4.3. Optimization and Inference

During training, for pairwise query and exemplar (X, Z) in the training set with step transition labels and action score labels, the final objective function for the video pair is:

$$J = \sum_{k=1}^{L} \mathcal{L}_{\text{BCE}}(\hat{p}_k, p_k) + \mathcal{L}_{\text{MSE}}. \tag{9}$$

During testing, for a test video X_{test}, we adopt a multi-exemplar voting strategy [47] to select M exemplars from the training set and then construct M video pairs $\{(X_{\text{test}}, Z_j)\}_{j=1}^{M}$ with exemplar score labels $\{y_{Z_j}\}_{j=1}^{M}$. The process of multi-exemplar voting can be written as:

$$\hat{y}_{X_{\text{test}}} = \frac{1}{M} \sum_{j=1}^{M} (\mathcal{P}(X_{\text{test}}, Z_j|\Theta) + y_{Z_j}). \tag{10}$$

5. Experiments

5.1. Evaluation Metrics

We comprehensively evaluate our approach on two aspects, namely procedure segmentation and action quality assessment, and compute the following three metrics.

Average Intersection over Union. When the procedure segmentation is finished, a set of predicted step transitions are obtained for each video sample. We rewrite these step transition predictions as a set of 1D bounding boxes, denoted as $\mathcal{B}_p = \{\hat{t}_{k+1} - \hat{t}_k\}_{k=1}^{L-1}$. Supposed that the ground-truth bounding boxes can be written as $\mathcal{B}_g = \{t_{k+1} - t_k\}_{k=1}^{L-1}$, we calculate the average Intersection over Union (AIoU) between two bounding boxes (i.e., $\hat{t}_{k+1} - \hat{t}_k$ and $t_{k+1} - t_k$) and determine the correctness of each prediction if IoU_i is larger than a certain threshold d. We describe the above operation as a metric AIoU@d for evaluating our approach:

$$\text{AIoU@}d = \frac{1}{N} \sum_{i=1}^{N} \mathcal{I}(\text{IoU}_i \geq d) \tag{11}$$

$$\text{IoU}_i = |\mathcal{B}_p \cap \mathcal{B}_g| / |\mathcal{B}_p \cup \mathcal{B}_g| \tag{12}$$

where IoU_i indicates the Intersection over Union for the i-th sample and $\mathcal{I}(\cdot)$ is an indicator that outputs 1 if $\text{IoU}_i \geq d$, whereas outputs 0. The higher of AIoU@d, the better of procedure segmentation.

Spearman's rank correlation. Following the previous work [27, 28, 30, 39, 47], we adopt Spearman's rank correlation (ρ) to measure the AQA performance of our approach. ρ is defined as:

$$\rho = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2 \sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}} \tag{13}$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ denote the ranking of two series, respectively. The higher ρ the better performance.

Relative ℓ_2-distance. Following [47], we also utilize relative ℓ_2-distance (R-ℓ_2) to measure the AQA performance of our approach. Given the highest and lowest scores of an action, namely $y_{\max}$ and $y_{\min}$, R-ℓ_2 can be defined as:

$$\text{R-}\ell_2 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{|y_i - \hat{y}_i|}{y_{\max} - y_{\min}} \right) \tag{14}$$

where y_i and $\hat{y}_i$ indicate the ground-truth and predicted scores for the i-th sample, respectively. The lower of R_{ℓ_2}, the better performance.

Table 2. Comparisons of performance with existing AQA methods on FineDiving. (w/o DN) indicates selecting exemplars randomly; (w/ DN) indicates using dive numbers to select exemplars; / indicates without procedure segmentation.

Method (w/o DN)	AIoU@ 0.5	AIoU@ 0.75	ρ	R-ℓ_2($\times$100)
USDL [39]	/	/	0.8302	0.5927
MUSDL [39]	/	/	0.8427	0.5733
CoRe [47]	/	/	0.8631	0.5565
TSA	**80.71**	**30.17**	**0.8925**	**0.4782**
Method (w/ DN)	AIoU@ 0.5	AIoU@ 0.75	ρ	R-ℓ_2($\times$100)
USDL [39]	/	/	0.8913	0.3822
MUSDL [39]	/	/	0.8978	0.3704
CoRe [47]	/	/	0.9061	0.3615
TSA	**82.51**	**34.31**	**0.9203**	**0.3420**

5.2. Implementation Details

Experiment Settings. We adopted the I3D model pretrained on the Kinetics [2] dataset as $\mathcal{F}$ with the initial learning rate 10^{-4}. We set the initial learning rates of $\mathcal{T}$ as 10^{-3}. We utilized Adam [15] optimizer and set weight decay as 0. Similar to [39, 47], we extracted 96 frames for each video, split them into 9 snippets, and then fed them into I3D, where each snippet contains 16 continuous frames with stride 10 frames. Following the experiment settings in [30, 39, 47], we selected 75 percent of samples are for training and 25 percent are for testing in all the experiments. We also specified network parameters for two blocks b_1 and b_2 in $\mathcal{S}$. In the block b_1, (m, n) in the sub-blocks equal to (1024, 12), (512, 24), (256, 48), and (128, 96), respectively. The block b_2 is a three-layer MLP. Furthermore, we set M as 10 in the multi-exemplar voting strategy and set the number of step transitions L as 2. More details about the criterion of selecting exemplars and the number of step transitions can be found in the supplementary materials.

Compared Methods. We reported the performance of the following methods including baseline and different versions of our approach:

• $\mathcal{F}$+$\mathcal{R}$ (Baseline), $\mathcal{F}$+$\mathcal{R}^\star$, and $\mathcal{F}$+$\mathcal{R}^\sharp$: The baseline uses I3D to extract visual features for each input video and predicts the score through a three-layer MLP with ReLU non-linearity, which is optimized by the MSE loss between the prediction and the ground truth. $\star$ indicates the baseline adopting the asymmetric training strategy and $\sharp$ denotes the baseline concatenating dive numbers to features.

• $\mathcal{F}$+$\mathcal{S}$+$\mathcal{R}$: The procedure segmentation component is introduced into the baseline, which is optimized by the combination of MSE and BCE losses.

• TSA, TSA†: The approach was proposed in Section 4. $\dagger$ indicates parsing action procedure using the ground-truth step transition labels instead of the prediction of procedure segmentation, which can be seen as an oracle for TSA.

Table 3. Ablation studies on FineDiving. / indicates the methods without segmentation and ✓ denotes the method using the ground-truth step transition labels.

Method (w/o DN)	AIoU@ 0.5	AIoU@ 0.75	ρ	R-ℓ_2($\times$100)
$\mathcal{F}$+$\mathcal{R}$	/	/	0.8504	0.5837
$\mathcal{F}$+$\mathcal{R}^\star$	/	/	0.8452	0.6022
$\mathcal{F}$+$\mathcal{R}^\sharp$	/	/	0.8516	0.5736
$\mathcal{F}$+$\mathcal{S}$+$\mathcal{R}$	77.44	26.36	0.8602	0.5687
TSA	**80.71**	**30.17**	**0.8925**	**0.4782**
TSA†	✓	✓	0.9029	0.4536
Method (w/ DN)	AIoU@ 0.5	AIoU@ 0.75	ρ	R-ℓ_2($\times$100)
$\mathcal{F}$+$\mathcal{R}$	/	/	0.8576	0.5695
$\mathcal{F}$+$\mathcal{R}^\star$	/	/	0.8563	0.5770
$\mathcal{F}$+$\mathcal{R}^\sharp$	/	/	0.8721	0.5435
$\mathcal{F}$+$\mathcal{S}$+$\mathcal{R}$	78.64	29.37	0.8793	0.5428
TSA	**82.51**	**34.31**	**0.9203**	**0.3420**
TSA†	✓	✓	0.9310	0.3260

5.3. Results and Analysis

Comparisons with the State-of-the-art Methods. Table 2 shows the experimental results of our approach and other AQA methods, trained and evaluated on the FineDiving dataset. We see that our approach achieves the state-of-the-art. Specifically, compared with the methods (USDL, MUSDL, and CoRe) without using dive numbers to select exemplars (i.e., w/o DN), our approach respectively obtained 6.23%, 4.98% and 2.94% improvements on Spearman's rank correlation. Meanwhile, our approach also achieved 0.1145, 0.0951, and 0.0783 improvements on Relative ℓ_2-distance compared to those methods, respectively. Similarly, compared with USDL, MUSDL, and CoRe that use dive numbers to select exemplars (i.e., w/ DN), our approach also obtained different improvements on both Spearman's rank correlation and Relative ℓ_2-distance.

Ablation studies. We conducted some analysis experiments under two method settings for studying the effects of dive number, asymmetric training strategy, and procedure-aware cross-attention learning. As shown in Table 3, the performance of $\mathcal{F}$+$\mathcal{R}$ was slightly better than that of $\mathcal{F}$+$\mathcal{R}^\star$, which verified the effectiveness of symmetric training strategy. Compared with $\mathcal{F}$+$\mathcal{R}$, $\mathcal{F}$+$\mathcal{R}^\sharp$ obtained 0.12% and 1.45% improvements on Spearman's rank correlation. Meanwhile, our approach also achieved 0.0101 and 0.026 improvements on Relative ℓ_2-distance. It demonstrated that concatenating dive numbers and I3D features can achieve a positive impact. Compared with $\mathcal{F}$+$\mathcal{R}$ (w/ DN), $\mathcal{F}$+$\mathcal{S}$+$\mathcal{R}$ (w/ DN) improved the performance from 85.76% to 87.93% on Spearman's rank correlation via introducing procedure segmentation. Compared with $\mathcal{F}$+$\mathcal{S}$+$\mathcal{R}$ (w/ DN), TSA (w/ DN) further improved the performance from 87.93% to 92.03% to on Spearman's rank correlation, which demonstrated the superiority of learning procedure-aware cross at-

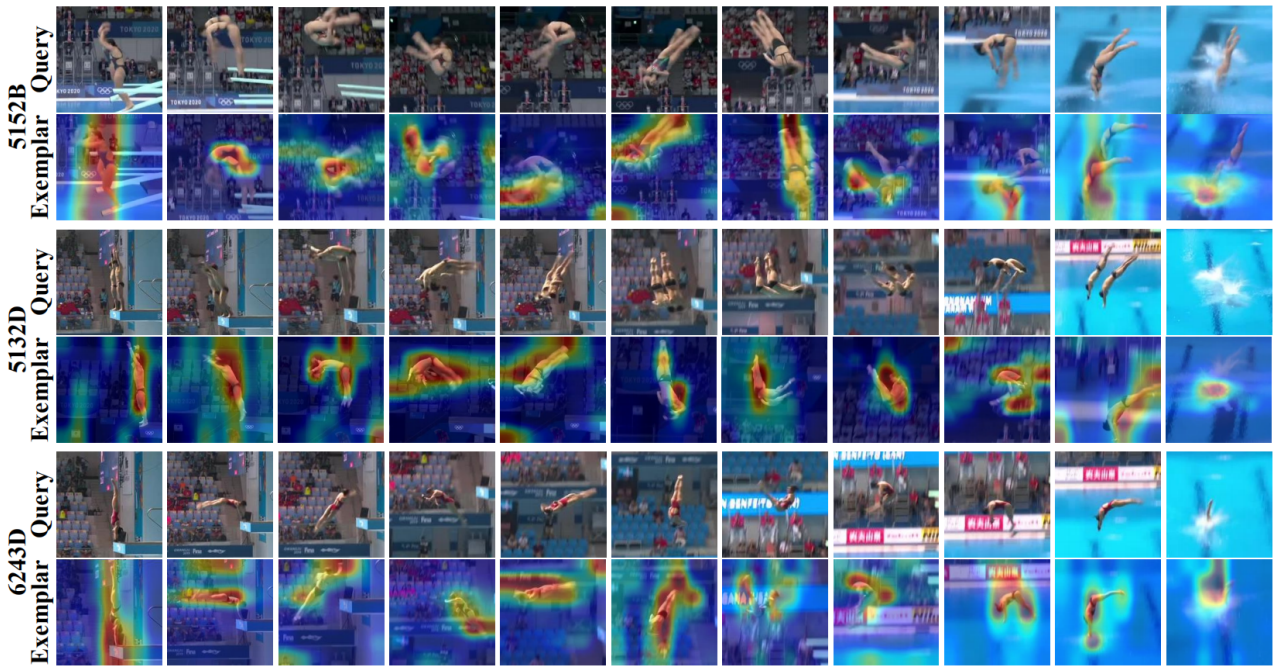

Figure 6. The visualization of procedure-aware cross attention between pairwise query and exemplar procedures. Our approach can focus on the exemplar regions that are consistent with the query step, which makes step-wise quality differences quantifying reliable. The presented pairwise query and exemplar contain the same action and sub-action types. (Best viewed in color.)

Table 4. Effects of the number of exemplars for voting.

M	AIoU@ 0.5	AIoU@ 0.75	ρ	R-$\ell_2(\times 100)$
1	76.01	26.56	0.9085	0.4020
5	80.64	31.78	0.9154	0.3658
10	**82.51**	**34.31**	**0.9203**	**0.3420**
15	82.52	34.31	0.9204	0.3419

tention between query and exemplar procedures.

Besides, for the multi-exemplar voting used in inference, the number of exemplars M is an important hyper-parameter that is a trade-off between better performance and larger computational costs. In Table 4, we conducted some experiments to study the impact of M on our approach TSA (w/ DN). It can be seen that with M increasing, the performance becomes better while the computational cost is larger. The improvement on Spearman's rank correlation becomes less significant when $M > 10$ and the similar trend on Relative ℓ_2-distance also can be found in Table 4.

5.4. Visualization

We visualize the procedure-aware cross-attention between query and exemplar on FineDiving, as shown in Figure 6. It can be seen that our approach highlights semantic, spatial, and temporal correspondent regions in the exemplar steps consistent with the query step, which makes the relative scores between query and exemplar procedures learned from fine-grained contrastive regression more interpretable.

6. Conclusion and Discussion

In this paper, we have constructed the first fine-grained sports video dataset, namely FineDiving, for assessing action quality. On FineDiving, we have proposed a procedure-aware action quality assessment approach via constructing a new temporal segmentation attention module, which learns semantic, spatial, and temporal consistent regions in pairwise steps in the query and exemplar procedures to make the inference process more interpretable, and achieve substantial improvements for existing AQA methods.

Limitations & Potential Negative Impact. The proposed method has an assumption that the number of step transitions in the action procedure is known. The fine-grained annotations need to be manually decomposed and professionally labeled.

Existing Assets and Personal Data. This work contributes a new dataset on the diving sport, where all the data is collected and downloaded on YouTube and bilibili websites. We are actively contacting the creators to ensure that appropriate consent has been obtained.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62125603, Grant 62106124, Grant U1813218, in part by China Postdoctoral Science Foundation under Grant 2020M680564, and in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI).

References

[1] 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 2177–2185, 2017. 2, 3

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2, 5, 7

[3] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *arXiv preprint arXiv:2104.11452*, 2021. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[5] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *CVPR*, pages 6057–6066, 2018. 2

[6] Hazel Doughty, Walterio 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. 2, 3

[7] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 2

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 2

[9] Srujana Gattupalli, Dylan Ebert, Michalis Papakostas, Fillia Makedon, and Vassilis Athitsos. Cognilearn: A deep learning-based interface for cognitive behavior assessment. In *IUI*, pages 577–587, 2017. 2

[10] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://www.thumos.info/, 2015. 2

[11] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 2

[12] James Hong, Matthew Fisher, Michaël Gharbi, and Kayvon Fatahalian. Video pose distillation for few-shot, fine-grained sports action recognition. In *ICCV*, pages 9254–9263, 2021. 3

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–231, 2012. 2

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014. 2

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 2

[17] Hongyang Li, Jun Chen, Ruimin Hu, Mei Yu, Huafeng Chen, and Zengmin Xu. Action recognition using visual attention with reinforcement learning. In *ICMM*, pages 365–376, 2019. 2

[18] Yongjun Li, Xiujuan Chai, and Xilin Chen. End-to-end learning for action quality assessment. In *PRCM*, pages 125–134, 2018. 2, 3

[19] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *ICCV*, pages 13536–13545, 2021. 2, 4

[20] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, pages 513–528, 2018. 2

[21] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019. 2

[22] Meredith Meyer, Dare A Baldwin, and Kara Sage. Assessing young children's hierarchical action segmentation. In *CogSci*, volume 33, 2011. 2

[23] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, pages 1–8, 2019. 2

[24] Alberto Montes, Amaia Salvador, Santiago Pascual, and Xavier Giro-i Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *arXiv preprint arXiv:1608.08128*, 2016. 2

[25] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, pages 392–405, 2010. 2

[26] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, pages 1817–1824, 2013. 2

[27] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *ICCV*, pages 6331–6340, 2019. 2, 3, 6

[28] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, pages 1468–1476, 2019. 2, 3, 4, 6

[29] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPRW*, pages 20–28, 2017. 2, 3, 4

[30] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *CVPR*, pages 304–313, 2019. 2, 3, 4, 6, 7

[31] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *ECCV*, pages 556–571, 2014. 2, 3, 4

[32] Charles F Schmidt. Understanding human action: Recognizing the plans and motives of other persons. 1976. 2

[33] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *ICPR*, pages 32–36, 2004. 2

[34] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 2, 4

[35] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra-and inter-action understanding via temporal action parsing. In *CVPR*, pages 730–739, 2020. 4

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[38] 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. 4

[39] 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*, pages 9839–9848, 2020. 2, 3, 6, 7

[40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 2

[41] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *TPAMI*, 40(6):1510–1517, 2017. 2

[42] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013. 2

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2

[44] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *TCSVT*, 30(12):4578–4590, 2019. 2, 3

[45] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, pages 591–600, 2020. 2

[46] 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. 2

[47] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *ICCV*, pages 7919–7928, 2021. 2, 3, 6, 7

[48] Qiang Zhang and Baoxin Li. Relative hidden markov models for video-based evaluation of motion skills in surgical training. *TPAMI*, 37(6):1206–1218, 2014. 2

[49] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*, pages 8668–8678, 2019. 2

[50] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017. 2