LOGO: A Long-Form Video Dataset for Group Action Quality Assessment

Shiyi Zhang[1,2,3], Wenxun Dai[1], Sujia Wang[1], Xiangwei Shen[1], Jiwen Lu[2,3], Jie Zhou[2,3], Yansong Tang[1,*]
[1] Shenzhen International Graduate School, Tsinghua University
[2] Department of Automation, Tsinghua University
[3] Beijing National Research Center for Information Science and Technology
{shiyi-zh19@mails.,lujiwen@,jzhou@,tang.yansong@sz.}tsinghua.edu.cn

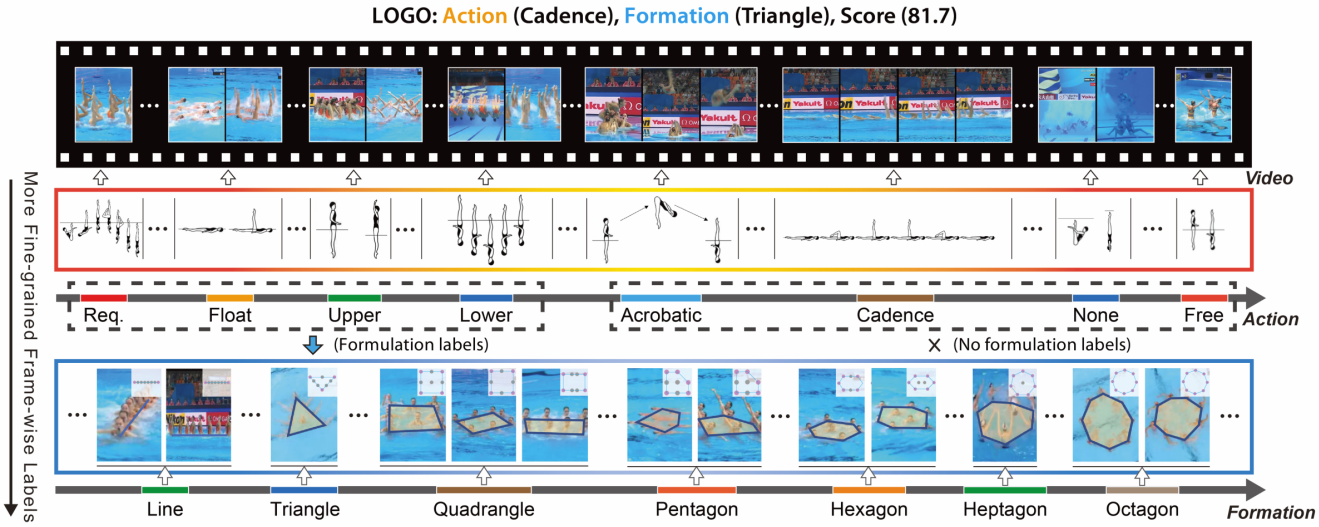

Figure 1. An overview of the ***LOGO*** dataset. LOGO is a multi-person long-form video dataset with frame-wise annotations on both action procedures (as shown in the second line) and formations (as shown in the third line, which reflects relations among actors) based on artistic swimming scenarios. It provides a potential for constructing an action quality assessment approach with the ability of modeling group information among actors. Longer video durations also challenge the ability of the method to aggregate long-term temporal information.

Abstract

Action quality assessment (AQA) has become an emerging topic since it can be extensively applied in numerous scenarios. However, most existing methods and datasets focus on single-person short-sequence scenes, hindering the application of AQA in more complex situations. To address this issue, we construct a new multi-person long-form video dataset for action quality assessment named LOGO. Distinguished in scenario complexity, our dataset contains 200 videos from 26 artistic swimming events with 8 athletes in each sample along with an average duration of 204.2 seconds. As for richness in annotations, LOGO includes formation labels to depict group information of multiple athletes and detailed annotations on action procedures. Furthermore, we propose a simple yet effective method to model relations among athletes and reason about the potential temporal logic in long-form videos. Specifically, we design a group-aware attention module, which can be easily plugged into existing AQA methods, to enrich the clip-wise representations based on contextual group information. To benchmark LOGO, we systematically conduct investigations on the performance of several popular methods in AQA and action segmentation. The results reveal the challenges our dataset brings. Extensive experiments also show that our approach achieves state-of-the-art on the LOGO dataset. The dataset and code will be released at https://github.com/shiyi-zh0408/LOGO.

1. Introduction

Action quality assessment (AQA) is applicable to many real-world contexts where people evaluate how well a specific action is performed such as sports events [13, 18, 29–

* indicates the corresponding author.

Table 1. Comparisons of LOGO and existing datasets of action quality assessment (upper part of the table) and group activity recognition (lower part of the table). *Score* indicates the score annotations; *Action* denotes action types and temporal boundaries; *Act.Label* indicates the action types of both individuals and groups. *Bbox* indicates bounding boxes for actors. *Formation* represents formation annotations. *Temp.* indicates temporal boundary, *Spat.* indicates spatial localization.

Dataset	Duration	Avg.Dur.	Anno.Type	Samples	Events	Form.Anno.	Year
MTL Dive [33]	15m,54s	6.0s	Score	159	1	✗	2014
UNLV Dive [31]	23m,26s	3.8s	Score	370	1	✗	2017
AQA-7-Dive [29]	37m,31s	4.1s	Score	549	6	✗	2019
MTL-AQA [30]	96m,29s	4.1s	Action,Score	1412	16	✗	2019
Rhythmic Gymnastics [53]	26h,23m,20s	1m,35s	Score	1000	4	✗	2020
FSD-10 [26]	-	3-30s	Action,Score	1484	-	✗	2020
FineDiving [50]	3h,30m,0s	4.2s	Action,Score	3000	30	✗	2022
Collective Activity [7]	-	-	Act.Label	44	-	✗	2009
NCAA [35]	16h,9m,52s	4s	Bbox,Act.Label	11436	-	✗	2016
Volleyball [16]	2h,12m,1s	1.64s	Bbox,Act.Label	4830	-	✗	2016
FineGym [38]	161h,1m,45s	45.7s	Act.Label,Temp.	12685	10	✗	2020
Multisports [25]	18h,34m,40s	20.9s	Act.Label,Temp.,Spat.	3200	247	✗	2021
LOGO(Ours)	11h,20m,41s	3m,24s	Action,Formation,Score	200	26	✓(15764)	

33, 46], healthcare [28, 39, 55–58], art performances, military parades, and others. Due to the extensive application of AQA, many efforts have been made over the past few years. Although some existing works have achieved promising performances in several simple scenarios, the application of AQA in many situations is still difficult to implement. In the data-driven era, the richness of the dataset largely determines whether the model can be applied to a wide range of scenarios or not. Inspired by this, we reviewed the existing datasets in AQA and concluded that they are not rich enough for the following two reasons:

Simplicity of Scenarios. We argue that the complexity of the AQA application scenes is reflected in two aspects, the number of people and the duration of the videos. In snowboarding, for example, there is only one performer, while there are multiple actors in a military parade. In diving, the duration is only about 5 seconds, while in artistic swimming and dance performances, the duration of the action reaches several minutes. However, most existing datasets contain a single performer in each sample and many of them collect videos of 3-8s [29–31, 33, 50], which makes it difficult for existing methods to model complex scenes with more actors and longer duration. In such cases, just focusing on how well actions are performed by each individual may be insufficient. Relations among actors should be built, and the potential temporal logic in long-term videos should be modeled.

Coarse-grained Annotations. Though there have been some long-form video datasets in AQA [49, 53] which provide more complex scenes, they typically contain the score as the only annotation. Such coarse-grained annotation makes it difficult for models to learn deeper information, especially in more complex situations. Simply judging action quality via regressing a score for a long-term video could be confusing since we cannot figure out how the model determines whether an action is well-performed or not.

To address these issues, we propose a multi-person long-form video dataset, LOGO (short for LOng-form GrOup). With 8 actors in each sample, LOGO contains videos with an average duration of 204.2s, much longer than most existing datasets in AQA, making the scenes more complex. Besides, as shown in Figure 1, LOGO contains fine-grained annotations including frame-wise action type labels and temporal boundaries of actions for each video. We also devise formation labels to depict relations among actors. Specifically, we use a convex polygon to represent the formation actors perform, which reflects their position information and group information. In general, LOGO is distinguished by its scenario complexity, while it also provides richer annotations compared to most existing datasets in AQA [29–31, 33, 49, 50, 53].

Furthermore, we build a plug-and-play module, GOAT (short for GrOup-aware ATtention), to bridge the gap between single-person short-sequence scenarios and multi-person long-sequence scenarios. Specifically, in the spatial domain, by building a graph for actors, we model the relations among them. The nodes of this graph represent actors' features extracted from a CNN, and the edges represent the relations among actors. Then we use a graph convolution network (GCN) [19] to model the group features from the graph. The optimized features of the graph then serve as "*queries*" and "*keys*" for GOAT. In the temporal do-

main, after feature extraction by the video feature backbone, the clip-wise features serve as "*values*" for GOAT. Instead of fusing the features simply using the average pooling as most previous works [44, 50, 52], GOAT learns the relations among clips and models the temporal features of the long-term videos based on the spatial information in every clip.

The contributions of this paper can be summarized as: (1) We construct the first multi-person long-form video dataset, LOGO, for action quality assessment. To the best of our knowledge, LOGO stands out for its longer average duration, the larger number of people, and richer annotations when compared to most existing datasets. Experimental results also reveal the challenges our proposed dataset brings. (2) We propose a plug-and-play group-aware module, GOAT, which models the group information and the temporal contextual relations for input videos, bridging the gap between single-person short-sequence scenarios and multi-person long-sequence scenarios. (3) Experimental results demonstrate that our group-aware approach obtains substantial improvements compared to existing methods in AQA and achieves the state-of-the-art.

2. Related Work

Action Quality Assessment. Assessing action quality is an increasingly popular trend in computer vision with wide applications such as video retrieval, instructional video analysis, *etc*. A wide variety of methods for formulating AQA tasks have been proposed [3, 10, 13, 24, 31, 33, 44, 48, 49, 52, 53]. In long video AQA tasks, actions performed at different times may significantly impact the scores given by experts. However, only a handful of works directly address the problem of long-term video AQA [3, 10, 48, 49, 53]. There are many datasets for AQA, including Diving [29–31, 33, 50], Figure Skating [26, 33], Gymnastic Vault [29, 31], Basketball [3], Fitness [43] and Rhythmic Gymnastics [53]. However, the problem of considering group information in multi-person sports has been relatively unexplored. As shown in Table 1, MIT-Dive [33], UNLV-Dive [31], AQA-7-Dive [29] and Rhythmic Gymnastics [53] only provide action scores, while MTL-AQA [30] FSD-10 [26] and FineDiving [50] only provide action types and scores. To perform AQA in long videos with multi-person, we construct a long-form artistic swimming video dataset, LOGO, with fine-grained frame-wise action and formation labels. Furthermore, we propose a method to model relations among athletes and reason about the potential temporal logic in long-form videos.

Group Activity Understanding. Group activity understanding, which concentrates on interpreting the collective activity from multi-person behavioral and interaction dynamics, has attracted plenty of work recently. The focus of existing methods has shifted from using shallow handcrafted features [6, 8, 20, 21, 36, 40] to using deep neural networks [2, 12, 15, 22, 34, 45]. Diverse datasets have been proposed to facilitate research in this area, such as the Collective Activity dataset [7], Volleyball dataset [16], etc. However, the methods and datasets mentioned above only stay at the level of group activity recognition; a more profound understanding of group activities should be able to not only find the high-level group activity but also evaluate the quality of group activities, which requires more attention to the quality of collaboration among actors and fine-grained actions of actors. In this paper, LOGO takes the first step towards quality assessment and sets a new benchmark for group activity understanding.

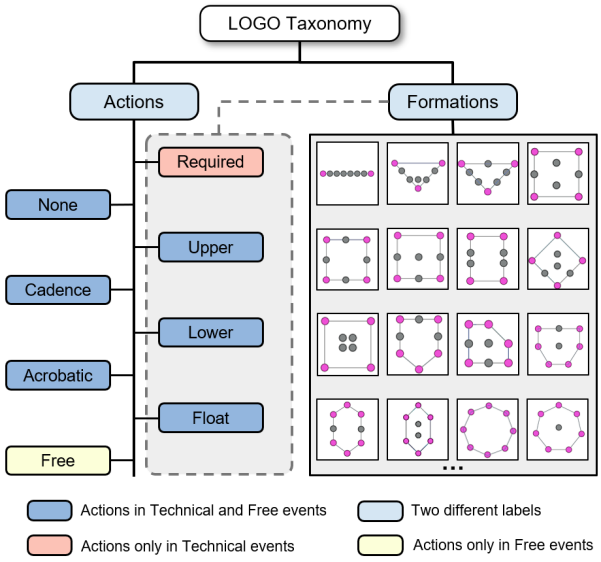

Figure 2. A tree structure of LOGO Taxonomy. LOGO organizes both the *Actions* and *Formations* annotations hierarchically. The left part shows the *Actions* categories of *Technical* and *Free* events. The right part depicts the formation annotation instances when the group is doing *Required*, *Upper*, *Lower* or *Float* actions (the right sub-tree of *Actions*) and not when the group is doing other actions, during which the formations are indistinguishable.

3. The LOGO Dataset

We propose a new multi-person long-form video dataset with detailed annotations on action and formation, LOGO, to set a new challenging benchmark for AQA. We will introduce it from its construction and annotation in this section.

3.1. Dataset Construction

Collection. According to the official FINA rules, the same rules, and prescribed movements are used from 2018 until the end of the 2022 World Championships. So we collected the officially designated *Technical* and *Free* artistic swimming competitions during this period (2 World Championships, 1 Olympic Games, and 4 World Series). We download competition videos with high resolution, e.g.,

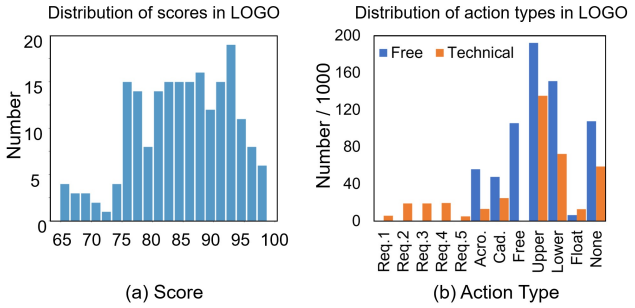

Figure 3. Statics of LOGO. (a) The score distribution of videos. (b) The action-type distribution of frames.

720p and 1080p, on online video platforms like YouTube. Then we sampled them at 25fps and 1fps sampling rates for fine-grained action and formation labeling. Each official videos provide various content, including three sub-scores, one final score, and frames from different views.

Lexicon. We construct a fine-grained video dataset organized by temporal structure, which contains action and formation manual annotations, shown in Figure 2. Herein, we design the labeling system with professional artistic swimming athletes to construct a lexicon for annotation, considering FINA rules and the actual scenario of the competitions. In the *Technical* event, the group size is eight people, the video length is 170±15s, and the actions include *Upper*, *Lower*, *Float*, *None*, *Acrobatic*, *Cadence*, and five *Required Elements*. Each competition cycle needs to complete five *Required Elements*, at least two *Acrobatic* movements, and at least one *Cadence* action. In the *Free* events, there are 8 people, the video length is 240±15s, and the actions include *Upper*, *Lower*, *Float*, *None*, *Acrobatic*, *Cadence*, and *Free* elements. When performing *Required*, *Upper*, *Lower*, and *Float*, the athletes form neat polygons as in Figure 2.

Annotation. Given an RGB artistic swimming video, the annotator utilizes our defined lexicon to label each frame with its action and formation. We accomplish the 25fps frame-wise action annotation stage utilizing the COIN annotation toolbox [42] and the 1fps frame-wise formation labels using Labelme. Specifically, we set strict rules defining the boundaries between artistic swimming sequences and the formation marking position and employ eight workers with prior knowledge in the artistic swimming domain to label the dataset frame by frame following the rules. The annotation results of one worker are checked and adjusted by another, which ensures annotation results are double-checked. Under this pipeline, the total time for the annotation process is above 600 hours.

3.2. Dataset Statics

The LOGO dataset consists of 200 video samples from 26 events with 204.2s average duration and above 11h total duration, covering 3 annotation types, 12 action types, and 17 formation types. Figure 3 shows the score and action type distributions among all the events. Table 1 reports more detailed information on the LOGO dataset and compares it with existing AQA datasets and other fine-grained sports datasets. Our dataset differs from existing AQA datasets in the annotation type and dataset scale. Specifically, our dataset includes formation types providing group information and longer videos on sports, helping to evaluate the quality of actions in multiplayer sports quantitatively. Other fine-grained sports datasets cannot be used for assessing action quality due to a lack of action scores. LOGO is the first fine-grained sports video dataset for Group AQA, filling the fine-grained group annotations void in AQA.

4. Approach

4.1. Problem Formulation

Most existing methods in AQA process the output of the video feature backbone using average pooling before regression. It can be represented as:

$$\hat{y} = \mathcal{R}[\frac{1}{T}\sum_{i=1}^{T}\mathcal{F}(X_i)|\Theta], \tag{1}$$

where X_i denotes the $i-$th clip of the video. $\mathcal{F}$ indicates the video feature backbone; T is the number of clips; $\mathcal{R}$ represents the regression algorithm; Θ is the learnable parameters of $\mathcal{R}$ and $\hat{y}$ is the predicted score. However, fusing temporal information simply by average pooling in scenes with longer duration may be confusing since there could be redundant or even invalid information in the video. Besides, when dealing with multi-person scenarios, we need to depict the relations among actors. We use group information to fuse the extracted features in the temporal axis, modeling the relations among clips. It can be represented as:

$$\hat{y} = \mathcal{R}[\frac{1}{T}\sum_{i=1}^{T} f_i|\Theta], \tag{2}$$

$$f_i = \sum_{j=1}^{T}\mathcal{G}(X_i, X_j)\mathcal{F}(X_j), i = 1...T, \tag{3}$$

where $\mathcal{G}(X_i, X_j)$ denotes the weight computed by group features of the i-th clip and the j-th clip; f_i indicates the reconstructed feature of the i-th clip.

4.2. Group-aware Attention

There are two components of GOAT, that is, group-aware GCN and temporal-fusion attention. The overview of our framework is illustrated in Figure 4.

Group-aware GCN. This component is shown in the right part of Figure 4. We extract the group features by modeling the relations among actors with graphs. Given a

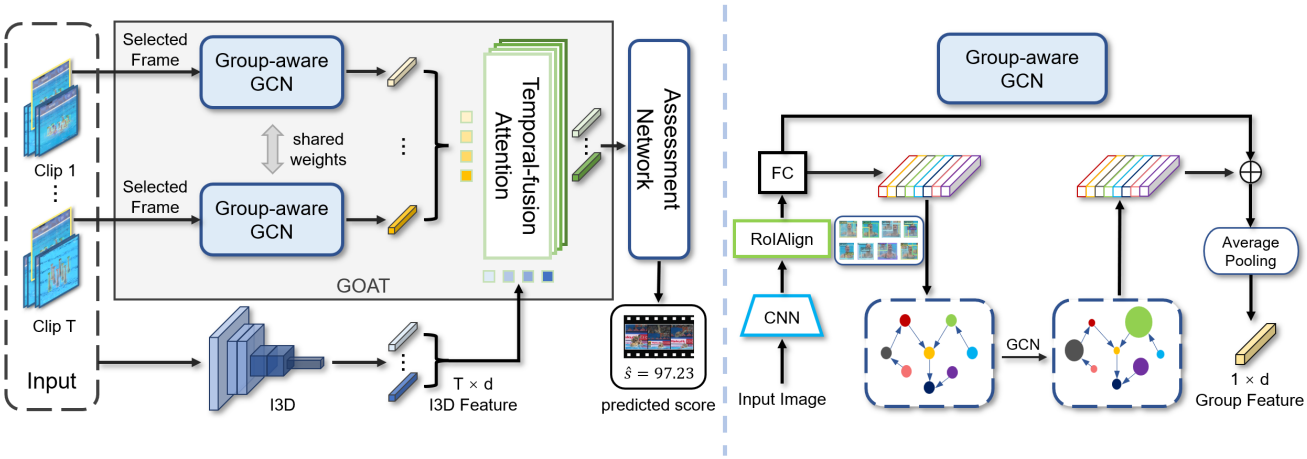

Figure 4. An overview of our group-aware approach for action quality assessment. First, we divide the video into several short clips of equal length. For each clip, we take the middle frame and perform object detection to get the bounding boxes of the actors. Then we send the frame into a CNN to extract the features of the actors. Then we use the feature vector of each actor as the node to construct the relation graph. We use Graph Convolutional Network to enhance the features in the graph, and send the output features into GOAT as "*queries*" and "*keys*". And the "*values*" are the features obtained from the clip by the video feature backbone such as I3D and video swin-transformer. Thus the aggregation in time can be completed. Finally, the output features are sent into the assessment network to predict the scores.

video, we fix it into the given duration via downsampling or upsampling and divide it into several clips of equal length. Then we take three steps to extract the group features. First, we select the middle frame from each clip and build features for the actors. Specifically, we adopt object detection [54] to get bounding boxes of actors. Then, we use Inception-V3 [41] on the frame to get the feature map, to which we apply RoIAlign [14] to extract features for each bounding box. We use a fully connected layer to get a d-dimensional vector from the feature map for each actor. Then we have a $N \times d$ representation denoted as G for each middle frame, where N denotes the number of actors and d indicates the dimension of feature vectors.

Second, we build a relation graph with the feature matrix G to model the relations among actors. Concretely, we regard each feature vector of an actor as a node in the graph. We follow the strategy in [47] to build edges among nodes by computing similarities of appearance features and relative location of actors.

Then, a Graph Convolutional Network (GCN) [19] is adopted to the relation graph to enhance node features by weighted aggregation, which can be represented as:

$$H^{(l+1)} = \sigma(AH^{(l)}W^{(l)}), \tag{4}$$

where A indicates the adjacent matrix; $W^{(l)}$ is the weight matrix in l-th layer; $H^{(l)}$ denotes the features of nodes in l-th layer and $H^{(0)} = G$. Finally, The group feature vector is extracted by adopting average pooling to the sum of the input and the output features of GCN. The group feature vector of the i-th clip is denoted as g_i.

Temporal-fusion Attention. This component fuses the video embeddings along the temporal axis based on the representations from the group-aware GCN. As shown in the left part of Figure 4, for each clip split from the video, we send it into an I3D [4] backbone as most previous works do [44, 50, 52, 53]. We leverage the strong capability of capturing global information of the multi-head attention encoder for learning group-aware video embeddings.

We denote the I3D features of the video as $\{\mathcal{F}(X_i)\}_{i=1}^{T}$, which serve as the "*value*" of the attention block. And the vectors $\{g_i\}_{i=1}^{T}$ from group-aware GCN serve as "*query*" and "*key*". The temporal-fusion attention module learns to discover the group feature correspondences among clips, and generates new temporal features in all of them. The module uses the relations of group features among different clips to assign different weights to different temporal features. The above attention learning can be represented as:

$$q^{(l)} = W_Q^{(l)} q^{(l-1)}, \tag{5}$$

$$k^{(l)} = W_K^{(l)} k^{(l-1)}, \tag{6}$$

$$W_{attn}^{(l)} = Softmax(\frac{q^{(l)} k^{(l)T}}{\sqrt{d}}), \tag{7}$$

$$v^{(l)} = BN[W_{attn}^{(l)} v^{(l-1)} + v^{(l-1)}], \tag{8}$$

where $W_Q^{(l)}, W_K^{(l)}, W_{attn}^{(l)}$ indicate the learnable weights in l-th layer; $q^{(l)}, k^{(l)}, v^{(l)}$ denote the "*query*", "*key*" and "*value*" in l-th layer; d is the dimension of feature vector and BN is the BatchNorm block. Besides, $q^{(0)} = k^{(0)} = \{g_i\}_{i=1}^{T}, v^{(0)} = \{\mathcal{F}(X_i)\}_{i=1}^{T}$.

4.3. Formation Detection

In 4.2, we propose a pipeline to fuse temporal representations based on a kind of spatial information, the group

features. However, such a fusion strategy can also be based on other kinds of spatial features and even without any features by simply using self-attention. To prove the effectiveness of the group features, we devise another kind of spatial features, formation features, to make a comparison.

Formation Detection Task. As mentioned in Section 3, we use a polygon to depict the formation of actors. As shown in Figure 2, the polygon is represented by several coordinates of some actors but not all of them. So in this task, we need to distinguish whether an actor is the vertex of the polygon or not.

Formation Detection Algorithm. Inspired by the "*anchor*" in object detection [37], we propose a feasible way to detect the vertexes. Given a picture, we split it into 1024 patches of equal size. For each patch, we set the center point as the "*anchor*". We send the image into Inception-V3 and extract the d-dimensional feature vector with RoIAlign [14] and a fully connected layer for each patch. To represent the absolute position of each patch, we use sine and cosine functions of different frequencies to encode the position of the "*anchor*". We add the d-dimensional positional encodings to the features and send them into a self-attention block to find the relations among different patches and build the d-dimensional formation features for them.

Then we take two steps to find vertexes. First, for each patch, we judge whether there is a vertex. Second, if there is a vertex, we calculate the relative coordinates of the vertex to the "*anchor*" point. So we detect vertexes by computing:

$$\hat{P} = [\hat{p}_1, ..., \hat{p}_n], \quad \hat{C} = [\hat{c}_1, ..., \hat{c}_n], \tag{9}$$

where $\hat{p}_i \in \mathbb{R}$ denotes the confidence of whether there is a vertex in the i-th patch and $\hat{c}_i \in \mathbb{R}^2$ is the relative coordinates of the vertex to the "*anchor*". In this way, the formation detection task is converted to a classification problem and a regression problem. And we respectively use a MLP to predict $\hat{P}$ and $\hat{C}$ upon the formation features mentioned above. The object function can be represented as:

$$\mathcal{L}_{BCE} = -\sum_i [p_i log\hat{p}_i + (1-p_i)log(1-\hat{p}_i)], \tag{10}$$

$$\mathcal{L}_{MSE} = ||C - \hat{C}||^2. \tag{11}$$

We minimize $\mathcal{L}_{BCE}$ and $\mathcal{L}_{MSE}$ to predict the vertexes. The d-dimensional formation features of middle frame in each clip can replace the group features in 4.2 to fuse the temporal information. Relative experiments are conducted in our ablation study.

5. Experiments

5.1. Implementation Details

Action Quality Assessment. Following [31, 49, 53], we take two stages to assess action quality: feature extraction and score prediction. In the feature extraction stage, we use the I3D model pretrained on Kinetics [4] dataset to extract the 1024-dimensional feature vector for each video clip. Specifically, we sample 5406 frames for each video, split them into 540 snippets, and fed them into I3D. Each snippet contains 16 continuous frames with a stride of 10 frames. In the ablation study, we also use the video swin-transformer [27] as the backbone, which is also pretrained on Kinetics. We follow the sampling strategy in [27] to use a temporal stride of 2. The Inception-V3 in 4.2 is pretrained on ImageNet [9] to extract the CNN features. Following [44, 50, 52], we split the dataset into 3:1 for training and testing in all the experiments. We also test our model using the Mindspore [1].

Action Segmentation. We conduct action segmentation on LOGO with several existing approaches (ASFormer [51], SSTDA [23], MS_TCN++ [5], ASRF [17])to provide a benchmark. We try two backbones to extract frame-wise features: I3D [4] model and video swin-transformer [27] pretrained on Kinetics. For SWIN, we expand the snippet size to 32 frames to extract 1536-dimensional features. In the training stage, we train action segmentation approaches on their default setting.

5.2. Evaluation Metrics

Action Quality Assessment. Following [44, 50, 52, 53], we comprehensively evaluate our approach to AQA under two metrics, Spearman's rank correlation (ρ) and relative ℓ_2-distance (R-ℓ_2). Spearman's rank correlation is used to evaluate the rank correlation between the prediction and ground-truth while the relative ℓ_2-distance (R-ℓ_2) focuses on the difference of the numerical values.

Action Segmentation. Following [5, 15, 23, 51], we adopt three widely used evaluation metrics including frame-wise accuracy (Acc), segmental edit distance and segmental F1 score at overlapping thresholds 10%, 25%, and 50%, denoted by F1@{10, 25, 50} to evaluate the performance of action segmentation approaches.

5.3. Comparison with the State of the Art

Action Quality Assessment. Table 2 shows the experimental results, which reveal great challenges to performing action quality assessment on the LOGO dataset. With the I3D backbone, the mainstream methods in single-person short-term AQA [44, 50, 52] attain the result of 0.4259, 0.4712, and 0.4518 on Spearman's rank correlation respectively. Besides, the results show that our approach achieves the state-of-the-art. We observe that USDL+GOAT, CoRe+GOAT, and TSA+GOAT consistently improve the performance over the original models. Specifically, our approach respectively obtained 8.48%, 4.73%, and 10.12% improvements on Spearman's rank correlation. Our method also outperforms the previous long-term AQA

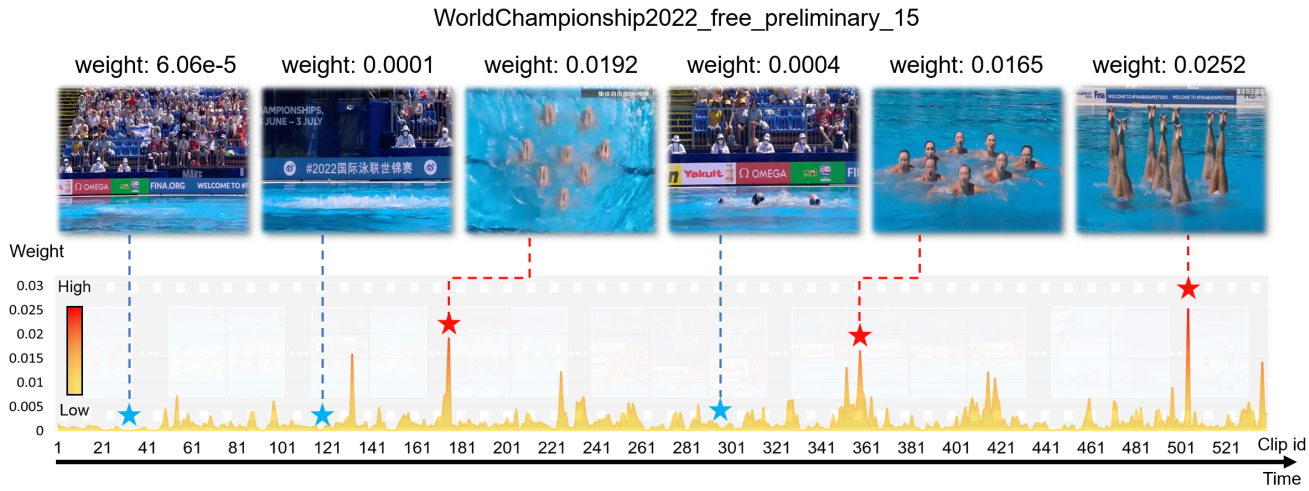

Figure 5. The visualization of the output of our proposed GOAT in action quality assessment. We use red stars to denote clips with high weight while using blue stars to represent clips with low weight. Our approach can focus on where the athletes perform effective movements with clear formations while it can also ignore the redundant part such as all actors are under-water.

Table 2. Comparisons of performance with existing AQA methods on LOGO. The higher ρ, the lower R-ℓ_2, the better performance.

Method	I3D		SWIN	
	$\rho\uparrow$	R-$\ell_2(\times100)\downarrow$	$\rho\uparrow$	R-$\ell_2(\times100)\downarrow$
USDL [44]	0.4259	5.7364	0.4725	5.0762
CoRe [52]	0.4712	5.4086	0.5002	5.9597
TSA [50]	0.4518	5.5326	0.4751	4.7778
ACTION-NET [53]	0.3057	5.8581	0.4101	5.5693
USDL [44]+GOAT	0.4620	**4.8739**	0.5349	5.0220
CoRe [52]+GOAT	**0.4935**	5.0716	**0.5599**	**4.7626**
TSA [50]+GOAT	0.4855	5.3943	0.4843	5.4086

Table 3. Action segmentation results on LOGO.

Method	Features	F1@{10,25,50}			Edit	Acc
ASFormer [51]	SWIN	81.7	79.8	71.1	75.3	81.0
ASFormer [51]	I3D	75.1	71.6	61.3	68.2	73.9
MS_TCN++ [23]	SWIN	80.8	78.8	70.2	73.5	80.1
MS_TCN++ [23]	I3D	72.6	69.2	58.5	66.0	70.6
SSTDA [5]	SWIN	77.8	75.9	66.3	70.0	79.2
SSTDA [5]	I3D	60.1	55.7	44.0	50.1	63.4
ASRF [17]	SWIN	80.8	79.1	72.1	73.2	80.0
ASRF [17]	I3D	73.6	70.5	59.8	66.8	69.8

method [53]. The experimental results illustrate the effectiveness of our proposed module when dealing with multi-person long-term scenarios.

Action Segmentation. Table 3 presents the experimental results. With the I3D feature, the previous action segmentation methods [5, 17, 23, 51] achieve the result of 73.9%, 70.6%, 63.4% and 69.8% frame accuracy respectively. And given an identical setting, we see that the SWIN-pretrained feature obtained significant improvement in all five metrics. Table 6 reveals that our method could improve the performance of MS_TCN++ and ASFormer and achieve the state-of-the-art with the group information.

Table 4. Comparisons of AQA results of CORE and CORE+GOAT based on existing short-term, two-player datasets. *@2* means AQA datasets with two-player scenes.

Dataset	CORE [52]		CORE [52]+GOAT	
	$\rho\uparrow$	R-$\ell_2(\times100)\downarrow$	$\rho\uparrow$	R-$\ell_2(\times100)\downarrow$
FineDiving [50]@2	0.8991	0.3751	0.9032	0.3529
TASD-2 [11]@2	0.9189	0.7863	0.9334	0.7054
AQA-7 [29]@2	0.9012	0.7302	0.9325	0.6423

5.4. Generalization of GOAT

Our method is currently suitable for multi-person scenarios. We select current AQA datasets with two-player scenes (there are only 2 players at most for now) and redivide them according to the number of people. TASD-2 [11] is a dataset collected from synchronized diving events with 2 players in each video. Based on the divided dataset, we obtain results in Table 4, which prove that GOAT can improve the AQA results of two-player scenes in other datasets. These results also demonstrate the generalization of GOAT. Our method can perform well in short-term, two-player scenes.

5.5. Ablation Study

We change the video feature backbone to verify the generality of our approach. Besides, we also try to use other spatial features or simply use self-attention for temporal information fusion to demonstrate the effectiveness of the group features. The experiments also illustrate that our temporal fusion strategy outperforms the average pooling.

Spatial Features. As shown in Table 5, we conduct experiments utilizing formation features to replace the group embeddings used in GOAT to fuse the temporal representations. The formation features are extracted from the attention block output in the formation detection pipeline men-

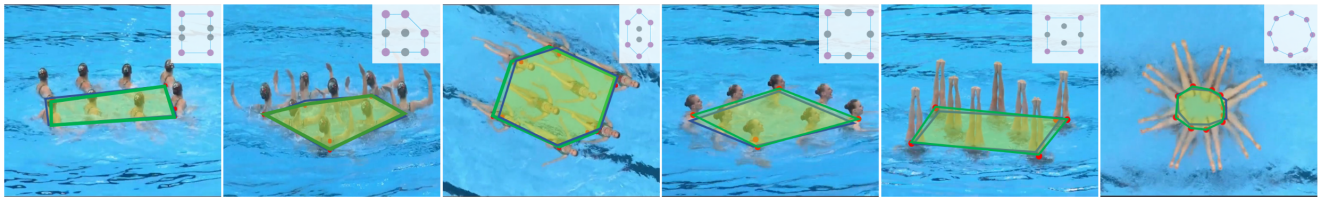

Figure 6. The visualization of the prediction results of our formation detector module. The green polygons represent prediction results and the yellow polygons with blue edges are the ground truth. The results show that our approach can detect the positions of actors and distinguish whether the athlete is the formation vertex or not, which guarantees the reliability of the formation features.

Table 5. Ablation studies on LOGO. *Self.* indicates self-attention; *Form.* indicates using formation features for temporal fusing; *AS.* means using the action segmentation features for temporal fusing.

Method	I3D		SWIN	
	$\rho\uparrow$	R-$\ell_2(\times100)\downarrow$	$\rho\uparrow$	R-$\ell_2(\times100)\downarrow$
USDL [44]+Self.	0.4335	4.8898	0.5078	4.1305
CoRe [52]+Self.	0.4348	5.6019	0.5269	4.5072
TSA [50]+Self.	0.4585	5.0336	0.4720	5.7280
USDL [44]+Form.	0.4355	5.7636	0.5370	5.0888
CoRe [52]+Form.	0.4820	5.3003	0.5401	5.1747
TSA [50]+Form.	0.4241	5.1821	0.4903	4.8627
USDL [44]+AS.	0.4512	4.8837	0.5108	5.0479
CoRe [52]+AS.	0.4825	5.2393	0.5364	4.9102
TSA [50]+AS.	0.4772	5.4203	0.4829	5.3561

Table 6. The action segmentation results of existing methods with GOAT on LOGO. *+GOAT* indicates the method with GOAT.

Method	Features	F1@{10,25,50}			Edit	Acc
MS_TCN++ [23]+GOAT	SWIN	80.9	79.0	70.5	74.0	81.1
MS_TCN++ [23]+GOAT	I3D	73.2	69.9	59.4	67.3	71.7
ASFormer [51]+GOAT	SWIN	82.2	80.3	73.6	75.9	81.7
ASFormer [51]+GOAT	I3D	75.8	72.0	62.1	68.8	74.6

tioned in 4.3, which is pretrained on the LOGO dataset by executing formation detection. We also adopt self-attention to replace GOAT, which means we perform our temporal fusion strategy without the assistance of any spatial information. The experimental results show that the formation features outperform the original models and self-attention models but perform worse than the group features in most cases. The results also show that the self-attention models perform better than the original models. Such results of our experiments prove that: (1) Our temporal fusion strategy outperforms the average pooling even without any spatial information. (2) By modeling the spatial information, the performance of our proposed approach is improved. (3) Building the relations among actors, the group features perform better than the formation features.

Video Feature Backbone. As shown in Table 5 and Table 2. We change the backbone to video swin-transformer (abbreviated as SWIN) and repeat all our experiments. Similar to the results shown in the action segmentation experiments, the methods based on SWIN perform much better than the methods based on I3D in most cases, which illustrates the effectiveness of SWIN. Besides, except for the case of TSA+Self, the results of our methods substantially outperform the original models based on SWIN, which proves the generality of our approach.

Asist AQA with Action Segmentation. To assist the AQA task with the action segmentation task (or action labels), we use the backbone features finetuned in the action segmentation task as the input of GOAT in the experiments shown in the lower part of Table 5. The results illustrate that with the help of the segmentation task, the performances of AQA are improved compared with the original methods.

5.6. Visualization

We visualize the output of GOAT, as shown in Figure 5. It shows that GOAT highlights the part where actors perform effective action with clear formations, demonstrating the effectiveness of our temporal fusion strategy. We also visualize the prediction results of our formation detection module, as shown in Figure 6. It can be seen that our approach can distinguish the vertexes of the formations.

6. Conclusion

In this paper, we construct the first multi-person long-form video dataset, LOGO, for action quality assessment. We also propose a group-aware module, GOAT, to build relations among multiple actors and fuse the temporal representations based on spatial information. Furthermore, the utilization of GOAT in action quality assessment and action segmentation both achieve substantial improvements compared to the existing methods.

Existing Assets and Personal Data. The videos in LOGO are downloaded from several websites such as YouTube. We are actively connecting with the authors to ensure that appropriate consent has been obtained.

Acknowledgments. This work was sponsored in part by the National Natural Science Foundation of China (Grant No. 62206153, 62125603), CAAI-Huawei MindSpore Open Fund, Deng Feng Fund, Young Elite Scientists Sponsorship Program by CAST (No. 2022QNRC001), and Shenzhen Stable Supporting Program (WDZC20220818112518001).

References

[1] Mindspore. https://www.mindspore.cn/. 6

[2] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *CVPR*, pages 4315–4324, 2017. 3

[3] 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. 3

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 5, 6

[5] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *CVPR*, pages 9454–9463, 2020. 6, 7

[6] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, pages 215–230, 2012. 3

[7] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCV*, pages 1282–1289, 2009. 2, 3

[8] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR*, pages 3273–3280, 2011. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[10] 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. 3

[11] Jibin Gao, Wei-Shi Zheng, Jia-Hui Pan, Chengying Gao, Yaowei Wang, Wei Zeng, and Jianhuang Lai. An asymmetric modeling for action assessment. In *ECCV*, pages 222–238, 2020. 7

[12] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *CVPR*, pages 839–848, 2020. 3

[13] Andrew S Gordon. Automated video assessment of human performance. In *AI-ED*, volume 2, 1995. 1, 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 5, 6

[15] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *ECCV*, pages 721–736, 2018. 3, 6

[16] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016. 2, 3

[17] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *WACV*, pages 2322–2331, 2021. 6, 7

[18] Marko Jug, Janez Perš, Branko Dežman, and Stanislav Kovačič. Trajectory based assessment of coordinated human activity. In *ICVS*, pages 534–543, 2003. 1

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2, 5

[20] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, pages 1354–1361, 2012. 3

[21] Tian Lan, Yang Wang, Weilong Yang, Stephen N Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *TPAMI*, 34(8):1549–1562, 2011. 3

[22] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *ICCV*, pages 13668–13677, 2021. 3

[23] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *TPAMI*, pages 1–1, 2020. 6, 7, 8

[24] Yongjun Li, Xiujuan Chai, and Xilin Chen. Scoringnet: Learning key fragment for action quality assessment with ranking loss in skilled sports. In *ACCV*, pages 149–164, 2018. 3

[25] 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

[26] Shenlan Liu, Xiang Liu, Gao Huang, Lin Feng, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Hong Qiao. Fsd-10: a dataset for competitive sports content analysis. *arXiv preprint arXiv:2002.03312*, 2020. 2, 3

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 6

[28] 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 *IPCAI*, pages 138–147, 2014. 2

[29] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, pages 1468–1476, 2019. 1, 2, 3, 7

[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. 1, 2, 3

[31] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPR*, pages 20–28, 2017. 1, 2, 3, 6

[32] Matej Perše, Matej Kristan, Janez Perš, and Stanislav Kovačič. *Automatic evaluation of organized basketball activity using bayesian networks*. 2007. 1

[33] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *ECCV*, pages 556–571, 2014. 1, 2, 3

[34] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *ECCV*, pages 101–117, 2018. 3

[35] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *CVPR*, pages 3043–3053, 2016. 2

[36] Vignesh Ramanathan, Bangpeng Yao, and Li Fei-Fei. Social role discovery in human events. In *CVPR*, pages 2475–2482, 2013. 3

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 28, 2015. 6

[38] 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

[39] Yachna Sharma, Vinay Bettadapura, Thomas Plötz, Nils Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Video based assessment of osats using sequential motion textures. 2014. 2

[40] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *CVPR*, pages 4576–4584, 2015. 3

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 5

[42] 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

[43] Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. *arXiv preprint arXiv:2212.04638*, 2022. 3

[44] 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. 3, 5, 6, 7, 8

[45] Haritha Thilakarathne, Aiden Nibali, Zhen He, and Stuart Morgan. Pose is all you need: The pose only group activity recognition system (pogars). *arXiv preprint arXiv:2108.04186*, 2021. 3

[46] Vinay Venkataraman, Ioannis Vlachos, and Pavan K Turaga. Dynamical regularity for action analysis. In *BMVC*, pages 67–1, 2015. 1

[47] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, pages 9964–9974, 2019. 5

[48] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *CVPR*, pages 3232–3241, 2022. 3

[49] 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, 6

[50] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *CVPR*, pages 2949–2958, 2022. 2, 3, 5, 6, 7, 8

[51] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021. 6, 7, 8

[52] 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. 3, 5, 6, 7, 8

[53] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *ACM MM*, pages 2526–2534, 2020. 2, 3, 5, 6, 7

[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 5

[55] Qiang Zhang and Baoxin Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In *MMAR*, pages 19–24, 2011. 2

[56] 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

[57] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Mark A Clements, and Irfan Essa. Automated assessment of surgical skills using frequency analysis. In *MICCAI*, pages 430–438, 2015. 2

[58] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, and Irfan Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *IJCARS*, 13(3):443–455, 2018. 2