Adaptive Spatiotemporal Graph Transformer Network for Action Quality Assessment

Jiang Liu, Huasheng Wang, Wei Zhou, Katarzyna Stawarz, Padraig Corcoran, Ying Chen and Hantao Liu

***Abstract*—Long video action quality assessment (AQA) aims to evaluate the performance of long-term actions depicted in a video and produce an overall assessment for action quality. A video of long-term actions often contains more complicated temporal and spatial information than that of short-term actions. However, existing approaches that segment a video into individual clips for independent analysis potentially disrupt the narrative flow and diminish contextual details within and across clips, impeding comprehensive video understanding. To address this challenge, we propose an adaptive spatiotemporal graph transformer network (ASGTN) that combines multiple graph structures and transformer attention mechanisms to capture both local and global contextual information within and across clips in a long video. Specifically, the adaptive spatiotemporal graph (ASG) combines a spatial graph branch, designed to enrich the local nuanced spatiotemporal relations within an individual clip, and a temporal graph branch, tailored to dynamically learn the semantic context across different clips. Furthermore, a transformer encoder is integrated to amplify the global dependencies across clips in the entire video. This structure is designed to preserve narrative coherence and maintain essential contextual details in video-level features. Finally, we employ a level-focused decoder to predict the action quality score distribution. Experiments demonstrate that our model achieves state-of-the-art results on popular AQA datasets. Our code is available at https://github.com/jiangliu5/ASGTN_AQA.**

***Index Terms*—Action quality assessment, Graph, transformer, deep learning, neural network.**

I. Introduction

VISION-BASED action quality assessment (AQA) aims to evaluate the execution quality of movement sequences performed by individuals through a given video. This emerging research subject has attracted growing attention in the computer vision community due to its application across various real-world scenarios such as sports event scoring [1], surgical skill evaluation [2], rehabilitation assessment [3] and many other areas [4], [5].

In comparison to human action recognition (HAR) [6], which entails discerning differences among diverse actions, action quality assessment (AQA) is considered more challenging due to the necessity of detecting subtle intra-action differences and the requirement for evaluation across the complete action sequence, as opposed to parts of video frames or segments [4].

Jiang Liu, Huasheng Wang, Wei Zhou, Katarzyna Stawarz, Padraig Corcoran, and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, CF244AG Cardiff, U.K. (email: liuj137@cardiff.ac.uk; lufei.whs@taobao.com; zhouw26@cardiff.ac.uk; stawarzk@cardiff.ac.uk; corcoranp@cardiff.ac.uk; liuh35@cardiff.ac.uk)

Huasheng Wang and Ying Chen are with Alibaba Group, Hangzhou, China. Corresponding author: Huasheng Wang (lufei.whs@taobao.com)

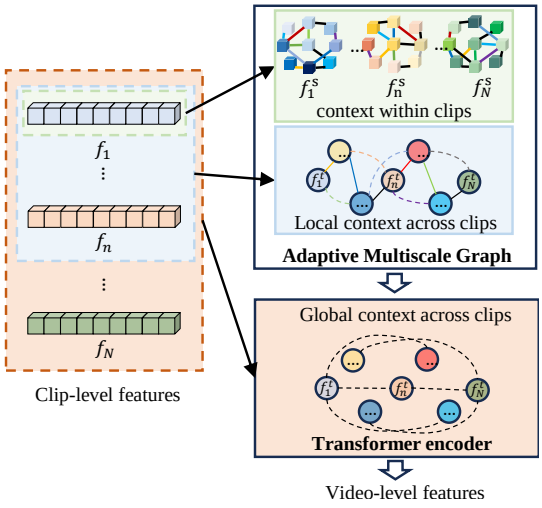


Fig. 1. Key concept of the proposed adaptive spatiotemporal graph transformer network: the method is to capture the intricate local interactions within an individual clip and across clips in an video, as well as the global contextual semantics information of the entire video. To achieve this, we propose an adaptive spatiotemporal graph specifically designed to capture the subtle interactions. In addition, a transformer encoder is integrated to enhance the long-range dependencies, enriching the video feature representation.

For instance, although it is achievable to recognise that an individual is skating from a single frame, assessing the quality of their performance depicted in the entire video presents a significant challenge. This challenge is further amplified in long-duration activities, such as rhythmic gymnastics and figure skating, which contain richer and more complex semantic information. In contrast to video quality assessment (VQA) [7], [8], which often employs a sampling strategy to evaluate overall quality, AQA requires a comprehensive understanding of the entire action sequence to effectively capture performance nuances. Thus, integrating the contextual semantic action information from long video sequences plays a significant role in AQA task. In existing research [9], [10], [11], [12] for assessing long-duration sports, the entire video is initially segmented into clips of equal length. Then, these clips are fed into neural networks such as 3D convolutional neural networks (C3D) [13], [14], inflated 3D convolutional neural networks (I3D) [15] and video swin transformer (VST) [16] to independently extract the spatiotemporal features of each clip. Finally, these clip-level features are averaged to aggregate a video-level representation and regress a final assessment score for action quality. However, the aforementioned segmentation and aggregation methods may face challenges in fully capturing both the local interactions between individual frames

within a clip and the global interactions across multiple clips. While some methods employ average pooling [1], long short-term memory (LSTM) [17], graph [18], [10], [19], [20] and transformer [11] to improve the aggregation and perception of contextual semantics among clip-level features, their performance is rather limited. The limitation mainly stems from how these models aggregate clip-level features over long video sequences. Average pooling treats each clip equally, ignoring the nuanced differences between clips. LSTM struggles with learning the relations across clips in long videos, which often encompass a broader array of actions and temporal cues. GCN-based methods excel at modelling local structured features by focusing on specific regions, but they lack the ability to capture the global dependencies across the entire video which is crucial for assessing overall action quality, particularly in long videos where context is distributed across multiple segments. On the other hand, transformers are adept at capturing global long-range dependencies, which is critical for learning the overall context of long videos. However, they tend to overlook local interactions and the nuanced spatial-temporal relationships within individual clips.

To address the above challenge, we propose an adaptive spatiotemporal graph transformer network to adaptively learn the local and global connections for predicting the performance of long-duration actions, as illustrated in Figure 1. Unlike traditional GCN-based methods that primarily focus on temporal information across clips, our approach can adaptively capture both semantic context across clips and detailed spatiotemporal relationships within each clip. There are two branches in the adaptive spatiotemporal graph module. For the spatial branch, inspired by [21], [22], we reshape the vector and construct a spatial graph to learn the subtle spatiotemporal contextual information within individual clips. To capture the local context across clips, we construct a temporal graph by setting each clip as a node and connecting the clips that are continuous in time. While transformer-based methods excel at capturing long-range dependencies, they encounter limitations in effectively extracting fine-grained local features, particularly within shorter durations. Our proposed approach integrates an adaptive spatiotemporal graph with a transformer encoder, enabling a more comprehensive representation of both local and global dependencies. By leveraging the spatial branch to refine fine-grained details within clips and the temporal branch to establish meaningful connections across clips, our model is particularly well-suited for long-duration video AQA, where both local and global contextual information are essential.

In summary, the contributions of our paper are listed below:

•To enrich the nuanced spatiotemporal relations within an individual clip and across different clips, we propose a novel adaptive spatiotemporal graph module, consisting of both temporal and spatial branches. Additionally, we propose a novel adaptive graph attention block embedded in both branches to further refine and capture the fine-grained relationships. This framework ensures the spatiotemporal interactions are effectively captured, providing a more detailed and robust video representation.

•To simultaneously capture local interactions within and across clips, as well as global dependencies throughout an entire video, we propose a novel spatiotemporal graph transformer framework. This framework combines our spatiotemporal graph with a transformer, providing a more comprehensive approach to model both local and global interactions for long video analysis.

•We conduct extensive experiments on popular AQA datasets, and our results demonstrate that the proposed method achieves state-of-the-art performance, showcasing its effectiveness in addressing the complexities of long video action quality assessment problem.

II. Related Work

A. Action quality assessment (AQA)

Existing work regards AQA as predicting a score from the video sequence of actions. According to the data modality, AQA can be classified into pose-based and video-based methods. The pose-based AQA methods have been developed in recent years but it's still challenging to obtain accurate skeleton information in complicated athletic actions through existing pose estimation algorithms and commercial depth sensors. In some circumstances, some vital clues e.g., splash for diving and props for rhythmic gymnastics which significantly contribute to the final AQA score could be ignored by these methods. Hence, we focus on assessing the action quality by video content in this paper. Many studies have been devoted to short-term video-based AQA. Parmar et al. [1] applied C3D to extract clip-level features, LSTM to aggregate video-level features and SVM for the final score regression. Zhang et al. [23] proposed to utilise the time-aware attention module to capture the relationship between clip-level features extracted by I3D networks. Yu et al. [24] proposed a Contrastive Regression (CoRe) framework by aggregating features of the input video and exemplar video for AQA. Tang et al. [25] proposed to predict the Gaussian distribution of scores rather than a single final AQA score.

While these methods have shown success with short-term videos, long videos encompass a broader array of actions and temporal cues, making score prediction increasingly complex. Xu et al. [9] proposed a self-attentive LSTM and a multi-scale skip LSTM network to jointly learn both local and global sequence information cross clips for assessing figure skating videos. Geng et al. [26] presented skating mixer, an MLP-based model to score figure skating with both auditory and visual information. Zhang et al. [27] proposed a self-supervised learning framework for action quality assessment. Ji et al. [28] proposed a Localization-assisted Uncertainty Score Disentanglement Network (LUSD-Net) that utilizes uncertainty regression to enhance feature disentanglement for scoring figure skating. Du et al. [29] proposed a semantics-guided network (SGN) to transfer knowledge from the semantic domain to the visual domain for scoring figure skating. Zeng et al. [10] proposed combining the static pose information feature in sampled frames and the motion feature in the video to assess the action quality of long videos. In addition, they employed a graph attention module to learn the cross-clips temporal relations in a long video. Nevertheless, they did not further explore the local temporal information among

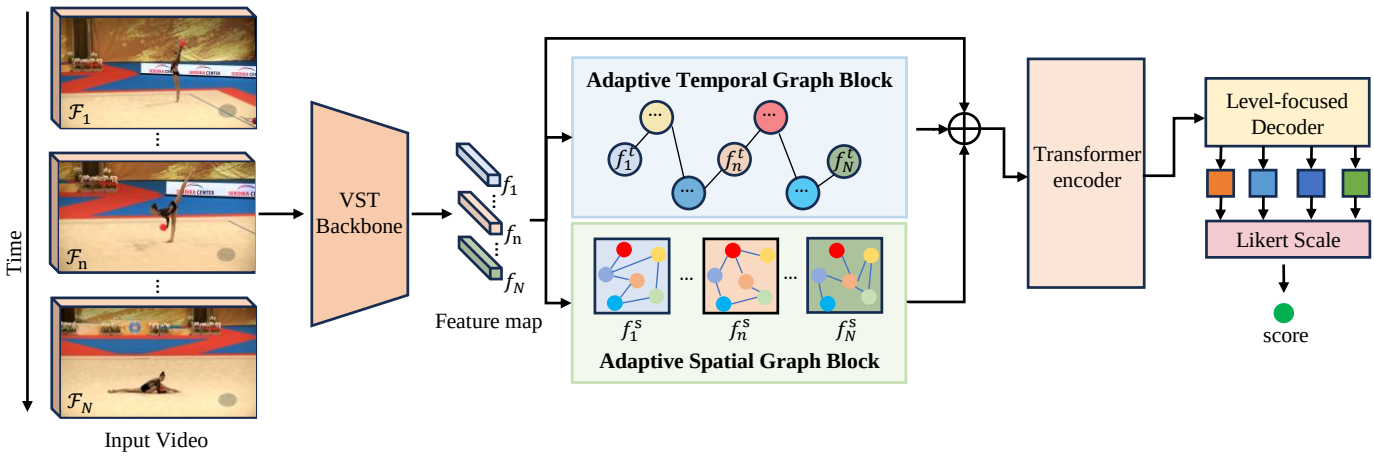


Fig. 2. Schematic overview of our proposed framework. An action video is uniformly segmented into N clips $F_1, \cdots, F_n, \cdots, F_N$. These clips are then input into VST for feature extraction. Next, we propose an adaptive spatiotemporal graph transformer by combining temporal and spatial graph modules and transformer encoder to enhance the local spatiotemporal and global semantic contextual information. Finally, We adopt the level-focused decoder and Likert scale to produce a final score for action quality assessment.

frames within a single clip. Xu et al. [11] assumed that athletes can perform different levels of skill at different parts of a long video, which represents different score contributions to the final AQA assessment. They introduce a transformer to learn the contextual information across clips and extract the representations of different levels. Then a Likert [12] scoring method is designed to combine the different levels for the final score prediction. In this network, while the transformer excels at grasping global long-range dependencies, it tends to overlook the subtleties of local interactions, both within clips and across clips.

B. Graph

In recent years, GCNs have been successfully applied to tasks in computer vision such as surveillance video understanding, action recognition and action quality assessment. Zhang et al. [30] proposed a Structural-Feature Adaptive Fusion Graph Convolutional Network (SFAGCN) consisting of GCN and TCN blocks for AQA. Li et al. [31], [32] employed ST-GCN to extract pose motion features for assessing action quality in long-duration actions such as figure skating. However, these methods are pose-based methods which construct graphs based on the extracted skeleton information instead of video content.

In the context of video-based methods, Zhou et al. [19] proposed a hierarchical GCN framework dedicated to learning the semantic context across video clips. The process begins with the creation of a motion graph, where each clip is treated as a node and connected to its adjacent clips, enhancing the feature representation of each clip. Subsequently, a fixed number of consecutive clips are connected as a scene graph, enabling the aggregation of a unified video-level representation. While this method has been effective for short videos, its effectiveness for videos longer than 1.5 minutes with more scene changes remains unknown.

In terms of long video AQA, Zeng et al. [10] employed GCN to enhance the temporal semantic information by setting all clips as the nodes and adopting an exponential kernel to calculate the adjacent connections for graph construction across clips. However, the method only explores the temporal information across clips and ignore the spatial-temporal information within a single clip.

C. Transformer

Transformers were introduced by Vaswani et al. [33] as a new attention-based mechanism for machine translation. Due to the advanced ability to model global relationships and the success in the NLP field, the transformer model has been widely applied in various computer vision tasks [34], [35], [36], [37]. Some studies [38], [11] have employed transformers for AQA. Bai et al. [38] applied DETR [39] structure to parse video features into a fixed number of temporally ordered temporal representations for short-term action assessment. However, they only employed the decoder module, as they found the encoder tends to smooth the temporal information by clip-level self-attention, thereby impairing the prediction performance. Conversely, Xu et al. [11] applied DETR structure and found the transformer encoder module can learn the global temporal context relationships and improve network performance in the long video AQA. Hence, we hypothesise that while transformers are effective in capturing global extensive temporal dependencies in long videos, their efficacy in extracting local fine-grained features within a shorter duration faces constraints.

III. Approach

In this section, we first describe the overall framework of the proposed adaptive spatiotemporal graph transformer network as illustrated in Figure 2. Then we provide detailed information regarding the proposed Adaptive spatiotemporal Graph (ASG) in Section III-B . The transformer encoder is presented in Section III-C. The level-focused decoder and Likert scale are presented in Section III-D. In Section III-E, we describe the loss function.

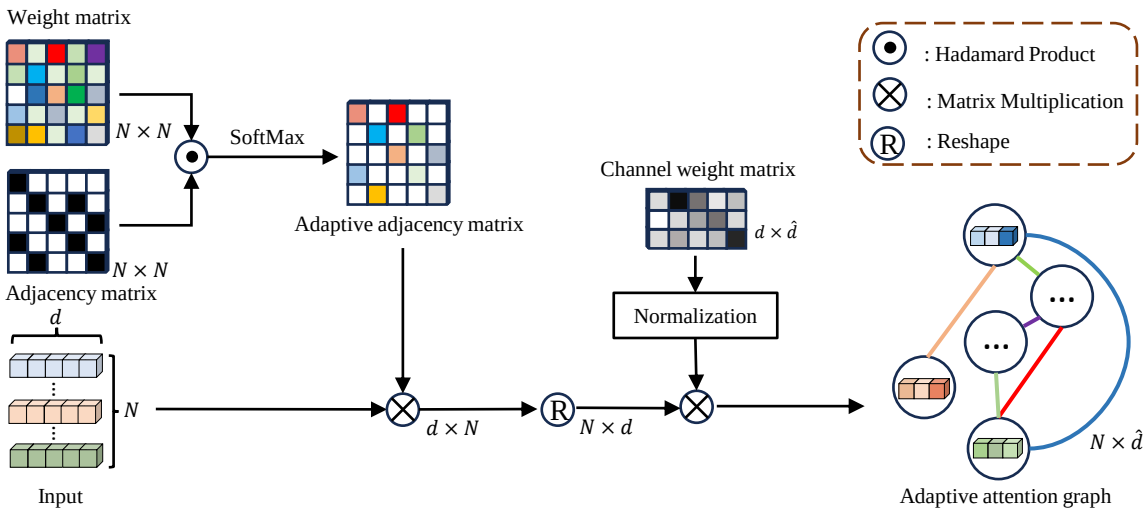


Fig. 3. Illustration of adaptive graph attention block.

A. Overall framework

As shown in Figure 2, following by the previous study [9], [10], [11], we first uniformly divide the whole video sequence F into non-overlapping clips which can be denoted as $F_1, \cdots, F_n, \cdots, F_N$, where N is the number of clips and each clip is composed of 32 consecutive frames. All clips are further sent into VST [16] pre-trained on Kinetics-600 [40], resulting in clip-level features $\{f_n\}_{n=1}^{N}$, where $f_n \in \mathbb{R}^d$.

To establish connections between clips, capture spatiotemporal dependencies, and preserve action continuity throughout the entire video, the features $\{f_n\}_{n=1}^{N}$ are fed into the proposed Adaptive Spatiotemporal Graph (ASG) module, which consists of both a spatial branch and a temporal branch. The spatial branch aims to learn the local spatial-temporal semantic information within each individual clip, resulting in $\{f_n^s\}_{n=1}^{N}$. The temporal branch is to learn the local and global contextual information across clips, resulting in $\{f_n^t\}_{n=1}^{N}$. Following the residual design, we aggregate these features as $\{f_n^{st}\}_{n=1}^{N}$ and send it to the transformer encoder for amplifying the global dependencies across clips in the entire video, resulting in $\{f_n^{gt}\}_{n=1}^{N}$.

More details of ASG and transformer encoder are described in Section III-B and Section III-C. Finally, a level-focused decoder and Likert scale are used for the final score prediction, which is described in Section III-D.

B. Adaptive Spatiotemporal Graph (ASG) Module

To effectively learn both the local fine-grained spatiotemporal information within clips and contextual temporal information across clips, we proposed an Adaptive Spatiotemporal Graph (ASG) module that consists of a spatial adaptive graph block and a temporal adaptive graph block, providing a more detailed and robust video representation.

Adaptive graph attention block. In conventional GCN, let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ denote the constructed graph of N nodes with nodes $v_i \in \mathcal{V}$ and edges $e_{ij} = (v_i, v_j) \in \mathcal{E}$. The edges denote the connection relations between nodes. The corresponding adjacent matrix can be presented as $A \in \mathbb{R}^{N \times N}$. However, the constructed graph considers all the relations as binary connections, which ignores the strength of the connections. To address this limitation and better model the varying strengths of relationships between connected nodes, we propose an adaptive graph attention block to automatically capture the strength of connections among these connected clips. As illustrated in Figure 3, given an input $X \in \mathbb{R}^{d \times N}$ and the adjacency matrix $A \in \mathbb{R}^{N \times N}$, we define a trainable weight matrix as $W \in \mathbb{R}^{N \times N}$ to learn the significance of the connectivity relationship adaptively.

Since each node vector has a 1024-dimensional channel space, fine-grained cross-channel interactions may not be fully captured. To enhance the modelling of these local relationships across nodes, we introduce a weight matrix $\bar{W} \in \mathbb{R}^{d \times \hat{d}}$ to assign importance scores to different channels. Let $\bar{W}_{ij}$ represent the probability of the importance of channel i relative to channel j. To ensure the probabilities sum up to 1 for each row, the $\bar{W}$ is normalised to $\hat{W} = (\hat{w}_{i,j}) = \frac{exp(w_{i,j})}{\sum_{k=1}^{C} exp(w_{k,j})}$. The adaptive graph block can be defined as:

$$F_{AGA} = ReLU((X \cdot SoftMax(A \odot W))^T \cdot \hat{W}), \quad (1)$$

where $\odot$ denotes the element-wise Hadamard product. We applied the proposed adaptive graph attention block in both spatial and temporal branches to capture local fine-grained spatiotemporal features within each clip and across clips. Then, we aggregated these features and sent it to the transformer encoder for amplifying the global dependencies across clips in the entire video.

Spatial branch construction. To capture the local fine-grained spatiotemporal information contained within individual clips, we construct a spatial graph for each clip-level feature $f_n \in \mathbb{R}^{1024}$. First, we reshape the 1024 dimensional vector into 32×32 matrix and treat each dimension as a graph node, then we build edges between adjacent nodes

and the nodes themselves according to the space position. Inspired by [21], [22], we reshape the feature to adaptively learn the potential local clues and fine spatiotemporal semantic information within each individual clip, which potentially contribute to the final prediction. The corresponding adjacent matrix $A_s^{ij} \in \mathbb{R}^{1024\times1024}$ is presented as:

$$\mathbf{A}_s^{ij} = \begin{cases} 0, & \text{if } |r_i - r_j| > 1 \text{ or } |c_i - c_j| > 1, \\ 1, & \text{otherwise,} \end{cases} \tag{2}$$

where r_i and c_i represent the row and column indices of the node i, which are computed by $r_i = \lfloor i/32 \rfloor$ and $c_i = i \mod 32$, respectively. The adjacency matrix $\mathbf{A}_s$ connects each node with its 8 neighbors (including itself), representing the spatially adjacent relations in a 32×32 reshaped matrix from a 1024 dimensional feature vector. Then the proposed adaptive graph attention block is used to learn the fine-grained spatiotemporal semantic information within each clip, which can be denoted as $\{f_n^s\}_{n=1}^N$, where $f_n^s \in \mathbb{R}^d$.

Temporal branch construction. As already mentioned in Section I, existing methods often segment a video into clips of equal length and employ a deep learning backbone to extract features for each clip independently. In this case, the semantic information across clips is ignored, leading to a negative impact on the overall AQA prediction of the entire video. To further learn the temporal semantic information across clips, we construct the adaptive temporal graph. Having obtained the clip-level feature of each clip denoted as $f_n \in \mathbb{R}^{1024}$, we treat each clip as a node. Since there is temporal continuity between adjacent clips, we connect the clips which are adjacent in temporal domain following the approach in [19]. The corresponding adjacent matrix $A_{\mathrm{t}}^{ij} \in \mathbb{R}^{N\times N}$ is defined as:

$$A_{\mathrm{t}}^{ij} = \begin{cases} 1, & \text{if } |i-j| \leq 1 \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where A_{t}^{ij} denotes the connections between the clip features f_i and f_j. Then we input the constructed temporal graph into our AGA block and produce the temporal contextual information across clips as $\{f_n^t\}_{n=1}^N$, where $f_n^t \in \mathbb{R}^d$.

C. Transformer encoder

To enhance the feature extraction of spatial and temporal information within and across clips, we combine $\{f_n\}_{n=1}^N + \alpha\{f_n^t\}_{n=1}^N + \beta\{f_n^s\}_{n=1}^N$ as the input for the transformer encoder, denoted as $\{f_n^{st}\}_{n=1}^N$, where $f_n^{st} \in \mathbb{R}^{32\times1\times1024}$. Here, α and β are adaptive trainable parameters that dynamically adjust their respective weights according to variations in model parameters. Each f_n^{st} captures the subtle spatiotemporal contextual information within individual clips and the local contextual information across clips. To further enhance the long-range dependencies, we employ a transformer encoder to enrich the global context in the whole video, resulting in the feature $\left\{\hat{f_n^{st}}\right\}_{n=1}^N$. The details of transformer encoder is defined as:

$$Output = softmax(\frac{QK^\top}{\sqrt{d_k}}V), \tag{4}$$

where $Q \in \mathbb{R}^{32\times N\times256}$, $K \in \mathbb{R}^{32\times N\times256}$, $V \in \mathbb{R}^{32\times N\times256}$ are derived from the linear projection of input $\{f_n^{st}\}_{n=1}^N$. This process involves leveraging the self-attention mechanism to gather contextual information for each clip. Through weighted aggregation across all clip-level features, the model can capture the global semantic correlations across all clips and capture long-range dependencies. Following contextual enrichment, the augmented contextual information is merged with the original f_n^{st}. Subsequently, these combined vectors undergo further refinement via a small feed-forward network, enhancing their interpretability and usefulness.

To iteratively refine contextual semantics, multiple encoders can be stacked in the model. It should be noted that, while this approach allows for deeper contextual understanding, it escalates computational demands and model complexity, potentially leading to a longer training time and higher resource requirements. The final output of the transformer encoder can be denoted as $\{f_n^{gt}\}_{n=1}^N$, where $f_n^{gt} \in \mathbb{R}^{32\times1\times256}$. These features which enrich the local and global context information within and across clips are pivotal for subsequent processing by the level-focused decoder as detailed.

D. Level-focused decoder

Direct regression typically outputs a single score, which may not adequately capture the distribution of quality features or the level-specific hierarchy inherent in human perception of action quality. To overcome this limitation, we draw inspiration from the DEtection TRansformer (DETR) [11] and introduce a level-focused decoder. This decoder refines the prototypes of different levels, denoted as $\{\hat{p}_k\}_{k=1}^K$, following the self-attention component. These updated prototypes are then utilised in the process known as the level decoupling, where relevant information is extracted from the video feature sequences through cross-attention. During the decoupling process, the module takes $\{f_n^{gt}\}_{n=1}^N$ and $\{p_k^K\}_{k=1}^K$ as input, and produces an output p_t^{agg} which can be interpreted as a "pure substance" containing information exclusively related to specific levels within the video. Subsequently, the obtained p_t^{agg} is utilised to activate video-agnostic prototypes $\{p_k^K\}_{k=1}^K$ by incorporating p_t^{agg} into each prototype. These vectors are then further processed through stacking and the feed-forward network. Multiple decoders may be stacked to refine the output, with each layer serving as input queries for the subsequent layer. The output of the final level-focused decoder layer is denoted as $\{p_t^{att}\}_{k=1}^K$, representing the level-focused features.

To link levels with the quality score, we employ discrete values to quantify each level and generate a quality score by combining the results. The combined weights are estimated from the level-focused features $\{p_t^{att}\}_{k=1}^K$, where each feature acts as a global representation of a specific performance level in the video. We determine a set of discrete values $\{s_k^g\}_{k=1}^K$ to represent individual levels, which remains fixed for a given dataset and are uniformly distributed within the valid score interval [0,1]. Next, we estimate response intensities u_k^g from level-focused features $\{p_t^{att}\}_{k=1}^K$ which are then normalised $\{u_k^g\}_{k=1}^K$ such that their sum is equal to 1, yielding new

TABLE I
PERFORMANCE COMPARISON OF OUR PROPOSED ASGTN MODEL VERSUS THE STATE-OF-THE-ART APPROACHES ON THE RG AND FIS-V DATASETS. AVG. IS THE AVERAGE SRCC VALUE ACROSS ALL CLASSES COMPUTED USING FISHER'S Z-VALUE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Dataset	RG					Fis-V		
Methods	Ball	Clubs	Hoop	Ribbon	Avg.	TES	PCS	Avg.
C3D+SVR [1]	0.375	0.551	0.495	0.516	0.483	0.400	0.590	0.501
MS-LSTM [9]	0.621	0.661	0.670	0.695	0.663	0.660	0.809	0.744
ACTION-NET [10]	0.684	0.737	0.733	0.754	0.728	0.694	0.809	0.757
Skating-Mixer [26]	-	-	-	-	-	0.680	0.820	0.759
GDLT [11]	0.746	0.802	0.765	0.741	0.765	0.685	0.820	0.761
LUSD-NET [28]	-	-	-	-	-	0.679	0.823	0.760
SGN [29]	-	-	-	-	-	0.700	0.830	0.773
PAMFN (only RGB) [41]	0.636	0.720	0.769	0.708	0.711	0.665	0.823	0.755
ASGTN (Ours)	**0.792**	**0.825**	**0.784**	**0.793**	**0.799**	**0.703**	**0.845**	**0.784**

TABLE II
THE PLCC PERFORMANCE OF OUR PROPOSED ASGTN MODEL ON THE RG AND FIS-V DATASETS. AVG. IS THE AVERAGE PLCC VALUE ACROSS ALL CLASSES COMPUTED USING FISHER'S Z-VALUE.

Dataset	RG					Fis-V		
Methods	Ball	Clubs	Hoop	Ribbon	Avg.	TES	PCS	Avg.
ASGTN (Ours)	**0.753**	**0.806**	**0.784**	**0.805**	**0.788**	**0.699**	**0.816**	**0.764**

weights $\{w_k^g\}_{k=1}^K$. Finally, the quality score s is calculated as the weighted sum of level-specific values: $s = \sum_{k=1}^K w_k^g \cdot s_k^g$.

E. Loss function

To regularise different level prototypes, we utilise a triplet loss [42] to ensure adequate separation between level-focused features of distinct levels. For a batch of B videos, the triplet loss for each video is computed as follows:

$$L_{trip} = \frac{1}{K}\sum_{k=1}^{K}\max(0, D_k^+ - D_k^- + \eta), \tag{5}$$

where $D_k^+ = \max_k \text{dist}(p_k^{att}, p_{k'}^{att})$, $D_k^- = \min_k \text{dist}(p_k^{att}, p_{k'}^{att})$, and $\text{dist}(\cdot,\cdot)$ represents the pairwise cosine distance metric. Here, η serves as a margin parameter.

The overall diversity loss, denoted as L_{div}, is then defined as the average of individual triplet losses for all videos in the batch: $L_{div} = \frac{1}{B}\sum_{i=1}^{B} L_{trip}$.

To directly minimise the errors between estimated scores and labels, we employ the mean-squared error (MSE) loss L_{MSE} to constrain the final score and ground truth score, along with the diversity loss term L_{div}:

$$L = L_{MSE} + \lambda L_{div}, \tag{6}$$

where λ is a trade-off hyper-parameter.

IV. EXPERIMENT

We carry out experiments on popular AQA databases,i.e., rhythmic gymnastics and figure skating video datasets to evaluate the proposed method. We first introduce the experiment setting, including datasets, evaluation metrics and implementation details. Then, we conduct a comparative analysis of our proposed model against state-of-the-art methods. Finally, an ablation study is conducted to analyse the effectiveness of our proposed two graph modules and channel attention modules.

A. Datasets and metrics

Figure Skating Video (Fis-V) contains 500 high-quality figure skating videos from international competitions. Each video is about 172 seconds, captured at a frame rate of 25 fps. In addition, Total Element Score (TES) and Total Program Component Score (PCS) given by nine judges to evaluate the performance of the skater at each stage over the whole competition are provided in the dataset.

Rhythmic Gymnastics (RG) contains 1000 videos of 4 types of gymnastic routines from the International Rhythmic Gymnastics Competition, including ball, clubs, hoop and ribbon. Each video is about 95 seconds, captured at a frame rate of 25 fps. Moreover, three types of scores from professional referees are provided: a difficulty score, an execution score and a total score.

Metrics. Following previous research, the Spearman's rank correlation coefficient (SRCC) ρ is adopted to evaluate the performance of AQA models by measuring the divergence between the ground truth scores and the predicted scores. SRCC is defined as follows:

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}}, \tag{7}$$

where p and q represent the ranking of ground truth and predicted scores. To provide a more comprehensive evaluation of our model's performance, we also employ the Pearson linear correlation coefficient (PLCC). A higher coefficient indicates better performance. Fisher's z-value [43] is used to compute the average performance across different classes.

B. Implementation details

We implement our proposed adaptive spatiotemporal graph transformer network, namely ASGTN by PyTorch, where both training and testing are conducted on a single NVIDIA GeForce RTX4090 GPU. Following previous work, we first divide the whole video into non-overlapping clips and each clip is composed of 32 consecutive frames. Then, we input

the clip into Video Swin Transformer (VST) [16] pre-trained on Kinetics-600 [40] for feature extraction and obtain 1024-dimensional clip-level features from the $avgpool$ layer of VST. Similar to [11], we set a fixed number of continuous clips for mini-batch training: 68 for RG and 124 for Fis-V.

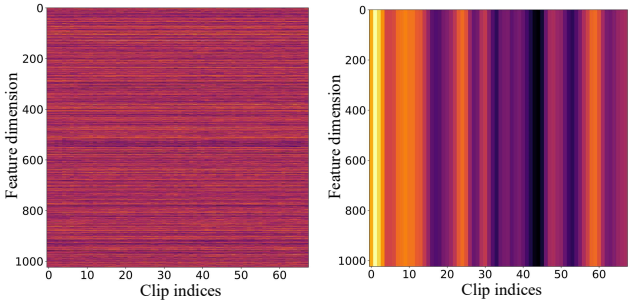
Next, the clip-level features are fed into our adaptive spatiotemporal graph transformer for feature aggregation. In the adaptive spatiotemporal graph model, the AGA block is used twice in both the temporal branch and the spatial branch. According to the clip numbers, we set the embedding dimensions of two AGA blocks in the spatial branch to $d_1^s = 136$, $d_2^s = 68$ for RG and $d_1^s = 248$, $d_2^s = 124$ for Fis-V. The embedding dimensions in the temporal branch are set to $d_1^t = 512$, $d_2^t = 1024$. The parameters α and β are both set to 0.1. The embedding dimension of transformer encoder is set to $D = 256$. We set the number of levels K to 4 for all types of actions. We set the batch size to 32. We use a stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 10^{-4}. The learning rate is set to 0.01 and gradually decreases to 0.0001 by cosine annealing strategy. For better convergence, we set different training epochs on four types of gymnastics routines and two types of skating scores (i.e., 250/400/500/150 for ball/clubs/hoop/ribbon on RG, 320/400 for TES/PCS on Fis-V). The η is set to 1.0 for all models. The λ is set to 1.0 for RG and 0.5 for Fis-V. To regularise the models, we use a dropout of 0.4/0.1 for RG/Fis-V. For splitting dataset, in the RG dataset, we follow the suggested evaluation protocol in [10], [11], we train individual models for distinct types of actions, employing 200 videos for training and 50 for testing. In the Fis-V dataset, we adopt the train-test split from [9], [11], utilizing 400 videos for training and 100 for testing.

C. Comparison with SOTA approaches

Tabe I illustrates the performance comparison of our proposed ASGTN model versus other state-of-the-art (SOTA) methods on the RG and Fis-V datasets. It can be seen that the ASGTN model outperforms the other SOTA methods, as evidenced by achieving the highest SRCC and average SRCC (highlighted in bold). Specifically, the proposed ASGTN model achieves an average SRCC of 0.799 and 0.784 on RG and Fis-V respectively. The performance of our model measured by PLCC is shown in Table II. Our ASGTN model achieves an average PLCC of 0.788 and 0.764 on RG and Fis-V, respectively. The results demonstrate the effectiveness of our model in capturing both local and global inherent dependencies for long-term video AQA.

D. Ablation study

The impact of Graph Transformer. To verify the effectiveness of the proposed adaptive spatiotemporal graph transformer framework, we carry out a series of ablation studies on the RG and Fis-V datasets, as detailed in Table III. We separately add the adaptive spatiotemporal graph (ASG) module which is described in Section III-B and transformer encoder which is described in Section III-C to the baseline model. It can be seen that both ASG and transformer can individually improve the prediction performance. For example, adding the ASG module increased the average SRCC by 5.81% on RG dataset and 5.92% on the Fis-V datasets. Additionally, integrating both components achieves the best and most reliable performance, demonstrating the superiority of the proposed adaptive spatiotemporal graph transformer network.



(a) (b)

Fig. 4. A visual example of temporal context feature maps. (a) is the original clip-level feature map, and (b) is temporal context feature map output by the temporal branch graph extraction in our proposed model.

The impact of Adaptive Spatiotemporal Graph. To show the effectiveness of the proposed Adaptive Spatiotemporal Graph, we create model variants by removing the temporal branch, the spatial branch, both temporal and spatial branches from the full ASGTN model. The corresponding results of this ablation study are shown in Table IV. It can be seen that the performance is reduced by removing any branch, indicating the necessity of including a temporal graph and a spatial graph in the proposed spatiotemporal graph structure. Table IV also shows that combining spatial branch and temporal branch consistently achieves best (1st/ 2nd) performance across both RG and Fis-V dataset and their sub-sets. However, this consistency is not evident for the model variant containing only spatial or temporal branch. Compared to the combined method, using a single branch can cause significant performance decline in some circumstances, e.g., ASGTN w/o SB exhibits 2% drop in "Clubs" and 2.5% drop in "TES". Furthermore, to verify the effectiveness of the adaptive graph attention block (AGAB), an ablation study is conducted by removing the AGAB from the ASG module. The resulting reduction in performance is observed across all actions, indicating the importance of including the AGAB component.

The impact of level-focused decoder. To evaluate the effectiveness of the level-focused decoder, we create a model variant by replacing the level-focused decoder with a direct regression method from the ASGTN model. The results, presented in Table V, show a performance decline when the level-focused decoder is replaced with direct regression. This demonstrates the effectiveness of the level-focused decoder in capturing nuanced quality features and improving model performance.

E. Qualitative analyses

Visualisation of feature maps from temporal branch In Figure 4, we visualise the original clip-level feature map (left)

TABLE III
ABLATION STUDY TO VERIFY THE CONTRIBUTION OF GRAPH AND TRANSFORMER ON THE RG AND FIS-V DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. '+ASG' IS THE MODEL VARIANT OF BASELINE ADDED BY THE ADAPTIVE SPATIOTEMPORAL GRAPH ONLY. '+TRANSFORMER' IS THE MODEL VARIANT OF BASELINE ADDED BY THE TRANSFORMER ENCODER ONLY. '+ASG+TRANSFORMER' IS THE MODEL VARIANT OF BASELINE ADDED BY THE SPATIOTEMPORAL GRAPH AND TRANSFORMER ENCODER COMBINED.

Dataset	RG					Fis-V		
Methods	Ball	Clubs	Hoop	Ribbon	Avg.	TES	PCS	Avg.
Baseline	0.735	0.698	0.680	0.771	0.723	0.612	0.786	0.710
+ ASG	0.770	0.769	0.722	0.793	0.765	0.671	0.815	0.752
+ Transformer	0.756	0.803	0.778	0.783	0.781	0.687	0.824	0.764
+ ASG + Transformer	**0.792**	**0.825**	**0.784**	**0.793**	**0.799**	**0.703**	**0.845**	**0.784**

TABLE IV
ABLATION STUDY TO VERIFY THE CONTRIBUTION OF TEMPORAL BRANCH AND SPATIAL BRANCH IN ASG ON THE RG AND FIS-V DATASETS. 'SB' MEANS SPATIAL BRANCH. 'TB' MEANS TEMPORAL BRANCH.

Dataset	RG					Fis-V		
Methods	Ball	Clubs	Hoop	Ribbon	Avg.	TES	PCS	Avg.
ASGTN	**0.792**	**0.825**	**0.784**	0.793	**0.799**	**0.703**	**0.845**	**0.784**
ASGTN w/o SB & TB	0.756	0.803	0.778	0.783	0.781	0.687	0.824	0.764
ASGTN w/o SB	0.791	0.805	0.771	**0.807**	0.794	0.678	0.840	0.771
ASGTN w/o TB	0.791	0.809	0.770	0.799	0.793	0.691	0.819	0.762
W/o AGAB	0.720	0.736	0.773	0.780	0.753	0.676	0.764	0.741

TABLE V
ABLATION STUDY TO VERIFY THE CONTRIBUTION OF LEVEL-FOCUSED DECODER ON THE RG AND FIS-V DATASETS.

Dataset	RG					Fis-V		
Methods	Ball	Clubs	Hoop	Ribbon	Avg.	TES	PCS	Avg.
ASGTN	**0.792**	**0.825**	**0.784**	**0.793**	**0.799**	**0.703**	**0.845**	**0.784**
ASGTN w/o level-focused decoder	0.699	0.798	0.742	0.712	0.740	0.688	0.827	0.766

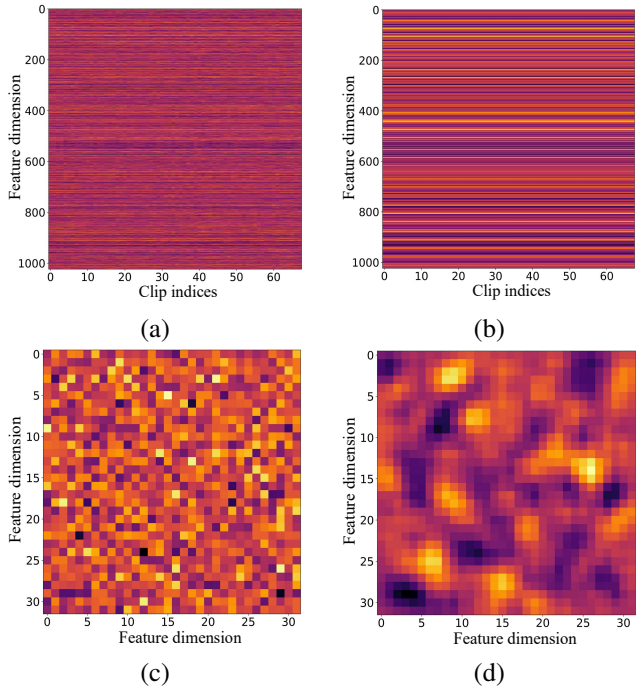
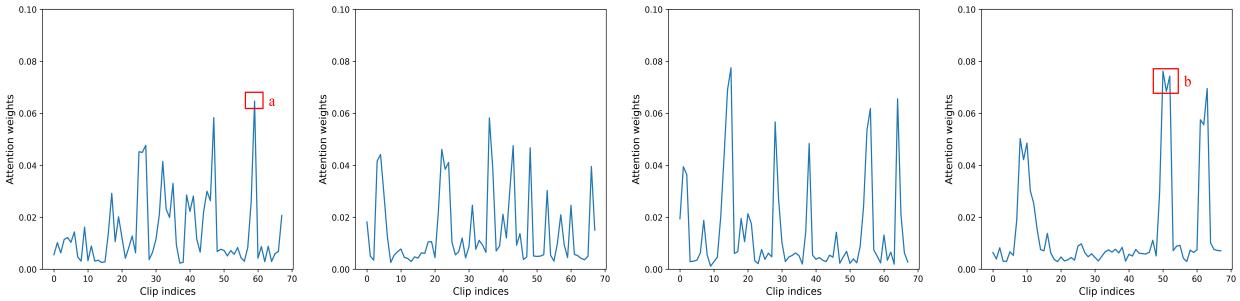


Fig. 5. A visual example of spatial context feature maps. (a) is the original clip-level feature map, and (b) is spatial context feature map. (c) is the original reshaped 32×32 feature map from a 1024-dimension clip. (d) is the reshaped 32×32 feature map extracted as the output of the spatial branch graph extraction in our proposed model.

and the contextual interaction feature map of temporal branch (right) on the RG dataset. The dimensions of the feature maps are 68×1024, where 68 represents the number of clips, and 1024 corresponds to the dimensional representation of each clip. The brighter regions denote areas of heightened attention. In Figure 4, it can be seen that compared with map (a), in map (b) the attention region has the continuity and difference in time domain. It reveals how the temporal branch leverages historical context from preceding clips to inform the analysis of subsequent clips. This is particularly crucial for long-term AQA where the coherence of actions can significantly influence the assessment quality. Thus, the temporal graph branch plays a pivotal role in learning both the individual characteristics of clips and their sequential dependencies, demonstrating its capability to extract temporal interaction information from the original feature map (a).

Visualisation of feature map from spatial branch The local interaction information extracted within each clip from spatial branch is illustrated in Figure 5. Map (a) is original clip-level feature map and map (b) is the feature map of spatial branch. Notably, this comparison reveals the model's ability to discern the significance of different dimensions, emphasising the spatial relationships inherent in the data. To further present the local information within each clip, we show the reshaped 32×32 original feature map from the first clip i.e., (c) and its corresponding reshaped 32×32 feature map extracted from spatial graph branch i.e., (d). The enhanced representation in map (d) highlights the ability of the spatial branch to facilitate interactions among adjacent pixels, thereby enriching the local interaction information within each individual clip. This additional local context strengthens the representation of fine-grained spatiotemporal information, ultimately contributing to the overall performance improvement.



(a) 1st level (b) 2nd level (c) 3rd level (d) 4th level

Fig. 6. Visualisation of cross-attention weights of each level prototype in the last level-focused decoder layer. The sample of #22 video on Ball of the RG dataset. The 1st level represents the poorest performance and the 4th level represents the best performance.

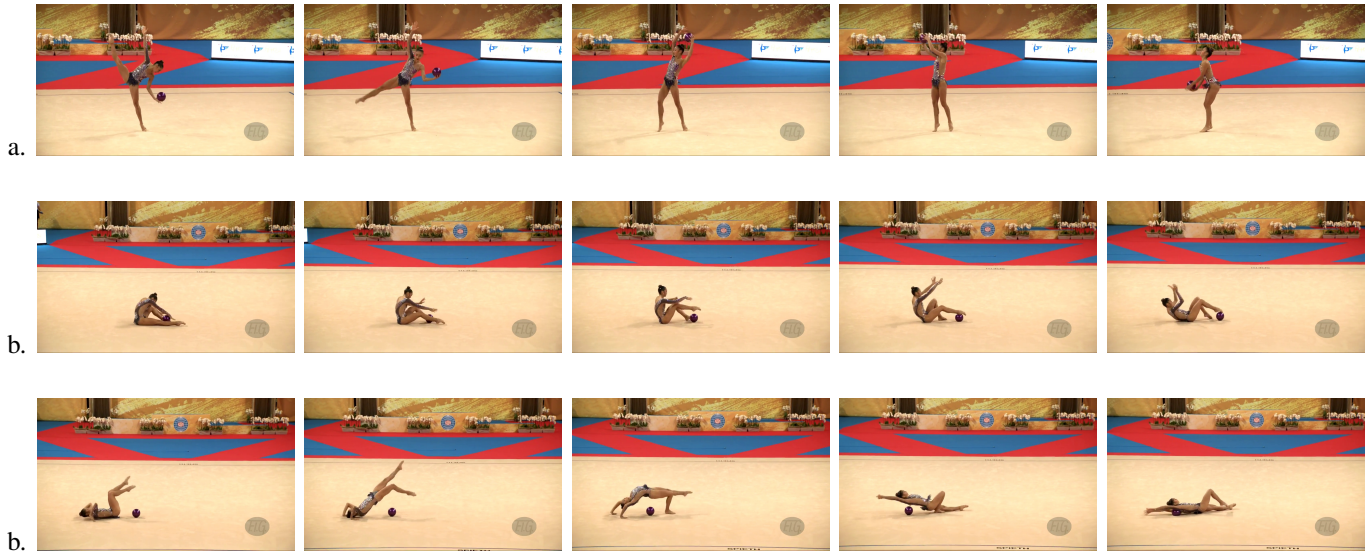


Fig. 7. Video clips with different performance levels. The first row shows a clip at point "a" in Figure 6 with the highest attention in the 1st (poorest performance) level. The second and third rows shows clips at point "b" in Figure 6 with the highest attention in the 4th (best performance) level.

Visualization of cross-attention weights. In Figure 6, we visualise the cross-attention weights of the four levels of prototypes in the last level-focused decoder layer, which shows the cross-attention weights of four levels on each clip. The 1st level represents the poorest performance and the 4th level represents the best performance. To show the video clip with different levels, in Figure 7, we visualise the point marked as "a" with highest attention in the 1st level and the point marked as"b" with highest attention in 4th level. We can see that movement "b" is more skilful than movement "a", which supports the fact that more skillful actions contribute more to higher performance scores.

TABLE VI
COMPARISON OF TRAINING TIME AND PARAMETER COUNT FOR THE PROPOSED ASGTN AND ITS MODEL VARIANTS. 'ASPATIALGTN' MEANS THE ASGTN CONTAINING ONLY THE SPATIAL BRANCH. 'ATEMPORALGTN' MEANS ASGTN CONTAINING ONLY THE TEMPORAL BRANCH.

Model	Training Time (per epoch)	Parameter Count
ASpatialGTN	21.0s	1.91M
ATemporalGTN	21.8s	2.90M
ASGTN (ours)	21.9s	2.96M

V. DISCUSSION

In our proposed ASGTN, we combine the spatial and temporal branches to achieve a more robust and holistic representation. While the combined model (ASGTN) consistently outperforms it model variant containing only spatial or temporal branch, this improvement comes at the cost of increased complexity. In practice, there is often a delicate balance between a model's complexity and its prediction performance, requiring careful consideration of trade-offs depending on the specific application. To analyse complexity, we present measurements of both time consumption and parameter count for our proposed ASGTN and its model variants on the Fis-V dataset, as shown in Table IV. Although the inclusion of dual branches increases the parameter count compared to a single-branch model, the additional time cost remains minimal, especially when weighed against the performance gains.

Recently, some studies [41], [29] have taken multimodal approaches, leveraging diverse data sources such as audio and narration to provide a more comprehensive understanding of actions. Without the need for additional modalities, our

proposed method focuses specifically on enhancing the extraction of action features from RGB video data. Our method aims to enhance the representation of actions, by leveraging the local and global contextual information within and across clips in the whole video. This allows us to effectively capture the intricate details necessary for assessing action quality, which can be beneficial in scenarios where multimodal is not applicable.

To critically evaluate the generalisation ability of our model, a cross-dataset evaluation serves as a valuable approach. However, conducting a reliable cross-dataset evaluation requires larger and more diverse datasets, which are currently limited in the literature. We conducted a preliminary cross-dataset evaluation on the RG dataset, where the ASGTN model was trained on the Ball action and evaluated on the other three actions including Clubs, Hoop, and Ribbon. The resulting SRCC values are 0.333 for Clubs, 0.462 for Hoop, and 0.618 for Ribbon. The relatively low SRCC scores, particularly for the Clubs action, indicate the challenges of generalising across actions with distinct motion dynamics and body postures. This difficulty is likely to be attributed to the significant differences between the motion patterns in Clubs, Hoop, and Ribbon compared to Ball. However, the higher performance on Hoop and Ribbon suggests that the model can still capture certain shared features or similarities between these actions. Future work will focus on expanding the dataset and developing strategies to enhance cross-action generalisation; and explore integrating our method into sports analytics and healthcare. In sports, it could provide objective performance assessments for coaches and athletes. In healthcare, it may aid physiotherapists in tracking rehabilitation and optimising recovery plans.

VI. Conclusion

In this paper, we propose an adaptive spatiotemporal graph transformer to effectively capture both the local and global contextual information within and across clips in a video for action quality assessment. This is achieved by an adaptive spatiotemporal graph, which contains a spatial branch and a temporal branch. The spatial branch can enrich the local fine-grained spatiotemporal information by constructing the spatial graph within each individual clip. The temporal branch can model the temporal relations of clips across the entire video to learn the semantic information. In addition, the transformer encoder is employed to further amplify the long-term relations across different clips. The enhanced video-level features can better predict the Likert score distribution and improve the assessment performance. Experimental results on two long-term AQA datasets demonstrate that our proposed ASGTN model outperforms the state-of-the-art methods.

References

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

[2] Jibin Gao, Jia-Hui Pan, Shao-Jie Zhang, and Wei-Shi Zheng. Automatic modelling for interactive action assessment. *IJCV*, 131(3):659–679, 2023.

[3] Vinay Venkataraman, Pavan Turaga, Nicole Lehrer, Michael Baran, Thanassis Rikakis, and Steven Wolf. Attractor-shape for dynamical analysis of human movement: Applications in stroke rehabilitation and action recognition. In *CVPR*, pages 514–520, 2013.

[4] Jiang Liu, Huasheng Wang, Katarzyna Stawarz, Shiyin Li, Yao Fu, and Hantao Liu. Vision-based human action quality assessment: A systematic review. *Expert Systems with Applications*, page 125642, 2024.

[5] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2021.

[6] Ziwei Zheng, Le Yang, Yulin Wang, Miao Zhang, Lijun He, Gao Huang, and Fan Li. Dynamic spatial focus for efficient compressed video action recognition. *IEEE TCSVT*, 34(2):695–708, 2024.

[7] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE TPAMI*, 45(12):15185–15202, 2023.

[8] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Discovqa: Temporal distortion-content transformers for video quality assessment. *IEEE TCSVT*, 33(9):4840–4854, 2023.

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

[10] 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 Multimedia*, pages 2526–2534, 2020.

[11] 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.

[12] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 1932.

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

[14] Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE TIP*, 30:3098–3112, 2021.

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

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

[17] 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.

[18] Simeng Sun, Tao Yu, Jiahua Xu, Wei Zhou, and Zhibo Chen. Graphiqa: Learning distortion graph representations for blind image quality assessment. *Trans. Multi.*, 25:2912–2925, January 2023.

[19] Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. Hierarchical graph convolutional networks for action quality assessment. *IEEE TCSVT*, 2023.

[20] Minglang Qiao, Mai Xu, Lai Jiang, Peng Lei, Shijie Wen, Yunjin Chen, and Leonid Sigal. Hypersor: Context-aware graph hypernetwork for salient object ranking. *IEEE TPAMI, year=2024, publisher=IEEE*.

[21] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pages 1191–1200, 2022.

[22] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *CVPR*, pages 1140–1149, 2022.

[23] Yu Zhang, Wei Xiong, and Siya Mi. Learning time-aware features for action quality assessment. *Pattern Recognition Letters*, 158:104–110, 2022.

[24] 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.

[25] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou. Uncertainty-aware score distribution learning for action quality assessment. pages 9836–9845, 2020.

[26] Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: long-term sport audio-visual modeling with mlps. AAAI, 2023.

[27] Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Semi-supervised action quality assessment with self-supervised segment feature recovery. *IEEE TCSVT*, 32(9):6017–6028, 2022.

[28] Yanli Ji, Lingfeng Ye, Huili Huang, Lijing Mao, Yang Zhou, and Lingling Gao. Localization-assisted uncertainty score disentanglement network for action quality assessment. In *ACM Multimedia*, MM

'23, page 8590–8597, New York, NY, USA, 2023. Association for Computing Machinery.

[29] Zexing Du, Di He, Xue Wang, and Qing Wang. Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia*, 26:4987–4997, 2024.

[30] Zhitao Zhang, Zhengyou Wang, Shanna Zhuang, and Fuyu Huang. Structure-feature fusion adaptive graph convolutional networks for skeleton-based action recognition. *IEEE Access*, 8:228108–228117, 2020.

[31] Hui-Ying Li, Qing Lei, Hong-Bo Zhang, and Ji-Xiang Du. Skeleton based action quality assessment of figure skating videos. In *International Conference on Information Technology in Medicine and Education (ITME)*, pages 196–200. IEEE, 2021.

[32] Huiying Li, Qing Lei, Hongbo Zhang, Jixiang Du, and Shangce Gao. Skeleton-based deep pose feature learning for action quality assessment on figure skating videos. *Journal of Visual Communication and Image Representation*, 89:103625, 2022.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[34] Huasheng Wang, Jiang Liu, Hongchen Tan, Jianxun Lou, Xiaochang Liu, Wei Zhou, and Hantao Liu. Blind image quality assessment via adaptive graph attention. *IEEE TCSVT*, 2024.

[35] Tie Liu, Shengxi Li, Mai Xu, Li Yang, and Xiaofei Wang. Assessing face image quality: A large-scale database and a transformer method. *IEEE TPAMI*, 2024.

[36] Xin Deng, Enpeng Liu, Chao Gao, Shengxi Li, Shuhang Gu, and Mai Xu. Crosshomo: Cross-modality and cross-resolution homography estimation. *IEEE TPAMI*, 2024.

[37] Fangyuan Gao, Xin Deng, Junpeng Jing, Xin Zou, and Mai Xu. Extremely low bit-rate image compression via invertible image generation. *IEEE TCSVT*, 2023.

[38] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *ECCV*, pages 422–438. Springer, 2022.

[39] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[40] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[41] Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE TIP*, 33:1600–1613, 2024.

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[43] Alan J Faller. An average correlation coefficient. *Journal of Applied Meteorology (1962-1982)*, pages 203–205, 1981.