Fine-Grained Spatio-Temporal Parsing Network for Action Quality Assessment

Kumie Gedamu, Yanli Ji, *Member, IEEE*, Yang Yang, *Senior Member, IEEE*, Jie Shao, *Member, IEEE*, and Heng Tao Shen, *Fellow, IEEE*

***Abstract*— Action Quality Assessment (AQA) plays an important role in video analysis, which is applied to evaluate the quality of specific actions, i.e., sports activities. However, it is still challenging because there are lots of small action discrepancies with similar backgrounds, but current approaches mostly adopt holistic video representations. So that fine-grained intra-class variations are unable to be captured. To address the aforementioned challenge, we propose a Fine-grained Spatio-temporal Parsing Network (FSPN) which is composed of the intra-sequence action parsing module and spatiotemporal multiscale transformer module to learn fine-grained spatiotemporal sub-action representations for more reliable AQA. The intra-sequence action parsing module performs semantical sub-action parsing by mining sub-actions at fine-grained levels. It enables a correct description of the subtle differences between action sequences. The spatiotemporal multiscale transformer module learns motion-oriented action features and obtains their long-range dependencies among sub-actions at different scales. Furthermore, we design a group contrastive loss to train the model and learn more discriminative feature representations for sub-actions without explicit supervision. We exhaustively evaluate our proposed approach in the FineDiving, AQA-7, and MTL-AQA datasets. Extensive experiment results demonstrate the effectiveness and feasibility of our proposed approach, which outperforms the state-of-the-art methods by a significant margin.**

***Index Terms*— Action quality assessment, fine-grained representation, multiscale transformer, action parsing.**

I. Introduction

ACTION Quality Assessment (AQA) has attracted lots of attention in several real-world applications, which is applied to evaluate the quality of specific professional actions such as sports activities [1], [2], [3], [4], [5], [6], and medical skill assessments [7], [8], [9], [10]. As an example, AQA systems can assist doctors in monitoring patients' daily tasks during medical care to evaluate their rehabilitation process. In sports, coaches can employ AQA to train athletes and improve their performance. AQA is different from conventional human action recognition [11], [12], [13], [14], [15] because it focuses on assessing how well the action is performed. These imply that learning coarse-grained action features is less important in AQA. Because, AQA depends on fine-grained sub-action sequences, subtle visual differences, action durations, and poses, which require fine-grained and detailed representations of action features. Therefore, due to a small discrepancy happening with similar environments and poor intra-class discrepancies from the same coarse action category, AQA is a more challenging problem.

Several approaches addressed the challenge of AQA as a regression problem or a pairwise comparison as a ranking problem [3], [4], [5], [8], [9], [10], [16]. Most existing approaches are trained on coarse-grained action features and use holistic video representations. However, these approaches ignore fine-grained features of sub-actions and fail to capture intra-class variations among them. Furthermore, coarse-grained descriptions are insufficient for AQA problems due to small action discrepancies happening in common backgrounds. Thus, it would be helpful to know how fine-grained sub-action sequences along with their temporal dependencies contribute to the final score estimation. Parmar and Morris [3] mentioned that all sub-action sequences contribute equally to the action assessment score prediction. However, due to different poses and subtle visual differences, sub-action sequences contribute differently. Moreover, most sub-actions share similar attributes and common backgrounds, which are biased towards the backgrounds [17], [18]. In this paper, we explore the contribution of contextual information, motion-oriented feature representation, and intra-sequence sub-action parsing for AQA tasks.

Sports activities such as diving action usually involve multiple fine-grained sub-actions as exhibited in Fig. 1(a). For example, the diving sequence can be separated into: *Forward*, *Somersaults*, *Twist*, and *Entry*. However, segmenting action sequences along with their temporal dependence remains a

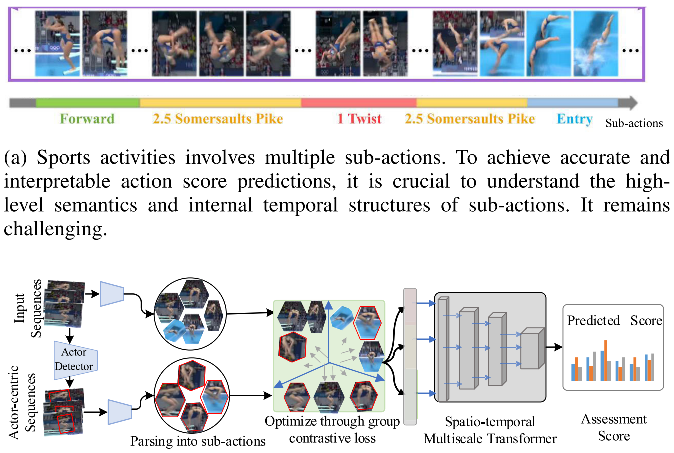
Manuscript received 29 January 2023; revised 8 October 2023; accepted 31 October 2023. Date of publication 14 November 2023; date of current version 20 November 2023. This work was supported in part by the Ministry of Science and Technology of China, Foreign Young Talents Program under Grant QN2023036003L; in part by the Science and Technology Innovation Committee of Shenzhen Municipality Foundation under Grant JCYJ20210324132203007; in part by the National Natural Science Foundation of China under Grant U20B2063; and in part by the Dongguan Songshan Lake Introduction Program of Leading Innovative and Entrepreneurial Talent. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lucio Marcenaro. *(Corresponding author: Yanli Ji.)*

Kumie Gedamu is with the Sichuan Artificial Intelligence Research Institute, Yibin 644000, China, and also with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

Yanli Ji and Jie Shao are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China (e-mail: yanliji@uestc.edu.cn).

Yang Yang and Heng Tao Shen are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

Digital Object Identifier 10.1109/TIP.2023.3331212



(a) Sports activities involves multiple sub-actions. To achieve accurate and interpretable action score predictions, it is crucial to understand the high-level semantics and internal temporal structures of sub-actions. It remains challenging.

(b) Solution in our proposed approach.

Fig. 1. Motivation of our proposed approach. We design an intra-sequence action parsing module to mine sub-actions at fine-grained levels and enable the correct description of action sequences. We further propose a spatiotemporal multiscale transformer module to learn motion-oriented action features and obtain their long-range dependencies among sub-actions at different scales. The proposed solution contributes to learning fine-grained semantic representation of sub-actions for reliable AQA.

challenging task due to: 1) The lack of predefined sub-action label classes and the association between action sequences. 2) The sub-action sequences are more finely granular and their transitions between consecutive segments are often smoother, which makes it difficult to distinguish their boundaries. Besides, sub-actions share similar attributes due to their fine-grained nature and common backgrounds. As a result, a model biased towards such backgrounds performs poorly when it comes to fine-grained tasks like scene-invariant AQA. These observations motivate us to develop an intra-sequence action parsing module that performs semantical sub-action parsing along with their temporal dependencies by mining sub-actions at fine-grained levels. The proposed module enables a correct description of action sequences involving multiple sub-actions. Furthermore, by parsing the given action sequence into separate sub-actions, we can effectively capture the high-level semantics and internal temporal structure of action sequences. To reduce the model's reliance on the video background, we adopted a pre-trained object detection model to extract actor-centric regions from the input videos [19] as shown in Fig. 1(b). Hence, we discover the distinct pattern of sub-actions and explore their internal temporal dependencies.

Fine-grained solutions of sub-actions usually involve temporal dependencies, which are the key to understanding the high-level semantics and estimating AQA scores more accurately. However, learning such temporal dependencies remains a major challenge because it requires modeling internal temporal dependencies between sub-actions. For example, the sub-action "Somersaults" tends to be followed by the sub-action "Twist", and "Twist" usually repeats multiple times in a diving video. Thus, effectively capturing and utilizing these temporal dependencies is crucial in improving the performance of the model. To model the temporal dependencies, the 1D temporal convolution and transformer network are adopted for single-scale feature representation [20]. However, low-level scales lack sufficient semantic descriptions of sub-actions and high-level scales can not provide fine-grained descriptions of sub-actions. Given the aforementioned concerns and to capture high-level semantics and internal temporal dependencies of sub-actions, we propose a spatiotemporal multiscale transformer module. The proposed module learns motion-oriented feature representations and obtains their temporal dependencies among sub-actions at multiple scales. As shown in Fig. 1(b), the actor-centric action features are served as queries and the whole scene action features are served as memories (keys and values). The proposed method expands channel capability in hierarchical order to learn sufficient semantic and internal temporal structure of sub-actions.

In this paper, we propose a Fine-grained Spatiotemporal Parsing Network (FSPN) for AQA, which effectively captures fine-grained action features and sufficient spatiotemporal information. The overall structure of the proposed FSPN is shown in Fig. 2, consisting of intra-sequence action parsing and spatiotemporal multiscale transformer modules. The intra-sequence action parsing module performs semantical sub-action parsing by mining sub-actions at fine-grained levels, enabling a correct description of subtle visual differences in the sub-action sequences. To optimize intra-sequence action parsing, we design a group contrastive loss to perform unsupervised training and learn more discriminative feature representations among sub-actions. The spatiotemporal multiscale transformer module learns motion-oriented action features and obtains their long-range dependencies at multiple scales. Through the multiscale temporal fusion, features are aggregated at each stage to generate a unified feature representation. The integrated representation is used for the final action assessment. In summary, our contributions are as follows:

- We propose a Fine-grained Spatiotemporal Parsing Network for AQA, consisting of intra-sequence action parsing and spatiotemporal multiscale transformer modules.
- The intra-sequence action parsing module mines sub-action sequences at fine-grained levels to obtain semantic sub-action features and enables a correct description of subtle visual differences between action sequences.
- The spatiotemporal multiscale transformer module learns motion-oriented action features and obtains their long-range dependencies among sub-actions at different scales to provide sufficient spatio-temporal information.
- We quantitatively analyze the effectiveness and demonstrate the superiority of the proposed FSPN over the state-of-the-art approaches in three AQA datasets, i.e., FineDiving, MTL-AQA, and AQA-7.

The remainder of this work is arranged as follows: In Section II, related work is presented, including action parsing, and vision transformers. The proposed FSPN is explained in Section III in detail, describing the theory of intra-sequence action parsing and spatiotemporal multiscale transformer. Section IV illustrates the experimental results including ablation studies of FSPN and a comparison with state-of-the-art approaches. Section V gives a conclusion of the work.

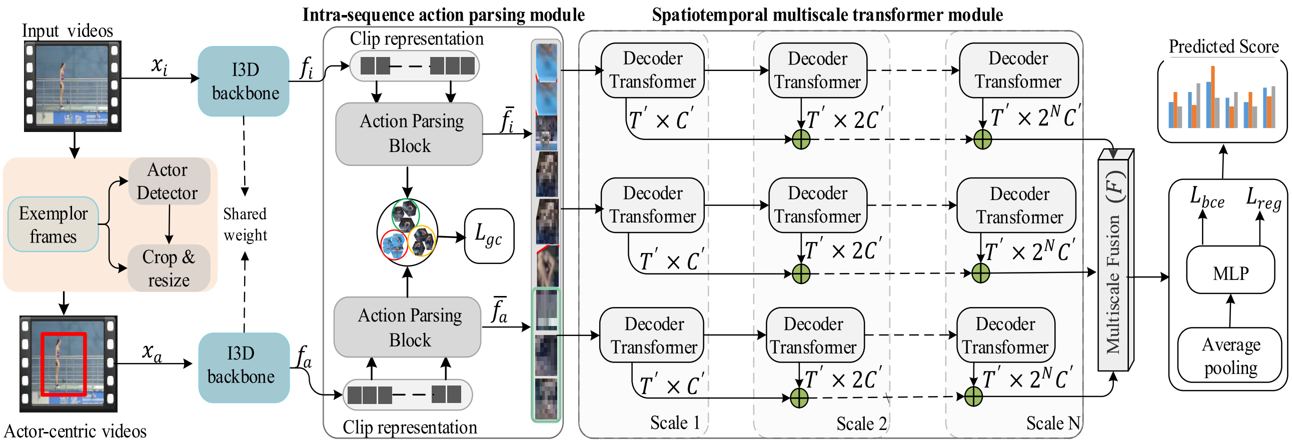


Fig. 2. The overall structure of our FSPN. Initially, an object detector is employed to extract the actor-centric region, denoted as x_a. Utilizing a Siamese I3D backbone [21], we extract spatiotemporal action features represented as f_i and f_a from pairwise inputs x_i and x_a, respectively. We perform intra-sequence action parsing which is performed in an unsupervised manner with group contrastive loss, L_{gc}. With sub-action features $\bar{f}_i$ and $\bar{f}_a$, we propose a spatiotemporal multiscale transformer to learn long-range dependencies of sub-actions at multiple scales. Then, we concatenate multiple scale features to get the final action feature representation $\mathcal{F}$, which will be used for AQA. The binary cross-entropy loss (L_{bce}) and score regression loss (L_{reg}) are set for the network training.

II. Related Work

A. Action Quality Assessment

The AQA is generally conducted based on reliable score labels provided by expert judges [2], [3], [4], [5], [14]. At the early stage, the AQA task is formulated as a classification task that classifies the performance of actions into several levels [22], [23]. Currently, unlike action recognition [13], [24], [25], [26], [27], [28], [29], [30], the AQA mainly follows two kinds of formulations:- 1) the regression formulation, 2) the pairwise ranking formulation.

1) Regression Formulation: The regression formulation is widely used in sporting events [2], [3], [4]. Existing approaches of regression formulation address various challenges including clip-level scoring [4], joint motion and adaptive learning [2], [31], uncertainty-aware [5], transfer learning [18], asymmetric modelling of primary and secondary motions [32] and pose estimation [33]. Specifically, Pirsiavash et al. [3] proposed clip-level scoring and used spatiotemporal action features to estimate AQA scores for sports activities. Pose+DCT [34] extracted joint locations and computed the discrete cosine transformation along the temporal dimension using SVR with fully-connected regressors in the C3D-SVR method [2]. Parmar and Morris [4] proposed another work by conceptualized AQA as a regression problem based on individual joint motion learning in gymnastics competitions and surgical procedures. The hybrid approach is presented by Zeng et al. [35], which combined static and dynamic action information, and considered the contributions of different stages. Xu et al. [16] proposed self-attentive and multiscale skip convolutional LSTM to aggregate information from the individual clips. By constructing the model with KL divergence loss, USDL [5] formulates the score regression as a distribution learning problem. By constructing a pairwise temporal segmentation attention module, Xu et al. [20] proposed a procedure-aware representation for AQA. Zhang et al. [36] introduced a group-aware attention approach, which utilizes graph CNN to incorporate contextual group information and temporal relations. In contrast to these approaches, the FSPN identifies distinct patterns of action features by mining sub-action sequences at fine-grained levels.

2) Pairwise Ranking Formulation: In some daily scenarios, where the performance scores are not available, the AQA problem is re-formulated as a pairwise ranking problem [7], [8], [10], [31], [37]. Doughty et al. [8] used a rank-aware loss function to attend to skill-relevant parts of a given video. An approach to estimate motions, behaviors, and performance assessment in basketball was also proposed. A novel loss function was proposed by Bertasius et al. [37] that learns discriminative features when videos exhibited variance in skill and shared features when videos displayed similar skills. Pairwise action assessment is also proposed using a siamese learning strategy [16]. However, they mainly focus on longer, more ambiguous tasks and only predict overall ranks, limiting their applications to AQA tasks that require some quantitative comparisons. Yu et al. [38] proposed the Contrastive Regression (CoRe) framework to learn the relative scores by pair-wise comparison, highlighting the differences between videos and guiding the models to learn the key hints for action assessment. Bai et al. [39] presented a temporal parsing transformer that decomposes the global features into fine-grained temporal hierarchical representations for AQA tasks. Li et al. [40] proposed pairwise contrastive learning to learn relative scores between pair videos to improve the performance of AQA. Zhou et al. [41] developed a hierarchical GCN for analyzing action procedures and motion units. Their method refines semantic features, reduces information confusion, and aggregates dependencies. Different from these methods, our approach identifies distinct patterns of action features and captures the internal temporal structure by parsing the given action into separate sub-actions without explicit supervision.

B. Action Parsing and Image Quality Assessment

Fine-grained action parsing is also studied in [6], [17], [42], [43], [44], and [45]. Zhang et al. [17] developed

Temporal Query Networks, which ensure relevant segments are addressed by the query. Dian et al. [6] proposed using TransParser to mine sub-actions without examining training data labels. Another aspect of quality evaluation is image quality assessment [46], [47], [48], [49]. The purpose of image quality assessment is different from AQA, which automatically measures the image quality that is highly correlated with the clarity of the image. However, the objective of AQA is to measure the quality of action execution, which includes correctness, completeness, and fluency. Additionally, image quality assessment is based on images rather than videos. Therefore, it is difficult to apply these methods to AQA which requires quantifying actions within videos where human movements are highly valued along the time dimension.

C. Vision Transformer

Transformers [50] play a major role in the current enthusiasm in applying them to vision tasks, particularly the Vision Transformers (ViT) [51] and Detection Transformers [52]. There is a growing interest in the application of transformers to vision tasks, such as object detection [52], [53], [54] and image classification [51], [52], [53], [55], [56], [57]. Some of the works built hierarchical transformer networks to generate pyramid features [54], [58], [59], [60]. By starting from a low-resolution image and a small channel dimension, the stages hierarchically increase channel capacity while reducing spatial resolution [58], [60]. Our proposed multiscale transformer module builds directly upon [52] and allows the expansion of channels in a staged manner to provide sufficient semantic and temporal information through motion-oriented feature representation.

III. Proposed Approach

The FSPN is proposed to understand the high-level semantics and internal temporal dependence of sub-action sequences. The pipeline of the FSPN approach is depicted in Fig. 2. In FSPN, we design intra-sequence action parsing module to capture the description of subtle differences between sub-actions. Then we present a spatiotemporal multiscale transformer module to learn motion-oriented features and obtain their long-range dependencies among sub-actions at different scales. Through the multi-scale temporal fusion, features are finally aggregated to generate a unified feature representation for score estimation.

A. Problem Formulation

Given an input video $x_i \in \mathbb{R}^{T\times H\times W\times C}$ (where T, H, W and C refer to the clip length, height, width, and number of channels, respectively) with action quality score label y_i, the AQA problem is formulated as a regression operation to the predicted score $\bar{y}_i$ based on the input videos,

$$\bar{y}_i = \mathcal{R}_\theta(E_\vartheta(x_i)) \tag{1}$$

where $\mathcal{R}$ and E are the regressor and feature extractor network parameterized by θ and ϑ, respectively. During training, the regression problem is usually solved by minimizing the MSE between the predicted and the ground-truth assessment scores. However, existing methods cannot capture fine-grained visual differences due to holistic video representations.

To this end, we propose a Fine-grained Spatiotemporal Parsing Network (FSPN), which enables us to leverage the fine-grained sub-action patterns without any explicit supervision and provide sufficient spatiotemporal information at different scales. We first employ a pre-trained object detector [19] $D(.)$ to learn action-centric and motion-oriented action features of each video by identifying the main actors to mitigate disturbance from the video background. Specifically, given an input clip x_i, we obtain an actor-centric region $x_a = D(x_i)$, where $a = i$ and $x_a \in \mathbb{R}^{T\times H\times W\times C}$. Then, we feed the pairwise input x_i and x_a into a shared I3D backbone [21] to extract spatiotemporal action features,

$$f_i = E_\vartheta(x_i) \quad f_a = E_\vartheta(x_a) \tag{2}$$

where f_i and f_a are extracted spatiotemporal action features. Then, our proposed FSPN is formulated as a regression-based AQA strategy to capture fine-grained sub-action features and intra-class representation. Specifically, given a mini-batch of whole scene input video $\{x_i^b, y_i^b\}_{b=1}^B$ and actor-centric region $\{x_a^b, y_a^b\}_{b=1}^B$, the AQA can be formulated as:

$$\bar{y}_i = \mathcal{R}_\theta\big(\mathbb{F}_\Theta(E_\vartheta(x_i)), \mathbb{F}_\Theta(E_\vartheta(x_a))\big) \tag{3}$$

where $\mathbb{F}$ is the overall end-to-end representation learning in our proposed approach parameterized by Θ, $\bar{y}_i$ is the predicted action quality score of input action sequences.

B. Intra-Sequence Action Parsing

1) Intra-Sequence Action Parsing (IAP): The main objective of action parsing is to identify the starting and ending frames of sub-action sequences. We design actor-centric intra-sequence action parsing and segment sub-actions from sequences with semantics and temporal correspondence, which is performed in an unsupervised manner. This allows us to extract valuable insights and discover their semantic spatial and temporal information. Specifically, the proposed action parsing can predict the probability distribution of S sub-actions and identify transition steps occurring at the t^{th} frame. We use up-sampling decoder block and MLP layers projection head to construct a distribution probability generator, which projects sampled features to a probability vector A_s, $s = 1, \cdots, S$. The probability vector A_s corresponds to the sub-action instance, which is a middle-level representation of s^{th} sub-action sequence. We define the operation by Intra-sequence Action Parsing (denoted as IAP).

$$[A_1, \cdots, A_S] = \mathrm{IAP}\big(f_i, f_a\big) \tag{4}$$

$$\bar{t}_s = \operatorname*{argmax}_{\frac{T}{S}(s-1)\leq\bar{t}\leq\frac{T}{S}s} A_s(\bar{t}) \tag{5}$$

Here $A_s \in R^S$ is the prediction probability score of s^{th} sub-action, $A_s(\bar{t})$ denotes the predicted probability distribution of sub-action at t^{th} frame, $\bar{t}_s$ is the prediction of the s^{th} sequence. As shown in Equation 5, the representative pattern of the predicted probability, i.e., $\mathrm{argmax}\big(A_s(\bar{t})\big)$ is selected. Thus, the two consecutive frames, i.e., t and $t+1$ have different

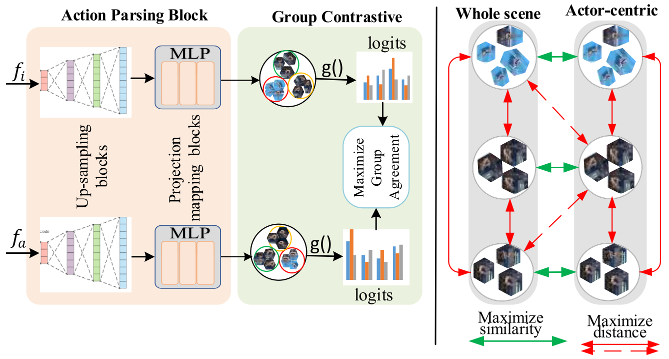


Fig. 3. Overview of intra-sequence action parsing and segment sub-actions from sequences with group contrastive loss. Groups are obtained by clustering features with the same semantic sub-action parsing pseudo-labels. Here, g() is the logits of action sequences.

representations. Following that, the action parsing marks the beginning of a new sub-action instance at the $(t+1)^{th}$ frame.

The proposed intra-sequence action parsing is composed of two blocks, the up-sampling block and MLP projection layers as exhibited in Fig. 3. The up-sampling block consists of four sub-blocks with (1024, 12), (512, 24), (256, 48), and (128, 96) spatial-temporal dimensions. In the blocks, I3D feature lengths in the temporal axis are extended using convolution layers, and spatial dimensions are reduced via max-pooling. The sampled features are projected to a probability vector A_s, using three MLP layers projection. Equation 5 ensures that the predicted sub-actions are being ordered, i.e., $\bar{t}_1 \leq \cdots \leq \bar{t}_s$. In this way, the action parsing problem can be converted into a dense classification and predicts which frames belong to the s^{th} sub-action instance. Meanwhile, existing AQA datasets do not provide fine-grained sub-action labels or step transition labels [2], [3], [4]. Taking advantage of the differences among a group of action sequences, we adopt contrastive learning to make a better representation of action features [61], [62], [63]. Thus, a group contrastive loss is designed to learn more discriminative representation among sub-actions and train the proposed module in an unsupervised manner.

2) Group Contrastive Learning: The actor-centric and whole scene action features, i.e., f_i and f_a, share the same semantic action information, and the similarity between the two input representations must be maximized. However, directly applying contrastive learning between f_a and f_i would push similar sub-action instances and learn different representations for action sequences with the same semantics. Thus, we design a group contrastive learning method to explore relationships within the neighborhood of videos by grouping similar sub-action sequences. Additionally, the problem of different representations of the same semantic sub-action sequences can be avoided by applying group contrastive learning approaches. Hence, the sub-action parsing probability distribution vector of the two inputs is assigned pseudo-labels that correspond to the class having the maximum activation and high-level semantic similarity. Specifically, p denotes the pseudo-label of sub-actions A_s^i and A_s^a. The features having the same pseudo-label in a mini-batch form a group as shown in Equation 6.

$$G_k^p = \frac{\sum_{i=1}^{B} \mathbb{1}_{\{p=A_s^k\}} g(A_s^k)}{T_B} \tag{6}$$

Where g() is the logits of the sequence A_s^k, $k \in \{i, a\}$. $\mathbb{1}$ is an indicator function that evaluates to 1 for the sequence with pseudo-label equal to $p \in S$ in the k^{th} sequences. T_B is the number of such sequences in the mini-batch B. Based on the high-class consistency between the two groups, the two groups should provide similar feature representations. We treat the average representation of groups with the same sub-action pseudo-labels as positive pairs (G_a^p, G_i^p) and sub-actions from the same group video with the varying pseudo-labels as negative pairs (G_a^p, G_k^q). Then, the group contrastive loss function is formulated as Equation 7.

$$L_{gc} = -\log \frac{h(G_a^p, G_i^p)/\tau}{h(G_a^p, G_i^p)/\tau + \sum_{q=1,k}^{S} \mathbb{1}_{p \neq q} h(G_a^p, G_k^q)/\tau} \tag{7}$$

Where $h()$ is the exponential of cosine similarity, τ is the temperature hyperparameters, $G_k^p \in S$, and $G_k^q \in S$. Compared with conventional contrastive learning in our solution the negative samples come from the same action sequence but from different sub-action sequences. Thus, the group contrastive loss optimizes the actor-centric intra-sequence action parsing and generates sub-action sequences $\bar{f}_i$ and $\bar{f}_a$.

C. Spatio-Temporal Multiscale Transformer

The proposed spatiotemporal multiscale transformer uses sub-action features $\bar{f}_i$ and $\bar{f}_a$ to learn long-range dependencies between sub-action sequences at different scales.

1) Actor-Centric Multiscale Transformer: Inspired by the previous methods [58], the channel resolution of our spatiotemporal multiscale transformer varies at different stages of the network, which progressively increases channel dimensions. We gradually increase the channel dimension together with the scale increasing in each stage. The proposed spatiotemporal multiscale transformer has three stages and all stages share a similar structure. Each scale contains three transformer blocks with 8 attention heads that are operated on the same scale with identical channel and spatiotemporal dimensions. Given an input feature with the size of $T' \times C'$, we first apply a 1D convolution layer with kernel $= 3$ and stride $= 1$ to maintain the same number of patches as the parsing output. After that, the intermediate tensors are passed through a multiscale transformer with L blocks, and the output of stage n is reshaped into a dimension of $T' \times 2^n C'$, $n \in \{1, \cdots, N\}$. At different scale stages, it has coarse-grained action features at early layers and fine-grained action features at deeper layers. By starting with a small channel dimension, multiple stages hierarchically expand the channel capacity when a stage transition occurs. Using the MLP, each scale stage's output channel is extended by the factor of 2^n, *i.e*, changing resolution from $T' \times 2^n C'$ to $T' \times 2^{n+1} C'$, $C' \in \{64, 128, 256\}$. Similarly, we obtain the feature maps of the subsequent stages by using features output from the previous

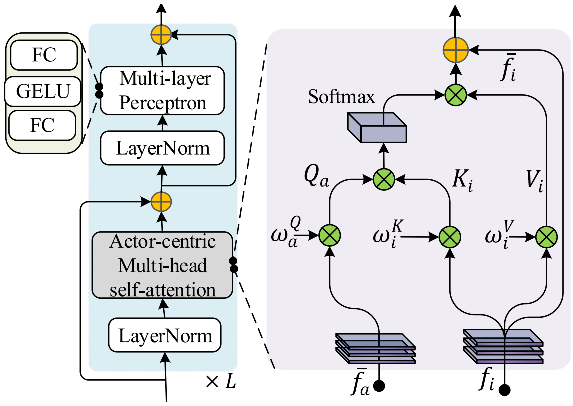


Fig. 4. One block of the spatiotemporal multiscale transformer. The actor-centric feature $\bar{f}_a$ serves as queries, and the input $\bar{f}_i$ serves as memories (keys and values).

stage as input. As illustrated in Fig. 4, the action-centric action features $\bar{f}_a$ are served as queries, and the input action features $\bar{f}_i$ are served as memories (keys and values). The action features $\bar{f}_a$ and $\bar{f}_i$ are projected into three matrices with learnable weight parameters ω_a^Q, ω_i^K, ω_i^V. After the projection, we obtain the intermediate tensor of Q_a, K_i, V_i with the linear operation as Equation 8. Subsequently, attention matrix $\mathbf{A}tten$ is obtained through Equation 9.

$$Q_a = \bar{f}_a\omega_a^Q, \quad K_i = \bar{\boldsymbol{f}}_i\omega_i^K, \ V_i = \bar{f}_i\omega_i^V \tag{8}$$

$$\mathbf{A}tten = \text{Multihead}(\text{LN}(\frac{Q_a(K_i)^\top}{\sqrt{d_k}})V_i) + \bar{f}_i \tag{9}$$

Where $\sqrt{d_k}$ is used to normalize the inner product. Similar to [50], [51], and [52], the proposed module consists of a Multi-head cross-attention learning and several MLP blocks. The learned attention enhances action features with the operation in Equation 10.

$$\hat{\boldsymbol{f}}_i = \text{MLP}(\text{LN}(\mathbf{A}tten)) + \bar{f}_i \tag{10}$$

The LayerNorm (LN) and residual connections are applied before and after the actor-centric multi-head self-attention block. The MLP block contains two linear layers with a GELU non-linearity. Thus, the queries and keys can be more aware of motion-oriented action features and obtain their long-range dependencies of sub-actions at different scales. Multiscale temporal fusion then takes the enhanced output action features from the last spatiotemporal multiscale transformer from each stage and integrates them to create a unified representation.

2) Multiscale Temporal Fusion: To model temporal relationships and generate a unified feature representation with different scales, the multiscale temporal fusion module aggregates features from multiple scales as illustrated in Fig. 5. However, the feature lengths of different scale stages are different. To address this issue, we need to interpolate and reshape them into a predefined size via the up-sampling operation, where different scale features are linearly projected into defined temporal and channel dimensions. Specifically, a linear convolution is applied to reshape upsampled features from the previous scale stages to balance resolution. Thus, the output F_n from the stage $n \in \{1, \cdots, N\}$ can be formulated as Equation 11 and upsampled $T' \times 2^n C'$ to $T' \times 2^{n+1} C'$, then sum with a feature of the next scale F_{n+1}, shown in Equation 12.

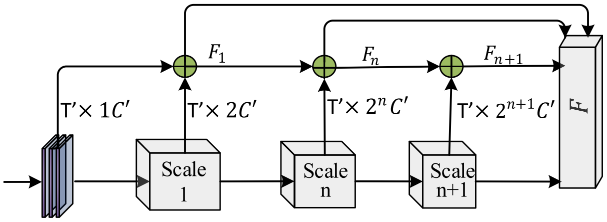


Fig. 5. Multiscale temporal fusion. The output F_n from the stage n is resized and upsampled to $T' \times 2^{n+1} C'$, then being summed with the next scale feature F_{n+1}. The integrated spatiotemporal feature $\mathcal{F}$ is used to estimate AQA score.

$$U_\varphi(F_n) = \text{Upsampling}(F_n \mathcal{W}^n) \tag{11}$$

$$F_{n+1} = U_\varphi(F_n) \oplus F_{n+1}\mathcal{W}^{n+1} \tag{12}$$

$$\mathcal{F} = \text{Concat}(F_1, \cdots, F_N) \tag{13}$$

Where $\mathcal{W}^n \in \mathbb{R}^{T' \times 2^n C}$ refers to an upsampling rate of n, $\oplus$ indicates the element-wise addition. Finally, all the refined features F_n have the same feature length, and we concatenate them to get the final multiscale feature representation $\mathcal{F}$ in Equation 13. In this way, our proposed method effectively fuses each stage of spatiotemporal action features and creates a unified feature representation. Then, the integrated feature representation is used to estimate the final AQA scores.

D. Optimization and Inference

1) Overall Training Loss: We adopt average pooling on the multiscale feature representation $\mathcal{F}$ and then utilize two MLP layers to estimate classification label Υ_i and regression target y_i. The L_{bce} loss is operated on each feature pair to generate a one-hot classification label Υ_i. Similarly to CoRe [38] without utilizing tree structure indicates whether the ground truth score y_i lies in the i^{th} interval. Then, the L_{bce} loss is shown in Equation 14. We optimize regressor R_θ by computing the MSE between the ground-truth y_i and prediction $\bar{y}_i$ in Equation 15.

$$L_{bce} = -\sum_{i=1}^{I}(\Upsilon_i \log(\bar{\Upsilon}_i) + (1 - \Upsilon_i)\log(1 - \bar{\Upsilon}_i)) \tag{14}$$

$$L_{reg} = \sum_{i=1}^{I} \mathbb{1}(\Upsilon_i = 1)(\|\bar{y}_i - y_i\|^2) \tag{15}$$

Where $\bar{\Upsilon}_i$ is the predicted classification probability. Therefore, the overall objective function of our proposed FSPN is formulated as Equation 16.

$$L_{total} = L_{gc} + L_{bce} + L_{reg} \tag{16}$$

2) Inference: We construct video pairs M, $\{x_t^m, x_a^m\}_{m=1}^M$ following [20], [38] for fair comparisons. Then, the overall objective function of our proposed FSPN can be written as:

$$\bar{y}_{it} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{R}_\theta(\mathbb{F}_\Theta(E_\vartheta(x_t^m)), \mathbb{F}_\Theta(E_\vartheta(x_a^m))) \tag{17}$$

where $\bar{y}_{it}$ is the predicted action quality score of test action sequence x_t and $\mathbb{F}$ is the feature representation learning in our proposed approach parameterized by Θ.

IV. Experimental Results

A. Datasets and Experiment Settings

1) Datasets: We evaluate our proposed FSPN approach on three AQA datasets, FineDiving dataset [20], MTL-AQA dataset [4], and AQA-7 dataset [3].

a) FineDiving: The FineDiving dataset consists of 3000 video samples, covering 52 action types, 29 sub-action types, and 23 Difficulty Degree types. This dataset differs from existing AQA datasets in terms of annotation type and scale. The dataset provides fine-grained annotations that include action types, sub-action types, coarse-grained and fine-grained temporal boundaries, as well as action scores, in contrast to MIT-Dive, UNLV, and AQA-7-Dive datasets. MTL-AQA [4] provides coarse-grained annotations, such as action types and temporal boundaries. However, FineDiving is the first fine-grained dataset with detailed annotations of different sports videos for AQA evaluation.

b) AQA-7: The AQA-7 dataset contains 1189 samples from seven different actions collected from the Winter and Summer Olympic Games. It is composed of two previously released datasets: UNLV-Dive [4] which is named single-diving-10m platform in AQA-7, containing 370 samples; UNLV-Vault [4] which is named gymnastic-vault in AQA-7, containing 176 samples. Other action classes are newly collected in this dataset: synchronous-diving-3m springboard involving 88 samples, and synchronous-diving-10m platform involving 91 samples. The big air skiing sample collection consists of 175 samples and the big air snowboarding sample collection consists of 206 samples.

c) MTL-AQA: The MTL-AQA dataset focuses on diving, covering a wide range of activities. There are 1412 samples collected from 16 different world events. Different annotations are available in this dataset to support research on different tasks, such as AQA, action recognition, and comment generation. Additionally, raw annotations of score and difficulty (DD) are available from multiple judges. Following the evaluation protocol in frameworks [4], [38], we separate the dataset into a training set with 1059 samples and a test set with 353 samples.

2) Implementation Details: We implement the proposed approach using the PyTorch deep learning [64]. The I3D backbone [21] is trained using the Kinetics-400 dataset and utilized as the feature extractor with a learning rate of 10^{-4}. We utilize Adam optimizer with weight decay as zero [65]. In the proposed approach, we have three stages with three spatiotemporal multiscale transformer blocks. In the spatiotemporal multiscale transformer, the number of attention heads is set to 8. The initial learning rate is set to $1e-3$ for the proposed FSPN. Our model is trained with a batch size of 8 with 200 epochs. Following [20], we extract 96 frames for each video in the FineDiving dataset and split them into 9 clips. In the experiment on AQA-7 and MTL-AQA datasets, we follow [38] to extract 103 frames for each video clip and segment them into 10 overlapping snippets, each containing 16 continuous frames. Since there are annotations about the Degree of Difficulty (DD) in the MTL-AQA dataset. To obtain the final result, we multiply the predicted score by the ground-truth DD in the inference time.

TABLE I

Evaluation on Components of FSPN in the FineDiving Dataset

Approaches	IAP	SMT	MTF	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}$ ($\times$100) ↓
I3D+MLP	-	-	-	0.850	0.583
I3D+IAP	✓	×	×	0.891	0.468
I3D+SMT	×	✓	×	0.928	0.332
I3D+IAP+SMT	✓	✓	×	0.934	0.300
I3D+SMT+MTF	×	✓	✓	0.936	0.289
FSPN (Ours)	✓	✓	✓	**0.942**	**0.278**

3) Evaluation Metrics: To keep alignment with existing approaches [5], [20], [38], we adopt two metrics to measure the performance of our approach, i.e., the Spearman's rank correlation (Spr. Corr. ρ, ranging from -1 to 1, higher is the better) and relative $R_{\mathcal{L}_2}$ distance (lower is the better). The Spr. Corr. was adopted as our main evaluation metric to measure the difference between the predicted and ground-truth series:

$$\rho = \frac{\sum_j (y_i - y)(\bar{y}_i - \bar{y})}{\sqrt{\sum_i (y_i - y)^2 \sum_i (\bar{y}_i - \bar{y})^2}} \tag{18}$$

where y_i and $\bar{y}_i$ represent the ranking for each sample of two series, respectively. The relative $R_{\mathcal{L}_2}$ measures the numerical precision of each sample compared with ground truth [38]. If y_{max} and y_{min} are the highest and lowest scores for an action, respectively, then $R_{\mathcal{L}_2}$ is defined as follows:

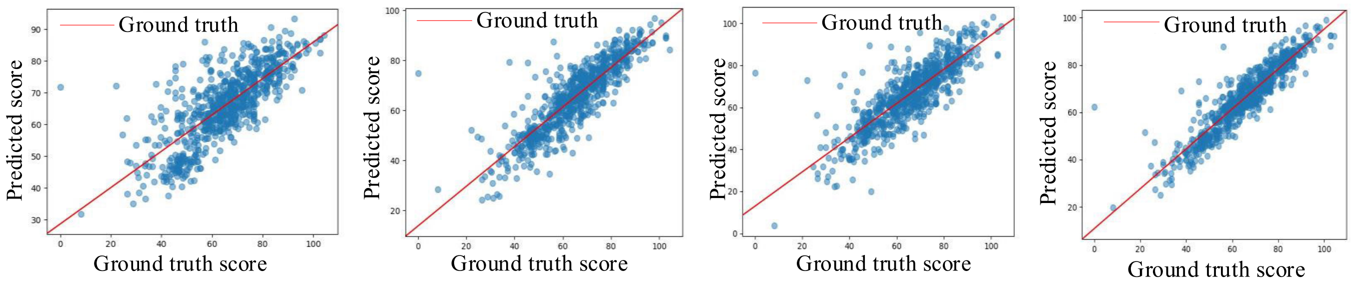
$$R_{\mathcal{L}_2} = \frac{1}{I} \sum_{i=1}^{I} \Big(\frac{|y_i - \bar{y}_i|}{y_{max} - y_{min}} \Big) \tag{19}$$

Here y_i and $\bar{y}_i$ refer to the ground truth score and predicted score for the i^{th} sample, respectively. The average value of Spr. Corr. for all action categories is calculated from the individual per-action correlations using Fisher's z-value, as described in [66].

B. Ablation Study

1) Evaluation on the Components of FSPN: We discuss the effectiveness of each proposed module of our proposed FSPN model in the FineDiving dataset [20]. These modules include the Intra-sequence Action Parsing (IAP), the Spatiotemporal Multiscale Transformer (SMT), and the Multiscale Temporal Fusion (MTF). In the experiment, there are N = 3 stages with L = 3 spatiotemporal multiscale transformer blocks, each consisting of 8 attention heads in all configurations. The I3D backbone [21] with three MLPs (I3D+MLP) is the baseline model. Optimization of the baseline model is based on the MSE loss between the prediction and the ground truth.

a) Intra-sequence Action Parsing (IAP): On top of the I3D backbone, adding our proposed IAP module (I3D+IAP) significantly improves the model performance by +4.08%, which is 0.891 on Spr. Corr. and 0.468 on relative $R_{\mathcal{L}_2}$ distance as shown in Table I. These results imply that the proposed intra-sequence action parsing effectively identifies distinct patterns of sub-actions, and captures the internal temporal structure by parsing the given action into separate sub-actions without explicit supervision.



(a) I3D+MLP (b) I3D+IAP (c) I3D+SMT (d) FSPN (Ours)

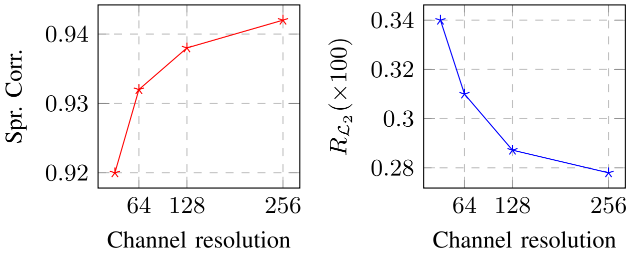
Fig. 6. Comparison of prediction score distributions obtained by different solutions, i.e., I3D+MLP, I3D+IAP, I3D+SMT and FSPN, in the FineDiving dataset. The red line refers to the ground truth scores, while the blue points represent the predicted scores.

b) Spatiotemporal Multiscale Transformer (SMT): We also observe that on top of the I3D backbone, adding our proposed SMT (I3D+SMT) improves the performance of AQA by 7%, which is 0.928% on Spr. Corr. and 0.332 on $R_{\mathcal{L}_2}$ compare to the baseline as shown in Table I. This is a testament to the effectiveness of the proposed module, indicating its potential utility in improving middle-level representations and capturing motion-oriented features to maintain consistent video representation. Moreover, the proposed module captures the coarse-grained features in the early stage and fine-grained features at a later stage, along with their temporal dependencies.

c) Multiscale Temporal Fusion (MTF): Adding our MTF module to aggregate multiple scale features (I3D+SMT+MTF) also gives a performance improvement, which achieves Spr. Corr. of 0.936 and relative $R_{\mathcal{L}_2}$ distance of 0.289 with increasing a minimal computational cost. While adding all the components, our FSPN achieves a significance of +9.17% performance improvement on Spr. Corr. validating that each component in FSPN contributes to the final AQA prediction score. The remarkable improvements underscore the significant contributions made by each module enabling our approach effectiveness to represent fine-grained action features.

d) Visualization of prediction score distributions for different solutions: We draw the scatter plots of prediction score distributions in Fig. 6 when we use different solutions for the AQA task. We compare three solutions with our proposed FSPN approach, i.e., I3D+MLP, I3D+IAP, I3D+SMT. As shown in Fig. 6, the scatter plot visually demonstrates the effectiveness of our approach and the contribution of two modules, IAP and SMT. It can be seen that our modules, IAP and SMT, have much more compact distributions of predicted scores than the baseline I3D+MLP, where prediction scores are much closer to the ground-truth line. By incorporating all proposed components, the resulting distribution became more evenly dispersed, indicating the successful capture of underlying patterns in the video. This led to a superior understanding of fine-grained features and a significant enhancement in performance.

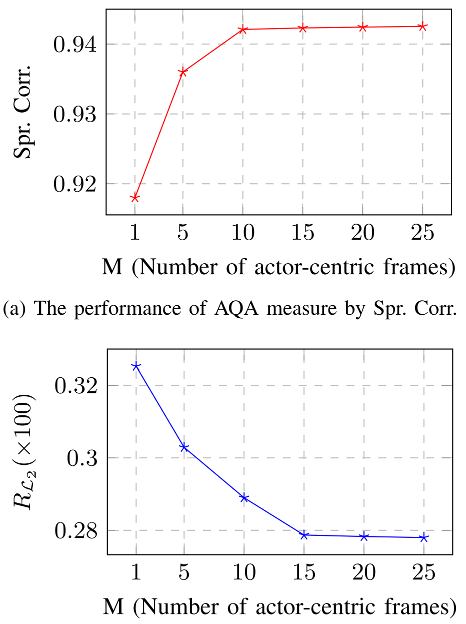
2) Effect of Channel Resolution for Action Feature Learning: In our proposed multiscale transformer approach, we extend [52] to provide sufficient spatial and temporal information by representing motion-oriented feature representations in different channel resolutions. We provide experimental results that evaluate the effectiveness of different channel resolutions as shown in Fig. 7. We design three variants of channel resolution, namely 64, 128, and 256. As shown in the figure, a multiscale transformer with two or more channel resolutions gives a significant performance improvement over a single-scale transformer. Specifically, the performance improves from 0.9182 to 0.942 in terms of the Spr. Corr. and the $R_{\mathcal{L}_2}$ distance decreases from 0.34 to 0.27 when transitioning from a channel resolution of 64 to 256. These results imply that incorporating multiple channel resolutions effectively captures spatiotemporal dependencies and subtle differences in sub-actions, allowing for the capture of coarse-grained features early on and fine-grained features later.



(a) The Spr. Corr. ρ score (b) The $R_{\mathcal{L}_2}$ distance

Fig. 7. Effect of channel resolution for action feature learning in the FineDiving dataset.

3) Effects of Frame Number in Actor-Centric Branch at Inference Phase: To evaluate the effectiveness of the actor-centric frame during the inference phase, we conducted a comprehensive analysis. The illustration presented in Fig. 8 shows the evaluation of the significance of the number of actor-centric frames for FSPN during the inference stage. The number of actor-centric frames is denoted as $M \in \{1, 5, 10, 15, 20, 25\}$. Specifically, we adopt a multi-exemplar voting strategy [38] for this evaluation at inference time, which balances better performance and computational cost trade-off. The Spr. Corr. and $R_{\mathcal{L}_2}$ results are shown in Fig. 8. Increasing M from 1 to 15 progressively improves the performance of the FSPN. However, the model tends to remain constant for M>15 (i.e., 25 being the upper limit). This phenomenon indicates that densely sampling M>15 significantly increases the computational cost since the model needs to process a larger number of actor-centric frames without a significant



(a) The performance of AQA measure by Spr. Corr.

(b) The performance of AQA measure by $R_{\mathcal{L}_2}$ distance

Fig. 8. Effects of the actor-centric frame number at inference phase in the FineDiving dataset.

TABLE II
EVALUATION ON THE NUMBER OF TRANSFORMER BLOCKS IN THE FINEDIVING DATASET

Blocks (L)	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}(\times 100)$ ↓
1	0.936	0.278
3	**0.942**	**0.278**
5	0.933	0.289
7	0.929	0.313

performance improvement. Moreover, sampling more exemplar frames could introduce redundant information due to larger temporal dependencies, thereby making the model more sensitive to temporal alignments across different videos.

4) Evaluation on the Number of Transformer Blocks: A multi-scale transformer with multiple stages significantly improves performance compared to a single-scale transformer. To verify our hypothesis, we conducted experiments on the performance of AQA depending on the number of transformer blocks for our proposed FSPN. Table II shows the effect of stacking different numbers of transformer blocks. It can be observed that when $L = 3$, the assessment performance reaches the highest action quality score of 0.942 on Spr. Corr. and a relative $R_{\mathcal{L}_2}$ distance of 0.278. When the number of blocks exceeds 5, the performance tends to plateau, with a slight decrease in the AQA performance. This could indicate an overfitting problem that arises from the increasing number of transformer blocks.

5) Evaluation on the Number of Attention Heads: Multi-head self-attention approaches are known for their superior performance in the spatial-temporal domains, which explore the potential dependencies between semantic and temporal information adaptively. Table III shows the comparative results on the different numbers of attention heads. A multi-head attention layer divides channels into different groups based on the number of attention heads. To model the spatiotemporal multiscale transformer, each set of features is sent to a multi-head attention head. The number of parameters remains the same despite changing the number of attention heads. As the number of heads increases, it becomes possible to model more complex relationships. The number of channels processed by each head decreases as the number of heads increases. We select even numbers of attention heads (these numbers have no relationship with results). The performance improves gradually when the multi-head number increases from 2 to 8. The prediction accuracy tends to be saturated when the multi-head number is greater than 8. Our proposed FSPN with 8 parallel attention heads achieves the best performance.

6) Evaluation on Loss Functions: To assess the effectiveness of the proposed loss functions, we conducted experiments on the Fine-Diving dataset. The obtained results are presented in Table IV. Notably, we observed a significant drop in performance from 0.932 to 0.891 on the Spr. Corr. and an increase in $R_{\mathcal{L}2}$ from 0.314 to 0.468 when the l_{gc} loss was removed. This result highlights the importance of the proposed loss in enabling the model to learn temporal diversity and discriminative action features among sub-actions without explicit supervision. Similarly, we found that removing the L_{bce} loss resulted in a performance reduction from 0.942 to 0.932. This result suggests that the proposed loss functions facilitate the model in learning spatiotemporal action features and acquiring an understanding of temporal dynamics. Consequently, our FSPN demonstrates enhanced learning of action features when both loss functions are simultaneously optimized.

7) Effect of Learning Action Features From Actor-Centric Regions: To demonstrate the significance of learning from the actor-centric region, we removed the actor-centric branch and learned the action features directly from the whole scene action information. The experimental results are shown in Table V. When the actor-centric branch is removed, the performance significantly drops from 0.942 to 0.918 on Spr. Corr. and an increase in $R_{\mathcal{L}_2}$ distance from 0.278 to 0.325. This drop can be attributed to the fine-grained nature of the action with a small discrepancy happening in similar backgrounds, which introduces bias from the video background. Our evaluation highlights the importance of learning from motion-oriented action features as a vital component for achieving superior performance. Additionally, it facilitates the learning of discriminative action features with internal temporal dependence on sub-action sequences.

8) Effect of Learning Action Features From Multiscale Transformer: In order to assess the impact of learning action features from our multi-scale transformer approach, we conducted a comparative analysis between different scale transformer. The purpose is to examine the effectiveness and advantages of incorporating multiple scales in capturing action information. The results, presented in Table VI, demonstrate a significant improvement in AQA performance achieved by the spatiotemporal multiscale transformer. Specifically, when transitioning from our lightweight model (FSPN-S) to the deeper model (FSPN), there is an increase in the AQA score from

TABLE III
EVALUATION ON THE NUMBER OF ATTENTION HEADS IN THE FINEDIVING DATASET

# Heads	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}(\times 100)$ ↓
2	0.931	0.317
4	0.929	0.313
6	0.933	0.289
8	**0.942**	**0.278**
10	0.938	0.289
12	0.935	0.299

TABLE IV
EVALUATION ON LOSS FUNCTIONS IN THE FINEDIVING DATASET

Approaches	l_{gc}	l_{bce}	L_{reg}	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}$ ($\times 100$) ↓
I3D+MLP	×	×	✓	0.850	0.583
I3D+SMT+MTF	×	✓	✓	0.891	0.468
I3D+IAP+MLP	✓	×	✓	0.932	0.314
FSPN (Ours)	✓	✓	✓	**0.942**	**0.278**

TABLE V
EFFECT OF LEARNING ACTION FEATURES FROM ACTOR-CENTRIC REGIONS IN THE FINEDIVING DATASET. W/O DENOTES WITHOUT ACTOR-CENTRIC BRANCH

Approaches	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}(\times 100)$ ↓
I3D+MLP	0.850	0.583
FSPN w/o	0.918	0.325
FSPN (Ours)	0.942	0.278

0.932 to 0.942 on the Spr. Corr. and a decrease from 0.300 to 0.278 on the $R_{\mathcal{L}_2}$ distance. The hierarchical multiscale transformer design demonstrates its effectiveness in understanding temporal action features, resulting in improved performance. Thus, the proposed approach captures fine-grained details and high-level action features, enhancing its ability to represent fine-grained action patterns in scene-invariant AQA scenarios. Furthermore, this design effectively captures spatiotemporal dependencies and subtle differences in sub-action sequences.

9) Evaluation on Model Complexity: A comparison of model complexity was conducted between our FSPN and related methods, as shown in Table VII. Our approach achieves a Spr. Corr. of 0.9601 and $R_{\mathcal{L}_2}$ distance of 0.221 on MTL-AQA dataset, outperforming HGCN [41] with fewer FLOPs. However, it is noted that our FSPN exhibits a slightly higher number of trainable parameters compared to HGCN [41]. This increase in complexity is due to the incorporation of multiple channel resolutions in our approach. The primary reason for this deliberate design choice is to effectively capture spatiotemporal dependencies and subtle differences in sub-actions at different scales, which are inherent challenges in AQA. It is crucial to address these challenges to achieve superior performance with the integration of multi-scale transformers. Our experiments have consistently demonstrated that the additional complexity of the model significantly improves its ability to capture fine-grained patterns. We have undertaken comprehensive experiments to ensure that the benefits obtained from this increased complexity justify the associated costs. Our experiments have consistently shown remarkable performance gains with this approach when compared to the SOTA methods on AQA tasks. Moreover, we have taken the necessary steps to optimize the computational efficiency of our model during training. We have used techniques such as batch normalization, and parameter sharing (via Siamese Networks) to mitigate any unnecessary computational burden and make the model more efficient.

TABLE VI
EFFECT OF LEARNING ACTION FEATURES FROM THE MULTISCALE TRANSFORMER IN THE FINEDIVING DATASET. FSPN-S AND FSPN-M REPRESENT SMALL-SCALE AND MEDIUM-SCALE, RESPECTIVELY

Approaches	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}(\times 100)$ ↓
I3D+MLP	0.850	0.583
TSA [20]	0.920	0.342
FSPN-S	0.932	0.300
FSPN-M	0.938	0.285
FSPN (Ours)	0.942	0.278

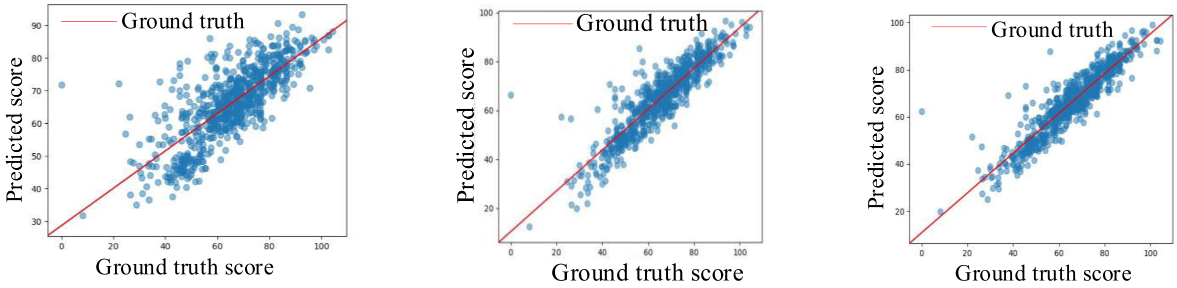
TABLE VII
EVALUATION ON COMPUTATION COMPLEXITY OF OUR PROPOSED FSPN AND RELATED APPROACHES IN THE MTL-AQA DATASET

Approaches	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}$ ($\times 100$) ↓	FLOPs	Param.(M)
MUSDL [5]	0.9273	0.451	1.50M	0.79M
CoRe [38]	0.9512	0.260	1.50M	2.21M
H-GCN [41]	0.9563	0.235	2.70M	2.10M
FSPN (Ours)	0.9601	0.221	1.51M	3.75M

10) Influence of Network Parameters on Multiscale Transformer Design: A comparison of the influence of network parameters according to the multiscale transformer design is shown in Table VIII. Our lightweight FSPN-S achieves 0.932 on Spr. Corr. and 0.300 on $R_{\mathcal{L}_2}$ distance, with only 1.49M FLOPs and 0.4M trainable parameters, outperforming the TSA [20]. Our deeper model, FSPN, achieves a performance improvement from 0.932 to 0.942 on Spr. Corr. and a decrease in the $R_{\mathcal{L}_2}$ distance from 0.300 to 0.278 with a slight increase in computation. Specifically, there is an increase in computational FLOPS by +0.1M and +3.15M trainable parameters compared to TSA [20]. These findings highlight the trade-off between performance improvement and computational cost. Overall, our results indicate that the performance continues to improve as more spatiotemporal channel resolutions are used in the proposed multiscale transformer design, with little increase in computational cost.

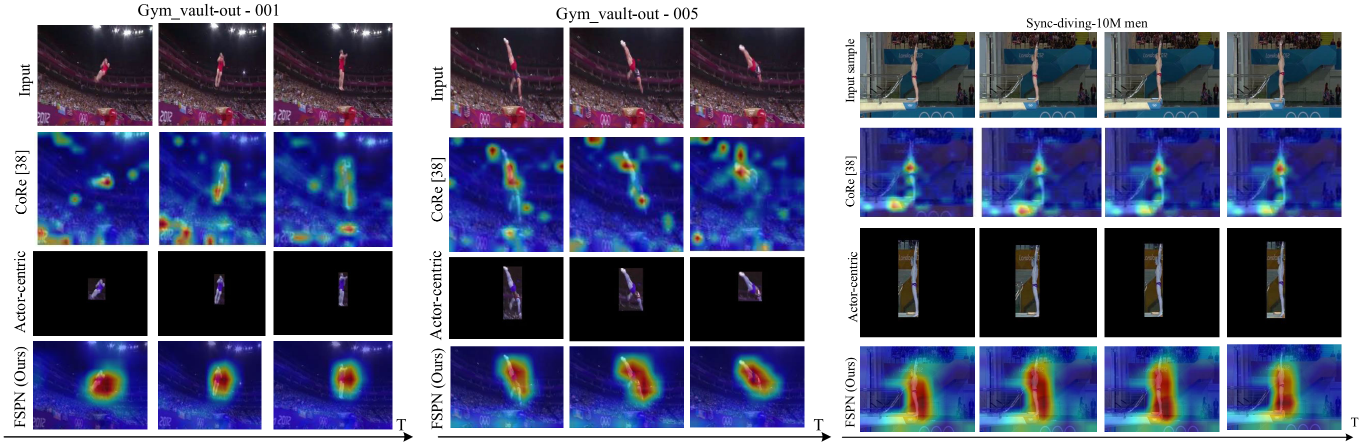
C. Qualitative Evaluation Results

1) Comparison of Scatter Plots of Score Prediction Distributions for Three Approaches: To have an intuitive comparison between our proposed FSPN and other AQA approaches, we visualize the predicted score distributions in the form of a scatter plot in Fig. 9. The y-coordinate and x-coordinate represent predicted and ground truth scores, respectively. The blue points represent the score predictions obtained by three models, i.e., I3D+MLP, TSA [20] and our



(a) I3D+MLP (b) TSA [20] (c) FSPN (Ours)

Fig. 9. Comparison among different approaches with scatter plots of prediction scores in the FineDiving dataset. The red line refers to the ground truth and the blue points represent prediction scores of samples in the test set.



(a) Gym-vault action class (b) Sync-diving-10m-out action class

Fig. 10. Visualization results of actor-centric regions and attention heat maps in the AQA-7 dataset. These visualization results imply that actor-centric detection contributes to learning motion-oriented action features and reducing background biases in the video.

TABLE VIII

INFLUENCE OF TRAINABLE PARAMETERS ON MULTISCALE TRANSFORMER DESIGN IN THE FIVEDIVING DATASET. FSPN-S AND FSPN-M DENOTE SMALL-SCALE AND MEDIUM-SCALE, RESPECTIVELY

Approaches	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}$ (×100) ↓	FLOPs (M)	Param.(M)
TSA [20]	0.920	0.342	1.50M	0.60M
FSPN-S	0.932	0.300	1.49M	0.48M
FSPN-M	0.938	0.285	1.50M	1.31M
FSPN (Ours)	0.942	0.278	1.51M	3.75M

TABLE IX

COMPARISON WITH THE STATE-OF-THE-ART APPROACHES IN THE FINEDIVING DATASET. HERE W/DN AND W/O DN REFER TO WITH AND WITHOUT DIVE NUMBERS, RESPECTIVELY

Approaches (w/o DN)	Year	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}(\times 100)$ ↓
USDL [5]	2020	0.8302	0.592
MUSDL [5]	2020	0.8427	0.573
CoRe [38]	2021	0.8631	0.556
TSA [20]	2022	0.8925	0.478
I3D+MLP	-	0.8504	0.583
FSPN (Ours)	2023	**0.9125**	**0.413**
Approaches (w/ DN)	Year	Spr. Corr. (ρ) ↑	$R_{\mathcal{L}_2}(\times 100)$ ↓
USDL [5]	2020	0.8504	0.583
MUSDL [5]	2020	0.8978	0.370
CoRe [38]	2021	0.9061	0.361
TSA [20]	2022	0.9203	0.342
I3D+MLP	-	0.8576	0.569
FSPN (Ours)	2023	**0.9421**	**0.278**

FSPN, and the red line refers to the ground-truth result. These figures show that FSPN obtains a very satisfactory performance for AQA since the predicted scores are well aligned with the ground truth line. The experiment results provide compelling evidence of the effectiveness of our motivation in enhancing the model's capability to generate accurate and discriminative action features from actor-centric regions.

2) Attention Heat-Map and Actor-Centric Regions: To further prove the effectiveness of our proposed method, we show the action-centric region detection and attention heat-map visualization using Grad-CAM [67]. The results of our proposed FSPN are illustrated in Fig. 10. The actor-centric detection contributes to focusing on motion-oriented features and ignoring noisy information in the background, leading to a more focused and accurate analysis of the actions in the videos. By incorporating actor-centric motion features, the proposed module can better infer and understand action features with small discrepancies happening in similar backgrounds. The attention results indicate our proposed

TABLE X

Comparison With the State-of-the-Art Approaches in the AQA-7 Dataset. These Results Indicate That Motion-Oriented Feature Representations and Sub-Action Parsing Enable the Model to Learn Discriminative Action Features for Scene-Invariant AQA

Metrics	Approaches	Year	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg. Corr.
Spr. Corr. (ρ) ↑	Pose+DCT [34]	2014	0.5300	0.1000	–	–	–	–	–
	C3D-LSTM [2]	2017	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
	C3D-SVR [2]	2017	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
	ST-GCN [11]	2018	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433
	JRG [2]	2019	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
	Asymmetric [32]	2020	0.7419	0.7296	0.5890	0.4960	0.9298	0.9043	0.7789
	USDL [5]	2020	0.8099	0.7570	0.6538	0.7109	0.9166	0.8878	0.8102
	CoRe [38]	2021	0.8824	0.7746	0.7115	0.6624	0.9442	0.9078	0.8401
	TSA-Net [9]	2021	0.8379	0.8004	0.6657	0.6962	0.9493	0.9334	0.8476
	Adaptive [31]	2022	0.8306	0.7593	0.7208	0.6940	0.9588	0.9298	0.8500
	TAP [39]	2022	**0.8969**	0.8043	0.7336	0.6965	0.9456	**0.9545**	**0.8715**
	PCLA [40]	2022	0.8697	**0.8759**	**0.7754**	0.5778	**0.9629**	0.9541	**0.879**
	UD-AQA [68]	2022	0.8532	0.7663	0.6836	0.5596	0.9281	0.9438	0.8318
	H-GCN [41]	2023	0.8867	0.7917	0.7326	0.6447	0.9213	0.9424	0.8501
	I3D+MLP	-	0.7438	0.7342	0.5190	0.5103	0.8915	0.8703	0.7472
	FSPN (Ours)	2023	0.8786	0.8343	0.6736	**0.7265**	0.9556	**0.9545**	**0.8724**
	Approaches	Year	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg. Corr.
$R_{\mathcal{L}_2}(\times 100)$ ↓	C3D-SVR [2]	2017	1.53	3.12	6.79	7.03	17.84	4.83	6.86
	USDL [5]	2020	0.79	2.09	4.82	4.94	0.65	2.14	2.57
	CoRe [38]	2021	0.64	1.78	3.67	3.87	0.41	2.35	2.12
	TAP [39]	2022	**0.53**	1.69	**2.89**	3.30	0.33	1.33	1.68
	H-GCN [41]	2023	0.59	1.85	3.59	3.61	0.82	1.40	1.98
	I3D+MLP	-	0.81	2.54	6.06	5.31	1.41	3.08	3.20
	FSPN (Ours)	2023	0.69	**0.82**	3.89	**2.28**	**0.25**	**1.31**	**1.56**

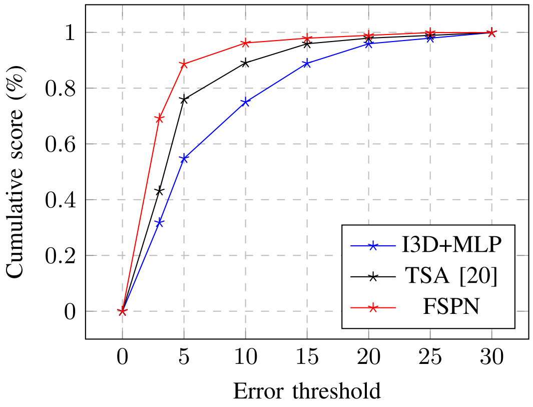


Fig. 11. Cumulative score curve in the FineDiving dataset. A larger area under the curve indicates a higher performance.

approach's effectiveness in learning semantically discriminative action features that can improve the performance of AQA. As shown, our approach reduces background noise and provides sufficient spatiotemporal action information, which is important for the AQA tasks.

3) Cumulative Score Curves (CS): The cumulative score curve prediction accuracy at the error ϵ is computed as $\mathrm{CS}(\alpha) = \frac{I_{\epsilon \leq \alpha}}{I} \times 100\%$. $I_{\epsilon \leq \alpha}$ is the number of videos on which the prediction error ϵ is not larger than the threshold α. In Fig. 11, we show the cumulative score curves of our proposed approach and state-of-the-art method TSA [20]. Given the error threshold ϵ, the samples whose absolute differences between their prediction and ground truth are less than ϵ will be regarded as positive samples. A larger area under the curve indicates a higher performance. Under any error threshold, FSPN (red line) shows a better performance in predicting accurate scores. These results demonstrate the effectiveness and feasibility of our proposed approach.

4) Failure Cases of the Proposed Approach: In the AQA-7 dataset, Fig. 12 presents the failure case analysis of the proposed approach FSPN in detecting actor-centric regions and attention heat-map. The figure reveals a noticeable mismatch between the actor-centric region and the attention heat map results, indicating the challenges faced by the proposed approach in learning consistent representations for the gym-vault-out action class. These challenges stem from very noisy and corrupted videos, as well as intra-action confusion and inter-action incoherence, which hinder the approach's ability to accurately represent complex actions. The presence of extremely noisy and corrupted videos within the dataset significantly contributes to the challenges encountered by the proposed approach. Moreover, the lack of clarity and visual coherence in these videos further hinders the accurate detection of actor-centric regions and impedes the learning process for discriminative action features. The intra-action confusion, where different instances of the same action exhibit variations in appearance and motion, and inter-action incoherence, where different actions might share similar visual cues, pose additional challenges for the approach.

D. Comparison With the State-of-the-Art Approaches

The comparison between our proposed FSPN with other AQA approaches in the FineDiving [20], AQA-7 [3], and

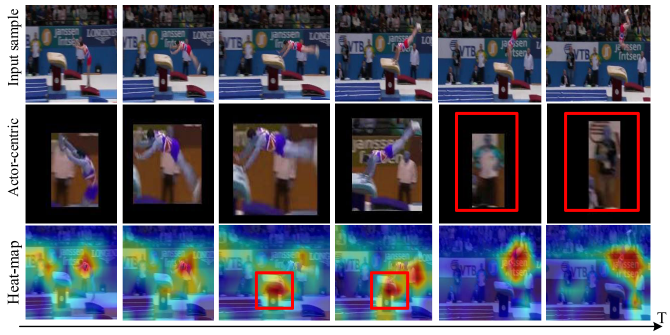


Fig. 12. The failure cases of our proposed approach for actor-centric region determination and attention heat-map in the AQA-7 dataset. The visualization results imply that the lack of visual and inter-action coherence hinders accurate actor-centric region detections, and learning discriminative action features.

MTL-AQA [4] datasets are shown in Table IX, Table X, and Table XI, respectively. As shown, our proposed FSPN achieves superior performance compared to state-of-the-art approaches. This is due to fine-grained and motion-oriented feature representation of action sequences, which improve the performance of AQA.

1) Comparison in the FineDiving Dataset: The comparison result between our proposed FSPN with existing state-of-the-art AQA approaches is shown in Table IX. We include the results of the I3D+MLP as the baseline, which clearly shows the performance improvement obtained by our proposed FSPN. Our approach achieves superior performance both with Dive Number (w/DN) and without Dive Number (w/o DN) among the compared approaches, namely, TSA [20], CoRe [38], and USDL [5]. Specifically, our proposed FSPN achieves 0.9125 Spr. Corr. and 0.413 in relative $R_{\mathcal{L}_2}$ distance w/o DN. Meanwhile, our proposed FSPN obtains 2.18% improvement on Spr. Corr., which is 0.9421, and 0.530 improvement on relative $R_{\mathcal{L}_2}$ distance, which is 0.2782 w/DN, compared with the TSA [20], clearly showing the effectiveness of the proposed approach. It is worth noting that TSA [20] utilizes step transition labels to extract procedure-aware representation. However, our FSPN identifies distinct patterns of sub-actions in the absence of ground truth.

2) Comparison in the AQA-7 Dataset: The comparison between the proposed approach FSPN with the state-of-the-art approaches is presented in Table X. Our proposed approach FSPN achieves the best average correlation (0.8724) compared with existing approaches, i.e., ST-GCN [11], C3D-LSTM [2], C3D-SVR [2], USDL [5], CoRe [38], adaptive approaches [9] and [31] approaches. Notably, FSPN achieves state-of-the-art results in three sports classes and demonstrates competitive performance in the remaining sports activities. Specifically, when compared to TSA-Net [9], FSPN exhibits an average improvement of +2.48% in Spr. Corr. Moreover, FSPN surpasses the tree-based approach of CoRe [38] by an average margin of +3.23% and outperforms the graph-based approach H-GCN [41] by +2.23%. These results highlight the superior performance of FSPN in capturing the underlying patterns and sub-action dependencies in sports activities. Additionally, FSPN demonstrates its effectiveness and feasibility by achieving a relative $R_{\mathcal{L}_2}$ distance improvement of +1.56 ($\times$100). This improvement shows our approach obtains sufficient semantic and temporal action features.

TABLE XI

COMPARISON WITH THE STATE-OF-THE-ART APPROACHES IN THE MTL-AQA DATASET. HERE W/DD AND W/O DD REFER TO WITH AND WITHOUT DEGREE OF DIFFICULTY LABELS, RESPECTIVELY

Approaches (w/o DD)	Year	Spr. Corr. (ρ)↑	$R_{\mathcal{L}_2}(\times 100)\downarrow$
Pose+DCT [34]	2014	0.2682	-
C3D-SVR [2]	2017	0.7716	-
C3D-LSTM [2]	2017	0.8489	-
MSCADC-MTL [44]	2019	0.8612	-
C3D-AVG-MTL [44]	2019	0.9044	-
USDL [5]	2020	0.9066	0.654
MUSDL [5]	2020	0.9158	0.609
CoRe [38]	2021	0.9341	0.365
TSA-Net [9]	2021	0.9422	-
UD-AQA [68]	2022	0.9545	0.259
TAP [39]	2022	0.9451	0.3222
PCLA [40]	2022	0.9230	-
H-GCN [41]	2023	0.9390	0.360
I3D+MLP	-	0.8921	0.707
FSPN (Ours)	2023	0.9382	0.370
Approaches (w/ DD)	Year	Spr. Corr. (ρ)↑	$R_{\mathcal{L}_2}(\times 100)\downarrow$
USDL [5]	2020	0.9231	0.468
MUSDL [5]	2020	0.9273	0.451
CoRe [38]	2021	0.9512	0.260
UD-AQA [68]	2022	0.9545	0.259
H-GCN [41]	2023	0.9563	0.235
I3D+MLP	-	0.9381	0.394
FSPN (Ours)	2023	**0.9601**	**0.221**

3) Comparison in the MTL-AQA Dataset: Table XI shows the comparison results of our proposed FSPN and existing AQA approaches in the MTL-AQA dataset [4]. Similar to [38] and HGCN [41], we also verify the contribution of Degree of Difficulty (DD) labels, i.e., with (w/DD) and without (w/o DD) in this dataset. As shown in Table XI, our proposed FSPN outperforms other approaches by large margins. We observe that the proposed approach achieves Spr. Corr. of 0.9382 and relative $R_{\mathcal{L}_2}$ distance of 0.370, which outperforms the tree-based approach CoRe [38] w/o DD labels. By training w/DD, the proposed FSPN becomes much better, which achieves the Spr. Corr. of 0.9601, and relative $R_{\mathcal{L}_2}$ distance of 0.221. Specifically, our proposed approach gained a 0.38% improvement in Spr. Corr. and a 1.4% improvement in $R_{\mathcal{L}_2}$ distance compared to the graph-based approach HGCN [41]. These results imply that motion-oriented feature representations and intra-sequence action parsing make it easier for the model to learn more discriminative regions and obtain sufficient semantic and temporal action features.

V. CONCLUSION

In this paper, we exhibited the effectiveness of the proposed Fine-grained spatiotemporal parsing Network (FSPN) for AQA. The proposed FSPN learned fine-grained sub-action patterns, correct description of subtle visual differences between sub-actions, and motion-oriented feature representations and obtained their long-range dependencies among sub-actions at varying scales. To verify the contribution of our proposed FSPN, extensive experiments, and ablation studies were conducted in three challenging AQA datasets. The proposed approach achieved substantial improvements compared with existing AQA approaches. The results of actor-centric region detection, scatter plots, cumulative score curves, and attention maps were also visualized to show the visual interpretability and feasibility of our proposed approach. In the future, there is a potential to enhance spatiotemporal parsing by maximizing the mutual information between pseudo-labels. These pseudo-labels are generated from feature representations in an unsupervised manner. To achieve this, it would be beneficial to incorporate a group of contrastive learning techniques that involve diverse samples. This approach will facilitate the learning of discriminative features.

REFERENCES

[1] Y. Zhang, W. Xiong, and S. Mi, “Learning time-aware features for action quality assessment,” *Pattern Recognit. Lett.*, vol. 158, pp. 104–110, Jun. 2022.

[2] J.-H. Pan, J. Gao, and W.-S. Zheng, “Action assessment by joint relation graphs,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6330–6339.

[3] P. Parmar and B. T. Morris, “What and how well you performed? A multitask learning approach to action quality assessment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 304–313.

[4] P. Parmar and B. T. Morris, “Learning to score Olympic events,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 76–84.

[5] Y. Tang et al., “Uncertainty-aware score distribution learning for action quality assessment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9839–9848.

[6] D. Shao, Y. Zhao, B. Dai, and D. Lin, “Intra- and inter-action understanding via temporal action parsing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 727–736.

[7] Z. Li, Y. Huang, M. Cai, and Y. Sato, “Manipulation-skill assessment from videos with spatial attention network,” in *Proc. IEEE Int. Conf. Comput. Visi. Workshops*, Oct. 2019, pp. 4385–4395.

[8] H. Doughty, D. Damen, and W. Mayol-Cuevas, “Who’s better? Who’s best? Pairwise deep ranking for skill determination,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6057–6066.

[9] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, “TSA-Net: Tube self-attention network for action quality assessment,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4902–4910.

[10] H. Doughty, W. Mayol-Cuevas, and D. Damen, “The pros and cons: Rank-aware temporal attention for skill determination in long videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7854–7863.

[11] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[12] M. Assefa et al., “Audio-visual contrastive and consistency learning for semi-supervised action recognition,” *IEEE Trans. Multimedia*, early access, Sep. 7, 2023, doi: 10.1109/TMM.2023.3312856.

[13] Y. Ji, Y. Zhan, Y. Yang, X. Xu, F. Shen, and H. T. Shen, “A context knowledge map guided coarse-to-fine action recognition,” *IEEE Trans. Image Process.*, vol. 29, pp. 2742–2752, 2020.

[14] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao, “Deep active learning with contaminated tags for image aesthetics assessment,” *IEEE Trans. Image Process.*, early access, Apr. 15, 2018, doi: 10.1109/TIP.2018.2828326.

[15] M. Assefa, W. Jiang, K. Gedamu, G. Yilma, B. Kumeda, and M. Ayalew, “Self-supervised scene-debiasing for video representation learning via background patching,” *IEEE Trans. Multimedia*, vol. 25, pp. 5500–5515, 2022.

[16] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, “Learning to score figure skating sport videos,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4578–4590, Dec. 2020.

[17] C. Zhang, A. Gupta, and A. Zisserman, “Temporal query networks for fine-grained video understanding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4484–4494.

[18] Y. Li, X. Chai, and X. Chen, “End-to-end learning for action quality assessment,” in *Advances in Multimedia Information Processing—PCM 2018*. Cham, Switzerland: Springer, 2018, pp. 125–134.

[19] W. Liu et al., “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[20] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, “FineDiving: A fine-grained dataset for procedure-aware action quality assessment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2949–2958.

[21] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[22] S. Gattupalli, D. Ebert, M. Papakostas, F. Makedon, and V. Athitsos, “CogniLearn: A deep learning-based interface for cognitive behavior assessment,” in *Proc. 22nd Int. Conf. Intell. User Interfaces*, Mar. 2017, pp. 577–587.

[23] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa, “Automated assessment of surgical skills using frequency analysis,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2015, pp. 430–438.

[24] L. Gao, Y. Ji, K. Gedamu, X. Zhu, X. Xu, and H. T. Shen, “View-invariant human action recognition via view transformation network (VTN),” *IEEE Trans. Multimedia*, vol. 24, pp. 4493–4503, 2022.

[25] K. Gedamu, Y. Ji, Y. Yang, L. Gao, and H. T. Shen, “Arbitrary-view human action recognition via novel-view action generation,” *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 108043.

[26] G. A. Kumie, G. Y. Abawatew, M. A. Habtie, and M. A. Mesfin, “Spatio-temporal dual-attention network for view-invariant human action recognition,” in *Proc. 14th Int. Conf. Digit. Image Process. (ICDIP)*, Oct. 2022, pp. 213–222.

[27] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[28] X. Shu, B. Xu, L. Zhang, and J. Tang, “Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7559–7576, Jun. 2023.

[29] B. Xu and X. Shu, “Pyramid self-attention polymerization learning for semi-supervised skeleton-based action recognition,” 2023, *arXiv:2302.02327*.

[30] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, “Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 10, 2023, doi: 10.1109/TNNLS.2023.3247103.

[31] J.-H. Pan, J. Gao, and W.-S. Zheng, “Adaptive action assessment,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8779–8795, Dec. 2022.

[32] J. Gao et al., “An asymmetric modeling for action assessment,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 222–238.

[33] M. Nekoui, F. O. T. Cruz, and L. Cheng, “FALCONS: Fast learner-grader for contorted poses in sports,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3941–3949.

[34] H. Pirsiavash, C. Vondrick, and A. Torralba, “Assessing the quality of actions,” in *Proc. Eur. Conf. Comput. Vis.*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham, Switzerland, 2014, pp. 556–571.

[35] L.-A. Zeng et al., “Hybrid dynamic-static context-aware attention network for action assessment in long videos,” in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2526–2534.

[36] S. Zhang et al., “LOGO: A long-form video dataset for group action quality assessment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2405–2414.

[37] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, “Am I a baller? Basketball performance assessment from first-person videos,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2196–2204.

[38] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, “Group-aware contrastive regression for action quality assessment,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7899–7908.

[39] Y. Bai et al., “Action quality assessment with temporal parsing transformer,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 422–438.

[40] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, “Pairwise contrastive learning network for action quality assessment,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 457–473.

[41] K. Zhou, Y. Ma, H. P. H. Shum, and X. Liang, "Hierarchical graph convolutional networks for action quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 30, 2023, doi: 10.1109/TCSVT.2023.3281413.

[42] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.

[43] P. Lei and S. Todorovic, "Temporal deformable residual networks for action segmentation in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6742–6751.

[44] A. Montes, A. Salvador, S. Pascual, and X. G.-I. Nieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst. Workshop*, 2016, pp. 1–5.

[45] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2613–2622.

[46] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1784–1791.

[47] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2394–2402.

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[49] X. Gao, W. Lu, D. Tao, and X. Li, "Image quality assessment based on multiscale geometric analysis," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1409–1423, Jul. 2009.

[50] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[51] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.

[52] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[53] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10073–10082.

[54] Y. Li et al., "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4794–4804.

[55] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3285–3294.

[56] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3464–3473.

[57] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[58] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6804–6815.

[59] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[60] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[61] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1–11.

[62] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[63] A. Singh et al., "Semi-supervised action recognition with temporal contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10389–10399.

[64] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. Adv. Neural Inf. Process. Syst. Workshop*, 2017, pp. 1–4.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[66] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1468–1476.

[67] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[68] C. Zhou and Y. Huang, "Uncertainty-driven action quality assessment," 2022, *arXiv:2207.14513*.