Self-supervised subaction Parsing Network for Semi-supervised Action Quality Assessment

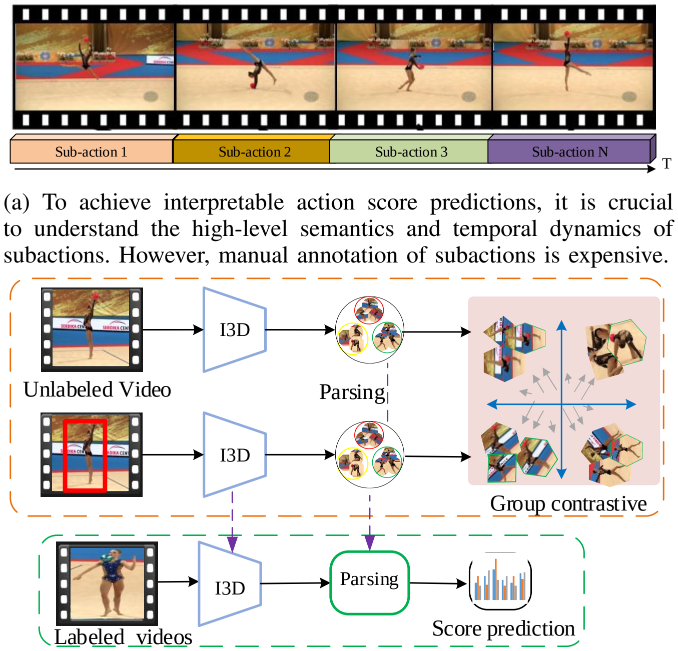
Kumie Gedamu, Yanli Ji*, Yang Yang, Jie Shao, Heng Tao Shen

***Abstract*—Semi-supervised Action Quality Assessment (AQA) using limited labeled and massive unlabeled samples to achieve high-quality assessment is an attractive but challenging task. The main challenge relies on how to exploit solid and consistent representations of action sequences for building a bridge between labeled and unlabeled samples in the semi-supervised AQA. To address the issue, we propose a Self-supervised subAction Parsing Network (SAP-Net) that employs a teacher-student network structure to learn consistent semantic representations between labeled and unlabeled samples for semi-supervised AQA. We perform actor-centric region detection, generating high-quality pseudo-labels in the teacher branch, which assists the student branch in learning discriminative action features. We further design a self-supervised subaction parsing solution to locate and parse fine-grained subaction sequences. Then, we present the group contrastive learning with pseudo-labels to capture consistent motion-oriented action features in the two branches. We evaluate our proposed SAP-Net on four public datasets: the MTL-AQA, FineDiving, Rhythmic Gymnastics, and FineFS datasets. The experiment results show that our approach outperforms state-of-the-art semi-supervised methods by a significant margin.**

***Index Terms*—Action analysis, Action quality assessment, Semi-supervised learning.**

I. Introduction

Action Quality Assessment (AQA) has garnered significant attention in various real-world applications. It is used to evaluate the quality of specific professional actions, such as sports activities [1]–[5], medical rehabilitation, and training skill assessments [6]–[11]. As an example, action assessment systems can help healthcare professionals monitor the daily activities of patients to evaluate rehabilitation progress, providing information to optimize treatment. Similarly, coaches can use AQA to improve athletes' training performance. However, AQA is more challenging because all the videos have the same action routine (*e.g.,* "take-off", "flight, and "entry") in similar backgrounds to aquatic centers [12]–[16]. Recently, fully supervised AQA has made remarkable progress, incorporating various techniques such as clip-level scoring [2], joint motion learning [17], uncertainty-aware [18], asymmetric modeling [19], group-aware attention [20], and pairwise contrastive learning [12]–[16]. The aforementioned approaches face limitations due to contextual biases and holistic representation, which restrict their ability to capture subtle variations and lack exploration of subaction temporal structures [21]–[24].



(a) To achieve interpretable action score predictions, it is crucial to understand the high-level semantics and temporal dynamics of subactions. However, manual annotation of subactions is expensive.

(b) Solution in our proposed approach.

Fig. 1: Motivation of our proposed approach. We set up a teacher-student network structure to learn consistent representations under a semi-supervised setting. We further design a self-supervised subaction parsing solution to guide valid consistent learning in the two branches.

The manual collection and annotation process of AQA samples in real-world applications is often impractical due to the requirement of domain-specific knowledge and expertise [25]–[27]. However, the approaches mentioned above have been highly based on human annotations, which can be costly to obtain [25]. Therefore, it is crucial to explore semi-supervised AQA with only a limited amount of labeled and a massive amount of unlabeled samples, which can reduce the model dependency on manual annotations [21]–[24], [28], [29]. Thus, we design a teacher-student network structure to learn consistent semantic representations of labeled and unlabeled samples for semi-supervised AQA. Initially, we adopt a pre-trained object detection model [30] to extract

Kumie Gedamu is with 1) Sichuan Artificial Intelligence Research Institute, Yibin, China. 2) School of Computer Science and Engineering, University of Electronic Science and Technology of China.

Yanli Ji is with School of Intelligent Systems Engineering, Sun Yat-sen University (Shenzhen Campus).

Jie Shao are with 1) School of Computer Science and Engineering, University of Electronic Science and Technology of China; 2) Shenzhen Institute for Advanced Study, UESTC.

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

*Corresponding author, jiyanli82@gmail.com

Manuscript received xx xx, 2023; revised xx xx, 2024.

motion-oriented action features, which can reduce the model's reliance on the video background for scene-invariant AQA. Thus, unlike previous works [31]–[33], our teacher branch receives the actor-centric region and generates high-quality pseudolabels with high confidence. These pseudo-labels, in turn, assist the student branch in learning and inferring motion-oriented action features through consistency regularization.

Fine-grained solutions of scene-invariant AQA usually involve understanding the semantic and internal temporal dependencies of subactions to achieve a comprehensive understanding of action execution. To clarify our hypothesis, we can examine the action sequence presented in Fig. 1(a). Thus, by parsing the given action sequence into separate subaction sequences, we can effectively capture the high-level semantics and internal temporal dependencies among subactions. However, parsing subaction sequences along with their temporal dependence remains a challenging task due to: 1) the lack of predefined subaction label classes. 2) Subactions are more finely granular, and their transitions between consecutive segments are often smoother [13], which makes it difficult to distinguish their boundaries. To address these challenges, we propose a self-supervised subaction parsing module to understand the high-level semantics and internal temporal dependencies of subactions. We identify and parse subaction sequences based on cluster-based subset selection in each branch. This involves selecting a subset of latent states as subaction states and determining the assignment of video frames to these states. Thus, the output of the subset selection provides supervisory information in the form of pseudo-labels during the subaction parsing. In this way, we are able to effectively capture high-level semantics and the internal temporal structure of subaction sequences.

Through self-supervised subaction parsing, we obtain subaction sequences along with their temporal dependencies. The parsed subaction feature representations in the two branches share the same intra-sequence action semantics. Thus, the semantic similarity between the two representations is maximized and the semantically different subactions is minimized [33]–[35]. However, applying direct contrastive loss between teacher and student action sequences would push apart similar subaction sequences and learn different representations for intra-sequence semantics [35]. To address these challenges, a group contrastive learning method is proposed to explore relationships within the video neighborhood by grouping semantically similar subactions in the two input branches, as shown in Fig. 1(b). The subaction groups are formed by clustering actions with the same pseudo-labels, and they are represented by averaging the representations of the constituent subactions. This allows the approach to maintain the semantic similarities between subactions, rather than learning divergent representations for similar action sequences. In doing so, the module enhances the temporal diversity of the learned features and ensures a consistent video representation.

In this paper, we propose a Self-supervised subAction Parsing Network (SAP-Net) that aims to learn consistent semantic representations and internal temporal structures by leveraging both labeled and unlabeled samples. We utilize a teacher-student network structure, where the teacher branch generates high-quality pseudo-labels and assists the student branch in learning discriminative features. In each branch, we design a self-supervised subaction parsing module to parse fine-grained temporal subactions. Furthermore, a group contrastive learning module is designed to capture consistent motion-oriented features and learn temporal diversity in the two branches. By jointly training self-supervised and semi-supervised learning with a limited number of labeled samples, SAP-Net achieves consistent video representation. In summary, the major contributions of our SAP-Net are listed below:

- We design a Self-supervised subAction Parsing Network (SAP-Net) that employs a teacher-student network structure to learn consistent semantic representations of action sequences by establishing a bridge between labeled and unlabeled samples for the semi-supervised AQA.
- We propose a self-supervised subaction parsing module that identifies and parses subactions in each branch, enabling a deeper understanding of the high-level semantics and internal temporal structure of subactions.
- We propose group contrastive learning to capture consistent motion-oriented action features along with their temporal dependencies in the two branches.
- We conduct extensive experiments to analyze the effectiveness of our approach over the state-of-the-art methods on the MTL-AQA [1], FineDiving [13], Rhythmic Gymnastics (RG) [36], and FineFS [37] datasets.

The remainder of the paper is organized as follows. We review related work in Section II, highlighting the gaps that our SAP-Net aims to address. We then present the proposed approach with its innovative features in Section III. In Section IV, we conduct a comprehensive evaluation of SAP-Net's performance through a series of experiments on benchmark datasets. Finally, Section V gives a conclusion of the work.

II. Related Work

A. Action Quality Assessment

Initially, AQA was approached as a classification task, where actions were categorized into different levels of performance [38] and [39]. As AQA research progressed, two main formulations emerged:

Regression Formulation: Existing approaches have employed regression-based techniques to tackle a range of challenges. Specifically, Pirsiavash *et al.* [1] used spatiotemporal action features and proposed clip-level scoring to estimate AQA scores for sports activities. Pose+DCT [40] utilized joint localization to extract the position of each joint and employed SVR to compute the discrete cosine transformation along the temporal dimension. Pirsiavash *et al.* [2] framed AQA as a regression problem, focusing on learning individual joint motion for gymnastics and surgical procedures. Zeng *et al.* [36] introduced a hybrid approach that combines static and dynamic action features, considering the contributions of different stages to the AQA score. Xu *et al.* [41] introduced a self-attentive and multiscale skip convolutional LSTM approach for aggregating information from individual clips. Tang *et al.* [18] construct the model using KL divergence to formulate the score regression as a distribution learning problem.

Xu *et al.* [13] introduced a procedure-aware representation by constructing a pairwise temporal segmentation attention module. Similarly, Gedamu *et al.* [14] introduced a fine-grained representation with a multiscale transformer to address the scene-invariant AQA problem. Zhou *et al.* [8] devised a hierarchical GCN approach that refines semantic features, reduces information confusion, and aggregates dependencies for analyzing action procedures and motion units. Zhang *et al.* [42] proposed the Distribution Auto Encoder module addresses aleatoric uncertainty by encoding videos into distributions using the VAE reparameterization trick. Zhang *et al.* [20] proposed a group-aware attention approach that incorporates contextual group information and temporal relations using graph CNN. Work in [43] utilized semantic attributes for query learning to enhance the assessment of gymnastic routines from video. Overall, the previously summarized methods relied on human annotations, which can be costly to obtain. Action segmentation from videos could provide a comprehensive representation of human actions [44]. Similar to our approach, Zhang *et al.* [25] proposed a semi-supervised AQA method that uses self-supervised learning on unlabeled videos to recover feature representations of masked segments. However, their approach has limitations due to contextual biases and holistic representations, which restrict its ability to capture subtle variations. In contrast, our approach employs a teacher-student network structure to learn consistent semantic representations and improve AQA performance on limited labeled videos. The teacher branch receives the actor-centric region, generating high-quality pseudo-labels, and assists the student branch in learning discriminative action features.

Pairwise Ranking Formulation. The performance scores are unavailable in certain daily scenarios, leading to the reformulation of the AQA problem as a pairwise ranking problem. For instance, Doughty *et al.* [9] employed a rank-aware loss function to focus on skill-relevant segments of a video for estimating motions and assessing performance in basketball. Furthermore, Doughty *et al.* [45] introduced a novel loss function that enables the learning of discriminative features in videos with varying skill levels. A siamese learning strategy is employed in [41], which primarily concentrates on longer sequences and provides predictions solely for overall ranks, thereby restricting the applicability of AQA to scenarios that entail quantitative comparisons. Xu *et al.* [46] introduced a Likert scoring paradigm inspired by psychometrics, enabling the quantification of grades and the generation of quality scores. Following this, Fang *et al.* [47] proposed an action parsing transformer to disintegrate the holistic feature into a more fine-grained step-wise representation. Yu *et al.* [15] proposed the Contrastive Regression (CoRe) framework, which employs pair-wise comparison to learn relative scores. Li *et al.* [16] put forth pairwise contrastive learning as a method to learn relative scores between pairs of videos. Similarly, Qi *et al.* [48] proposed a multi-stage contrastive regression framework for AQA that efficiently extracts spatial-temporal action features. Recently, Bai *et al.* [12] presented a temporal parsing transformer that decomposes global features into a fine-grained temporal hierarchical representation. In contrast to these approaches, our method exploits solid and consistent representations of action sequences. This allows us to build a bridge between the labeled and unlabeled samples, enabling a semi-supervised approach to AQA.

In addition, the following works have explored action parsing [49]–[51]. Zhang *et al.* [50] introduced Temporal Query Networks, which ensure that relevant segments are targeted by the query network. Dian *et al.* [49] proposed the use of TransParser to extract subactions without relying on training data labels. In contrast, the proposed method identifies distinct subaction patterns and captures middle-level representations without explicit supervision, enabling semi-supervised AQA.

B. Semi-supervised Learning

The semi-supervised learning has gained popularity in recent years due to its ability to leverage massive amounts of unlabeled data under a limited label regime. The two important pipelines were consistency regularization [29] and pseudo-labeling [22]–[24]. Pseudo-labeling refers to the conversion of model predictions into one-hot labels, which is often based on the confidence threshold that retains unlabeled samples for which the classifier is confident [31]. In contrast, consistency regularization measures the discrepancy between model predictions of two perturbed unlabeled samples [29], [31]. Recent self-supervised learning methods have used contrastive loss as an auxiliary loss for action recognition [52], with techniques such as [35] and [53]. Wang *et al.* [54] incorporated self-supervised learning into a semi-supervised temporal action proposal by designing two temporal perturbations with pretext tasks. VideoSSL [55] proposed semi-supervised action classification by training the encoder with ImageNet pre-trained models. Some frameworks used a single frame labeled with a bounding box within the temporal boundary of the fully supervised counterpart as the supervisory signal for Video Grounding [56], and masked pseudo-labeling autoencoder was designed for semi-supervised point cloud action recognition [57]. Ding *et al.* [21] proposed an action affinity loss to integrate action priors for semi-supervised learning by exploring the correlation of actions between labeled and unlabeled procedural videos. Unlike previous works, our teacher branch generates high-quality pseudo-labels with high confidence that assist the student branch in learning motion-oriented action features through consistency regularization.

C. Contrastive learning

Contrastive representation learning has shown its power in a broad range of computer vision [22], [33], [34], [58]. The main goal is to capture the underlying structure and properties of the data that are invariant to specific transformations or augmentations. [33], [34], [58]. As an example, TCL [35] uses both individual and group contrastive learning as a self-supervised auxiliary task. Zhai *et al.* [26] proposes image rotation prediction and transformation pretext tasks for learning from unlabeled images. The work in [59] proposes skeleton inpainting and neighborhood consistency modeling to learn discriminative representations from unlabeled data [59]. Work in [60] argues that a single representation to

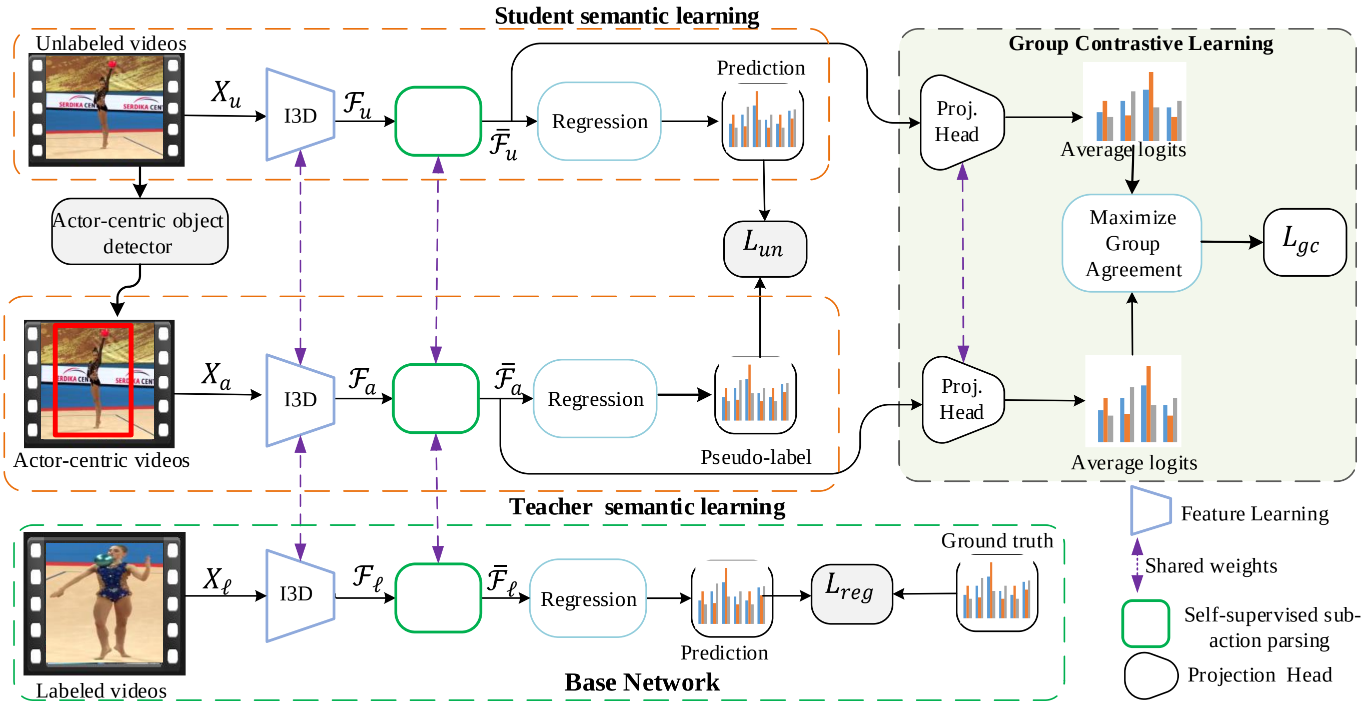


Fig. 2: Overview of our proposed SAP-Net. Initially, an object detector [30] extracts the actor-centric region, then we use a siamese I3D backbone [61] to extract spatiotemporal action features. The teacher branch receives the actor-centric region, generating high-quality pseudo-labels, and assists the student branch in learning and inferring motion-oriented action features through consistency regularization. We design a self-supervised subaction parsing module in each branch to parse subaction sequences. Then, the group contrastive learning with pseudo-labels captures consistent motion-oriented features and learns temporal diversity in the two branches. The subaction groups are formed by clustering actions with the same pseudo-labels.

capture both types of features is sub-optimal. In response, [60] proposed decomposing the representation space into stationary and non-stationary features using contrastive learning from long and short views. These mentioned approaches are not suitable for AQA tasks. Unlike the previous approach, we group similar subactions and employ negative samples from different subactions within the same action class. This approach effectively captures discriminative high-level internal temporal boundaries of subaction sequences, ensuring that similar subaction instances are appropriately grouped.

III. Proposed Approach

We aim to exploit a solid and consistent semantic representation by building a bridge between labeled and unlabeled samples. Fig. 2 shows the overall structure of our proposed approach. In SAP-Net, we design a teacher-student network structure to learn consistent representations for labeled and unlabeled samples for semisupervised learning. In each branch, we design a self-supervised subaction parsing module to parse fine-grained subactions. Once we obtain subaction sequences along with their temporal dependencies, we present a group contrastive learning module to maintain consistent video representation by grouping semantically similar subactions.

A. Preliminary definition

In semi-supervised AQA, the training set consists of a limited labeled set, $X_\ell = \{x_i^\ell, y_i\}_{i=1}^{\mathbb{N}_\ell}$, $x_i^\ell \in \mathbb{R}^{T\times H\times W\times C}$ (where T, H, W, and C refer to clip length, height, width and number of channels, respectively), and a massive unlabeled sample set, $X_u = \{x_i^u\}_{i=1}^{\mathbb{N}_u}$, $x_i^u \in \mathbb{R}^{T\times H\times W\times C}$. Here, $\mathbb{N}_u \gg \mathbb{N}_\ell$, and y_i refers to the semantic score label corresponding to x_i^ℓ. Then, we design a teacher-student network structure to learn consistent representations for labeled and unlabeled samples in semi-supervised setting, where the teacher branch pre-trains an object detector [30] to locate human action regions on each frame. These regions are extracted and compose a new set $X_a = \{x_i^a\}_{i=1}^{\mathbb{N}_u}$. The X_a and X_u are used as pairwise inputs for the teacher-student network to train subaction parsing and group contrastive learning modules. We use the I3D backbone E, parameterized by θ to extract spatio-temporal visual features, as illustrated in Eqn. 1.

$$\mathcal{F}_u = E_\theta(X_u), \quad \mathcal{F}_a = E_\theta(X_a), \quad \mathcal{F}_\ell = E_\theta(X_\ell) \quad (1)$$

Through a weight-sharing I3D backbone network [61], we extract spatiotemporal action features represented as $\mathcal{F}_u$ and $\mathcal{F}_a$ from pairwise inputs X_u and X_a, respectively.

B. Teacher-student network for semi-supervised AQA

In the context of AQA, obtaining a large amount of labeled data can be challenging and time consuming due to the requirement of domain-specific knowledge. To address this issue, we propose a semi-supervised learning approach that leverages both labeled and unlabeled data to learn consistent semantic representations between samples. The core of our semi-supervised AQA method is the teacher-student network architecture. The teacher branch processes the most discriminative regions in each video, such as actors. By focusing on these actor-centric regions, the teacher branch captures and extracts motion-oriented action features that are crucial for scene-invariant AQA problem. Using these features, the teacher branch generates high-quality pseudo-labels for the unlabeled data. The student branch learns to infer motion-oriented features by minimizing the discrepancy between its predictions and the pseudo-labels generated by the teacher branch. This process, known as consistency regularization, encourages the student branch to make predictions consistent with the teacher's pseudo-labels for the unlabeled data. Thus, the semi-supervised regression loss function for this teacher-student network can be conceptualized as follows:

$$L_{un} = -\frac{1}{\mathbb{N}_u}\sum_{i=1}^{\mathbb{N}_u} ||R_\vartheta\left(\bar{\mathcal{F}}_u\right) - \left(\mathbb{1}\max\left(R_\vartheta\left(\bar{\mathcal{F}}_a\right)\right) \geq \tau\right)||^2 \quad (2)$$

where τ is the predefined threshold. $\mathbb{1}$ is the indicator function when the maximum class probability exceeds τ otherwise 0. This semi-supervised regression loss function is designed to leverage the teacher's high-quality pseudo-labels to guide the student branch's learning process. By minimizing the discrepancy between the student's predictions and the teacher's pseudo-labels, the student branch is encouraged to learn consistent and motion-oriented features, effectively utilizing both labeled and unlabeled data for semi-supervised AQA.

C. *Self-supervised subaction parsing*

We have devised a self-supervised data-driven approach to determine the granularity of subactions that can be parsed from a given action sequence with semantic and temporal correspondences. The goal is to identify a distinct pattern of subactions with their temporal dependencies and learn the spatio-temporal structure of actions to achieve a better representation. Assuming that there are K subactions, we aim to identify all video frames belonging to each of the K subactions. This allows us to extract valuable insights and discover spatial and temporal correspondences. The module is composed of two components that work together to enhance the temporal diversity and improve the representation of semi-supervised action performance assessment.

1) Pseudo-label generation: To identify a concise and representative set of subsets from a provided dataset, we employ subset selection to create pseudo-labels for subaction parsing. To simplify the explanation, we use $\mathcal{F}_w^t$, where $w \in \{u, a\}$, to represent the input features at time t. We employ attention refinement, inspired by [24] and [62] to determine the region of the video frame that provides the most representative features. The process of attention refinement helps us identify crucial visual cues within the video frames, as depicted in Fig. 3. This attention refinement guides the selection of the region with the most representative features. By identifying the salient regions and enabling the back-propagation of gradients through the differentiable module [24], we allow the network to learn effectively from the refined features and improve its performance in subsequent tasks.

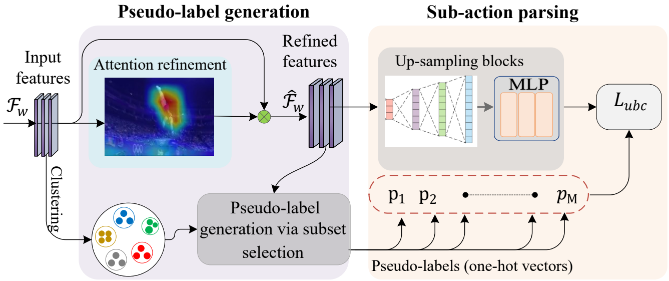


Fig. 3: Illustration of self-supervised subaction parsing module. The proposed module sets binary cross-entropy loss, L_{ubc}. The symbol $\otimes$ denotes channel-wise multiplication.

We first apply Global Average Pooling (GAP) to each channel and reduce the input feature maps to the $C \times 1 \times 1$ vector. This pooling operation helps capture the overall information within each channel. Next, we perform convolution with sigmoid activation, which computes channel-wise attention. Applying the sigmoid function, we obtain attention scores for each channel, highlighting the important regions. Using these attention scores, we then perform channel-wise multiplication, resulting in a refined feature map $\hat{\mathcal{F}}_w^t$. This refined feature map focuses on the most relevant regions, enhancing the discriminative features for subaction parsing. The process of generating pseudo-labels utilizing attention features and clusters of latent states. First, we carefully select a subset of the latent states, treating them as subactions. Then, it assigns video frames to these selected states. This subset selection process leads to the generation of pseudo-labels, which are subsequently used as training labels for subaction parsing.

To localize subactions, the generation of latent states is accomplished by running the k-means algorithm with M centers on the input feature $\mathcal{F}_w^t$. The process of selecting subsets based on clustering involves using the states M within the set $\mathcal{S}$ together with the refined attention feature $\hat{F}_w^t$ in the subset selection component. This results in a set of latent states $\mathcal{S} = \{s_1, \cdots, s_M\}$, which represent distinct patterns or clusters in the input features. The clustering-based subset selection is then used to accomplish this task, as defined in Eqn. 3.

$$Z(\hat{\mathcal{S}}) \triangleq \frac{1}{T}\sum_{t=1}^{T} \min_{i=\{1,\cdots,M\}} \|\hat{\mathcal{F}}_w^t - s_i\|_2 \quad (3)$$

The outputs consist of the $\hat{K}$ selected states that correspond to subaction sequences. To optimize Eqn. 3, we use a greedy algorithm to minimize the objective function $Z(\hat{\mathcal{S}})$, as shown in Algorithm 1. We start with an empty set Γ and iteratively add states from $\mathcal{S}$ to Γ, selecting the states that minimize the cost function compared to the elements already in Γ. We continue this process up to a maximum of $\hat{K}$ states. These $\hat{K}$ states in Γ are then used as pseudo-labels, denoted as $P = \{p_1, p_2, \cdots, p_M\}$. Thus, Eqn. 3 is responsible for generating these pseudo-labels, assigning each frame to one of the $\hat{K}$

Algorithm 1 Pseudo-label generation via subset selection

```
Input: (1) The set of all possible states: ${1, ..., \hat{K}}$ and the input feature $\hat{\mathcal{F}}_w^t$, (2) Cost function: $Z(\Gamma)$ representing the loss, (3) gain function: $\delta\Gamma(z^*) = Z(\Gamma) - Z(\Gamma \cup \{z^*\})$, decrease the loss function when including state $z^*$ in $\Gamma$.
Output: Select at most $\hat{K}$ states as subaction sequences.
Initialize the active set: $\Gamma \leftarrow \emptyset$
for $i = 1$ to $\hat{K}$ do
    Find the state $z^* \in \{1, \ldots, M\} \setminus \Gamma$ that minimizes the cost function the most:
    $z^* \leftarrow \arg\min_{z \in \{1,\ldots,M\} \setminus \Gamma} Z(\Gamma \cup \{z\})$
    Calculate the gain $\delta\Gamma(z^*)$ as the difference in the cost function.
    $\delta\Gamma(z^*) \leftarrow Z(\Gamma) - Z(\Gamma \cup \{z^*\})$
    Include $z^*$ in the active set $\Gamma$:
    Update $\Gamma \leftarrow \Gamma \cup \{z^*\}$
end for
return The resulting active set $\Gamma$, containing at most $\hat{K}$ states, is used for pseudo-labels in subaction parsing.
```

states in P. Finally, these pseudo-labels P serve as supervisory information during the subaction parsing process.

2) subaction parsing: During the subaction parsing stage depicted in Fig. 3, each frame generates an M-dimensional vector representing the probabilities of belonging to the M states, using the input feature $\hat{\mathcal{F}}_w^t$. Similar to [13], the subaction parsing process consists of an up-sampling decoder block and MLP projection layers. The up-sampling decoder block is composed of four sub-blocks with varying spatial-temporal dimensions: $(1024, 12)$, $(512, 24)$, $(256, 48)$, and $(128, 96)$. Each sub-block employs convolution layers to expand the attention-enhanced feature, $\hat{\mathcal{F}}_w^t$, along the temporal axis. Furthermore, max-pooling is applied to reduce the spatial dimensions. After the up-sampling decoder block, the obtained features are projected into a probability vector using MLP projection layers. This projection is performed by applying Eqn. 4 and Eqn. 5. The resulting probability vector denoted as $\mathcal{A} = \{a_1, \cdots, a_M\}$, represents the likelihood of subactions occurring within the video sequence.

$$[a_1, \cdots, a_M] = \mathbb{F}_\emptyset(\hat{\mathcal{F}}_w^t) \tag{4}$$

$$\bar{t}_k = \underset{\frac{T}{M}(k-1) \le t \le \frac{T}{M}k}{\text{argmax}} \; a_M(t) \tag{5}$$

$$L_{ubc} = -\sum_t p_M(t) \log_M(t) + (1 - p_M(t)) \log(1 - a_M(t)) \tag{6}$$

where $\mathbb{F}$ represents the self-supervised subaction parsing parameterized by $\emptyset$. The $a_M \in \mathbb{R}^T$ is the prediction probability of the M^{th} latent state. The pseudo-label ground truth at the t^{th} frame is denoted by $p_M(t)$, and the predicted probability distribution of the subaction is represented by $a_M(t)$. The prediction of the M^{th} sequence is $\bar{t}_M$, which acts as a mid-level representation of the given action sequence. This probability vector corresponds to the subactions. To ensure consistency between the pseudo-labels and the subaction parsing predicted by $a_M(t)$, we formulate a binary cross-entropy loss, L_{ubc}, as shown in Eqn. 6. We use this loss as a supervision signal for subaction parsing, enabling a fine-grained high-level internal structure along with their temporal dependencies.

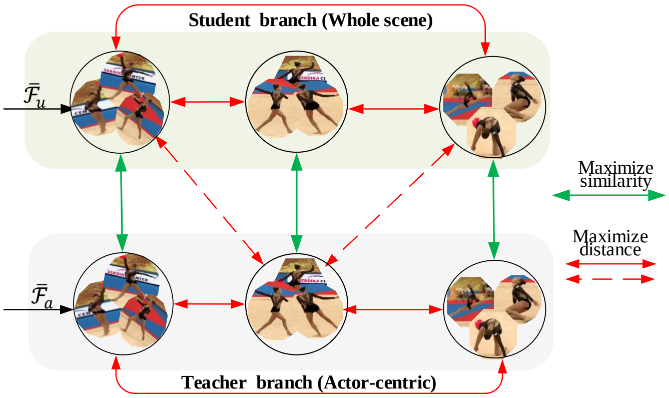


Fig. 4: Illustrating subaction-based group contrastive learning. By forming a group of subactions, we maximize the distance with different semantic subactions and minimize the distance between action features that share the same semantic.

D. Group contrastive learning

Through self-supervised subaction parsing, we obtain subaction sequences along with their temporal dependencies, as shown in Fig. 4. We define feature representations of parsed subactions in the teacher and student branches, as $\bar{\mathcal{F}}_u$ and $\bar{\mathcal{F}}_a$, sharing the same action semantic. To achieve representative representations, we design a group contrastive learning approach. This approach minimizes the distance between action features that share the same semantic, and maximizes the distance with different semantic subactions, learning the fine granularities of subactions and capturing consistent motion-oriented action features. Thus, the parsed subaction features are transformed for group contrastive learning using an MLP projection head h_φ. This projection head contains three linear layers followed by a spatiotemporal GAP, resulting in $h_u = h_\varphi(\bar{\mathcal{F}}_u)$ and $h_a = h_\varphi(\bar{\mathcal{F}}_a)$. The subaction probability distribution vectors of these inputs are then assigned pseudo-labels corresponding to the class with the maximum activation and high semantic similarity. The features with semantically similar pseudo-labels are grouped together in mini-batch, as shown in Eqn. 7.

$$G_w^p = \frac{\sum_{i=1}^{B} \mathbb{1}_{\{p=h_w\}} \mathrm{g}(h_w)}{T_B} \tag{7}$$

where $\mathrm{g(h_w)}$ is the average logits of the sequence h_w, $w \in \{u, a\}$. $\mathbb{1}$ is an indicator function that evaluates to 1. T_B refers to the number of sequences in the mini-batch B.

It is expected that the two groups would exhibit similar feature representations. We consider the mean representation of groups that share the same subaction pseudo-labels as positive pairs (G_a^p, G_u^p), while subactions from the same group video with different pseudo-labels are regarded as negative pairs (G_a^p, G_w^q). Consequently, the loss function for subaction-based group contrastive learning is defined in Eqn. 8.

$$L_{gc} = -\log \frac{\mathcal{H}(G_a^p, G_u^p)/\iota}{\mathcal{H}(G_a^p, G_u^p)/\iota + \sum_{q=1,k}^{S} \mathbb{1}_{p \neq q} \mathcal{H}(G_a^p, G_w^q)/\iota} \tag{8}$$

where $\mathcal{H}(,)$ calculates the cosine similarity, and ι is the temperature hyperparameter. In this way, we are shaping learned representations and distinguishing temporally related subactions. This ensures that the module learns to differentiate between various subactions with their temporal correspondent representation.

E. Loss functions

Given a limited labeled video X_ℓ in the base branch, the AQA problem is formulated as a regression problem to predict the AQA score $\bar{y}_i$ for the labeled samples:

$$\bar{y}_i = \mathcal{R}_\vartheta \left(\bar{\mathcal{F}}_\ell \right) \tag{9}$$

where $\mathcal{R}$ is the regressor parameterized by ϑ. To perform supervised training for the AQA on the limited labeled samples, we employ the MSE loss, which is defined as follows:

$$L_{reg} = \frac{1}{\mathbb{N}_\ell} \sum_{i=1}^{\mathbb{N}_\ell} \left(\bar{y}_i - y_i \right)^2 \tag{10}$$

Thus, the proposed approach consists of a supervised regression module, a self-supervised subaction parsing module, and a group contrastive learning module. Hence, the overall training loss of the proposed approach is given by:

$$L_{all} = L_{reg} + \lambda_1 L_{ubc} + \lambda_2 L_{un} + \lambda_3 L_{gc} \tag{11}$$

where L_{reg} is the regression loss for supervised AQA, L_{un} is the pseudo-label consistency regularization loss, L_{ubc} is a self-supervised binary cross-entropy loss, and L_{gc} is the subaction-based group contrastive loss. The λ_1, λ_2, and λ_3 are the hyperparameter weights of subaction parsing, consistency regularization, and group contrastive learning, respectively.

IV. Experiment

In this section, we present the experimental setup and analyze the results. We first describe the datasets, implementation details, and evaluation metrics of our proposed approach. Then, we conduct an extensive ablation study and compare the results to state-of-the-art full-supervised and semi-supervised AQA methods. Finally, we provide qualitative visualizations to demonstrate the effectiveness of our proposed approach and discuss failure case scenarios of our proposed method.

A. Datasets and Experiment Settings

1) Datasets: We evaluate our approach on four large-scale AQA datasets: the MTL-AQA [1], the FineDiving [13], the Rhythmic Gymnastics [36] and the FineFS datasets [37].

MTL-AQA [1]: The dataset focuses on diving, covering a wide range of actions. There are 1412 samples collected from 16 different world events. Different annotations are available in this dataset such as AQA, action recognition, and comments. Additionally, raw score annotations and Degree of Difficulty (DD) are available from multiple judges. Following [2], we divide the dataset into 1059 training and 353 test samples.

FineDiving [13]: The FineDiving dataset consists of 3000 video samples, covering 52 action types, 29 subaction 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, subaction types, temporal boundaries, as well as action scores. Similar to [13], we select 75% of samples for training and 25% for testing in all the experiments.

Rhythmic Gymnastics dataset (RG) [36]: The dataset contains a total of 1,000 videos with four types of gymnastics: ball, clubs, hoop, and ribbon. Each routine type consists of 250 videos. We follow the evaluation protocol in [36] and partition the dataset into 200 training videos and 50 test videos for each gymnastics routine type.

Fine-grained Figure Skating dataset (FineFS) [37]: The dataset contains 1167 samples where 729 are from short program (SP) and 438 samples from free skating (FS). It includes RGB videos, estimated skeleton data, fine-grained score labels, and technical subaction categories. Following [37], the dataset is split into 933 training and 234 test samples.

2) Implementation: We pre-trained the I3D backbone [61] in the Kinetics dataset to extract visual features with an initial learning rate of 10^{-4}. We utilize the Adam optimizer with weight decay set to zero. We search for the optimal ω from the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, and set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ as default setting in all experiments. The initial learning rate of our SAP-Net is set to $1e-3$. Following [13], we select 96 frames from each video in the FineDiving dataset and split them into 9 clips. Following [15], we extract 103 clips in MTL-AQA, and multiply predicted scores by the Difficulty Degree (DD). To maintain consistency with the training data, we uniformly sample labeled and unlabeled samples based on action scores. To achieve better convergence, we assign different numbers of epochs for each category in the RG and FineFS datasets following [36], [37], [46]. We set 300, 400, 500, and 300 epochs for Ball, Clubs, Hoop, and Ribbon, respectively. Additionally, we set 400 epochs for the model training with SP, and 500 epochs for the FS samples.

3) Evaluation metrics: To keep alignment with existing approaches [13], [15], we adopt Spearman's rank correlation (Spr. Corr. ρ, ranges from -1 to 1, the larger value is the better) to measure the difference between predicted and ground-truth scores. This metric can be formulated as:

$$\rho = \frac{\sum_i (y_i - y)(\bar{y}_i - \bar{y})}{\sqrt{\sum_i (y_i - y)^2 \sum_i (\bar{y}_i - \bar{y})^2}} \tag{12}$$

where y_i and $\bar{y}_i$ indicate the rankings of two score sequences, respectively.

B. Ablation Study

1) Evaluation on components of our SAP-Net: We explore the effectiveness of the components in the MTL-AQA and FineDiving datasets through quantitative and qualitative analysis. To demonstrate the effectiveness of each module, we

TABLE I: Evaluation on components of our SAP-Net in the MTL-AQA dataset. Thus, our approach exploits solid and consistent representations of action sequences.

	Semi-supervised				Supervised
Approach	Spr. Corr. (ρ)				Spr. Corr. (ρ)
	10%	20%	30%	40%	
I3D+MLP [25]	0.618	0.643	0.681	0.703	0.8921
I3D+SSP	0.702	0.742	0.762	0.782	0.9312
I3D+GCL	0.692	0.723	0.751	0.772	0.924
SAP-Net (Ours)	**0.729**	**0.762**	**0.782**	**0.801**	**0.9638**

TABLE II: Evaluation on components of our SAP-Net in the FineDiving dataset. Thus, our module captures the high-level semantics and internal temporal structure of subactions.

	Semi-supervised				Supervised
Approach	Spr. Corr. (ρ)				Spr. Corr. (ρ)
	5%	10%	15%	20%	
I3D+MLP [25]	0.693	0.708	0.735	0.748	0.8576
I3D+SSP	0.749	0.781	0.791	0.827	0.9283
I3D+GCL	0.757	0.775	0.793	0.815	0.9171
SAP-Net (Ours)	**0.779**	**0.801**	**0.821**	**0.845**	**0.9482**

attempt to remove them one by one from our proposed SAP-Net and evaluate their performance. As a baseline, we use the I3D backbone network [61] to extract visual features from videos, which are then fed into a three-layer MLP (I3D+MLP) for score regression. The model is optimized using MSE loss. For component evaluation, we combine the I3D backbone with the Self-supervised subaction Parsing (I3D+SSP) module and subaction-based Group Contrastive Learning module (I3D+GCL) for score estimation. The experimental results are shown in Table I and Table II.

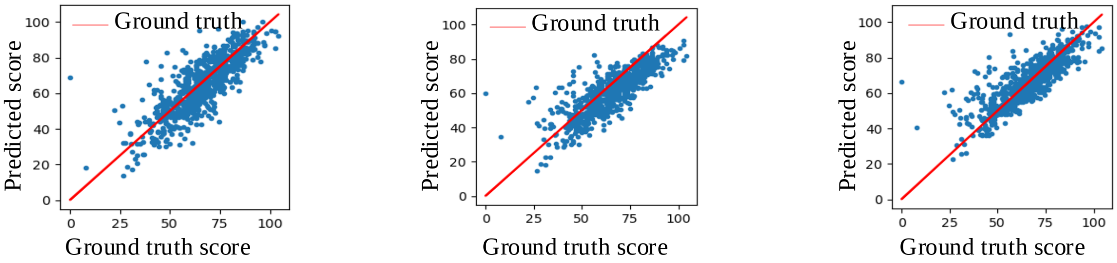
Effect of self-supervised subaction parsing (I3D+SSP). Our proposed SSP module, trained using a limited number of labeled samples and a larger number of unlabeled videos with the L_{ubc} loss, demonstrates encouraging improvements in score prediction. Our SSP module achieves significant enhancements with various percentages of labeled samples compared to the baseline I3D+MLP in the MTL-AQA and FineDiving datasets. For example, our SSP module demonstrates superior performance score predictions compared to the baseline I3D+MLP, achieving impressive performance scores of 0.742 and 0.827 on the MTL-AQA and FineDiving datasets, respectively, with 20% labeled samples. These results highlight the effectiveness of our self-supervised learning approach in capturing distinct patterns and the fine-grained temporal structure of action features, leading to enhanced representation for semi-supervised AQA performance score prediction. Compared to supervised approaches, our approach achieves remarkable scores of 0.9312 and 0.9283 on the MTL-AQA and FineDiving datasets. The proposed module captures both the high-level semantics and internal temporal structure of subactions in unlabeled samples for semi-supervised AQA.

Effect of subaction-based group contrastive learning (I3D+GCL). Our proposed module (I3D+GCL) is trained with a limited number of labeled videos and a group contrastive loss with a massive amount of unlabeled videos. As shown in Table I and Table II, our proposed module shows notable enhancements over the baseline I3D+MLP. The training approach allows us to effectively leverage both labeled and unlabeled samples, enhancing the performance of our proposed approach. For example, with only 10% of labeled training samples, our module shows a remarkable performance score of 0.692 and 0.775 in the MTL-AQA and FineDiving datasets, which are higher than the SOTA approaches. These experimental results validate the effectiveness of our proposed approach, emphasizing its potential to improve middle-level representation and capture motion-oriented features. Among full-supervised AQA, our proposed module achieves performance of 0.9240 and 0.9171 in the MTL-AQA and FineDiving datasets. According to experiments, the SSP and GCL have their contribution and advantages, thus we combine the SSP and GCL for capturing the internal dependencies in addressing fine-grained scene-invariant AQA problems.

Combining SSP and GCL in our SAP-Net. While incorporating all the components, the proposed SAP-Net achieves a remarkable improvement in two datasets. For example, when using only 20% labeled videos for model training, our proposed model demonstrates a significant performance score prediction of 0.762 and 0.845 in the MTL-AQA and FineDiving datasets. This emphasizes the synergistic effect of combining SSP and GCL modules, resulting in a powerful framework for enhancing unlabeled sample representation. Furthermore, it boosts the overall performance of AQA and enables our approach to effectively represent unlabeled videos. In a fully supervised, the combined model SAP-Net achieves even higher performance, obtaining scores of 0.9638 and 0.9482 in the MTL-AQA and FineDiving datasets. Through joint optimization, our approach achieves a superior understanding of fine-grained action sequences.

Qualitative analysis on predicted score distribution. In Fig. 5, we present the distribution of predicted scores in different models, *i.e.,* I3D+SSP, I3D+GCL, and SAP-Net, using 20% labeled samples. The y- and x-coordinates represent the predicted score and the ground truth scores, respectively. By examining the scatter plot, we can gain insight into the comparative distribution of the predicted scores in the various models. Furthermore, the scatter plot provides a visual representation, enabling a comprehensive comparison between the models and highlighting the effectiveness of our approach. Specifically, our proposed approach incorporates all the proposed components, leading to a more uniformly scattered distribution of predicted scores, which is close to the ground truth. This improved scattering can be attributed to the alignment of predicted points with the ground truth, indicating that our proposed approach successfully captures the underlying patterns in the videos.

2) Sensitivity to the proportion of labeled samples: To verify the robustness of our approach with different labeled samples, we conduct an analysis of sensitivity to the number of labeled samples on the MTL-AQA dataset, as depicted in Fig. 6. Our proposed SAP-Net demonstrates promising results, even when trained with a significantly smaller amount of labeled samples. Specifically, utilizing only the data labeled with 10%, our approach achieves a comparable Spr. Corr.



(a) I3D+SSP (20% labels) (b) I3D+GCL (20% labels) (c) **SAP-Net** (20% labels)

Fig. 5: A comparison of the proposed module's predicted score distribution in a scatter plot with 20% labeled samples in the FineDiving dataset. The red line refers to the ground truth score, while the blue scatter points refer to predicted scores.

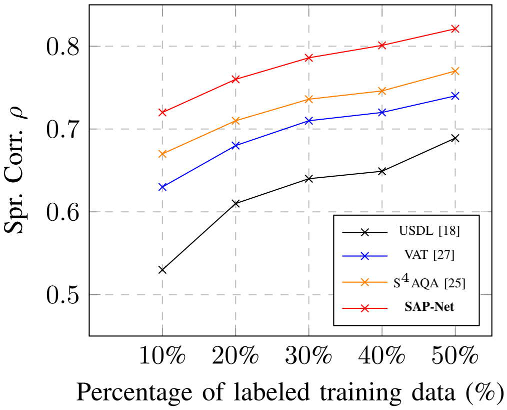


Fig. 6: Comparison of experiment results when we use different proportions of labeled samples in the MTL-AQA dataset.

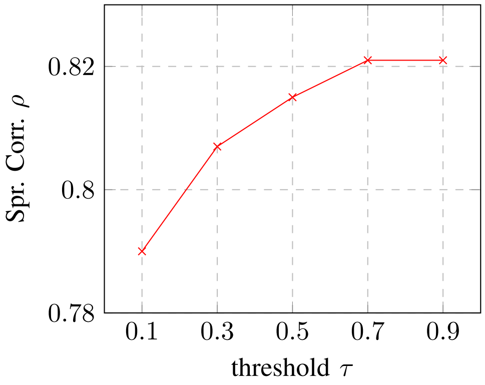


Fig. 7: Results of varying threshold τ with only 15% labeled training samples in the FineDiving dataset. The best result is achieved when setting τ = 0.7.

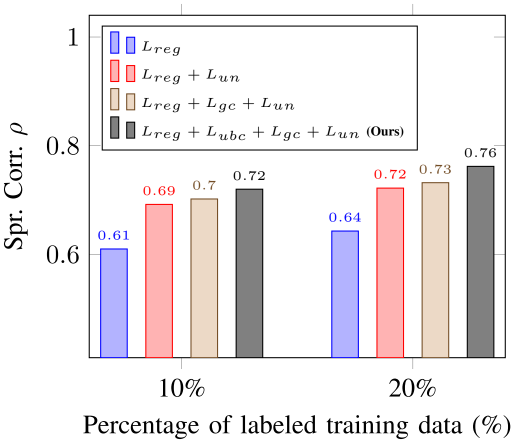


Fig. 8: Evaluation on loss functions in the MTL-AQA dataset. These results imply that proposed losses play a crucial role in learning fine-grained features along with temporal diversity.

performance with S^{4}AQA [25], which employed 40% labeled samples. The effectiveness of our method lies in its ability to leverage fine-grained and middle-level representations from unlabeled videos. With an increase in the number of labeled videos, our approach consistently enhances the prediction of the AQA performance score. These results imply that the proposed modules exploit solid and consistent representations between labeled and unlabeled samples in the semi-supervised AQA.

3) Effect of threshold value τ: We analyze the effect of different threshold values τ by conducting experiments using only 15% of labeled samples on the FineDiving dataset. We present the results in Fig. 7. Adjusting the threshold value plays a crucial role in determining the performance of the AQA. When the threshold value is set too low, it causes a decrease in the overall performance of score prediction. However, by increasing the threshold value, we observe that the performance score prediction is increased. These experimental results clearly show that the optimal results are achieved when the value of τ is set to 0.7. With this threshold, our approach identifies relevant fine-grained features and improves the overall performance score prediction.

4) Evaluation on loss functions: The experimental results depicted in Fig. 8 demonstrate a clear trend in the contribution of the proposed loss functions. An interesting observation is that the performance substantially deteriorates dramatically when excluding either the L_{gc} or L_{ubc} losses. These results emphasize that the proposed losses maintain high-quality pseudo-labels and learn fine-grained action features. The modules leverage temporal diversity consistent video representation, and improved performance score prediction. Thus, it is worth noting that both the L_{gc} and L_{ubc} losses play a crucial role in learning fine-grained features with their temporal diversity.

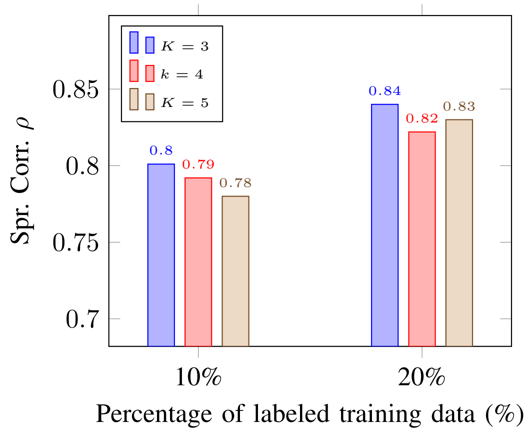


Fig. 9: Evaluation on the optimal number K of subaction sequences in the FineDiving dataset. The comparison shows that the optimal value for K is 3.

5) Evaluation on the optimal number K of subactions: To explore the effect of the number of subaction sequences on hyperparameter K, we conduct an experiment on the FineDiving dataset as shown in Fig. 9. We varied the value of K and analyzed its impact on the performance of our approach. The experiments revealed the significance of choosing a suitable K value for achieving optimal performance score prediction. A small value of K may not adequately capture the complexity and nuances of the actions, leading to suboptimal results. On the other hand, a high K runs the risk of overfitting, where the model becomes too specific to the training data and fails to generalize well to unlabeled samples. The ideal K captures the characteristics of subaction sequences. As shown in Fig. 9, the three subactions provide the optimal value of K, enhancing the overall AQA performance score prediction.

6) Effect of generating pseudo-labels from actor-centric region: To demonstrate the contribution of generating pseudo-labels from the actor-centric region, we replaced the actor-centric input with an augmented version featuring varying pixel-level distributions from the input clips while preserving the semantic meaning of the video unchanged. The experimental results are shown in Fig. 10. Our evaluation demonstrates that learning from motion features and generating pseudo-labels are vital components for achieving superior performance score prediction. Our approach not only enhances the quality of pseudo-labels but also facilitates the learning of discriminative action features from unlabeled videos and an overall performance boost. The experimental results provide compelling evidence of the effectiveness of our motivation in enhancing the model's capability to generate accurate and discriminative pseudo-labels from actor-centric regions.

7) Effect of attention refinement for subaction parsing: By incorporating attention refinement, we identify the salient regions and enable the back-propagation of gradients through the differentiable module. This is the dual advantage of the module [24]. To verify the contribution, we conduct experiments and summarize results in Table III, offering quantitative evidence of attention refinement on subaction parsing and pseudo-label generation. Through our experiments, we observe that refining attention features leads to improved subaction parsing and pseudo-label generation, enhancing the overall performance score prediction of AQA.

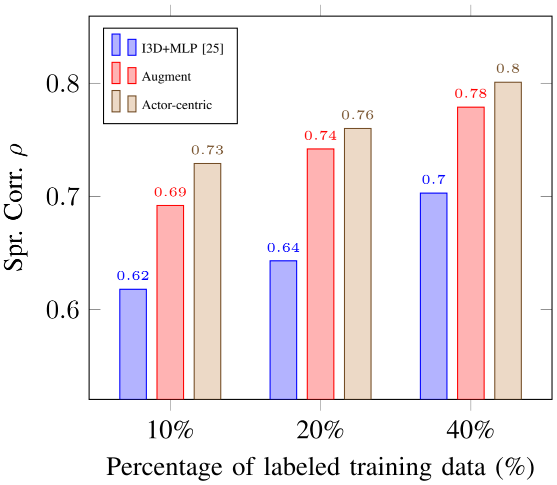


Fig. 10: Effect of generating pseudo-label from actor-centric region in the MTL-AQA dataset. By leveraging motion information, the model gains a deeper understanding of temporal dynamics.

TABLE III: Effect of attention refinement for subaction parsing and pseudo-label generation in the MTL-AQA dataset. “w/o” stands for without attention refinement.

Approach	Spr. Corr. (ρ)			
	10%	20%	30%	40%
I3D+MLP [25]	0.618	0.643	0.681	0.703
Pseudo-label w/o	0.702	0.742	0.762	0.782
subaction parsing w/o	0.702	0.728	0.748	0.762
SAP-Net w/o	0.689	0.702	0.732	0.752
SAP-Net (Ours)	**0.729**	**0.762**	**0.782**	**0.801**

C. Comparison with the state-of-the-art approaches

The comparison between our proposed approach and the state-of-the-art methods implemented on the MTL-AQA, the FineDiving, the RG, and the FineFs datasets are shown in Table IV, Table V, Table VI, and Table VII, respectively.

1) Comparison in the: MTL-AQA dataset The comparison between the proposed approach with the state-of-the-art approaches in the MTL-AQA dataset is shown in Table IV. Among semi-supervised learning methods, COREG [63] has the worst performance because it struggles to learn discriminative action representations with its non-parametric model [25]. This is attributed to the inherent challenge of learning discriminative action representation using a non-parametric approach. However, our approach effectively overcomes this limitation and obtains a more robust video representation with limited labeled samples. Comparing our approach with S^{4}AQA [25], we achieve an improvement of up to 5.3% and 5.5% on the performance of the MTL-AQA dataset with 10% and 40% labeled samples, respectively. These results imply the motion-oriented

TABLE IV: Comparison with the state-of-the-art approaches in the MTL-AQA dataset. We compare fully-supervised and semi-supervised approaches (10% and 40% labled samples are used for model training).

	Approach	Year	Sp. Corr. (ρ)
Supervised	USDL [18]	2020	0.9066
	MUSDL [18]	2020	0.9158
	CoRe [15]	2021	0.9512
	TSA-Net [10]	2021	0.9422
	TAP [12]	2022	0.9607
	PCLN [16]	2022	0.9230
	DAE-CoRe [42]	2023	0.9589
	HGCN [8]	2023	0.9563
	FSPN [14]	2023	0.9601
	SGN [43]	2023	0.9607
	APT [47]	**2023**	**0.9678**
	SAP-Net (Ours)	-	0.9638

	Approach	Year	Sp. Corr. (ρ) 10%	40%
Semi-supervised	I3D+MLP [18]	2020	0.618	0.703
	C3D-AVG-MTL [1]	2018	0.584	0.656
	COREG [63]	2019	0.487	0.526
	Pseudo-labels [22]	2017	0.622	0.716
	VAT [27]	2017	0.635	0.724
	S^{4}L [26]	2019	0.621	0.721
	S^{4}AQA [25]	2022	0.676	0.746
	SAP-Net (Ours)	-	**0.729**	**0.801**

TABLE V: Comparison with the state-of-the-art approaches in the FineDiving dataset. We compare with fully supervised and semi-supervised methods (10% and 20% labeled samples are used for model training).

	Approach	Year	Sp. Corr. (ρ)
Supervised	USDL [18]	2020	0.8913
	MUSDL [18]	2020	0.8978
	CoRe [15]	2021	0.9061
	TSA [13]	2022	0.9203
	APT [47]	2023	0.9246
	FSPN [14]	2023	0.9421
	SAP-Net (Ours)	-	**0.9482**

	Approach	Year	Sp. Corr. (ρ) 10%	20%
Semi-supervised	I3D+MLP [13]	2022	0.708	0.748
	USDL [18]	2020	0.732	0.762
	MUSDL [18]	2020	0.772	0.810
	TSA [13]	2022	0.772	0.810
	SAP-Net (Ours)	-	**0.801**	**0.845**

TABLE VI: Comparison with the state-of-the-art approaches with 40% labeled samples for semi-supervised prediction in the Rhythmic Gymnastics dataset.

Method	Year	Ball	Clubs	Hoop	Ribbon	Avg
COREG [63]	2005	0.230	0.338	0.331	0.268	0.292
SVR [64]	2014	0.175	0.243	0.261	0.309	0.248
Action-Net [36]	2020	0.196	0.403	0.319	0.305	0.308
Pseudo-labels [22]	2017	0.183	0.330	0.346	0.305	0.292
VAT [27]	2019	0.208	0.355	0.345	0.292	0.301
S^{4}L [26]	2019	0.209	0.325	0.324	0.290	0.288
S^{4}AQA [25]	2022	0.248	0.388	0.372	0.357	0.342
SAP-Net (Ours)	-	**0.339**	**0.421**	**0.420**	**0.392**	**0.393**

TABLE VII: Comparison with the state-of-the-art approaches in the FineFS dataset with 50% labeled samples. The † indicates the results of our implementation with the backbone of I3D [61] and vision transformer [65].

Approach	Year	SP (ρ)		FS (ρ)	
		PCS	TES	PCS	TES
I3D+MLP†	-	0.586	0.621	0.552	0.613
VST+MLP†	-	0.613	0.631	0.672	0.675
Action-Net [36]	2022	0.594	0.550	0.574	0.626
GDLT [46]	2022	0.576	0.491	0.792	0.675
SAP-Net (Ours)	-	**0.648**	**0.653**	**0.815**	**0.721**

feature representation, combined with capturing semantic and temporal structures of subactions, enhances the overall score prediction. In comparison to the fully supervised approaches, our proposed approach achieves comparable results compared to APT [47] as shown in Table IV. These results imply that our model learns more discriminative features and improves performance under a low-label regime.

2) Comparison in the FineDiving dataset: The comparison between our proposed approach with other AQA approaches in the FineDiving dataset [13] is shown in Table V. As shown in the table, our proposed approach shows superior performance compared to state-of-the-art approaches in both fully supervised and semi-supervised AQA. With a Spr. Corr. score of 0.9482, our approach surpasses fully-supervised methods. This result implies that the teacher branch plays a crucial role in guiding the student branch to learn motion-oriented action features by reducing the model's reliance on the video background. Furthermore, our approach surpasses existing semi-supervised AQA approaches, achieving Spr. Corr. scores of 0.801 and 0.845 with only 10% and 20% of training label samples, respectively, as shown in Table V. In general, the proposed approach not only exhibits superior performance but also offers the benefit of adaptability to various learning scenarios. By using a limited amount of labeled data, this semi-supervised approach can reduce model dependency on manual annotations, potentially saving time for real-time scenarios.

3) Comparison in the Rhythmic Gymnastics (RG) dataset: We assessed the effectiveness of our proposed approach by comparing it with semi-supervised approaches on the RG dataset. The results are presented in Table VI. Specifically, our method shows an improvement of up to 5.1% compared to the state-of-the-art semi-supervised approach S^{4}AQA [25]. Our approach provides a comprehensive framework for capturing fine-grained action information and accurately assessing AQA performance. This improvement is achieved by effectively utilizing unlabeled videos to learn discriminative representations and understand the temporal structures of action sequences.

4) Comparison in the Fine-grained Figure Skating dataset (FineFS): We conduct a comprehensive comparison between our proposed SAP-Net and semi-supervised baselines implemented on the FineFS dataset as shown in Table VII. The table demonstrates the superiority of our approach over the state-of-the-art GDLT [46] method. It achieves an impressive average score of 0.650, outperforming GDLT [46] on SP (Short Program) and 0.768, showcasing a 3.5% improvement on FineFS (Free Skating). These results imply that our proposed modules have the potential to enhance the discriminative nature of representation and leverage the unlabeled samples, ultimately

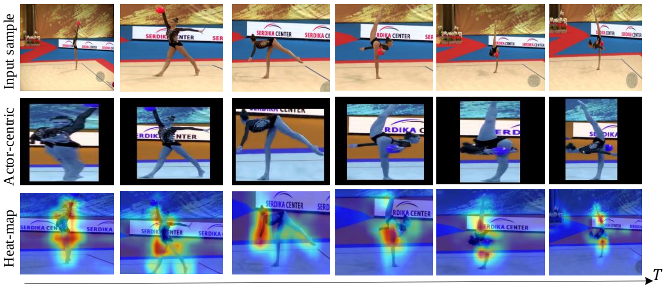


Fig. 11: Visualization results of attention heat-map (student branch) and actor-centric (teacher branch) in the RG dataset.

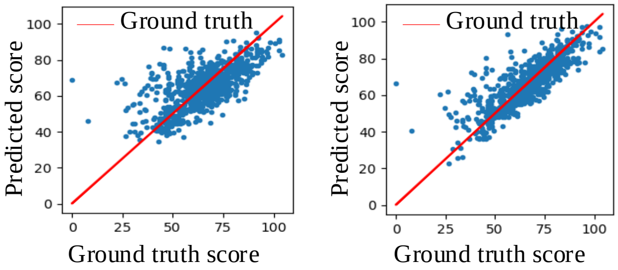
improving the overall performance score prediction.

D. Visualization Results

1) Attention heat-map and actor-centric regions: The teacher and student branches learn a consistent video representation. To provide a visual understanding, Fig. 11 shows the action-centric region detection performed by the teacher branch and attention heatmap visualization using Grad-CAM [66] generated by the student branch. Our actor-centric representation prioritizes motion features while effectively ignoring background noises, leading to more focused and consistent representations. By utilizing motion-oriented action features, the student branch can better infer and understand action features with small discrepancies happening in similar backgrounds. The attention results obtained from the student branch further validate the effectiveness of our proposed approach. Furthermore, the consistency between the attention patterns of the student and teacher branches reinforces the robustness of our proposed approach.

2) Scatter plot of score prediction results with different proportions of labeled samples: In Fig. 12, we visualize the prediction results in the form of a scatter plot to enable an intuitive comparison between prediction results when using different proportions of labeled training samples for semi-supervised prediction. As the number of labeled videos increases, predicted scores become more uniformly scattered. Because the predicted score points are close to the ground truth. This is due to our approach, which makes the action feature representation of unlabeled videos more representative and learns consistent motion-oriented feature representations.

3) Failed case scenarios of our proposed approach: While our proposed approach generally performs well, it is important to analyze the failed case scenarios to identify areas for improvement. The predicted score error histogram on the MTL-AQA test samples is shown in Fig. 13, where blue, orange, and green bars are errors of using 10%, 20%, and 40% labeled samples for semi-supervised model training, respectively. As shown in the figure, only one sample has errors greater than 40, namely, #340 with 40% of the labeled samples. Specifically, during the diving attempt with a difficulty of 3.1. The diver was unable to maintain a proper rhythm on the springboard and failed to execute the necessary movements after takeoff as shown in Fig. 14, and ultimately made a heavy splash upon entering the water [1], [8]. As a result of these errors, the dive has been judged as a score of 0. Our model, along with other approaches such as [8], [12], [25], [67], performs poorly for this particular type of action sequence. Thus, the challenge is addressed by incorporating more discriminative semantic auxiliary information to improve the model's performance in specific scenarios.



(a) SAP-Net (5% labels) (b) SAP-Net (20% labels)

Fig. 12: The comparison of predicted score distribution in a scatter plot under different proportions of labeled samples in the FineDiving dataset.

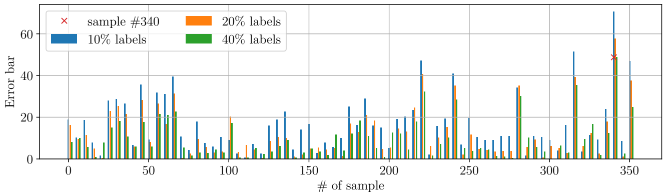


Fig. 13: The failure case study of our model predicted score error in the MTL-AQA test set. The blue, orange, and green colors refer to error bars of experiments using 10%, 20%, and 40% labeled samples for model training.

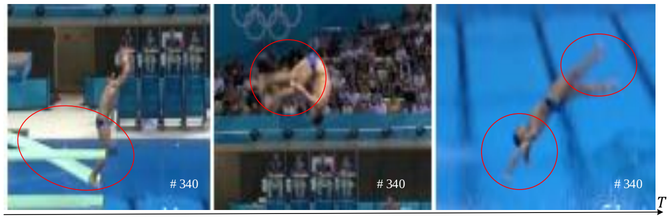


Fig. 14: The failure case scenarios of our proposed approach on the diving action, *i.e.,* #340, showed a large error in the MTL-AQA dataset, where only 40% labeled samples are used.

V. Conclusion

In this paper, we propose a self-supervised subaction parsing network that aims to learn consistent motion-oriented feature representations and internal temporal structures of subaction sequences, bridging labeled and unlabeled samples for semi-supervised AQA. In the proposed approach, a teacher-student network structure is employed, where the teacher branch receives the actor-centric region to generate high-quality pseudo-labels, and assists the student branch in learning motion-oriented action features along with temporal dependencies. In each branch, a self-supervised subaction parsing module is incorporated to locate and parse subaction sequences. Furthermore, the presented group contrastive learning approach captured consistent motion-oriented action features and temporal dynamics across the two branches. Experimental results and ablation studies demonstrated the effectiveness of our proposed approach on four challenging AQA datasets. Our experiments suggest that our approach holds promise for enhancing AQA performance in practical scenarios. We are currently focusing our future research on improving our model's performance in specific scenarios by incorporating more discriminative auxiliary information. This includes athlete poses, detection of foul actions, and semantic description of action sequences, which can effectively mitigate and model the spatiotemporal dynamics of scene-invariant action sequences.

Acknowledgments

This work was supported by the Ministry of Science and Technology of China Foreign Young Talents Program under grant No. QN2023036003L, Sichuan Science and Technology Program under grant No. 2024JDHJ0059, Natural Science Foundation of China under grant No. W2433164 and the Science and Technology Innovation Committee of Shenzhen Municipality Foundation (No.JCYJ20210324132203007 and JCYJ20230807120000001).

References

[1] P. Parmar and B. Tran Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *CVPR-Workshops*, 2019, pp. 304–313.

[2] P. Parmar and B. T. Morris, "Learning to score olympic events," in *CVPR-Workshops*, 2017, pp. 76–84.

[3] Y. Zhang, W. Xiong, and S. Mi, "Learning time-aware features for action quality assessment," *Pattern Recognition Letters*, vol. 158, pp. 104–110, 2022.

[4] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *ICCV*, 2011, pp. 1784–1791.

[5] Y. Tian, H. Zeng, J. Hou, J. Chen, and K.-K. Ma, "Light field image quality assessment via the light field coherence," *IEEE Transactions on Image Processing*, vol. 29, pp. 7945–7956, 2020.

[6] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

[8] K. Zhou, Y. Ma, H. P. H. Shum, and X. Liang, "Hierarchical graph convolutional networks for action quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[9] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *CVPR*, 2018, pp. 6057–6066.

[10] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *ACM MM*, 2021, pp. 4902–4910.

[11] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *CVPR*, 2019, pp. 7854–7863.

[12] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *ECCV*, 2022, pp. 422–438.

[13] 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 *CVPR*, 2022, pp. 2949–2958.

[14] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Fine-grained spatio-temporal parsing network for action quality assessment," *IEEE Transactions on Image Processing*, vol. 32, pp. 6386–6400, 2023.

[15] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *ICCV*, 2021, pp. 7899–7908.

[16] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *ECCV*, 2022, pp. 457–473.

[17] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *ICCV*, 2019, pp. 6330–6339.

[18] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *CVPR*, 2020, pp. 9839–9848.

[19] J. Gao, W.-S. Zheng, J.-H. Pan, C. Gao, Y. Wang, W. Zeng, and J. Lai, "An asymmetric modeling for action assessment," in *ECCV*, 2020, pp. 222–238.

[20] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *CVPR*, 2023, pp. 2405–2414.

[21] G. Ding and A. Yao, "Leveraging action affinity and continuity for semi-supervised temporal action segmentation," in *ECCV*, 2022.

[22] P. Hou, X. Geng, Z.-W. Huo, and J.-Q. Lv, "Semi-supervised adaptive label distribution learning for facial age estimation," in *in Proc. AAAI Conf. Artif. Intell*, vol. 31, no. 1, 2017.

[23] J. Ji, K. Cao, and J. C. Niebles, "Learning temporal action proposals with fewer labels," in *ICCV*, 2019, pp. 7072–7081.

[24] E. Elhamifar and D. Huynh, "Self-supervised multi-task procedure learning from instructional videos," in *ECCV*, 2020, pp. 557–573.

[25] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, "Semi-supervised action quality assessment with self-supervised segment feature recovery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6017–6028, 2022.

[26] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *ICCV*, 2019, p. 476–1485.

[27] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019.

[28] X. Wang, S. Zhang, Z. Qing, Y. Shao, C. Gao, and N. Sang, "Self-supervised learning for semi-supervised temporal action proposal," in *CVPR*, 2021, pp. 1905–1914.

[29] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *ICLR*, 2020.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016, pp. 21–37.

[31] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *NeurIPS*, 2020, pp. 596–608.

[32] J. Li, C. Xiong, and S. C. H. Hoi, "Comatch: Semi-supervised learning with contrastive graph regularization," in *ICCV*, 2021, pp. 9455–9464.

[33] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1–11.

[34] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9726–9735.

[35] A. Singh, O. Chakraborty, A. Varshney, R. Panda, R. Feris, K. Saenko, and A. Das, "Semi-supervised action recognition with temporal contrastive learning," in *CVPR*, 2021, pp. 10 389–10 399.

[36] L.-A. Zeng, F.-T. Hong, W.-S. Zheng, Q.-Z. Yu, W. Zeng, Y.-W. Wang, and J.-H. Lai, "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *ACM MM*, 2020, pp. 2526–2534.

[37] Y. Ji, L. Ye, H. Huang, L. Mao, Y. Zhou, and L. Gao, "Localization-assisted uncertainty score disentanglement network for action quality assessment," in *ACM MM*, 2023, p. 8590–8597.

[38] S. Gattupalli, D. Ebert, M. Papakostas, F. Makedon, and V. Athitsos, "Cognilearn: A deep learning-based interface for cognitive behavior assessment," in *ICIUI*, 2017, p. 577–587.

[39] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa, "Automated assessment of surgical skills using frequency analysis," in *ICMICCAI*, 2015, pp. 430–438.

[40] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*, Cham, 2014, pp. 556–571.

[41] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4578–4590, 2020.

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

[43] Z. Du, D. He, X. Wang, and Q. Wang, "Learning semantics-guided representations for scoring figure skating," *IEEE Transactions on Multimedia*, vol. 26, pp. 4987–4997, 2024.

[44] Q. Liu, X. Liu, K. Liu, X. Gu, and W. Liu, "Sigformer: Sparse signal-guided transformer for multi-modal human action segmentation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 8, pp. 1–22, 2024.

[45] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "Am i a baller? basketball performance assessment from first-person videos," in *CVPR*, 2017, pp. 2196–2204.

[46] A. Xu, L.-A. Zeng, and W.-S. Zheng, "Likert scoring with grade decoupling for long-term action assessment," in *CVPR*, 2022, pp. 3222–3231.

[47] H. Fang, W. Zhou, and H. Li, "End-to-end action quality assessment with action parsing transformer," in *VCIP*, 2023, pp. 1–5.

[48] Q. An, M. Qi, and H. Ma, "Multi-stage contrastive regression for action quality assessment," in *ICASSP*, 2024, pp. 4110–4114.

[49] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Intra- and inter-action understanding via temporal action parsing," in *CVPR*, 2020, pp. 727–736.

[50] C. Zhang, A. Gputa, and A. Zisserman, "Temporal query networks for fine-grained video understanding," in *CVPR*, 2021, pp. 4484–4494.

[51] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *CVPR.*, 2020, pp. 2613–2622.

[52] 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 Transactions on Multimedia*, vol. 25, pp. 5500–5515, 2023.

[53] J. Xiao, L. Jing, L. Zhang, J. He, Q. She, Z. Zhou, A. Yuille, and Y. Li, "Learning from temporal gradient for semi-supervised action recognition," in *CVPR*, 2022, pp. 3242–3252.

[54] X. Wang, S. Zhang, Z. Qing, Y. Shao, C. Gao, and N. Sang, "Self-supervised learning for semi-supervised temporal action proposal," in *CVPR*, 2021, pp. 1905–1914.

[55] L. Jing, T. Parag, Z. Wu, Y. Tian, and H. Wang, "Videossl: Semi-supervised learning for video classification," in *WACV*, 2021, pp. 1110–1119.

[56] K. Liu, M. Qu, Y. Liu, Y. Wei, W. Zhe, Y. Zhao, and W. Liu, "Single-frame supervision for spatio-temporal video grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. doi: 10.1109/TPAMI.2024.3415087, 2024.

[57] X. Chen, W. Liu, X. Liu, Y. Zhang, J. Han, and T. Mei, "Maple: Masked pseudo-labeling autoencoder for semi-supervised point cloud action recognition," in *ACM Multimedia*, 2022, pp. 708–718.

[58] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, "Videomoco: Contrastive video representation learning with temporally adversarial examples," in *CVPR*, 2021, pp. 11 200–11 209.

[59] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3d action recognition," in *ECCV*, 2020, pp. 35–51.

[60] N. Behrmann, M. Fayyaz, J. Gall, and M. Noroozi, "Long short view feature decomposition via contrastive video representation learning," in *ICCV*, 2021, pp. 9224–9233.

[61] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 4724–4733.

[62] N. Alsudays, J. Wu, Y.-K. Lai, and Z. Ji, "Afpsnet: Multi-class part parsing based on scaled attention and feature fusion," in *WACV*, 2023, pp. 4033–4042.

[63] Z.-H. Zhou and M. Li, "Semi-supervised regression with co-training," in *IJCAI*, 2005, p. 908–913.

[64] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014, pp. 556–571.

[65] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *ICCV*, 2021, pp. 6816–6826.

[66] 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 *CVPR*, 2017, pp. 618–626.

[67] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *ECCV*, 2022, pp. 457–473.