Learning Semantics-Guided Representations for Scoring Figure Skating

Zexing Du, Di He, Xue Wang and Qing Wang, *Senior Member, IEEE*

***Abstract*—This paper explores semantic-aware representations for scoring figure skating videos. Most existing approaches to sports video analysis only focus on reasoning action scores based on visual input, limiting their ability to depict high-level semantic representations. Here, we propose a teacher-student-based network with an attention mechanism to realize an adaptive knowledge transfer from the semantic domain to the visual domain, which is termed semantics-guided network (SGN). Specifically, we use a set of learnable atomic queries in the student branch to mimic the semantic-aware distribution in the teacher branch, which is represented by the visual and semantic inputs. In addition, we propose three auxiliary losses to align features in different domains. With aligned feature representations, the adapted teacher is capable of transferring the semantic knowledge to the student. To verify the effectiveness of our method, we collect a new dataset OlympicFS for scoring figure skating. Besides action scores, OlympicFS also provides professional comments on actions for learning semantic representations. By evaluating four challenging datasets, our method achieves state-of-the-art performance.**

***Index Terms*—Figure skating videos, sports video analysis, multi-modality representation learning, teacher-student network, action quality assessment.**

I. Introduction

BENEFITING from the healthy and graceful characteristics of figure skating, an increasing number of people are participating in this sport. And there are a great number of figure skating videos uploaded online with the development of digital cameras and media-sharing platforms. Therefore, it has become increasingly important to accurately analyze various performance indicators in sports videos, which have a great range of applications in automatically scoring the players, highlighting shot generation, and video summarization [1]. Unlike action recognition focuses on classifying actions within a few seconds [2]–[4], long-term figure skating analysis is more challenging since they contain richer and more complicated correlations [5]. Although a great progress has been achieved in figure skating analysis with the development of neural networks and large-scale datasets [1], [6], [7], how visual processing depicts and interacts with semantic representations remains unclear.

Cognitive neuroscience research has found that the human capability of recognizing actions involves two main processes, the rapid visual analysis of the action in posterior regions along the ventral stream and the activation of semantic knowledge in anterior regions [8]. Visual processing leads to the automatic activation of the conceptual knowledge [9], while the semantic representations are activated along the ventral stream. However, existing methods for sports video analysis mainly focus on exploiting visual context, without exploring semantic information in videos, which limits their ability to cope with high-dimensional semantic representations. Accordingly, to address the question of how the visual properties of actions elicit semantic information, we collect a new multimodal dataset and propose a semantics-guided network (SGN) to bridge the gap between semantic and visual domains.

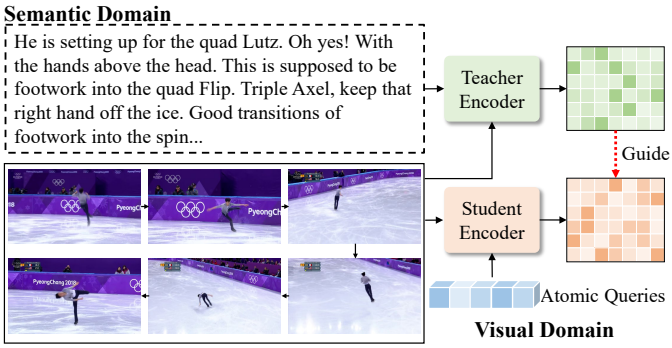


Fig. 1. An overview of our method. In the semantic domain, we encode comments and video features jointly to generate semantic-aware representations. Then these semantic-aware representations are used to guide the training of visual features. The semantic information is only used for training, and we get the figure skating score only based on the visual input during inference.

A professional figure skating commentator can point out key moments in the competition, such as impressive *jumps* or *falls*, which could be collected in international figure skating competitions and provide rich semantic information for visual representations. Based on this insight, we collect a new dataset, named OlympicFS, from Olympic Winter Games in Pyeongchang 2018 and Beijing 2022. Our OlympicFS contains four categories of figure skating competitions, *i.e.*, men/ladies short program and men/ladies free skating. For annotations, we provide scores from professional judges in competitions. Importantly for our purposes, we also collect detailed feedback from sports commentators, which provide rich semantic information for sports analysis that was overlooked in previous works [6], [10].

To explore semantics-guided representations, we introduce a teacher-student strategy specific to Transformers. Our model aims at using semantic-aware representations to guide the training of visual features in the visual domain, as illustrated in

Manuscript received May 11, 2023; revised July 16 and September 1, 2023. Accepted October 25, 2023. This work was supported by NSFC under Grant 62031023 and Grant 61801396. *(Corresponding author: Qing Wang.)*

The authors are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: duzexing@mail.nwpu.edu.cn; xwang@nwpu.edu.cn; qwang@nwpu.edu.cn).

Fig. 1. Especially, in the semantic domain, we first aggregate semantic descriptions and visual features by conducting cross-attention [11] between semantic comments and videos to generate semantic-aware representations. Then, these features serve as the teacher to provide supervision for the visual domain. In the student branch, we define a set of learnable atomic queries to describe different components in videos. For example, a figure skating action consists of several key parts, such as *spin*, *sequence*, *jump*, *etc.* These fine-grained queries enable the model to identify the key atomic actions in videos. The training of these atomic queries is guided by the teacher in the semantic domain. Besides comments in our OlympicFS, the semantic representations in the teacher branch could also be extracted from other modalities, such as the music in [10].

To bridge the feature space in both domains, we propose three auxiliary losses in this work, which use the semantic-aware distribution to guide the visual representations. Firstly, we use a loss term to have the student network mimic the teacher's cross-attention distribution. In this way, the adapted teacher is capable of transferring semantic knowledge to visual representations. Then, we propose a contrastive loss [12] to align features in these two different domains. Furthermore, a score consistency constraint is defined for teacher and student branch to align the learned feature representations. After training, the obtained atomic queries can independently extract semantic information from sports videos, without relying on labeled comments. Consequently, these semantic comments are not used during inference, which ensures the feasibility of our method in real-world scenarios.

We evaluate our proposed method on our OlympicFS and other two public datasets for scoring figure skating, *i.e.*, FS1000 [10] and Fis-v [1]. Moreover, we also conduct experiments on the MTL-AQA dataset [13], which is designed for diving and also annotates action quality scores and description of the dive. As a result, our method could outperform previous works, demonstrating its effectiveness. We hope our exploration will provide significant insights concerning knowledge transferring across different modalities to grasp a full understanding of sports videos. In short, our contributions are summarized as follows:

- We propose a teacher-student-based network SGN, which extracts semantic-aware representations to guide the training of visual features for scoring figure skating videos.
- Three auxiliary losses are proposed to align features in semantic and visual domains and transfer semantic information across these two feature spaces.
- We propose a new dataset OlympicFS, which is annotated with detailed comments for learning semantic representations in figure skating videos.
- Extensive experimental results on OlympicFS, FS1000, Fis-v, and MTL-AQA verify the effectiveness of our method.

II. Related Work

This section reviews closely related work on sports video analysis. Furthermore, we will discuss some literatures on video analysis of figure skating. Finally, we discuss the relevant methods based on multimodal representation learning.

A. Sports Video Analysis

Sports video analysis has recently been topical in the research communities. This technology has important applications in professional sports [14], [15], such as football [16], [17], basketball [18], volleyball [19], figure skating [1], [6], [7], [10] and other fields [20]–[22]. Using deep learning technology, computer vision systems can be trained to automatically identify different types of objects and actions in sports games and provide more accurate analysis and predictions for sports understanding.

Besides detecting and recognizing actions in sports videos, there are also a great number of works focusing on action quality assessment (AQA) [5], [13], [23]–[32]. Compared to action recognition [2]–[4], [33] focuses on correctly classifying the action sequences from different categories, AQA is more challenging as it requires dealinges with the videos from the same category with poor intra-class discriminant. The mainstream methods treat AQA as a regression task, relying on reliable score labels provided by expert judges. Early works [7], [34] in this field used support vector regression to perform regression, with input features consisting of either hand-crafted discrete cosine transform or deep C3D [35] features. There were also works using LSTM [1] and graph neural networks [36] to explore spatio-temporal correlations in videos. Parmar and Morris [13] introduced the concept of multi-task learning to enhance the model capacity for AQA. Tang *et al.* [30] presented a novel approach called uncertainty-aware score distribution learning, which aimed to address the inherent ambiguity in action score labels assigned by human judges. More recently, Yu *et al.* [24] developed a group-aware regression tree (CoRe) to replace the traditional score regression. Xu *et al.* [5] designed a Likert scoring paradigm to quantify the grades explicitly. Li *et al.* [23] proposed a pairwise contrastive learning network to focus on the subtle difference between videos. Bai *et al.* [27] introduced a temporal parsing transformer (TPT) to decompose the holistic feature into temporal part-level representations.

B. Figure Skating Analysis

In computer vision, the analysis of figure skating videos can be traced back to [7], which trained a regression model from spatio-temporal pose features to scores obtained from expert judges and gathered Olympic videos for action assessment. Similarly, Xu *et al.* [1] collected 500 figure skating videos from ladies single program for action scoring. They also developed an architecture, containing self-attentive LSTM and multi-scale LSTM, to learn the local and global sequential information in videos. Moreover, another fine-grained classification dataset FSD-10 was introduced in [6], which consisted of 10 different actions in men/ladies programs. For classification, they further proposed a key-frame-based temporal segment network. Additionally, several dedicated models have been proposed for figure skating analysis. Nakano *et al.* [37] detected the highlight in figure skating programs with people's reactions. ACTION-Net [38] learned the video dynamic information and static postures of the detected athletes in specific frames to strengthen the specific postures in videos.

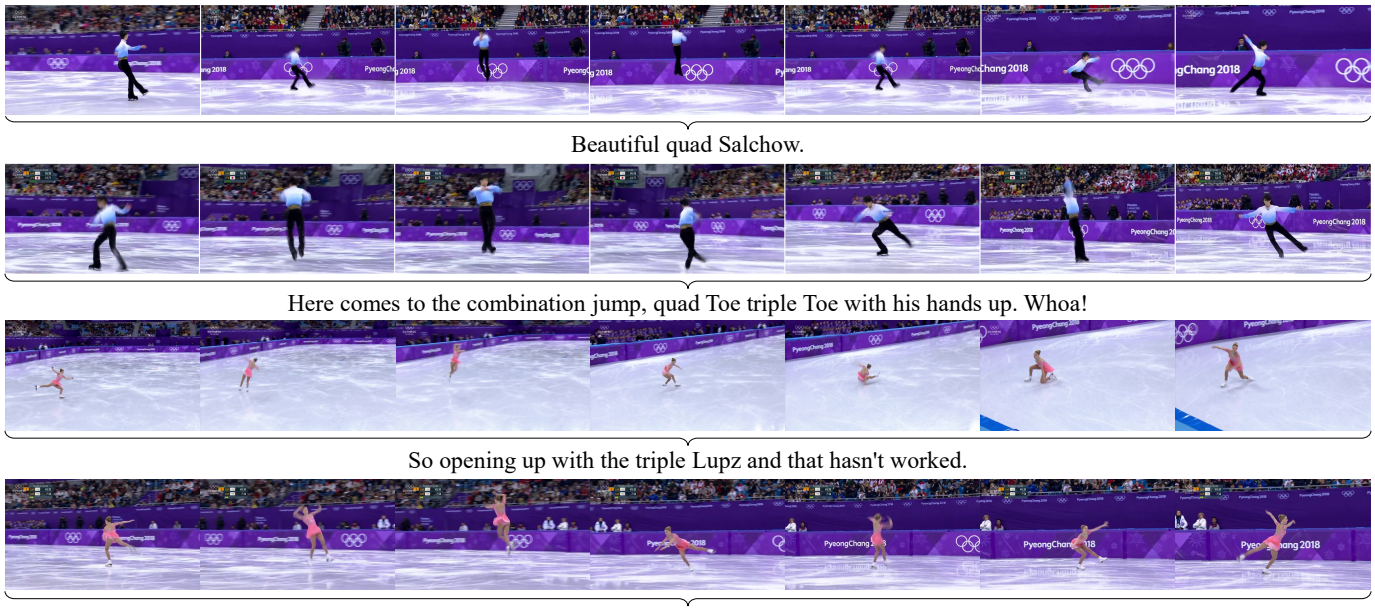


Fig. 2. Some examples of our OlympicFS dataset. Each row shows a complete figure skating action and we sample some important frames. We collect this dataset from Olympic Winter Games in Pyeongchang 2018 and Beijing 2022, containing men/ladies short program and men/ladies free skating. Besides actions scores, we also provide professional comments for each video.

EAGLE-Eye [39] built a two-stream network to reason about the coordination among the joints and appearance dynamics throughout the performance. More recently, Xia *et al.* [10] extended the MLP framework into a multimodal fashion MLP-Mixer and effectively learns long-term representations through the designed memory recurrent unit. Moreover, they also collected an audio-visual FS1000 dataset, containing over 1000 videos for scoring figure skating videos.

In summary, the major technical differences between previous methods and ours include the following three aspects. (1) We focus on a new problem, *i.e.*, learning semantics-guided representations for scoring figure skating, which is not explored in [10], [24], [27]. (2) CoRe [24] and TPT [27] rely only on visual features and do not delve into the semantic information in videos. Instead, we aggregate semantic and visual features jointly to generate semantic-aware representations. (3) Although MLP-Mixer [10] learns multimodal representations by modeling audio and visual features, one major difference is that we only utilize semantic features to guide the training of atomic queries. By learning semantics through atomic queries, we solely rely on video input during the inference.

C. Multimodal Learning

There exists a rich exploration in multimodal learning, especially in the deep learning era [40]–[43]. Not only is this task integral to advancing machine perception of our world where information often comes in different modalities, but also has important implications in fundamental research such as robotics, visual question answering, video captioning, and retrieval. More recently, Transformers [11] are prevalent in natural language processing and have also shown promising performance in computer vision [44]–[46]. Therefore, more and more works adopt Transformers architectures to predict contextualized latent representations from different views. While these approaches rely on large-scale datasets and employ multimodal self-supervision tasks for pretraining, we focus on transferring semantic knowledge to the visual domain to improve the understanding of figure skating actions. Some methods [47], [48] focus on aggregating multimodal information from pre-trained large-scale models. However, in the field of action assessment, there have not yet been related large-scale models developed. Therefore, our method and collected dataset could serve as a solid starting point for future research on multimodal learning in this domain.

III. OlympicFS Dataset

To further facilitate the study of learning semantic-aware representations for figure skating scoring, we collected a new dataset OlympicFS with high-quality videos. All of the videos used in this study were captured by professional camera devices during high-level competitions in Olympic Winter Games. The dataset is designed to predict scores in figure skating competitions, including rich annotations such as action scores, program categories and professional comments, which may further advance research in this field.

A. Data Collection

To construct the dataset, we searched and downloaded a large quantity of figure skating videos from professional, high-standard international skating competitions, including the Olympic Winter Games in Pyeongchang 2018 and Beijing 2022. Normally, OlympicFS consists of four categories: men/ladies short program and men/ladies free skating. We collected official video records of them from the Internet, ensuring these video records are complete, distinctive and of high-resolutions, *e.g.* 1280×720.

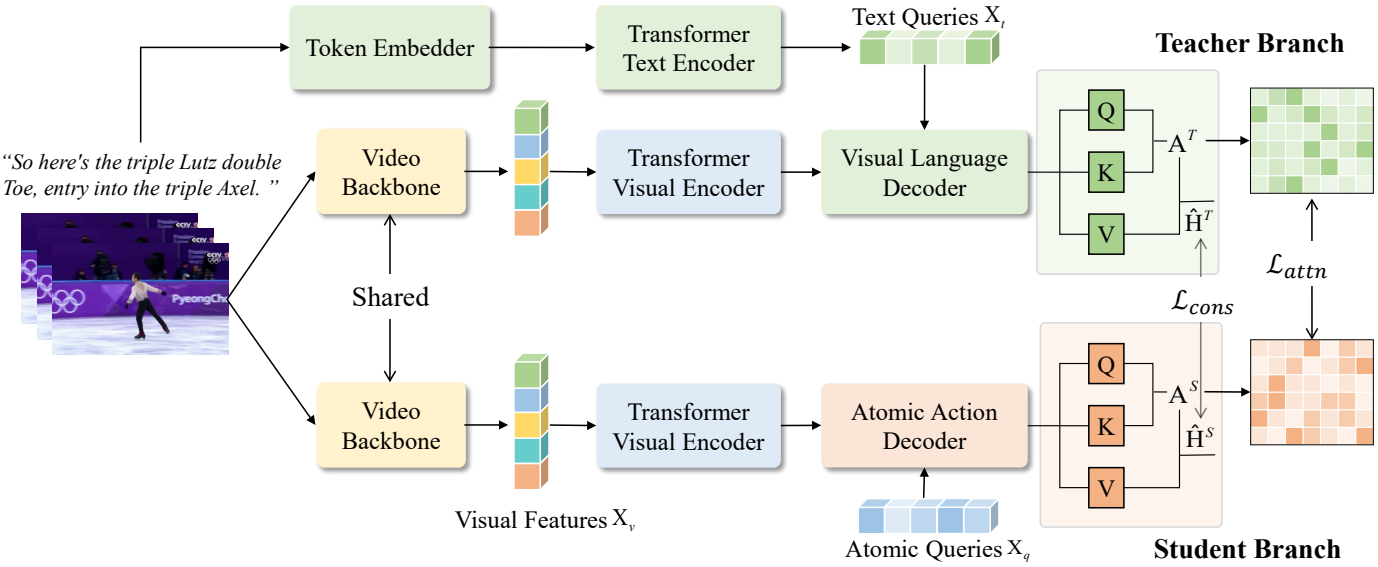


Fig. 3. The pipeline of our proposed method. Firstly, we extract video features using the pre-trained video backbone. Then, the teacher branch learns semantic-aware representations by constructing cross-attention between visual features and text queries. The student branch builds the attention mechanism using a set of learnable atomic queries, which is guided by the teacher's distributions. During inference, only the student branch is used with visual inputs, ensuring the feasibility of our method in real-world scenarios.

TABLE I
DATASET COMPARISON. FS REFERS TO THE LONG-TERM FIGURE SKATING TASK. V, T, AND A REPRESENT VISUAL, TEXT, AND AUDIO FEATURES.

Dataset	MTL-AQA [13]	Fis-v [1]	FS1000 [10]	OlympicFS
Field	Diving	FS	FS	FS
Feature	V+T	V	V+A	V+T

The raw videos collected from competitions are typically untrimmed and capture the entire procedure, including the performances of all players, highlight replays, warm-up parts, and waiting-for-score at the Kiss&Cry. However, these redundant parts may not be necessarily useful in judging the figure skating performance. And we aim to predict the figure skating scores from the competition performance of each player, instead of these "backgrounds". Therefore, we manually process all videos, reserving pure competition performance clips of players from the exact beginning to the ending moment of actions. Some clips of OlympicFS are shown in Fig. 2.

B. Annotations

After collecting videos, we carefully annotated each video with two scores, namely, Technical Element Score (TES) and Program Component Score (PCS). These scores are given by the mark scheme of the figure skating competition. The TES is calculated based on the difficulty and execution of the technical elements performed by the skater, such as jumps, spins, and step sequences. The PCS evaluates the overall performance of the skater in terms of their skating skills, performance/execution, choreography, interpretation, and musicality. Both the TES and PCS are given by different referees who are experts on figure skating competition.

In addition to scores, we further collected professional commentary during figure skating competitions as shown in TABLE I. Similarly, we only collected commentary during the skating process and did not use any post-scoring comments. To our best knowledge, our dataset is the first one to utilize the commentary feature in this area. Both score and commentary annotation stages adopt a cross-validating method. Specifically, we employ two workers who have prior knowledge in the figure skating domain and divide data into two parts without overlap. The annotation results of one worker are checked and adjusted by another, which ensures annotation results are double-checked. Under this pipeline, the total time of the whole annotation process is about 100 hours. The dataset will be released for further research purposes in this community.

IV. METHOD

Our proposed method is tailored for figure skating scoring involving multiple individuals. In Section IV-A, we would first show the details of feature extraction. Next, Section IV-B will shed light on the details of extracting semantic-aware representations. Then, we would elaborate on the structure of SGN in Section IV-C, which learns semantic information from the teacher branch. Moreover, the scoring loss will be introduced in Section IV-D.

A. Feature Extraction

The pipeline of our method is illustrated in Fig. 3. Given a long figure skating video, which usually has thousands of frames, we first follow [4] to divide the input video into T_v segments. Each segment contains multiple frames and we input these video segments into well-designed projection models [46] to extract visual features. Then, an MLP is applied for reducing the dimension of backbone features. Obtained feature sequences are denoted as $\mathrm{X}_v \in \mathbb{R}^{T_v \times D}$, where D is the feature dimension. For the text input, we extract features $\mathrm{X}_t \in \mathbb{R}^{T_t \times D}$ by a token embedder followed by a pre-trained Transformer [49].

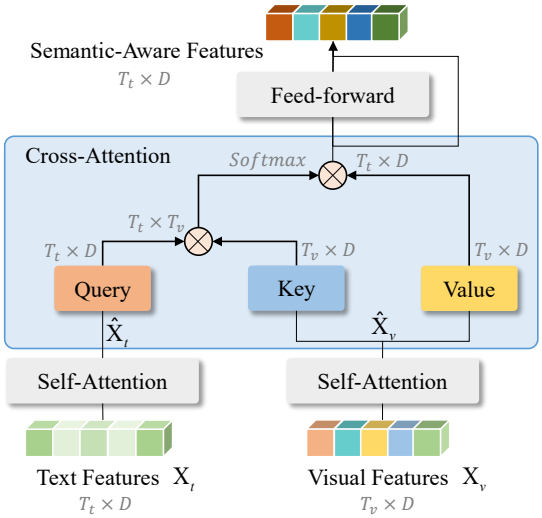


Fig. 4. The structure of the teacher branch, which learns semantic-aware representations by cross-attention. The shapes of important tensors are shown in gray. $\otimes$ denotes matrix multiplication. *Query*, *Key* and *Value* are three different linear projections.

B. Learning Semantic-Aware Representations

Since the features are independently extracted from the video sequence, each clip only contains information of current segment and lacks global context information. Therefore, we first use the self-attention encoder to enrich segment-wise representations. The self-attention mechanism involves the weighted aggregation of segment features to obtain the context information of each segment. The weights used in this process are determined by the correlations between the current segment and others,

$$\mathrm{H}_0 = Softmax\left(\frac{W_{qs}\mathrm{X}_v(W_{ks}\mathrm{X}_v)^\top}{\sqrt{D}}\right)W_{vs}\mathrm{X}_v + \mathrm{X}_v, \quad (1)$$

where W_{qs}, W_{ks} and W_{vs} are trainable matrices. Then, the H_0 is passed into a feed-forward network (FFN) for further fusion, which are represented as $\hat{\mathrm{X}}_v$. Fig. 4 also shows this process, where the FFN is omitted for simplicity.

In the teacher branch, we aim to extract semantic-aware representations from provided commentaries. For this purpose, we construct cross-attention between visual and text features to learn semantic correlations of them. Inspired by DETR [50], the Transformer decoder used in our model includes three parts, *i.e.,* self-attention, cross-attention and FFN, as illustrated in Fig. 4. Especially, the self-attention mechanism is applied for mining the relationship among text features. Similarly, we denote the updated text representations after the self-attention as $\hat{\mathrm{X}}_t$.

The context-aware representations are learned by cross-attention between the extracted $\hat{\mathrm{X}}_v$ and $\hat{\mathrm{X}}_t$. Firstly, the *query* is generated by $\hat{\mathrm{X}}_t$, while the *key* and *value* are transformed from $\hat{\mathrm{X}}_v$ via three different linear layers:

$$\mathrm{Q}_t = W_q\hat{\mathrm{X}}_t,\ \mathrm{K}_v = W_k\hat{\mathrm{X}}_v,\ \mathrm{V}_v = W_v\hat{\mathrm{X}}_v, \quad (2)$$

where W_q, W_k and W_v are the trainable weights. The semantic correlations between text and visual features are measured by the dot-product similarity between the corresponding *query-key* pair, which is formulated as

$$\mathrm{A}^T = Softmax\left(\frac{\mathrm{Q}_t\mathrm{K}_v^\top}{\sqrt{D}}\right), \quad (3)$$

where $\sqrt{D}$ serves as a scaling factor. $\mathrm{A}^T \in \mathbb{R}^{T_t \times T_v}$ shows how much the text features are related to the visual representations. Finally, the output of cross-attention is obtained by aggregating information between A^T and V_v, followed by FFN,

$$\mathrm{H}^T = \mathrm{FFN}(\mathrm{A}^T\mathrm{V}_v), \quad (4)$$

where $\mathrm{H}^T \in \mathbb{R}^{T_t \times D}$ is the output of the teacher branch.

C. Semantics-Guided Network

Unlike previous works [10], [13] that processed visual and semantic information together for scoring actions, we use semantic-aware representations to guide the learning of visual features. In detail, we first define a set of learnable atomic queries $\mathrm{X}_q \in \mathbb{R}^{K \times D}$ in the student branch, where K is the number of queries. These queries are used to represent the key semantic information for scoring, such as the *glorious jump* or *terrible fall* in figure skating. Then, the implementation details are defined analogously with the teacher branch while we replace the text features with these learnable queries, which also include self-attention, cross-attention and FFN. We denote the updated queries after the self-attention as $\hat{\mathrm{X}}_q$, and the following operations are

$$\mathrm{Q}'_q = W'_q\hat{\mathrm{X}}_q,\ \mathrm{K}'_v = W'_k\hat{\mathrm{X}}_v,\ \mathrm{V}'_v = W'_v\hat{\mathrm{X}}_v, \quad (5)$$

$$\mathrm{A}^S = Softmax\left(\frac{\mathrm{Q}'_q\mathrm{K}'^\top_v}{\sqrt{D}}\right), \quad (6)$$

$$\mathrm{H}^S = \mathrm{FFN}(\mathrm{A}^S\mathrm{V}'_v), \quad (7)$$

where $\mathrm{A}^S \in \mathbb{R}^{K \times T_v}$ and $\mathrm{H}^S \in \mathbb{R}^{K \times D}$ are the attention map and output of the student branch. After that, we use the semantic knowledge in the teacher branch to guide the learning of the student.

To learn semantic-aware representations, we first transfer the self-attention matrices between two branches to explore co-reference relationships between input tokens. The semantic knowledge is implicitly encoded and has promising potential for figure skating scoring. We formulate the distillation loss of the attention distribution by minimizing the divergence between the self-attention matrices of the teacher and the student, *i.e.,* A^T and A^S. However, $\mathrm{A}^T \in \mathbb{R}^{T_t \times T_v}$ and $\mathrm{A}^S \in \mathbb{R}^{K \times T_v}$ usually have different dimensions, where $T_t >> K$. Therefore, we apply *max-pooling* along the T_t/K dimension to generate $\hat{\mathrm{A}}^T/\hat{\mathrm{A}}^S$, which extracts the most salient feature and enhances the feature representation capability, without losing useful information. The distillation loss is therefore formulated as

$$\mathcal{L}_{attn} = \frac{1}{T_v h}\sum_{i=1}^{T_v}\sum_{j=1}^{h}\mathrm{MSE}(\hat{\mathrm{A}}^T_{i,j}, \hat{\mathrm{A}}^S_{i,j}), \quad (8)$$

where MSE is the mean squared error, h is the attention heads in the Transformer. $\hat{\mathrm{A}}_{i,j}$ is the normalized attention for i-th clip at j-th head.

Besides constraining the attention distribution of teacher and student branches, we also employ a contrastive loss [12] between the output features. Especially, the noise contrastive estimation (NCE) loss is used to align the teacher & student's feature representations by contrasting the target instance (H^S) with more negative samples and aligning with its positive sample (H^T). We use the *average-pooling* along the T_t/K dimension to map the student and teacher features to the identical dimension, *i.e.,* $\hat{\mathrm{H}}^T$ and $\hat{\mathrm{H}}^S$. And the objective loss is defined as

$$\mathcal{L}_{cons} = -\log \frac{\exp(\mathrm{sim}(\hat{\mathrm{H}}_i^S, \hat{\mathrm{H}}_i^T)/\tau)}{\sum_{j=1}^{N} \mathbb{1}_{[j\neq i]} \exp(\mathrm{sim}(\hat{\mathrm{H}}_i^S, \hat{\mathrm{H}}_j^T)/\tau)}, \quad (9)$$

where $\mathbb{1}_{[j\neq i]} \in \{0,1\}$ is an indicator function evaluating to 1 iff $j \neq i$, $\mathrm{sim}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/||\boldsymbol{u}||\,||\boldsymbol{v}||$ denotes the dot product between ℓ_2 normalized $\boldsymbol{u}$ and $\boldsymbol{v}$ (*i.e.,* cosine similarity), τ denotes the temperature hyper-parameter. The final loss is computed across all positive pairs in a mini-batch (N). Although contrastive learning loss has been employed in previous works [23], [24] for action assessment, we utilize it to learn semantic-aware representations from the teacher network, guaranteeing the consistency between the teacher & student's learned feature representations.

D. Figure Skating Scoring Loss

Here, we propose the loss function for figure skating scoring. A common solution is to tackle this problem as a regression task that maps the input video to the final score provided by referees. As most of the same sports events are competed in similar environment, the differences between the same competition videos are often very subtle, and there are slight differences in how the athletes perform the same actions. Based on this insight, we follow [23], [24] to reformulate the problem of action quality assessment as regressing the relative scores with reference to another video that has a shared category. We especially map the input video into the score space where the differences between the action qualities can be measured by the relative score. Therefore, for each video pair $\langle \mathrm{X}_{v,p}, \mathrm{X}_{v,q}\rangle$, $\langle \hat{S}_p, \hat{S}_q\rangle$ representing the ground truth of action quality score, it needs to minimize the error between the predicted relative score and the corresponding ground truth, which is defined as

$$\mathcal{L}_{score} = (\Delta S - |\hat{S}_p - \hat{S}_q|)^2, \quad (10)$$

where ΔS is based on the output of features of these two videos,

$$\Delta S = \mathcal{R}_\Theta(\hat{\mathrm{H}}_p, \hat{\mathrm{H}}_q), \quad (11)$$

where $\mathcal{R}_\Theta$ is the score regressor [24] parameterized by Θ. Our method has two branches, therefore, the relative score losses are calculated in two branches based on the corresponding extracted features, denoted as $\mathcal{L}_{score}^T$ and $\mathcal{L}_{score}^S$ for teacher and student branches respectively.

Furthermore, a score consistency constraint is defined for the teacher and student branches to align the learned feature representations. The consistency constraint confines the predicted relative score at the teacher branch is equal to the calculated score in the student branch. Therefore, a consistency loss function for the relative score is defined as

$$\mathcal{L}_{c\text{-}score} = (\Delta S^T - \Delta S^S)^2, \quad (12)$$

where ΔS^T and ΔS^S represent the predicted relative scores for teacher and student branches respectively.

Finally, the overall loss function of the proposed AQA model is summarized as

$$\mathcal{L}_{total} = \mathcal{L}_{score}^T + \mathcal{L}_{score}^S + \mathcal{L}_{attn} + \mathcal{L}_{cons} + \mathcal{L}_{c\text{-}score}, \quad (13)$$

where the weights for different parts are the same for simplicity.

In the testing phase, we only use the feature representations in the student branch, which are extracted by video backbone, Transformer visual encoder, and atomic action decoder, to predict the figure skating score. The teacher branch is only used during training to transfer semantic information, which guarantees efficiency and feasibility in a real-world deployment.

V. Experiment

A. Datasets and Implementation Details

1) Datasets: Besides our OlympicFS, the proposed method is also evaluated on FS1000 [10] and Fis-v [1] datasets for figure skating and MTL-AQA [13] dataset for diving to verify the effectiveness of our method.

FS1000 has a training set of 1000 videos and a validation set of 247 videos. There are totally eight categories of figure skating competitions in this dataset, namely, men/ladies/pairs short program, men/ladies/pairs free skating, and ice dance rhythm dance/free dance. Besides TES and PCS, five additional scores [10] are reported for FS1000 dataset, including Skating Skills (SS), Transitions (TR), Performance (PE), Composition (CO), and Interpretation of music (IN). **Fis-v** contains 400 videos for training and 100 for testing, which are trimmed from the ladies single short program. The TES and PCS are collected from the mark scheme of the figure skating competition. **MTL-AQA** comprises 1412 diving videos from 16 distinct events, featuring both male and female athletes performing on the 10m platform and 3m springboard. Each video is labeled with various metrics, including the final score, difficulty degree, and execution score assigned by the referees. As recommended by [13], we adopt a split configuration where 1059 videos are used for training and 353 videos are reserved for testing purposes. **OlympicFS** collects 200 (160 for training and 40 for testing) videos from Olympic Winter Games in Pyeongchang 2018 and Beijing 2022. It provides professional commentary in addition to scores to explore the impact of semantics on scoring performance.

2) Feature Extraction: As **FS1000** has visual and music inputs, we use the extracted features in [10] for a fair comparison. For **Fis-v**, which only contains video input without additional text or audio annotations, we fine-tune our trained student model on this dataset to show the effectiveness of our method. For **MTL-AQA**, we follow the settings in [24], [27] to make a fair comparison, which uses the I3D model [33] pre-trained on Kinetics as the backbone. In addition, the text

TABLE II
COMPARISON WITH STATE-OF-THE-ART ON FS1000. FOR MSE, THE LOWER THE BETTER; FOR SPEARMAN CORRELATION, THE HIGHER THE BETTER.

Methods	**Mean Square Error** (↓)							**Spearman Correlation** (↑)						
	TES	PCS	SS	TR	PE	CO	IN	TES	PCS	SS	TR	PE	CO	IN
C3D-LSTM [34]	308.30	25.85	0.92	0.99	1.21	0.97	1.01	0.78	0.53	0.50	0.52	0.52	0.57	0.47
MSCADC [28]	148.02	15.47	0.51	0.57	0.78	0.55	0.60	0.77	0.70	0.69	0.69	0.71	0.68	0.71
MS-LSTM [1]	94.55	11.03	0.45	0.49	0.76	0.43	0.47	0.86	0.80	0.77	0.78	0.76	0.79	0.78
M-BERT(Late) [51]	131.28	15.28	0.44	0.43	0.67	0.47	0.55	0.79	0.75	0.80	0.81	0.80	0.80	0.76
CoRe [24]	103.5	9.85	0.41	0.37	0.81	0.38	0.41	0.88	0.84	0.81	0.83	0.81	0.83	0.80
TPT [27]	80.00	8.88	0.34	0.37	0.63	0.34	0.39	0.88	0.83	0.82	0.82	0.81	0.82	0.81
MLP-Mixer [10]	81.24	9.47	0.35	0.35	0.62	0.37	0.39	0.88	0.82	0.80	0.81	0.80	0.81	0.81
Our SGN	**79.08**	**8.40**	**0.31**	**0.32**	**0.61**	**0.33**	**0.37**	**0.89**	**0.85**	**0.84**	**0.85**	**0.82**	**0.85**	**0.83**

feature of MTL-AQA dataset is extracted by the BERT model. We extract video features by Video Swin Transformer [46], and text features by BERT [49] for **OlympicFS**. The number of segments is set as $T_v = 200$ with 32 frames for each segment.

3) Evaluation Protocols: To make a fair comparison with previous works [1], [10], we adopt Spearman's rank correlation as an evaluation metric, which is defined as

$$\rho = \frac{\sum_i (p_i - p)(q_i - q)}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}}, \tag{14}$$

where p and q represent the ranking for each sample of two series respectively. Additionally, to give more insights into our model, the Mean Square Error (MSE) is also used to evaluate our model. Meanwhile, we further report the relative L2-distance (R-ℓ_2) [24] for MTL-AQA to measure the performance more precisely. Given the highest and lowest scores $s_{\max}$ and $s_{\min}$, R-ℓ_2 is defined as

$$\text{R-}\ell_2 = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{|s_k - \hat{s}_k|}{s_{\max} - s_{\min}} \right)^2, \tag{15}$$

where s_k and $\hat{s}_k$ represent the ground-truth and prediction scores for the kth sample. Spearman's correlation focuses more on the ranks of the predicted scores while MSE and R-ℓ_2 focus on the numerical values.

4) Implementation Details: For all experiments, we set the feature dimension $D = 128$. We adopt the Adam optimizer with the initial learning rate $1e$-3, and the weight decay is set to zero. We select ten exemplars for an input test video during inference and vote for the final score using the multi-exemplar voting strategy [24]. The attention layer in both student and teacher branches is set as one with four heads. We set the number of atomic queries in the student branch as four. We conduct experiments on a machine with two NVIDIA GeForce RTX 3090 GPUs and one 2.40GHz CPU.

B. Comparison with State-of-the-Art

1) FS1000: In TABLE II, we report the performance comparison to the state-of-the-art methods on FS1000 dataset, which include CNN-based [28], [34], LSTM-based [1], Transformer-based [51] and MLP-based [10] approaches. FS1000 dataset consists of different types of figure skating videos, which highly tests the robustness of the model. We replace text with audio, which is the same as [10]. Our SGN outperforms all previous methods. In particular, our method gets a lower MSE score and a higher Spearman correlation than MS-LSTM [1], which demonstrates extracting semantic-aware representations does work in understanding

TABLE III
COMPARISON WITH STATE-OF-THE-ART ON FIS-V DATASET. SP. CORR. IS SHORTED FOR SPEARMAN CORRELATION. "–" INDICATES NOT REPORTED.

Methods	**MSE** (↓)		**Sp. Corr.** (↑)	
	TES	PCS	TES	PCS
C3D-LSTM [34]	39.25	21.97	0.29	0.51
MSCADC [28]	25.93	11.94	0.50	0.61
MS-LSTM [1]	22.64	9.84	0.59	0.73
M-BERT (Late) [51]	27.73	12.38	0.53	0.72
GDLT [5]	–	–	0.69	0.82
CoRe [24]	23.50	9.25	0.66	0.82
TPT [27]	27.50	11.25	0.57	0.76
MLP-Mixer [10]	19.57	**7.96**	0.68	0.82
Our SGN	**19.05**	**7.96**	**0.70**	**0.83**

TABLE IV
COMPARISON WITH STATE-OF-THE-ART ON MTL-AQA DATASET. "W/O DD" MEANS THAT TRAINING AND TESTING PROCESSES DO NOT UTILIZE DIFFICULTY DEGREE LABELS, "W/ DD" MEANS EXPERIMENTS UTILIZING DIFFICULTY DEGREE LABELS.

Methods (w/o DD)	**Sp. Corr.** (↑)	R-ℓ_2($\times$100) (↓)	Year
Pose+DCT [7]	0.2682	–	2014
C3D-LSTM [34]	0.8489	–	2017
C3D-AVG-MTL [28]	0.9044	–	2019
I3D+MLP [30]	0.8921	0.707	2020
USDL [30]	0.9066	0.654	2020
MUSDL [30]	0.9158	0.609	2020
CoRe [24]	0.9341	0.365	2021
TSA-Net [52]	0.9422	–	2021
FSP [23]	0.8798	–	2022
ESP [23]	0.9230	–	2022
TPT [27]	0.9451	0.322	2022
Our SGN	**0.9500**	**0.274**	
Methods (w/ DD)	**Sp. Corr.** (↑)	R-ℓ_2($\times$100) (↓)	Year
USDL$_{DD}$ [30]	0.9231	0.468	2020
MUSDL [30]	0.9273	0.451	2020
I3D+MLP [24]	0.9381	0.394	2021
CoRe [24]	0.9512	0.260	2021
TPT [27]	**0.9607**	0.238	2022
Our SGN	**0.9607**	**0.232**	

TABLE V
COMPARISON WITH STATE-OF-THE-ART ON OLYMPICFS DATASET. WE EVALUATED THE PERFORMANCE WITH THE PUBLICLY AVAILABLE SOURCE CODE. # PARAMS REPRESENTS THE NUMBER OF PARAMETERS.

Methods	MSE (↓)		Sp. Corr. (↑)		# PARAMS
	TES	PCS	TES	PCS	
MS-LSTM [1]	212.23	214.00	0.8085	0.7916	2.66M
GDLT [5]	204.89	216.38	0.8213	0.8240	3.16M
MLP-Mixer [10]	240.50	251.82	0.8008	0.8233	5.65M
TPT [27]	138.92	65.41	0.8880	0.8928	15.76M
CoRe [24]	169.51	59.51	0.8912	0.9034	2.05M
Our SGN	**104.29**	**57.86**	**0.9088**	**0.9230**	1.47M

TABLE VI
ABLATION STUDIES ON THE MODEL COMPONENTS. EXPERIMENTS ARE CONDUCTED ON OLYMPICFS.

Methods	MSE (↓)		Sp. Corr. (↑)	
	TES	PCS	TES	PCS
Baseline	185.51	79.20	0.8617	0.8593
Encoder only	169.21	65.06	0.8780	0.8694
Student branch	152.73	62.27	0.8879	0.8774
Two branches	122.61	59.65	0.8934	0.9021
SGN	104.29	57.86	0.9088	0.9230

figure skating actions better. Meanwhile, we still achieve better performance when compared to the strong MLP-Mixer [10].

2) Fis-v: The performance by different models in terms of Spearman correlation and MSE are demonstrated in TABLE III. We evaluate the performance with respect to TES and PCS as in previous works. It is observed that our method achieves comparable or better performance than existing methods. These experimental results verify our analysis that guided by the semantic information, the student branch could also obtain semantic-aware representations. Moreover, our finding certainly opens a door for bringing multi-modality representation learning in video sports understanding.

3) MTL-AQA: Besides datasets for figure skating, we also conduct experiments on the MTL-AQA dataset, which also includes multi-modality inputs (*i.e.,* visual and text) and is designed for scoring diving actions. We summarize the performance of our method on MTL-AQA in TABLE IV. Since the MTL-AQA dataset includes degree of difficulty (DD) annotations for diving actions, we also examine the impact of DD on this dataset. We categorize all methods into two groups: those that utilize the DD labels during the training phase (bottom section of the table) and those that do not (upper section of the table) as [27]. The experimental results show that the proposed model achieves the best Spearman correlation and R-ℓ_2 regardless of whether or not DD is used. These results indicate that exploiting semantic information is conducive to understanding sports videos better.

4) OlympicFS: We summarize the performance of different methods in TABLE V. Besides our method, we also perform experiments on recently proposed approaches [1], [5], [10], [24], [27] with the publicly available source code. All training recipes are kept the same for a fair comparison. All methods use the same feature backbone [46], which is frozen during training. For [10], we replace its audio information with our textual information. It is observed that regardless of the type of score, our framework delivers better results than others. Especially, our method has a higher Spearman's rank correlation and lower MSE than CoRe [24], indicating that exploring semantic-aware representations does work in scoring figure skating.

For complexity, our method needs fewer trainable parameters compared to previous approaches. Note that the complexity of the backbone and the score regressor is not included when calculating the parameters. In summary, the proposed method achieves the best trade-off between accuracy and model complexity.

TABLE VII
ABLATION STUDIES ON LOSS FUNCTIONS. THE METRIC MSE IS REPORTED.

	$\mathcal{L}_{score}$	$\mathcal{L}_{attn}$	$\mathcal{L}_{cons}$	$\mathcal{L}_{c\text{-}cons}$	TES	PCS
1	✓	✗	✗	✗	152.73	62.27
2	✓	✓	✗	✗	140.14	61.73
3	✓	✓	✓	✗	122.61	59.61
4	✓	✗	✓	✓	125.55	60.64
5	✓	✓	✓	✓	104.29	57.86

C. Ablation Studies

1) Different Model Components: In this section, we perform a set of ablation studies to evaluate the effectiveness of our proposed model components and designs. All experiments use [46] for feature extraction. In detail, we mainly analyze the following models:

- **Baseline:** It directly pools the video features without Transformer and uses the regressor [24] for scoring.
- **Encoder only:** The visual encoder is used for feature extraction. The atomic queries are not used here.
- **Student branch:** We define a set of atomic queries in the student branch. The Transformer visual encoder and atomic action decoder are also used. Only scoring loss is used as supervision.
- **Two branches:** Besides the student branch, the teacher is also used, but the fusion is made by the add operation.
- **SGN:** Our proposed method in Section IV.

Experimental results are shown in TABLE VI. Firstly, it is observed that by using the Transformer, more useful features could be extracted, leading to improved scoring accuracy. The performance improvement achieved (from "Encoder only" to "Student branch") by using the atomic queries validates the effectiveness of atomic queries. Such an improvement is further enlarged when introducing both text and video clues in figure skating (*i.e.,* two branches), which further demonstrates that semantic learning is really important in this field. Furthermore, a considerable boost is also achieved when using our SGN, which agrees with our analysis that compared to directly adding semantic and text features, SGN could more effectively explore the semantic correlations in sports videos.

2) Loss Function: TABLE VII shows the ablation studies on the loss functions in Section IV-D. Note that the results generated by only using $\mathcal{L}_{score}$ (the 1st row) are the same

TABLE VIII
ABLATION STUDIES ON THE HYPER-PARAMETERS, INCLUDING THE ATTENTION LAYERS (l) AND ATOMIC QUERIES (n).

	MSE (↓)		**Sp. Corr.** (↑)	
	TES	PCS	TES	PCS
$l = 1$	104.29	57.86	0.9088	0.9230
$l = 2$	123.27	56.28	0.8965	0.9261
$l = 3$	128.69	69.89	0.8843	0.8994
$n = 4$	104.29	57.86	0.9088	0.9230
$n = 8$	121.98	58.44	0.9033	0.9234
$n = 12$	110.61	59.39	0.9108	0.9184

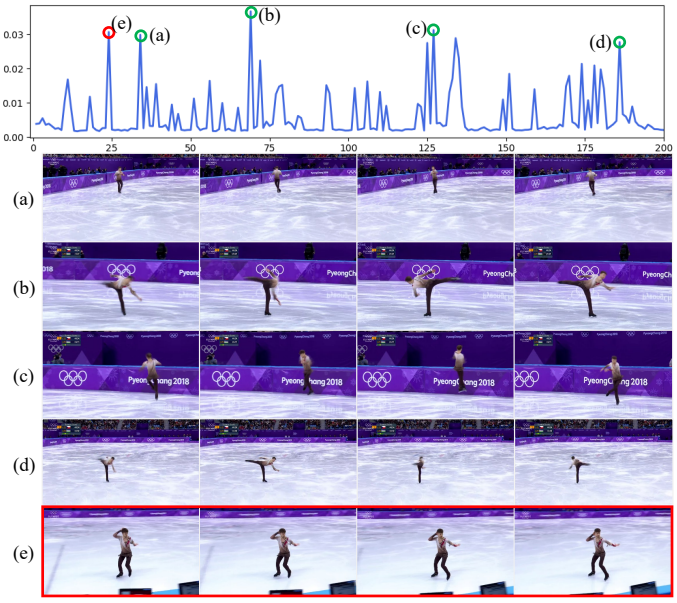


Fig. 5. Visualization of cross-attention weights in the student branch during testing. The first row shows the weight curve of atomic queries on videos. The next five rows are video segments corresponding to five markers on the curves, *i.e.,* (a), (b), (c), (d), and (e).

as the student branch in TABLE VI. By mimicking the distribution of attention (the 2nd row), improvements are achieved compared to the baseline, which demonstrate the effectiveness of attention distillation. The feature ($\mathcal{L}_{cons}$) and score ($\mathcal{L}_{c\text{-}cons}$) consistency constraints in the 3rd and 5th rows further verify the efficacy of the alignment schema. In the end, the combination of all the loss terms gives the best performance, confirming that our proposed attention distillation loss, hidden embedding contrastive loss and scoring consistency constraint are complementary to each other.

3) Hyper-parameters: This section will study the effect of some important hyper-parameters, which include the attention layers and atomic queries. Firstly, it is observed in TABLE VIII (upper) that using more attention layers would import negligible or no improvement in performance. Considering the effectiveness and efficiency, we use one attention layer for all experiments. In addition, we summarize the effect of queries in TABLE VIII (lower). It is observed that there is no significant improvement when increasing n. We conjecture it is because too many queries may bring ambiguity to the model. Therefore, the number of queries is set to four.

D. Visualization

Fig. 5 shows the cross-attention weights computed by Eq. (6) of atomic queries on a video sequence in the student branch during testing. The different fluctuations in the weight curve demonstrate different attention patterns. It is observed that our method pays more attention to the important moments in the video (such as *spin* and *jump*), which verifies our analysis that our method could extract semantic-aware representations in figure skating. In Fig. 5(e), it is observed that the weight of cross-attention is relatively high, but there are no crucial actions in this timestamp. By analyzing the preceding sequence, we find that this timestamp corresponds to the end of a *step sequence*. We conjecture that it is because the commentator's narration occurs after the *step sequence*.

VI. CONCLUSION

We have proposed an effective teacher-student network to learn semantics-guided representations for scoring figure skating. Firstly, we define a set of atomic queries to mimic the attention distribution in the teacher branch, where the teacher branch uses visual and text inputs to learn semantic-aware representations. In addition, we also propose three auxiliary losses to align features in two branches. Experimental results on public (Fis-v, FS1000, and MTL-AQA) and newly collected (OlympicFS) datasets verify the effectiveness and efficiency of our method.

REFERENCES

[1] 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, 2019.
[2] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6202–6211.
[3] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.
[4] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*, 2016, pp. 20–36.
[5] A. Xu, L.-A. Zeng, and W.-S. Zheng, "Likert scoring with grade decoupling for long-term action assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3232–3241.
[6] S. Liu, X. Liu, G. Huang, H. Qiao, L. Hu, D. Jiang, A. Zhang, Y. Liu, and G. Guo, "Fsd-10: A fine-grained classification dataset for figure skating," *Neurocomputing*, vol. 413, pp. 360–367, 2020.
[7] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *European Conference on Computer Vision*. Springer, 2014, pp. 556–571.
[8] B. J. Devereux, A. Clarke, and L. K. Tyler, "Integrated deep visual and semantic attractor neural networks predict fmri pattern-information along the ventral object processing pathway," *Scientific reports*, vol. 8, no. 1, p. 10636, 2018.
[9] A. Clarke and L. K. Tyler, "Understanding what we see: how we derive meaning from vision," *Trends in cognitive sciences*, vol. 19, no. 11, pp. 677–687, 2015.
[10] J. Xia, M. Zhuge, T. Geng, S. Fan, Y. Wei, Z. He, and F. Zheng, "Skating-mixer: Multimodal mlp for scoring figure skating," *arXiv preprint arXiv:2203.03990*, 2022.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[13] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 304–313.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[15] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 961–970.

[16] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2018, pp. 1711–1721.

[17] A. Deliege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4508–4519.

[18] Yan, Rui and Xie, Lingxi and Tang, Jinhui and Shu, Xiangbo and Tian, Qi, "Social adaptive module for weakly-supervised group activity recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 208–224.

[19] H. Yuan, D. Ni, and M. Wang, "Spatio-temporal dynamic inference network for group activity recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7476–7485.

[20] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, pp. 1–25, 2022.

[21] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011, 2018.

[22] M. Merler, K.-N. C. Mac, D. Joshi, Q.-B. Nguyen, S. Hammer, J. Kent, J. Xiong, M. N. Do, J. R. Smith, and R. S. Feris, "Automatic curation of sports highlights using multimodal excitement features," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1147–1160, 2019.

[23] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *European Conference on Computer Vision*. Springer, 2022, pp. 457–473.

[24] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7919–7928.

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2949–2958.

[26] J. Gao, J.-H. Pan, S.-J. Zhang, and W.-S. Zheng, "Automatic modelling for interactive action assessment," *International Journal of Computer Vision*, vol. 131, no. 3, pp. 659–679, 2023.

[27] 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 *European Conference on Computer Vision*. Springer, 2022, pp. 422–438.

[28] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1468–1476.

[29] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Towards unified surgical skill assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9522–9531.

[30] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9839–9848.

[31] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6057–6066.

[32] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7862–7871.

[33] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[34] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[36] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6331–6340.

[37] T. Nakano, A. Sakata, and A. Kishimoto, "Estimating blink probability for highlight detection in figure skating videos," 2020.

[38] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2526–2534.

[39] M. Nekoui, F. O. T. Cruz, and L. Cheng, "Eagle-eye: Extreme-pose action grader using detail bird's-eye view," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 394–402.

[40] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 166–18 176.

[41] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, "Compressing visual-linguistic model via knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1428–1438.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[43] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," *arXiv preprint arXiv:2302.14115*, 2023.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[45] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[46] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3202–3211.

[47] Z. Ma, G. Luo, J. Gao, L. Li, Y. Chen, S. Wang, C. Zhang, and W. Hu, "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 074–14 083.

[48] M. Wu, J. Gu, Y. Shen, M. Lin, C. Chen, and X. Sun, "End-to-end zero-shot hoi detection via vision and language knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2839–2846.

[49] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

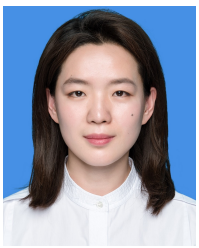
[50] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[51] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, and Y. Song, "Parameter efficient multimodal transformers for video representation learning," in *International Conference on Learning Representations*.

[52] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4902–4910.

Zexing Du is a Ph.D. candidate at the School of Computer Science, Northwestern Polytechnical University. His research interests include video analysis, action understanding, and multi-modal representation.

Di He received the B.S. degree from Shaanxi Normal University in 2021. She is currently a postgraduate with the School of Computer Science, Northwestern Polytechnical University. Her research interests are visual action interaction and group activity recognition.



Xue Wang received the B.S. and Ph.D. degrees from Northwestern Polytechnical University in 2007 and 2017, respectively. From 2012 to 2014, she studied at the University of Pennsylvania as a Visiting Ph.D. Student financed by the China Scholarship Council. She is currently an Associate Research Fellow with the School of Computer Science, Northwestern Polytechnical University. She focuses on building machines that understand the social signals and events that multiview/light field videos portray. Her research interests include computer vision, computational photography, and machine learning.

Qing Wang (Senior Member, IEEE) graduated from the Department of Mathematics, Peking University, in 1991. He received the Ph.D. degree from the Department of Computer Science, Northwestern Polytechnical University in 2000. He is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. He worked as a Research Scientist at the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, from 1999 to 2002. He also worked as a Visiting Scholar at the School of Information Engineering, The University of Sydney, Australia, in 2003 and 2004. In 2009 and 2012, he visited the Human-Computer Interaction Institute, Carnegie Mellon University, for six months, and the Department of Computer Science, University of Delaware, for one month. He has published more than 180 papers in international journals and conferences. His research interests include computer vision and computational photography, such as 3D vision, light field imaging and processing, and novel view synthesis. He is also a member of ACM. In 2006, he was awarded the Outstanding Talent Program of the New Century by the Ministry of Education, China.