

Semi-Supervised Action Quality Assessment with Self-Supervised Segment Feature Recovery

Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng

Abstract—Action Quality Assessment aims to evaluate how well an action performs. Existing methods have achieved remarkable progress on fully-supervised action assessment. However, in real-world applications, with expert’s experience, it is not always feasible to manually label all samples. Therefore, it is important to study the problem of semi-supervised action assessment with only a small amount of samples annotated. A major challenge for semi-supervised action assessment is how to exploit the temporal pattern from unlabeled videos. Inspired by the temporal dependencies of the action execution, we propose a self-supervised learning on the unlabeled videos by recovering the feature of a masked segment of an unlabeled video. Furthermore, we leverage adversarial learning to align the representation distribution of the labeled and the unlabeled samples to close their gap in the sample space since unlabeled samples always come from unseen actions. Finally, we propose an adversarial self-supervised framework for semi-supervised action quality assessment. The extensive experimental results on the MTL-AQA and the Rhythmic Gymnastics datasets will demonstrate the effectiveness of our framework, achieving the state-of-the-art performances of semi-supervised action quality assessment.

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

I. INTRODUCTION

THE purpose of the Action Quality Assessment (AQA) is to evaluate how well an action is performed in a video. While related to action recognition, action assessment is different from action recognition in the aspect that action assessment aims to tell how well an action is performed and where to pay attention, rather than classifying samples by their action types. Action assessment has good applications potential in various real-world scenarios. For example, in medical rehabilitation, the patients can complete the physical training with the feedback of an action assessment system [1], [2]. In sport, an action assessment model can help athletes achieve better training effects [3], [4], [5], [6]. Additionally, medical students can train their surgical skills with the help of an action assessment model [4], [5].

Recently, many existing works have achieved remarkable progress on fully-supervised action assessment in videos [3], [7], [4], [5], [8], [6]. These methods rely heavily on human annotations that are costly to obtain. However, in real-scene

Shao-Jie Zhang, Jia-Hui Pan and Jibin Gao are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China. E-mail: {zhangshj56, panjh7, gaojb5}@mail2.sysu.edu.cn.

Wei-Shi Zheng is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China, with Peng Cheng Laboratory, Shenzhen 518005, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China. E-mail: wszheng@ieee.org /zhwshi@mail.sysu.edu.cn. (Corresponding author: Wei-Shi Zheng)

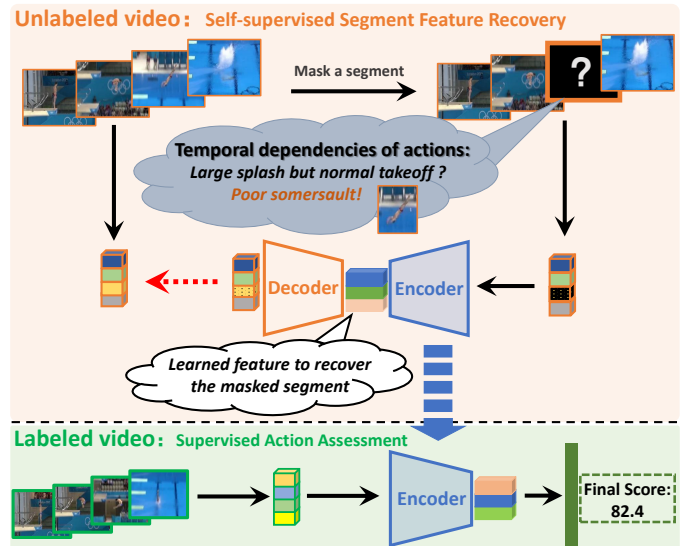


Fig. 1. Our semi-supervised learning for action assessment. Inspired by the temporal dependencies in action executions, we exploit the unlabeled videos by self-supervised segment feature recovery, that is, predicting the feature of a masked video segment. Meanwhile, the labeled videos are trained with their score annotations for action assessment with the help of the same encoder which is utilized to learn implicit feature for masked video segment recovery.

applications, for a tremendous amount of action samples, the assessment annotations of the action performance are difficult to obtain, since it is not practical to ask experts, who are in small numbers for a specific field, to annotate every sample. Therefore, it is an important challenge to develop a model that can learn to assess the performance of human action with only a small amount of samples annotated.

However, there is few relevant work that explores the semi-supervised action quality assessment. While it is common to address a major challenge of how to leverage the unlabeled samples, semi-supervised action quality assessment has its own characteristics for the semi-supervised learning. Firstly, the temporal dependencies of the unlabeled videos can be exploited for action assessment, since the action performance of a stage influences that of other stages. For example, in diving, if the athlete makes a slight mistake in the take-off posture, it will possibly lead to flaws in the subsequent stages and a bigger splash in the end. Similarly, if a dive ends with a large splash but starts with a good take-off posture, it is very likely that a mistake was made in the somersaults. The temporal dependencies are the key to understanding the execution of an action and how to learn such temporal dependencies remains a major challenge. Secondly, for utilizing

the unlabeled videos, an intuitive way is to obtain intermediate representations of labeled videos with the same encoder as the unlabeled videos, and predict the performance scores with an assessment module. However, different optimization objectives and sampling deviation between the labeled and the unlabeled data may lead to a representation misalignment in the feature space.

In this work, we form the Self-Supervised Semi-Supervised Action Quality Assessment (S^4 AQA), which overcomes the challenges mentioned above for semi-supervised action quality assessment. Firstly, in our S^4 AQA, we first exploit the temporal dependencies of the action execution in unlabeled videos for action assessment by means of a self-supervised segment feature recovery learning, as shown in Figure 1. For each unlabeled sample, we randomly mask a segment and try to recover the feature of the segment with an encoder and a decoder. In this way, we learn intermediate representations to depict the action execution for action assessment in videos. Secondly, we further propose an adversarial training mechanism to align the distribution of the representations of the labeled and the unlabeled videos to close their gap in the sample space since unlabeled samples always come from unseen actions. Therefore, our S^4 AQA consists of three modules: a masked segment feature recovery module to learn representations of unlabeled videos, an action assessment module to learn representations of labeled videos, and a representation distribution alignment module to align the feature distributions of the labeled and the unlabeled data. These three modules are jointly trained to achieve semi-supervised action quality assessment.

Compared to existing works, to the best of our knowledge, we are the first to explore the semi-supervised action quality assessment. In summary, the characteristics of our work are:

- 1) We propose to explore semi-supervised action assessment which aims at learning action assessment with only a small amount of labeled data.
- 2) We propose to exploit the temporal dependencies of the action execution on the unlabeled videos for action assessment by self-supervised masked segment feature recovery.

Experiments conducted on the MTL-AQA dataset and the Rhythmic Gymnastics dataset clearly demonstrate the effectiveness of our model, and we establish the state-of-the-art performances of semi-supervised action quality assessment.

II. RELATED WORK

A. Action Assessment

Some recent works on video analysis attempt action assessment [9], [3], [7], [4], [5], [6], [8], [10]. Action assessment is related to the research topic of action recognition [11], [12], [13], [14], [15], but they have different purposes. Action recognition focuses on classifying samples into varied action types, while action assessment focuses on evaluating how well an action is performed.

The existing methods on action assessment are mainly divided into three categories based on the problem formulation, namely, the classification-based, the pairwise-comparison-based and the regression-based methods. The classification-based methods [16], [17] try to divide the action videos

into several performance levels (*e.g.*, novice, intermediate and expert). The pairwise-comparison-based [18], [10] formulate action assessment as pairwise comparisons among samples. Jain *et al.* [19] propose a metric learning way to compare the given action video with a reference video to get the final score of action, which focus on the efficient utilization of score annotations. In comparison, our method focuses on leveraging the unlabeled data to improve the performance of the AQA model in the semi-supervised learning way. Besides, the regression-based methods [3], [7], [8], [4], [5], [6] assess the quality of actions by predicting the performance score for each input video. In order to predict the fine-grained scores of samples and attain the relative performance among all samples, this work follows the regression-based setting.

Several works focus on regression-based action assessment. An early work [9] proposes to predict the scores of the Olympic actions with discrete cosine transformation (DCT) to extract motion feature and support vector regression (SVR) to predict the performance scores. Then Parmar *et al.* [3] develop a model with 3d convolutional network and long short-term memory (LSTM) for action assessment which directly consumes RGB videos. After that, Pan *et al.* [4] and Gao *et al.* [5] model the interactions of human joints and the moving agents for action assessment, respectively. Zeng *et al.* [6] try to address action assessment in long videos. Tang *et al.* [8] propose to formulate action assessment as a distribution learning task considering the ambiguity in action assessment. Parmar *et al.* propose another work [7] to address action assessment in a multi-task training manner with video captioning and fine-grained action recognition. Parmar *et al.* [20] propose a feature prediction task under each segment to distill the spatial-temporal knowledge of a video segment from a pretrained 3D-CNN to a lighter-weight 2D-CNN that only observes the first frame, which reduces the memory and compute cost with a light-weight student 2D-CNN model without excessive performance degradation. In comparison with this work, we focus on mitigating the shortage of action videos with score annotations in AQA, and the target of our work is to learn the spatial-temporal representation from unlabeled videos to improve the AQA performance under the limited labeled videos.

However, the existing methods only explore fully-supervised action assessment that relies on heavy annotation effort, and none has attempted non-fully-supervised action assessment. This work instead, attempts semi-supervised action assessment in which only some videos in the training set are labeled.

B. Semi-supervised Learning

Many semi-supervised learning works have been proposed in various domains [21], [22], [23], [24], [25]. Recently, there is a growing interest in training deep neural networks in semi-supervised learning. For a more comprehensive survey, we refer interested readers to Ouali *et al.*'s review [26] about semi-supervised learning.

These works can be coarsely cast into several groups, including pseudo-label-based approaches [21], [27], [28], gen-

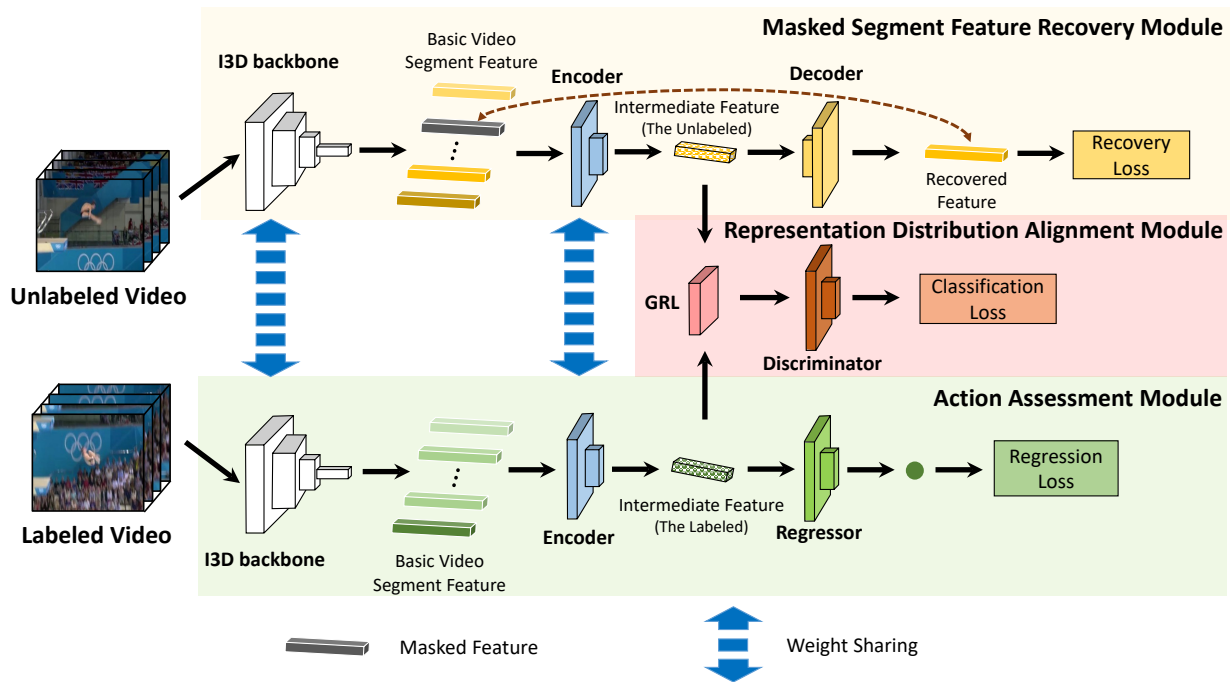


Fig. 2. The framework of our self-supervised semi-supervised action quality assessment (S^4AQA). The S^4AQA couples the self-supervised masked segment feature recovery task with the supervised score regression task in a multi-task learning way. Then, the representation distribution alignment module with a gradient reverse layer is used to align the representation distributions of the labeled and the unlabeled videos in an adversarial training way. GRL denotes the gradient reverse layer.

erative models [29], [30], and consistency regularization methods [31], [22], [23], *etc.*

For the pseudo-label-based approaches, self-training [21] produces pseudo labels for unlabeled data by the model trained on labeled data, and the multi-view training approaches [32], [33] produce pseudo labels by multiple models trained on different views of the labeled data. In [28], Jia *et al.* exploit pseudo-labels of the unlabeled data for semi-supervised action recognition. However, the Pseudo-labels method cannot correct the error in pseudo labels, which hinders the performance of semi-supervised learning. In comparison with this, we construct a masked segment feature recovery self-supervised task on the unlabeled action videos based on the temporal dependencies of action execution, which is more suitable for semi-supervised action assessment. Differently, consistency regularization is utilized to construct a corresponding semi-supervised learning (*e.g.*, II-Model [31], Mean Teacher [22], Virtual Adversarial Training [23] and Unsupervised Data Augmentation [34]). The basis of these methods is that the concept that the prediction results should not change significantly if a small perturbation is applied to unlabeled samples. Another group of semi-supervised approach is based on generative models, *e.g.*, variational autoencoders [29] and generative adversarial network [30].

While there has been remarkable progress in semi-supervised image analysis (image classification [31], [22], age estimation [35], *etc.*), semi-supervised learning for AQA is still a novel problem which has been rarely explored. In this work, we explore the problem of semi-supervised action assessment. We leverage the self-supervised task for masked

segment feature recovery to learn discriminative features from unlabeled videos to assist the assessment model.

C. Self-supervised Learning

Self-supervised representation learning has shown its power in a broad range of computer vision tasks [36], [37]. Huang *et al.* [37] propose the sampling rate regression and video segment order prediction pretext task for the self-supervised video representation learning while our method extends the self-supervised learning to the semi-supervised action assessment by establishing the mask segment feature recovery pretext task on the feature level and makes full use of unlabeled data to assist labeled data in a joint training way. Some recent works also leverage the self-supervised learning in the semi-supervised tasks. In [24], Zhai *et al.* propose a method called self-supervised semi-supervised learning which leverages the image rotation prediction and image transformations pretext tasks to learn representations from unlabeled images. The self-supervised tasks of skeleton inpainting and neighborhood consistency modelling [25] are leveraged to learn discriminative representation from unlabeled 3D skeleton data. These semi-supervised tasks are constructed on the images or low-dimensional coordinate positions of key joints which are not suitable for action quality assessment to learn discriminative action representations from videos. In AQA tasks, there are temporal dependencies between the performance of different stages of an action. Inspired by such particularity of the action videos, we propose the self-supervised masked segment feature recovery pretext task by recovering a masked video segment feature to learn implicit feature of action performance.

Moreover, the pretext task constructed on embedding space has more discriminative and less uncertain information than that constructed on low-level video frames, since the actions with similar performance can be encoded into feature vectors with closer distance in embedding space while there may be a big difference on the RGB values in video frames.

III. METHOD

Existing action quality assessment methods rely heavily on human annotations that are not easily acquired, neglecting the potential of a mass of unlabeled videos. Therefore, we explore semi-supervised action assessment and propose a self-supervised semi-supervised action quality assessment framework (S⁴AQA), to address this challenge. As shown in Figure 2, our S⁴AQA consists of three modules: a masked segment feature recovery module to learn representations of unlabeled videos, an action assessment module to learn representations of labeled videos and a representation distribution alignment module to align the feature distributions of the labeled and the unlabeled data. The three modules are jointly trained for semi-supervised action assessment.

A. Problem Formulation

Different from fully-supervised action assessment, in semi-supervised action assessment, only a small amount of training samples are annotated. The training samples are divided into two subsets: the labeled subset and the unlabeled subset, with the sizes of L and U , respectively. For each video, we divide it into T segments and obtain the basic video features with the Inflated 3D ConvNets (I3D) backbone [38] as in previous works [4], [5], [8]. Therefore, the basic features of a labeled video and an unlabeled video are written as $V^l = [v_1^l, v_2^l, \dots, v_t^l, \dots, v_T^l]$ and $V^u = [v_1^u, v_2^u, \dots, v_t^u, \dots, v_T^u]$, respectively. Our goal is to learn to assess the quality of the action performance by exploiting both the labeled and the unlabeled samples.

B. Learning Action Assessment on the Labeled Videos

After obtaining basic features of the videos, we learn to predict the performance scores on the labeled data. We first use an encoder $E(\cdot)$ to obtain the intermediate feature f^l (or f^u) of a video V^l (or V^u). The encoding process is written as

$$\begin{aligned} f^l &= E([v_1^l, v_2^l, \dots, v_t^l, \dots, v_T^l]), \quad l = 1, 2, \dots, L, \\ f^u &= E([v_1^u, v_2^u, \dots, v_t^u, \dots, v_T^u]), \quad u = 1, 2, \dots, U, \end{aligned} \quad (1)$$

where v_t^l and v_t^u are the basic video features of the t -th segment of the l -th labeled sample and the u -th unlabeled sample, respectively. L and U are the total numbers of labeled and unlabeled training samples, respectively. $E(\cdot)$ is an encoder that consists of three one-dimensional temporal convolution layers with ReLU activation functions, batch-normalization layers [39] and dropout layers, respectively.

After obtaining the intermediate representations, we predict an action performance score for the labeled samples, which is written as

$$\hat{y}^l = R(f^l), \quad l = 1, 2, \dots, L, \quad (2)$$

where f^l is the intermediate feature for the l -th labeled sample, and \hat{y}^l is the predicted score. $R(\cdot)$ is our regression module that consists of two fully connected layers with a ReLU activation.

To perform supervised assessment training on the labeled samples, we employ the Mean-Squared Error (MSE) loss, which is written as

$$\mathcal{L}_{reg} = \frac{1}{L} \sum_{l=1}^L (y^l - \hat{y}^l)^2, \quad (3)$$

where y^l and \hat{y}^l represent the ground-truth score and the predicted score, respectively. L is the total number of labeled training samples.

C. Learning Temporal Dependencies on the Unlabeled Videos

Currently, only the labeled samples are used for the assessment training, and a large number of unlabeled samples are left and unexploited. Inspired by the temporal dependencies in human action and the self-supervised semi-supervised learning scheme [24], we introduce the self-supervised masked segment feature recovery on the unlabeled videos to learn masked features for the action videos. More specifically, for an unlabeled video represented in basic feature $V^u = [v_1^u, v_2^u, \dots, v_t^u, \dots, v_T^u]$, we randomly select k -th basic segment feature v_k^u ($1 \leq k \leq T$) to mask and form \tilde{V}^u . Then, as in Equation (1), we feed the masked basic feature to our encoder to obtain an intermediate feature,

$$\tilde{f}^u = E(\tilde{V}^u), \quad u = 1, 2, \dots, U, \quad (4)$$

where \tilde{f}^u is the intermediate feature obtained from the masked feature \tilde{V}^u , and $E(\cdot)$ is the feature encoder as in Equation (1). U denotes the total number of the unlabeled videos.

After that, we try to recover the masked segment feature v_k^u from the intermediate feature \tilde{f}^u with a feature decoder $\Phi(\cdot)$. Mean-Absolute-Error (MAE) loss is employed to perform our self-supervised feature recovery learning, which is written as

$$\mathcal{L}_{rcvr} = \frac{1}{U} \sum_{u=1}^U |\Phi(\tilde{f}^u) - v_k^u|, \quad (5)$$

where U is the number of unlabeled samples. \tilde{f}^u represents the intermediate representation obtained from a masked video. $\Phi(\cdot)$ is the decoder for feature recovery, which is implemented with two fully connected layers and a ReLU activation. The masked segment feature recovery task leverages the intrinsic temporal dependency contained in a large number of unlabeled action videos as the supervisory signal to train the encoder and decoder, which can mitigate the shortage of the labeled data and provide richer supervisory signals beside the score annotations. The labeled and unlabeled data are learned via the encoder with shared weight, which guides the encoder to learn the intermediate representations compatible with score assessment and context-based segment feature recovery. The understanding of the context of action plays a key role in action assessment. Therefore, such a self-supervised training mechanism can guide our model to learn discriminative representations from the unlabeled videos.

D. Aligning the Representation Distributions

The intermediate features of the labeled and unlabeled samples could be misaligned due to the sampling bias [40] and the different training objectives in our S⁴AQA framework. Inspired by the domain adaptation work [41], we propose an adversarial training mechanism to align the intermediate representation distributions between the labeled and the unlabeled data. We denote the intermediate features of the labeled samples as class 1, and that of the unlabeled samples as class 0. We introduce a discriminator $D(\cdot)$ to distinguish the feature domain while our encoder tries to confuse the discriminator. This can be represented as such a min-max adversarial optimization problem:

$$\min_E \max_D \left\{ \frac{1}{2U} \sum_{f^u} (\log(1 - D(f^u)) + \log(1 - D(\tilde{f}^u))) + \frac{1}{L} \sum_{f^l} \log(D(f^l)) \right\}, \quad (6)$$

where f^l and f^u denote the features of a labeled and an unlabeled sample, respectively. \tilde{f}^u represents the feature obtained from a masked unlabeled video. Our discriminator module is implemented as a binary classifier.

Instead of optimizing the encoder and the discriminator iteratively, we implement this optimization by a gradient reverse layer (GRL) which can automatically reverse the gradient between the discriminator and the encoder to enable joint training of all modules. With such a joint training mechanism, our adversarial loss is written as

$$\mathcal{L}_{adv} = \sum_{f^i \in \{f^u, \tilde{f}^u, f^l\}} -[(1 - c_i) \log(1 - D(f^i)) + c_i \log(D(f^i))], \quad (7)$$

where c_i is the class label of a sample which is 1 for a labeled sample and 0 for an unlabeled sample. By optimizing the cross-entropy loss function above, the discriminator attempts to predict 1 for a labeled sample and 0 for an unlabeled sample. However, with the gradient reverse layer (GRL), the encoder tries to confuse the discriminator and thereby aligning the representation distributions of the labeled and the unlabeled data during our S⁴AQA training.

E. Optimization

Unlike the existing AQA methods that focus on fully-supervised action assessment, this work attempts semi-supervised action assessment. With only a few video samples annotated with action performance scores, a major challenge is to exploit the unlabeled video samples. We introduce a simple yet general framework called the S⁴AQA, for semi-supervised action assessment, with an encoder $E(\cdot)$ (in Equation (1)) to obtain intermediate features of the input videos, a regression module $R(\cdot)$ (in Equation (2)) to perform supervised training on the labeled samples, a decoder $\Phi(\cdot)$ (in Equation (5)) to perform self-supervised training on the unlabeled samples and a discriminator $D(\cdot)$ (in Equation (7)) that performs adversarial training to align the labeled and unlabeled features. The overall training loss of our S⁴AQA training is given by

$$\mathcal{L} = \mathcal{L}_{reg} + \lambda_1 \mathcal{L}_{rcvr} + \lambda_2 \mathcal{L}_{adv}, \quad (8)$$

where \mathcal{L}_{reg} is the loss for supervised assessment training of the labeled data, \mathcal{L}_{rcvr} is the loss for the self-supervised training of the unlabeled data, and \mathcal{L}_{adv} is the loss for the adversarial training to align the labeled and unlabeled features. λ_1 and λ_2 are non-negative scalar weights for the masked segment feature recovery and representation distribution alignment, respectively.

IV. EXPERIMENTS

In this section, we first introduce two experimental datasets and the implementation details of our S⁴AQA framework, and then report the comparison results of our model with the state-of-the-art fully-supervised AQA methods and semi-supervised methods implemented in AQA on two datasets, respectively. Finally, we demonstrate the effectiveness of each module in our framework through the ablation study, and perform further analysis of our model.

A. Experimental Setup

- Dataset. Two large-scale action quality assessment datasets, the *MTL-AQA dataset* [42] and the *Rhythmic Gymnastics dataset* [6], and a relatively small-scale dataset, the *JIGSAWS dataset* [43], are used in our experiments. Figure 3 presents samples of these three datasets.

The *MTL-AQA dataset* is the largest dataset to date for action quality assessment. There are 1,412 diving videos collected from 16 different international competitions. The diving samples include both male and female athletes, include the 3m Springboard as well as 10m Platform, individual and pairs of synchronized divers. The videos in the MTL-AQA dataset have not only the annotated scores, but also the category of actions that can be used for action recognition, and the commentary information of the on-site commentator that can be used for text analysis. The length of each video is approximately 103 frames and the frame rate is 30 fps. We follow the evaluation protocol suggested in [42] to divide the dataset into a training set with 1,059 samples and a test set with 353 samples. All frames of the video are leveraged to train the model in our experiments.

The *Rhythmic Gymnastics dataset* contains 1,000 videos from four different types of gymnastics routines: ball, clubs, hoop and ribbon, with 250 videos per routine type. All of the videos were collected from high-standard international competitions. Each video is annotated with a performance score which is provided by the referee on the spot. The length of each video is approximately 1 minute and 35 seconds, and the frame rate is 25 fps. We follow the evaluation protocol suggested in [6] to split the dataset into 200 training videos and 50 test videos for each gymnastics routine type. In our experiments, we evenly sample 2 frames per second for each original video, and the length of the sampled video is approximately 180 frames.

The *JIGSAWS dataset* contains 103 videos from three different types of surgical activities, including Suturing, Needle

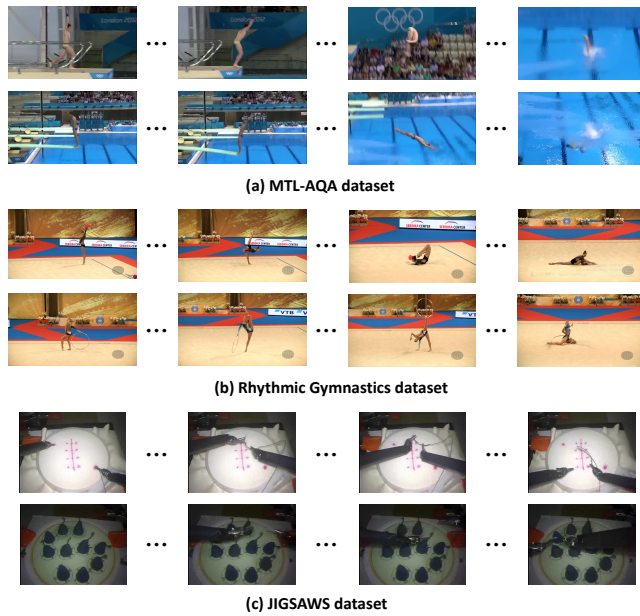


Fig. 3. Samples of (a) the MTL-AQA dataset, (b) the Rhythmic Gymnastics dataset and (c) the JIGSAWS dataset.

Passing and Knot Tying. The number of samples for each type of activity is 39, 28, and 36, respectively. To avoid severe overfitting under the limited labeled samples, we should ensure that the amount of labeled data is large enough to approximate the real data distribution. Therefore, we utilize both left-view and right-view action videos and perform horizontal flipping data augmentation. We follow [4] and regard the master tool manipulators and patient-side manipulators as nodes and extract discriminative DCT features of the 3D kinetics as basic video segment features. We take 50% of the training data as labeled data and the remaining data as unlabeled data. We follow the research [4], [8] and adopt four-fold cross-validation on the experimental results.

- **Metric.** Following the existing methods [9], [4], [5], we use Spearman's Rank Correlation (Sp.Corr) to evaluate the ranking correlation between the ground-truth scores and the predictions. Considering the fact that different referees may give a little different scores for a specific action performance, we tend to pay more attention to the relative performance between samples and less attention to the absolute scores of each sample to measure the performance of action assessment model. This evaluation metric can be formulated as

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}, \quad (9)$$

where x and y indicate the rankings of two score sequences respectively. The value of Sp.Corr ranges from -1 to 1. The larger the value of ρ , the better the assessment effect of the model. Following the existing works [4], [5], the average Spearman's Rank Correlation (Avg Sp.Corr) across multiple actions is computed by Fisher's z-value as in [7].

- **Baseline.** There are two main types of baselines used in our experiments. One type is the fully-supervised model

that achieves the best performance in the benchmarks for action quality assessment, and the other type is the semi-supervised model that can be naturally transferred to AQA. The introduction and processing details of all baseline models are as follows:

◦ Fully-supervised AQA methods

- 1) Support Vector Regression (SVR) [9]. We train a linear support vector regression by video segment features extracted by the I3D over the labeled training samples to predict the action scores. We leverage libsvm [44] to implement SVR in our experiments.
- 2) C3D-AVG-STL/C3D-AVG-MTL [42]. We utilize the I3D backbone to extract the basic video segment features for C3D-AVG-STL/C3D-AVG-MTL, which is the same as other baselines. We train the C3D-AVG-STL and C3D-AVG-MTL models on the labeled videos of the MTL-AQA dataset. C3D-AVG-MTL model leverages the action category and commentary text annotations as additional supervisory signals in a multi-task learning way, which introduces more manual annotations.
- 3) Uncertainty-aware Score Distribution Learning (USDL) [8]. USDL achieves the state-of-the-art performance on the MTL-AQA dataset by distribution learning. We train the USDL model on the labeled videos of the MTL-AQA dataset in our experiments.
- 4) DynAmic-static Context-aware attentiON NETwork (ACTION-NET) [6]. We train ACTION-NET on the labeled videos of the Rhythmic Gymnastics dataset. In order to make a fair comparison, we only use the dynamic information branch to train ACTION-NET, which is consistent with the input information of other models in our experiments.

◦ Semi-supervised methods implemented in AQA

- 1) Co-Training Semi-Supervised Regression (COREG) [32]. COREG is a non-parametric co-training style semi-supervised regression algorithm, which uses two k-nearest neighbor regressors with different distance metrics to generate the scores for unlabeled videos by estimating the influence of the labeling for unlabeled videos on the labeled videos. We utilize the average of video segment features extracted by I3D as the input of COREG.
- 2) Pseudo-labels [35]. In order to adapt to the semi-supervised regression, we use the k-nearest neighbor algorithm to generate the pseudo labels for unlabeled videos by the distance metric between labeled videos and unlabeled videos in the embedding space as [35]. We set the training weight of the unlabeled videos as 0.5 in the experiment.
- 3) Virtual Adversarial Training (VAT) [23]. VAT defines the adversarial direction without label annotations to make the model robust with the local perturbation on unlabeled data. We use the output vector of the penultimate fully connected layer of the regression module to approximate the virtual labels of unlabeled videos. In our evaluation, the weight of the regularization loss term and norm constraint are set to 1.0 and 2.5, respectively.

4) Self-Supervised and Semi-Supervised Learning (S^4L) [24]. In our experiment, we transfer this method from image to the video, which couples the video rotation prediction self-supervised task ($0^\circ, 90^\circ, 180^\circ, 270^\circ$) with score regression task in a joint training way.

- Implementation. We use the I3D network pre-trained on the Kinetics dataset [45] to extract basic video segment features of labeled and unlabeled videos. All video frames are center cropped from original frames and resized into 224×224 . And the dimension of each basic video segment feature extracted by the I3D backbone is 1024. The number of basic video segments in the MTL-AQA dataset and the Rhythmic Gymnastics dataset are 10 and 12 respectively. We normalize the scores of all samples to $0 \sim 5$. Following previous works [8], [6], we use Adam optimizer [46] to train models on the MTL-AQA dataset with a learning rate of 0.5×10^{-3} , a weight decay of 10^{-4} and 1000 epochs, and use the mini-batch stochastic gradient descent (SGD) optimizer to train models on the Rhythmic Gymnastics dataset with a learning rate of 10^{-4} , a weight decay of 10^{-4} and 500 epochs. We set $\lambda_1 = 40.0$ and $\lambda_2 = 0.1$ in Equation (8). The further analysis of the different weight combinations (λ_1, λ_2) can be found in section 4.4. For a fair comparison, all parametric semi-supervised methods use the same encoder and score regression network as our S^4AQA . We implement all baselines with PyTorch based on the source codes released by the authors. To ensure the distribution of sampled data is consistent with that of original training data, we uniformly sample labeled and unlabeled data according to the distribution of sample scores.

B. Comparison with SOTA Methods

To evaluate our method, we compare S^4AQA with the state-of-the-art fully-supervised AQA methods and semi-supervised baselines implemented on the MTL-AQA dataset, the Rhythmic Gymnastics dataset and the JIGSAWS dataset. The results are shown in Table I, Table II, and Table III. As shown in the tables, our method outperforms all the other methods and obtains state-of-the-art performances of semi-supervised action quality assessment. Specifically, compared with the state-of-the-art fully-supervised methods (USDL, ACTION-NET), our method brings an improvement of up to 0.146, 0.034 and 0.150 on the MTL-AQA dataset, the Rhythmic Gymnastics dataset, and the JIGSAWS dataset respectively, by effectively leveraging the unlabeled videos. Among all the semi-supervised learning methods, COREG has the worst performance because it is intractable for the non-parametric model to learn discriminative action representations. S^4L achieves relatively poor results, which suggests the video frame rotation prediction self-supervised task is inefficient to learn discriminative representation from unlabeled videos, and the assessment model derives little benefit from unlabeled samples. Pseudo-labels model obtains additional supervised information from the pseudo labels of unlabeled data and achieves similar performance to S^4L . However, the Pseudo-labels method cannot correct the error in pseudo labels, which hinders further performance improvement. Additionally, VAT is robust to small perturbation on input data, which prevents

TABLE I
THE TEST SP.CORR ON THE MTL-AQA DATASET WITH 10%/40% LABELS OF TRAINING SET. THE METHODS MARKED WITH * ARE THOSE TRAINED WITHOUT USING AN END-TO-END TRAINING STRATEGY.

| Method | 10% of labeled data | 40% of labeled data |
|--------------------|---------------------|---------------------|
| SVR [9] | 0.427 | 0.565 |
| USDL* [8] | 0.530 | 0.646 |
| C3D-AVG-STL* [42] | 0.561 | 0.632 |
| C3D-AVG-MTL* [42] | 0.584 | 0.656 |
| COREG [32] | 0.487 | 0.526 |
| Pseudo-labels [35] | 0.622 | 0.716 |
| VAT [23] | 0.635 | 0.724 |
| S^4L [24] | 0.621 | 0.721 |
| S^4AQA (Ours) | 0.676 | 0.746 |

TABLE II
THE TEST SP.CORR ON THE RHYTHMIC GYMNASTICS DATASET WITH 40% OF LABELS OF TRAINING SET. AVG DENOTES THE AVERAGE SPEARMAN'S RANK CORRELATION ACROSS MULTIPLE ACTION TYPES.

| Method | Ball | Clubs | Hoop | Ribbon | Avg |
|--------------------|--------------|--------------|--------------|--------------|--------------|
| SVR [9] | 0.175 | 0.243 | 0.261 | 0.309 | 0.248 |
| ACTION-NET [6] | 0.196 | 0.403 | 0.319 | 0.305 | 0.308 |
| COREG [32] | 0.230 | 0.338 | 0.331 | 0.268 | 0.292 |
| Pseudo-labels [35] | 0.183 | 0.330 | 0.346 | 0.305 | 0.292 |
| VAT [23] | 0.208 | 0.355 | 0.345 | 0.292 | 0.301 |
| S^4L [24] | 0.209 | 0.325 | 0.324 | 0.290 | 0.288 |
| S^4AQA (Ours) | 0.248 | 0.388 | 0.372 | 0.357 | 0.342 |

the model from overfitting to the labeled videos through regularization term and achieves better results than other semi-supervised baselines on three datasets. From the experimental results, we can observe that the smaller the proportion of labeled data, the greater the benefits of our model. It is worth noting that ACTION-NET is a strong baseline which uses well-designed GCN and attention mechanism to model the relationship between video segments on the Rhythmic Gymnastics dataset. Therefore, it is expected that the performances of other semi-supervised baselines with simple network architectures do not exceed ACTION-NET. However, our method with a simple architecture can outperform ACTION-NET and other baseline methods on the Rhythmic Gymnastics dataset by effectively leveraging unlabeled videos to learn discriminative representations.

C. Ablation Study

We investigate the effectiveness of the masked segment feature recovery module and representation distribution alignment module in our proposed S^4AQA on the MTL-AQA dataset through quantitative and qualitative analysis.

To show the effectiveness of our proposed masked segment feature recovery module and adversarial representation distribution alignment module, we try to remove either the masked segment feature recovery module or adversarial representation distribution alignment module from our architecture. The experimental results of ablation study are shown in Table IV and Table V. We use the abbreviations *AA*, *MSFR* and *RDA* to represent our action assessment module, our masked segment feature recovery module and our representation distribution alignment module, respectively. The model *AA* is trained with the score regression loss for labeled videos, *AA+RDA* model

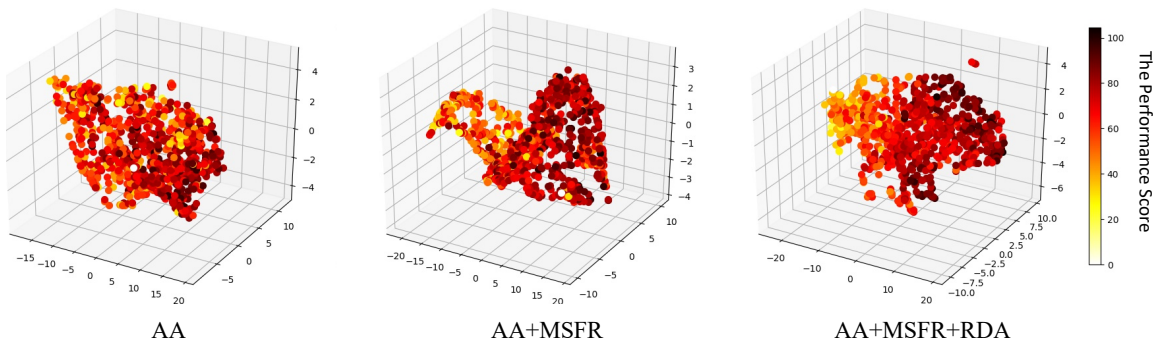


Fig. 4. The t-SNE visualization of representation distribution of unlabeled data on the MTL-AQA dataset with 10% of labels learned by AA , $AA + MSFR$ and $AA + MSFR + RDA$ models. The darker the color of the sample is, the higher its ground truth score will be. With the help of our proposed modules, the unlabeled videos with different scores are more uniformly scattered into different regions of representation space, making the representation of unlabeled videos more discriminative.

TABLE III

THE TEST SP.CORR ON THE JIGSAWS DATASET WITH 50% OF LABELS OF TRAINING SET. AVG DENOTES THE AVERAGE SPEARMAN'S RANK CORRELATION ACROSS MULTIPLE ACTION TYPES.

| Method | Suturing | Needle Passing | Knot Tying | Avg |
|--------------------|--------------|----------------|--------------|--------------|
| USDL [8] | 0.439 | 0.351 | 0.680 | 0.505 |
| Pseudo-labels [35] | 0.445 | 0.501 | 0.714 | 0.566 |
| VAT [23] | 0.524 | 0.526 | 0.749 | 0.612 |
| S^4L [24] | 0.455 | 0.529 | 0.730 | 0.585 |
| S^4AQA (Ours) | 0.533 | 0.552 | 0.813 | 0.655 |

TABLE IV

THE ABLATION STUDY RESULTS ON THE MTL-AQA DATASET. THE TEST SP.CORR IS REPORTED ON THE MTL-AQA DATASET WITH 10%/40% LABELS. AA , $MSFR$ AND RDA REPRESENT OUR ACTION ASSESSMENT MODULE, OUR MASKED SEGMENT FEATURE RECOVERY MODULE AND OUR REPRESENTATION DISTRIBUTION ALIGNMENT MODULE, RESPECTIVELY.

| Module | 10% of labeled data | 40% of labeled data |
|---------------|---------------------|---------------------|
| AA | 0.618 | 0.703 |
| $AA+RDA$ | 0.634 | 0.726 |
| $AA+MSFR$ | 0.661 | 0.732 |
| $AA+MSFR+RDA$ | 0.676 | 0.746 |

is trained with the score regression loss for labeled videos and the adversarial loss for the representations alignment, $AA + MSFR$ model is trained with the score regression loss for labeled videos and the masked feature recovery loss for unlabeled videos, and $AA + MSFR + RDA$ is our full model which introduces the adversarial representation distribution alignment into $AA + MSFR$ model. As shown in Table IV and Table V, $AA + MSFR$ brings an improvement of up to 0.043 on the MTL-AQA dataset, and 0.046 on the Rhythmic Gymnastics dataset over AA model. In addition, adding RDA ($AA + MSFR + RDA$) brings an improvement of up to 0.015 on the MTL-AQA dataset, and 0.013 on the Rhythmic Gymnastics dataset over $AA + MSFR$ model. This illustrates that the intermediate representation learned by our masked segment feature recovery module is complementary to the representations learned by our score regression module. And the adversarial training mechanism can further align the representation distributions of unlabeled videos and labeled videos, which further improves the performance of our model.

To further explore the effectiveness of each module, we

TABLE V

THE ABLATION STUDY RESULTS ON THE RHYTHMIC GYMNASTICS DATASET. THE TEST SP.CORR IS REPORTED ON THE RHYTHMIC GYMNASTICS DATASET WITH 40% OF LABELS.

| Module | Ball | Clubs | Hoop | Ribbon | Avg |
|---------------|--------------|--------------|--------------|--------------|--------------|
| AA | 0.192 | 0.322 | 0.327 | 0.289 | 0.283 |
| $AA+RDA$ | 0.204 | 0.327 | 0.324 | 0.290 | 0.287 |
| $AA+MSFR$ | 0.236 | 0.378 | 0.361 | 0.338 | 0.329 |
| $AA+MSFR+RDA$ | 0.248 | 0.388 | 0.372 | 0.357 | 0.342 |

visualize the representation distributions of unlabeled videos of AA , $AA + MSFR$ and $AA + MSFR + RDA$ by t-SNE [47] in Figure 4. Lighter colors indicate samples with lower scores, and darker colors indicate samples with higher scores. For the AA model trained with only the labeled videos, the feature distribution of unlabeled data is confusing, since the representations of many samples with lower scores are intertwined with that of samples with higher scores. To $AA + MSFR$ model, the representation distribution of unlabeled data has been improved to some extent. However, there are still some samples with higher scores around the area of samples with lower scores in the representation space (left part of the sub-picture at the middle of Figure 4). Compared with the former two models, the unlabeled samples with different scores in model $AA + MSFR + RDA$ (S^4AQA) are more uniformly scattered into different regions of representation space.

D. Further Analysis

We also evaluate the effect of the different weight combinations of overall training loss and compare the performance of different masking strategies of our masked segment feature recovery module. Finally, we explore the impact of different training mechanisms on the performance of our model.

- Evaluation on the weights of training loss. The effect of the different weight combinations (λ_1 , λ_2) of the self-supervised segment recovery loss and adversarial learning loss in Equation (8) is shown in Figure 5. As the λ_1 increases, our model learns more discriminative features that assist the assessment model from the self-supervised task for masked segment feature recovery. When λ_1 is greater than 40, the benefit of assessment model from self-supervised tasks begins

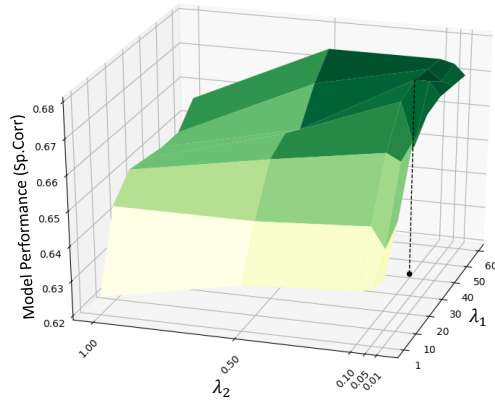


Fig. 5. The test Sp.Corr on the MTL-AQA dataset with 10% labels of training set under different weight combinations (λ_1 , λ_2) of overall training loss.

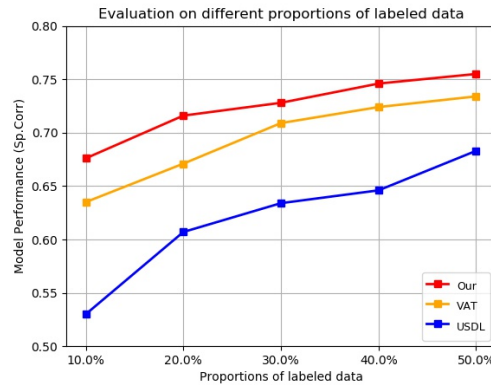


Fig. 6. The experimental comparison results of our model and other baselines under different proportion of labeled data on the MTL-AQA test set.

to decrease steadily. For the λ_2 of the representation distribution alignment adversarial loss, when its value is 0.1, the assessment model achieves relatively good performance. When λ_2 is too small, the intermediate feature alignment between labeled and unlabeled videos is not obvious, and when λ_2 is too large, excessive feature alignment will affect the training of the assessment model negatively. By comparing λ_1 and λ_2 , we can find that the self-supervised task for masked segment feature recovery occupies a more important position in the overall loss function than distribution alignment task.

- **Different masking strategies.** In order to explore the influence of different numbers of masked segment features, we conduct experiments on three different masking strategies (1 ~ 3 masked segment(s)) to evaluate our masked segment feature recovery module. In order not to change the structure of our masked segment feature recovery module, we simplify the recovery target of multiple features to the average of multiple target features. Additionally, we implement the feature prediction strategy in [20] on our 3D-CNN backbone, and construct the still segment by replacing each frame with the first frame in each segment to approximate a still frame. Then, we establish a self-supervised task to predict all the segment features of the original video based on the features of all still segments.

TABLE VI
THE TEST SP.CORR ON THE MTL-AQA DATASET WITH 10%/40% OF LABELED DATA IN TRAINING SET.

| Masking strategy | 10% of labeled data | 40% of labeled data |
|-----------------------------------|---------------------|---------------------|
| Mask one segment | 0.676 | 0.746 |
| Mask two segments | 0.660 | 0.730 |
| Mask three segments | 0.657 | 0.724 |
| Keep first frame per segment [20] | 0.629 | 0.706 |

TABLE VII
THE TEST SP.CORR RESULTS ON THE MTL-AQA DATASET WITH MULTI-MODAL DATA. AA REPRESENTS OUR ACTION ASSESSMENT MODULE.

| Model | 10% of labeled data | | 40% of labeled data | |
|-------------------|---------------------|--------------|---------------------|--------------|
| | RGB | RGB+Flow | RGB | RGB+Flow |
| base model (AA) | 0.618 | 0.633 | 0.703 | 0.715 |
| Full model | 0.676 | 0.708 | 0.746 | 0.765 |

We report the experimental results of different masking strategies on the MTL-AQA dataset with 10% and 40% of labels in Table VI. As shown in the table, as the number of the masked segment features increases, it will become more difficult for our model to recover masked segment features, resulting in the self-supervised learning task being unable to learn discriminative representations. The masking strategy in [20] masks most of the frames in the whole video whose predicted targets have serious uncertainty. Therefore, the assessment model attains a little benefit from the self-supervised task under this masking strategy.

- **Sensitivity to the number of labeled data.** In order to verify the robustness of our S^4 AQA with different sampling sizes, we conduct experiments on the MTL-AQA dataset under the settings where the proportion of labeled samples is 10%, 20%, 30%, 40% and 50%, respectively. The experimental comparison results of fully-supervised USDL [8], semi-supervised VAT [23] and our model are shown in Figure 6. We can observe that our model's Sp.Corr on test set can surpass USDL and VAT under different ratios of labeled data, which fully demonstrates that our proposed method is robust to different sampling sizes. It is worth mentioning that the test Sp.Corr of our S^4 AQA with 10% of labeled data close to that of USDL model with 50% of labeled data.

- **Extension to multi-modal data.** To confirm that our self-supervised segment feature recovery can be extended to data of other modalities with temporal dependencies, we additionally obtain the optical flow on the MTL-AQA dataset by TV-L1 algorithm [48]. In our modelling, we utilize RGB and optical flow features of unlabeled videos through our masked segment feature recovery module with different encoders and decoders, concatenate the intermediate features extracted from RGB and optical flow of labeled videos by respective encoders, and train our S^4 AQA model in a multi-modal manner. The experimental results are shown in Table VII. When 10% samples are annotated, the Sp.Corr of AA model (using only labeled data) is 0.633 without our self-supervised semi-supervised learning. With our S^4 AQA learning (using labeled and unlabeled data), the performance improves to 0.708. When 40% samples are annotated, the Sp.Corr of AA model is 0.715

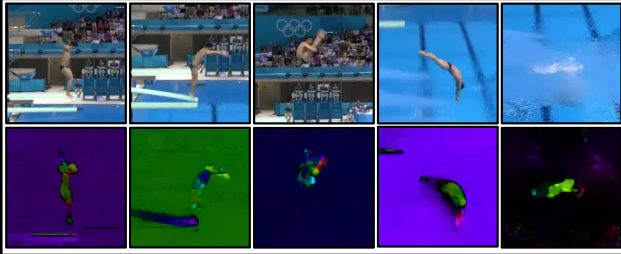

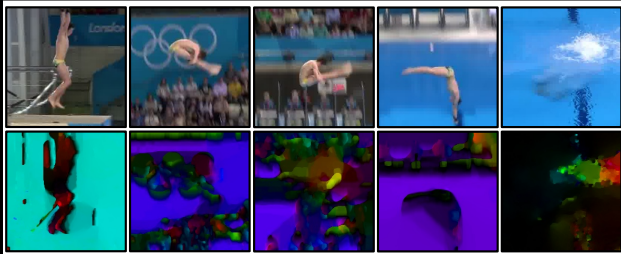
| | Video cases | Motion Complexity | Pred of Base Model | Pred of Full Model | Ground Truth Score |
|------------------|--|-------------------|-----------------------------|-----------------------------------|--------------------|
| Case A (Success) |  | <u>4.03</u> | 71.48 ($\Delta 3.28$) | 68.34 ($\Delta 0.14$) | 68.20 |
| Case B (Success) |  | <u>3.85</u> | 80.05 ($\Delta 17.05$) | 70.85 ($\Delta 7.85$) | 63.00 |
| Case C (Failure) |  | <u>12.34</u> | 73.14 ($\Delta 7.14$) | 72.73 ($\Delta 6.73$) | 66.00 |

Fig. 7. The case study of our model on the MTL-AQA datasets. Two successful cases and a failure case are presented. The motion complexity is measured by the intensity of optical flows [48] across the video. For each video, we present the predicted score by our base model (the AA model in our ablation study) and our full model, together with the ground-truth score. The prediction errors between the predicted scores and the ground truth are also provided.

TABLE VIII
THE PERFORMANCES OF DIFFERENT ALIGNMENT STRATEGIES WITH 10%/40% OF LABELED DATA ON THE MTL-AQA DATASET.

| Masking strategy | 10% of labeled data | 40% of labeled data |
|------------------------|---------------------|---------------------|
| <i>w/o</i> Alignment | 0.661 | 0.732 |
| GRL | 0.676 | 0.746 |
| MMD (Laplacian Kernel) | 0.675 | 0.739 |
| MMD (RBF Kernel) | 0.671 | 0.745 |

TABLE IX
THE TEST SP.CORR RESULTS ON THE MTL-AQA DATASET OF TRAINING MASKED SEGMENT FEATURE RECOVERY LOSS WITH LABELED DATA (DENOTED AS *w/* LABELED DATA) AND USING THE VANILLA SELF-SUPERVISED LEARNING MECHANISM WITH OUR PROPOSED SELF-SUPERVISED TASK (DENOTED AS VANILLA SSL).

| Method | 10% of labeled data | 40% of labeled data |
|------------------------|---------------------|---------------------|
| <i>w/</i> labeled data | 0.674 | 0.748 |
| Vanilla SSL | 0.616 | 0.707 |
| Ours | 0.676 | 0.746 |

without utilizing unlabeled data, and 0.765 with our S^4AQA . The experimental results demonstrate that our S^4AQA can be naturally transferred to multi-modal data.

- **Different alignment strategies.** We evaluate our model under different alignment strategies to verify the robustness

of our model to different representation distribution alignment strategies. We compare the common Gradient Reverse Layer (GRL) and the Maximum Mean Discrepancy (MMD) [49] alignment strategies under our self-supervised semi-supervised learning framework with 10% and 40% of labeled data on the MTL-AQA dataset. For the kernel function of MMD, we use the most common Laplacian Kernel and Radial Basis Function (RBF) Kernel. The experimental results are shown in Table VIII. We can observe that our method does not rely on a specific distribution alignment strategy, and the performances of the two alignment strategies are similar. However, GRL does not depend on the specific nonlinear kernel mapping to compute distribution discrepancy between the representation of labeled data and unlabeled data compared to MMD.

- **Different training mechanisms.** In this experiment, we evaluate our model on other training schemes, and the results are shown in Table IX. First, we introduce labeled data in our self-supervised masked feature recovery (shown as “*w/* labeled data” in Table IX). Adding labeled data has little effect on the performance of our model, indicating that using the unlabeled data is sufficient to learn the temporal dependencies for action assessment. Second, we compare our joint training mechanism with the vanilla self-supervised training mechanism (pre-training the model on unlabeled data with self-supervised loss

and then fine-tuning it on labeled data with score regression loss). It is observed that our joint training scheme outperforms the vanilla self-supervision training mechanism. This is because our joint training mechanism fits both the self-supervised task and the score regression task simultaneously, and it better exploits the unlabeled data.

- Case study. As shown in Figure 7, we perform case studies on the diving samples in the MTL-AQA dataset [42] and show two successful cases and one failure case of our model with the intensity of optical flow to quantify the motion complexity. For each video, we present the predicted scores of our base model (AA model in the ablation study, which is trained only with the labeled samples), the predicted scores of our full model, and the ground-truth scores. The models used for the case study are trained on only 10% labeled samples. From the results, we can see that our methods greatly reduce the score prediction error over the base model (Case A: from 3.28 to 0.14; Case B: from 17.05 to 7.85). This is because the proposed method can better leverage the unlabeled data for action assessment training, with masked segment feature recovery to learn the motion dynamics and the adversarial training to align the motion representations between the labeled and the unlabeled videos. However, in Case C both the base model and our full model have imperfect performances, with a prediction error of 7.14 and 6.73, respectively. Although our full model still performs a bit better, Case C attains a little benefit from the self-supervised task under the unlabeled data compared to successful cases. We note that this case is much more difficult, since it has a much higher motion complexity (more than 3 times compared with Case A and Case B). We believe that this problem can be mitigated by introducing more discriminative auxiliary information to assist the modelling of dynamic action evolution (such as poses, tracking bounding boxes of athletes, etc.)

V. CONCLUSIONS

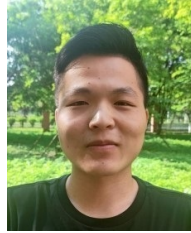
In this work, we explore semi-supervised learning for action quality assessment, in which a major challenge is to exploit the unlabeled data. To address this challenge, we propose to learn the temporal dependencies of the action execution by means of a self-supervised masked feature recovery module on the unlabeled data. Moreover, we propose an self-supervised semi-supervised framework for action quality assessment (S^4AQA). Additionally, an adversarial training module is applied to align the representation distributions of labeled and unlabeled videos. The extensive experiments on the MTL-AQA and the Rhythmic Gymnastics datasets illustrate the effectiveness of our model and the contribution of each module in our framework. With the proposed S^4AQA framework, we establish the state-of-the-art performance of semi-supervised action quality assessment.

REFERENCES

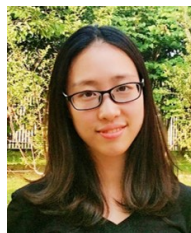
- [1] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriù, L. Romeo, and F. Verdini, "The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1436–1448, 2019.

- [2] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, pp. 468–477, 2020.
- [3] 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.
- [4] 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.
- [5] J. Gao, W.-S. Zheng, J.-H. Pan, C. Gao, Y. Wang, W. Zeng, and J. Lai, "An asymmetric modeling for action assessment," in *European Conference on Computer Vision*. Springer, 2020, pp. 222–238.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *European Conference on Computer Vision*. Springer, 2014, pp. 556–571.
- [10] 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.
- [11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.
- [12] S. Dai and H. Man, "Mixture statistic metric learning for robust human action and expression recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2484–2499, 2018.
- [13] J. Kong, H. Deng, and M. Jiang, "Symmetrical enhanced fusion network for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.
- [14] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2405–2415, 2019.
- [15] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 591–600.
- [16] A. Zia and I. Essa, "Automated surgical skill assessment in rmis training," *International journal of computer assisted radiology and surgery*, vol. 13, no. 5, pp. 731–739, 2018.
- [17] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *International journal of computer assisted radiology and surgery*, vol. 13, no. 3, pp. 443–455, 2018.
- [18] 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.
- [19] H. Jain, G. Harit, and A. Sharma, "Action quality assessment using siamese network-based deep metric learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2260–2273, 2020.
- [20] P. Parmar and B. Morris, "Hallucinet-ing spatiotemporal representations using a 2d-cnn," *Signals*, vol. 2, no. 3, pp. 604–618, 2021.
- [21] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013.
- [22] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.
- [23] 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, 2018.

- [24] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4I: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1476–1485.
- [25] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3d action recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–51.
- [26] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020.
- [27] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 393–400.
- [28] C. Jia, Z. Ding, Y. Kong, and Y. Fu, "Semi-supervised cross-modality action recognition by latent tensor transfer learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2801–2814, 2020.
- [29] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," *arXiv preprint arXiv:1406.5298*, 2014.
- [30] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.
- [31] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.
- [32] Z.-H. Zhou and M. Li, "Semi-supervised regression with co-training," in *IJCAI*, vol. 5, 2005, pp. 908–913.
- [33] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.
- [34] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," *arXiv preprint arXiv:1904.12848*, 2019.
- [35] P. Hou, X. Geng, Z.-W. Huo, and J.-Q. Lv, "Semi-supervised adaptive label distribution learning for facial age estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [36] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 334–10 343.
- [37] J. Huang, Y. Huang, Q. Wang, W. Yang, and H. Meng, "Self-supervised representation learning for videos by segmenting via sampling rate order prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.
- [38] 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.
- [39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
- [40] Q. Wang, W. Li, and L. V. Gool, "Semi-supervised learning by augmented distribution alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1466–1475.
- [41] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [42] 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.
- [43] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmadi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, 2014, p. 3.
- [44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [45] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
- [46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [47] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [48] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo, "Tv-11 optical flow estimation," *Image Processing On Line*, vol. 2013, pp. 137–150, 2013.
- [49] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.



Shao-Jie Zhang received his Bachelor's degree in Software Engineering from Northwestern Polytechnical University in 2020. He is now a Master's student in the School of Computer Science and Engineering at Sun Yat-Sen University. His research interests include computer vision and machine learning.



Jia-Hui Pan received her Bachelor's degree in Computer Science and Technology from Sun Yat-Sen University in 2018. She is now a Master's student in the School of Computer Science and Engineering at Sun Yat-Sen University. Her research interests include computer vision and machine learning.



Jibin Gao received his Bachelor's degree in Software Engineering from Sun Yat-Sen University in 2019. He is now a Master's student in the School of Computer Science and Engineering at Sun Yat-Sen University. His research interests include computer vision and machine learning.



Wei-Shi Zheng is now a full Professor with Sun Yat-sen University. Dr. Zheng received his Ph.D. degree in Applied Mathematics from Sun Yat-sen University in 2008. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. Especially, Dr. Zheng has active research on person re-identification in the last five years. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He has ever served as area chairs of CVPR, ICCV, BMVC and IJCAI. He is an IEEE MSA TC member. He is an associate editor of the Pattern Recognition Journal. He is a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of the Royal Society-Newton Advanced Fellowship of the United Kingdom.