Information Sciences 664 (2024) 120347

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Two-path target-aware contrastive regression for action quality assessment

Xiao Ke [a,b], Huangbiao Xu [a,b,*], Xiaofeng Lin [a,b], Wenzhong Guo [a,b]

[a] Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou, 350116, China

[b] Key Laboratory of Spatial Data Mining & Information Sharing, Ministry of Education, Fuzhou, 350003, China

ARTICLE INFO

Keywords:
Action quality assessment
Multi-view information
Video understanding

ABSTRACT

Action quality assessment (AQA) is a challenging vision task due to the complexity and variance of the scoring rules embedded in the videos. Recent approaches have reduced the prediction difficulty of AQA via learning action differences between videos, but there are still challenges in learning scoring rules and capturing feature differences. To address these challenges, we propose a two-path target-aware contrastive regression (T^2CR) framework. We propose to fuse direct and contrastive regression and exploit the consistency of information across multiple visual fields. Specifically, we first directly learn the relational mapping between global video features and scoring rules, which builds occupational domain prior knowledge to better capture local differences between videos. Then, we acquire the auxiliary visual fields of the videos through sparse sampling to learn the commonality of feature representations in multiple visual fields and eliminate the effect of subjective noise from a single visual field. To demonstrate the effectiveness of T^2CR, we conduct extensive experiments on four AQA datasets (MTL-AQA, FineDiving, AQA-7, JIGSAWS). Our method is superior to state-of-the-art methods without elaborate structural design and fine-grained information.

1. Introduction

Action Quality Assessment (AQA) is an emerging video understanding task that aims to automatically assess the performance quality of action sequences (*e.g.*, diving, gymnastics, surgery) and leads to increasing interest in its wide range of applications. AQA empowers AI to analyze and deliver objective assessments under a uniform standard, reducing individual subjectivity and social controversy over fairness and providing feedback that helps improve human skill levels. AQA finds utility in diverse scenarios, such as analyzing sports videos [12,27,42], assisting with healthcare technology assessments [44,10], determining specific skill levels [8,46], and others [50,49]. AQA is more challenging than traditional human action recognition. Unlike human action recognition [19,28,20,21], which recognize which action is performed in the current video among different classes, AQA aims to quantitatively assess how well the actions from the same class are performed. These action videos from the same specific domain have poor intra-class discrimination, with subtle action differences often resulting in significant score changes. Therefore, the main challenge of AQA is to build reliable score assessment models that can capture and understand subtle action variations.

* Corresponding author at: Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou, 350116, China.

E-mail addresses: huangbiaoxu0905@gmail.com, 231010005@fzu.edu.cn (H. Xu).

https://doi.org/10.1016/j.ins.2024.120347
Received 21 September 2023; Received in revised form 23 January 2024; Accepted 19 February 2024

Available online 28 February 2024

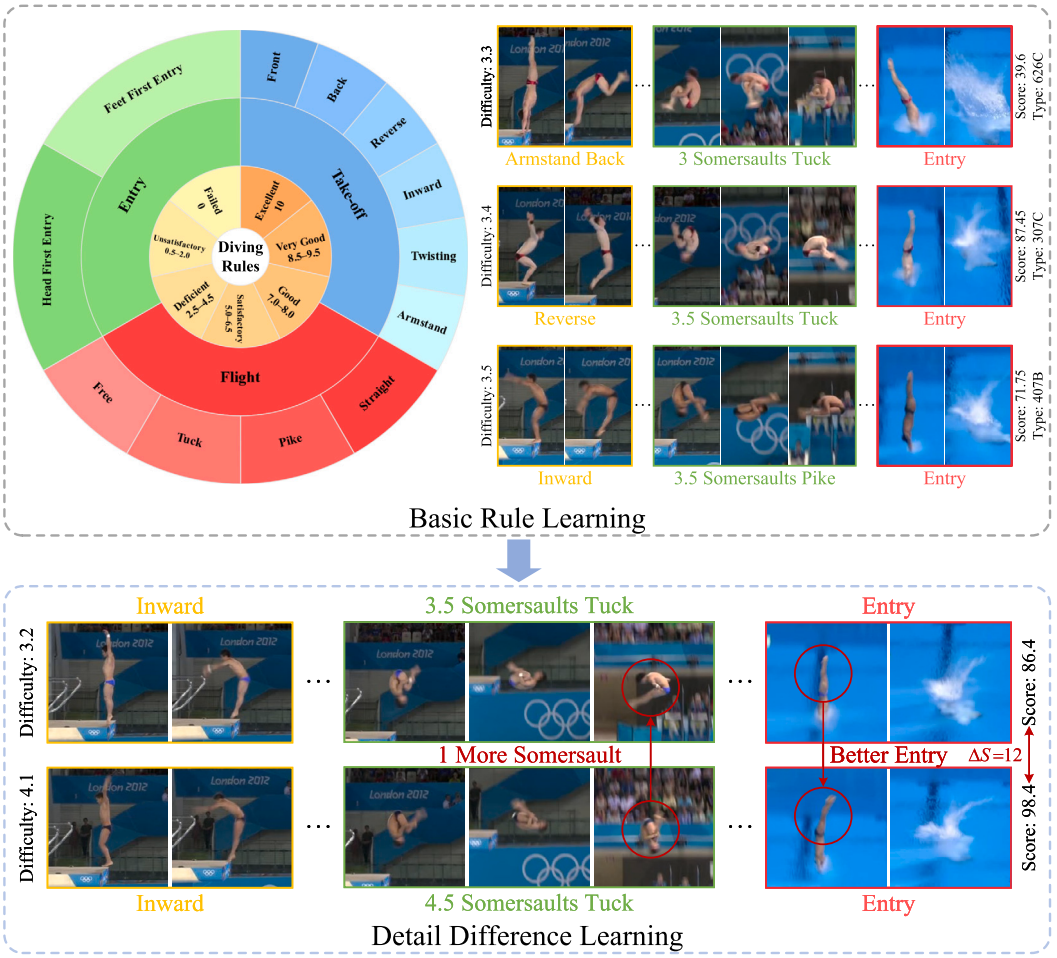

Fig. 1. An example of our idea of target-aware contrastive regression. Establishing occupational basis rules as the prior knowledge to better guide the comparison of differences in local actions. All figures in this paper are best viewed in color.

Most existing works [26,38,35] directly regress features on individual videos to obtain quality scores, which makes it challenging to capture the subtle variations of actions and the assessment effect seems to be bottlenecked. Thanks to the recently proposed CoRe [42] framework, learning action differences between videos to predict difference scores has become a new research idea for AQA. Unlike directly regressing the final score, learning the difference scores can convert the range of predicted scores to a smaller range of difference scores, reducing the margin of errors and the difficulty of regression (*e.g.*, the score range level for diving can be reduced from hundreds to tens).

However, is learning difference scores alone enough? AQA aims to give AI knowledge and insight close to that of human experts in a given occupation, which requires AI to discover and learn the complex rules of that occupation. Though learning the difference score reduces the difficulty of prediction, it also narrows the range of rules that can be learned. This often models the assessment of differences in local features only while ignoring the global feature mapping. On the other hand, effectively learning differences in the deep features is also a challenge. The query video has a different information distribution from the exemplar video. Using different videos as the exemplar will teach different information about feature differences, *i.e.*, there are different mappings of rules between actions and scores in different videos. These different mappings are essential expertise in the professional domain, which can lead to confusing and irregular feature learning. Moreover, the existing datasets are constructed from the subjective scores of the judges, which leads to noise that disturbs the learning of the samples. Similarly, these noises have different effects in different video samples, further increasing the difficulty of feature learning.

Considering these reflections, we summarize the challenges of AQA and difference score learning: (1) How to better model the relation between action features and scores to match the profession's rules? (2) How can the irregular information distribution be captured in different (query, exemplar) video pairs? (3) How to eliminate the effects of subjective noise inherent in datasets?

To address these challenges, we propose a new target-aware contrastive regression framework. We propose to merge the ideas of direct regression and contrastive regression. Specifically, we directly regress the global spatial-temporal domain features to the final score to model the "feature-score" relation, modeling the basis rule mapping between action variation and scores. Additionally, in diving competitions, the dive numbers represent the type and difficulty of the athletes' specific actions, which are known to the

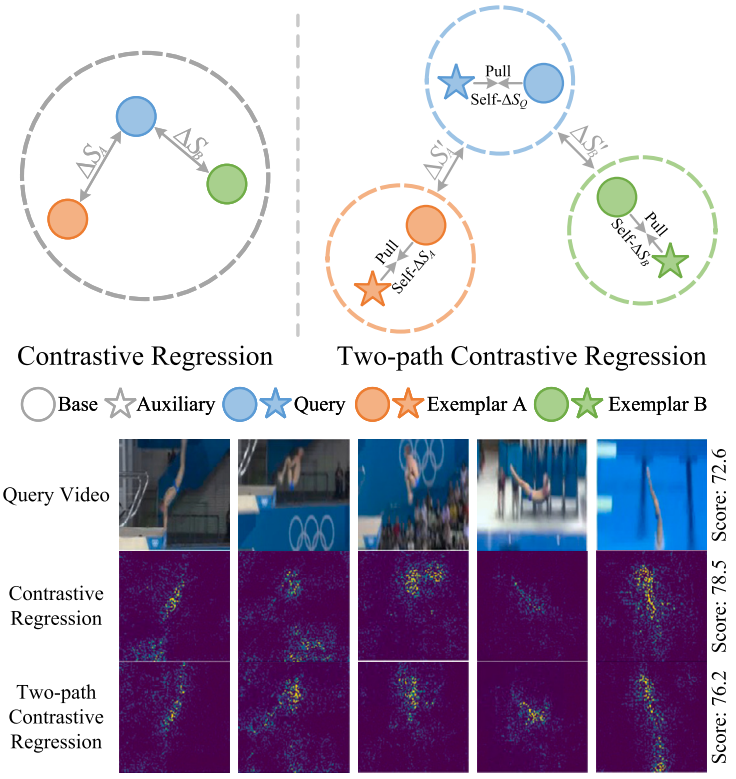

Fig. 2. An example of our idea of two-path contrastive regression. Our method fuses information from multiple visual fields to capture the commonality of features and eliminate the subjective noise from the single visual field. As a result, our method filters out noise and focuses more on the athlete.

judges in advance. We introduce the dive numbers as prior knowledge of the "action type" occupational rules, enabling global feature learning for classification and understanding the rule-based information associated with video actions. For example, the type "109B" represents the actions that include "Forward, 4.5 Somersault Pike, and Entry". We then learn the differences in local sub-actions between the query and exemplar videos, regressing the accurate difference scores that match the basic rules. This approach enables the model to construct a comprehensive basic prior knowledge of the occupational domain, which is then used to learn more detailed action differences. We aim to make the network understand the basic principles of assessment, leading to a more accurate and stable network [3,29] rather than just memorizing figures [43]. As shown in Fig. 1, this process is akin to humans accumulating a vast amount of basic knowledge to learn more advanced professional skills.

To better learn the difference scores, we further propose a two-path contrastive regression framework. Recently, sampling video at different rates has been beneficial for action recognition [9,32]. We process the input video with two different samples and learn the feature representation. For AQA, the different samples of the video should imply the same quality score. Thus, we pull these representations closer and maximize their similarity. Compared to directly capturing the difference information distribution in a video pair (query, exemplar), our approach captures the inherent distribution by fusing the two path fields of a single video. Then, it compares the information differences in the video pairs. This leads to a generic model which extracts the difference information between different video pairs and excludes the disruption of irregular distributions. Furthermore, our two-path contrastive regression framework incorporates self-difference information learning by performing self-difference contrast between two paths of the same video. The difference score between different views of a video should be zero, which is vital supervisory information. Leveraging this characteristic, we facilitate the difference transformer and difference score regression to acquire richer information and guide more accurate difference contrast between the query and exemplar videos.

Moreover, the different visual fields enhance objectivity in assessment, aligning with the realistic scoring model of multiple judges, who often score with subjective noise due to different views, focus, and perception degrees of actions. In contrast, our method learns information from multiple visual fields, eliminating the effect of subjective noise from a single visual field. As shown in Fig. 2, our method more effectively focuses on motion information, filtering out much irrelevant background noise.

To verify the effectiveness of our method, we conduct extensive experiments on four AQA datasets, including MTL-AQA [25], FineDiving [39], AQA-7 [24], and JIGSAWS [11]. The experimental results show that our method outperforms the state-of-the-art on Spearman's Rank Correlation and R-ℓ_2 metric. The source code of this research will be available. In summary, the main contributions of this work are three folds:

- To build a generic AQA model that conforms to occupational rules, we propose a novel target-aware contrastive regression framework. We directly model the mapping relation between global features and scores to learn the occupational basis rules and further assist in comparing the difference scores of local sub-actions.
- To eliminate the effect of irregular information distribution and subjective noise, we propose a new two-path contrastive regression framework. We fuse information from different visual fields of a video to learn a generic feature representation. Then, we further compare differences between different video pairs and remove the effect of subjective noise from a single visual field.
- We conduct extensive experiments on four public datasets to verify the effectiveness of our method. We achieve state-of-the-art performance on Spearman's Rank Correlation and R-ℓ_2 metric without fine-grained information and elaborate structural design.

The rest of the paper is organized as follows: In Section 2, we review related work. In Section 3, we present the specific framework and details of implementing our proposed T^2CR. And in Section 4, we conduct extensive experiments to validate the effectiveness of our approach. Finally, we conclude in Section 5.

2. Related work

2.1. Action quality assessment

There has been significant social interest in objectively assessing the degree of human action performance (*e.g.*, competitive sports and professional operations). Gordon [12] pioneered the exploration of the feasibility of automatic assessment of action videos. Pirsiavash et al. [27] first formulated the AQA task, introducing a learning approach to extract and regress spatial-temporal pose features of the human body into scores. They took the first step towards advances in action quality assessment.

Since then, AQA has acquired wide attention and developed rapidly, with mainstream works formulating it as a direct modeling regression between action features and judges' scores. For example, Parmar and Tran Morris [26] first utilized deep learning models, proposing C3D-SVR and C3D-LSTM to predict scores and a multiple-category dataset to explore full motion models applicable to multiple motion scenarios. Xu et al. [38] proposed an architecture containing two complementary LSTMs, combining self-attention to learn multi-scale video features. Pan et al. [23] built joint relation graphs in spatial and temporal terms to model the interaction between human joints, using posture information to assess human motion changes. Parmar and Morris [24] proposed a new multi-motion scenario dataset (AQA-7), introducing that knowledge transfer between different action scenarios is feasible. Parmar and Morris [25] proposed a larger multi-task dataset (MTL-AQA), suggesting that AQA learns to explain three related tasks (fine-grained action recognition, commentary generation, and estimation of scores). Tang et al. [35] proposed an uncertainty-aware score distribution learning (USDL) that describes AQA as the probability of different scores to reduce the underlying ambiguity in judges' subjective assessments. They further use MUSDL to simulate multiple judges' scoring rules in a realistic event. Dong et al. [6] introduced a multiple hidden substage learning and fusion network to assess athletes' performance by segmenting the video into five sub-stages. Wang et al. [36] introduced a single object tracker and a tube self-attention module (TSA) to focus on the athletes' trajectories, effectively generating rich spatial-temporal contextual information. Zhang et al. [45] used self-supervised learning of unlabeled video to recover masked segment features and adversarial learning to align labeled and unlabeled sample representations. Zhou et al. [48] proposed a hierarchical graph convolutional network (GCN) to improve assessment performance by correcting semantic information confusion, capturing local dynamics, and clustering continuous actions.

Recently, several methods compared input videos with exemplar videos to think about AQA from a new perspective. Jain et al. [14] proposed a reference guided regression (RGR) that uses the Siamese network to compare the similarity of input motion video pairs to assess the final score. Yu et al. [42] proposed a new contrastive regression framework (CoRe) to learn and predict the difference scores between paired input videos. Li et al. [16] proposed a new pairwise contrastive learning network (PCLN) to learn the subtle differences between videos. Xu et al. [39] first proposed a large-scale fine-grained diving action dataset (FineDiving) that uses fine-grained information to understand the action process and detect differences between paired videos at each step. Bai et al. [1] proposed a novel temporal parsing transformer with two novel loss functions to extract and compare the differences in fine-grained temporal part-level representational between videos. Unlike the above methods, which directly regress scores or just compare differences, our approach models the relation between global spatial-temporal features and scores and understands the scoring rules of the occupational domain to achieve local differences assessment in line with the rules.

2.2. Multi-view information in action videos

Multi-view learning has been widely discussed in the community and has been applied to many computer vision tasks. In recent years, there has been great success in employing multi-view information in human action recognition and understanding [37,13,15].

Simonyan et al. [31] first proposed a two-stream network consisting of two 2D convolutional neural networks (CNN) that fuse RGB and optical flow frames for high-performance recognition. Zheng et al. [47] encouraged consistent sparse representations across videos from different camera views to achieve cross-view action recognition. Shahroudy et al. [30] proposed a novel deep autoencoder based on a shared specific feature factorization network that exploits the complementary properties of RGB and depth modalities to fuse multimodal information. Feichtenhofer et al. [9] used fast and slow pathways to extract different video views, capturing spatial semantics at low frame rates and fine temporal motion at high frame rates. Inspired by this, several works [32,21] have benefited from using multi-view information from different sampled versions of the video in various tasks. Yan et al. [40] proposed a multi-view transformer (MTV) for video recognition, modeling different spatial-temporal resolutions and using independent

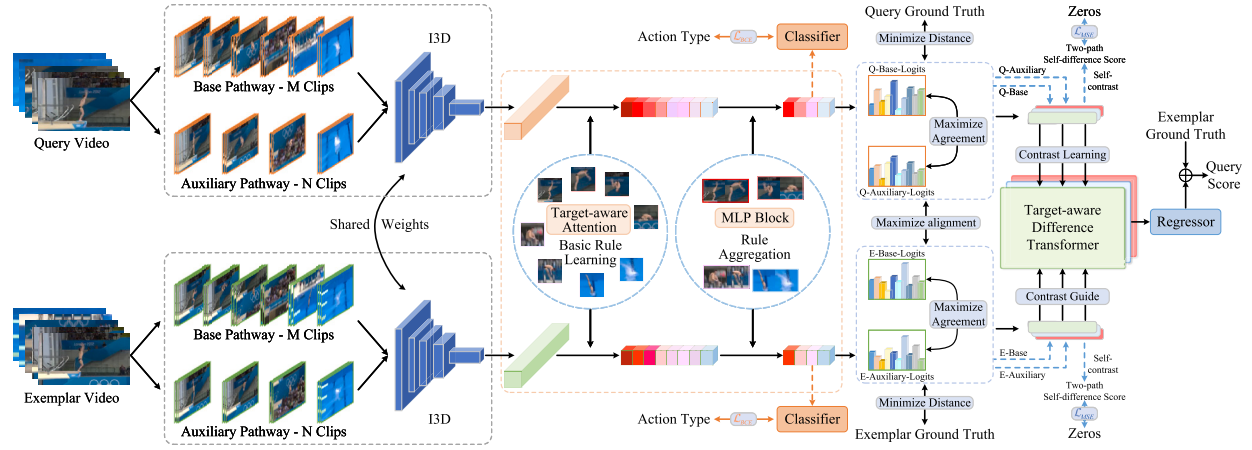

Fig. 3. The framework of our proposed two-path target-aware contrastive regression (T^2CR). We sample the input (query, exemplar) video pairs using two visual field paths with shared weights. The base path samples video clips normally, while the auxiliary path samples sparsely to obtain fewer clips. The I3D backbone is adopted to extract video global spatial-temporal features. We propose a target-aware attention module and introduce a prior knowledge of "action type" to build a "feature-score" model to learn the basic rules directly. We then maximize the agreement between the two paths and fuse different visual field information. Finally, we propose a target-aware difference transformer and use self-difference learning between the two paths to understand rich, subtle differences, further guiding the model in focusing on local action differences between video pairs and predicting the difference scores.

encoders of different views with lateral connections to fuse information. Liang et al. [18] proposed a novel View Knowledge Transfer Network (VKTNet) for multi-view action recognition, leveraging conditional generative adversarial networks (cGAN) to transfer view knowledge and a Siamese Scaling Network (SSN) for decision result fusion.

The fusion of multi-view information enables the capture of common video features. It leverages the complementarity among different views, eliminating noise interference from a single view and effectively recognizing video motion information. For AQA, multi-view learning aligns more with the assessment process involving multiple judges in real sports scenarios, avoiding the formation of inherent representations limited to a single view. Our method learns from two different sampled views of videos, extracting commonalities between the two paths while reducing noise interference. Additionally, we further leverage the self-difference information between the two paths to guide the model in better distinguishing the differences between the query and exemplar videos. Moreover, this approach embodies the concept of self-supervised learning, allowing for the acquisition of richer action information from the currently limited dataset.

3. Proposed method

In this section, we describe our T^2CR in detail. The main framework is shown in Fig. 3. Our method constructs a new two-path target-aware Contrastive Regression framework, which aims to propose a more compliant method for action quality assessment in line with occupational rules.

3.1. Target-aware contrastive regression

Problem formulation. For a given input video $\mathcal{Q}$, an exemplar video $\mathcal{E}$ is randomly selected to form a video pair with $\mathcal{Q}$. Our target-aware contrastive regression can be formulated as two types, namely direct regression and contrastive regression, which can be formulated as follows:

$$\hat{s}_{\mathcal{Q}}^{1} = \mathcal{R}\left(\mathcal{F}\left(\mathcal{Q}|\Theta\right)|\mathcal{W}\right), \tag{1}$$

$$\hat{s}_{\mathcal{Q}}^{2} = \mathcal{R}\left(\mathcal{C}\left(\mathcal{F}\left(\mathcal{Q}|\Theta\right), \mathcal{F}\left(\mathcal{E}|\Theta\right)|\phi\right)|\mathcal{W}\right) + s_{\mathcal{E}}, \tag{2}$$

$$\hat{s}_{\mathcal{Q}} = \frac{\hat{s}_{\mathcal{Q}}^{1} + \hat{s}_{\mathcal{Q}}^{2}}{2}, \tag{3}$$

where $\mathcal{F}$, $\mathcal{C}$, and $\mathcal{R}$ are the feature extraction, contrast difference, and regression model parameterized by Θ, ϕ, and $\mathcal{W}$, respectively. $\hat{s}_{\mathcal{Q}}^{1}$ is the score gained from regressing $\mathcal{Q}$ directly, $\hat{s}_{\mathcal{Q}}^{2}$ is the score based on the action quality label $s_{\mathcal{E}}$ of $\mathcal{E}$ and the contrastive regression of the differences between $\mathcal{Q}$ and $\mathcal{E}$, and $\hat{s}_{\mathcal{Q}}$ is the final prediction score of $\mathcal{Q}$.

Target-aware attention. We first introduce the feature extraction model, which consists of an I3D [2] backbone and target-aware attention. We follow [35,42] to segment the video into M overlapping clips containing 16 frames and use I3D to extract features from the clips. The extraction process for different clips shares the same parameters. Afterward, unlike traditional AQA methods that use pooling layers to fuse features to regress scores directly, our approach aims to further enhance the features by incorporating attention mechanisms to discern the significance and types of actions, introducing the prior knowledge to learn the corresponding career rules.

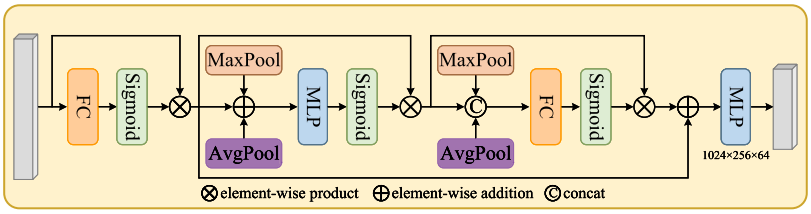

Fig. 4. The detailed architecture of the target-aware attention module.

To effectively learn career rules and model the relationship between global features and assessment scores, we propose a target-aware attention module. This module utilizes a hierarchical attention mechanism to explore the "action-rule" associations and capture the importance of different actions and their corresponding rules. Specifically, we extract video features with dimensions of $T \times C$, where T represents the global spatial-temporal features and C represents the number of video clips.

In the hierarchical attention module, we employ global attention along the T dimension to focus on the actions at all time steps and extract the temporal dynamics among them. This enables our model to comprehend actions' sequential nature and relationships. Furthermore, attention is applied along the C dimension to emphasize the importance of different video clips, allowing the model to extract fundamental action-related information at the clip level. It aims to allow the model to identify key moments and patterns contributing to a comprehensive understanding of career rules by paying attention to key clips. We lastly employ attention once again along the T dimension to capture subtle action details and further deepen the understanding of career rules. This multi-level attention mechanism aids the model in comprehending the fundamental principles and underlying regularities of career rules, ultimately enhancing the model's performance in action recognition and understanding. For an input feature $f_{T\times C}$, the process of obtaining the enhanced feature $f'_{T\times C}$ can be expressed as:

$$f'_{T\times C} = \sum_{t=1}^{T} \alpha_t^{(3)} \left(\sum_{c=1}^{C} \alpha_c^{(2)} \left(\sum_{t=1}^{T} \alpha_t^{(1)} f_{T\times C} \right) \right), \tag{4}$$

where $\alpha_t^{(1)}$, $\alpha_c^{(2)}$, and $\alpha_t^{(3)}$ represent the weight coefficients of the hierarchical attention at each of the three levels. Then, the feature-length is gradually reduced by an MLP module, which aggregates rules to further refine the action features into the latent embeddings $\left\{ f_{\mathcal{Q}_M}, f_{\mathcal{E}_M} \right\}$. The specific architecture of the target-aware attention module is shown in Fig. 4. After enhancing the features and aggregating the rules through this module, we further introduce the dive numbers of videos as the "action type" labels. By employing a fully connected layer and Softmax, we classify $\left\{ f_{\mathcal{Q}_M}, f_{\mathcal{E}_M} \right\}$ to guide the model in understanding the types and difficulty levels of the video actions. Additionally, we directly regress this feature to obtain the complete assessment score, establishing a rule-based mapping between global features and scores. Guided by this prior knowledge, the model is instructed to understand the career rules embedded in the video actions.

Unlike the latest AQA method [1] that extracted fine-grained temporal part-level representations by introducing new loss functions, our method focuses on global spatial-temporal feature representations, learning the attention weights of each action and aggregating representative rules. Our constraint between global features and assessment scores and the prior knowledge of "action type" help the model learn to understand career rules.

Contrast of differences. We now introduce the contrast difference model. The contrast difference model correlates spatial-temporal features between the query and exemplar instances to explore the degree of similarity and difference. We propose a target-aware difference transformer with multi-head cross-attention and MLP blocks to mine the correspondence of contextual semantic information between features. Specifically, we obtain M fine-grained action clips $\left\{ f_{\mathcal{Q}_M}, f_{\mathcal{E}_M} \right\} \in \mathbb{R}^d$ through target-aware attention, which are aligned in the category of rule aggregation. We then learn the corresponding spatial and temporal relations between the query and the exemplar, focusing on the consistent rule targets and exploring the action differences to generate new features. We utilize $f_{\mathcal{E}_M}$ to provide contrastive guidance for $f_{\mathcal{Q}_M}$ to predict the difference score $\Delta\hat{s}_Q$. The learning process can be described as:

$$f_{\mathcal{Q},\mathcal{E}} = Softmax\left(\frac{f_{\mathcal{Q}_M} W_Q \left(f_{\mathcal{E}_M} W_K \right)^T}{\sqrt{d}} \right) f_{\mathcal{E}_M} W_V, \tag{5}$$

$$\Delta\hat{s}_Q = \mathcal{R}\left(f_{\mathcal{Q},\mathcal{E}} | \mathcal{W} \right), \tag{6}$$

where $f_{\mathcal{Q}_M}$ and $f_{\mathcal{E}_M}$ serve as query and key-value pairs, respectively, with weight matrices W_Q, W_K, and W_V, and normalization factor $\sqrt{d}$. $f_{\mathcal{Q},\mathcal{E}}$ is the generated new difference feature. Then, we based on Eq. (2) to obtain the contrastive regression scores $\hat{s}_Q^2 = \Delta\hat{s}_Q + s_{\mathcal{E}}$. During training, $f_{\mathcal{Q}_M}$ also guides $f_{\mathcal{E}_M}$.

In our framework, we only use the transformer decoder [7] to compare feature differences without using the encoder module to encode global career rules. The reason why the encoder module is ineffective may be that it smooths temporal representations when

capturing global context information, which ignores the career rules implied by the temporally connected actions. In contrast, our target-aware attention retains temporal representations and focuses on the importance of each action more simply and effectively.

Optimization and inference. We regress features into scores using three successive combinations of fully connected layers and ReLU. Each clip gets a score, and the mean of the scores of M clips is the total prediction score of that video. During training, we optimize the model by minimizing the mean square error between the prediction scores and the ground truth labels and the cross-entropy loss J_{cls} for action category classification. Furthermore, to better explore the differences between videos, we incorporate the Soft-DTW [5] loss J_{aln} to align the latent embeddings of the spatial-temporal features of the two videos. This enables the model to learn the adaptive alignment of the temporal structures between actions, resulting in a more robust and accurate contrast of action differences. The objective function of the target-aware contrastive regression can be represented as:

$$\mathcal{L}_{MSE}(\hat{y}, y) = \|\hat{y} - y\|^2, \tag{7}$$

$$J_{reg} = \mathcal{L}_{MSE}\left(\hat{s}_Q^1, s_Q\right) + \mathcal{L}_{MSE}\left(\Delta\hat{s}_Q, s_Q - s_{\mathcal{E}}\right), \tag{8}$$

$$J_{cls} = -\sum_{\mathfrak{n}=1}^{\mathbb{N}} P_{\mathfrak{n}} \log \hat{P}_{\mathfrak{n}}, \tag{9}$$

$$J_{aln} = dtw_{\gamma}\left(f_{\mathcal{Q}_M}, f_{\mathcal{E}_M}\right) = -\gamma \log \sum_{A \in A_{\mathfrak{q},\mathfrak{e}}} e^{-\langle A, D\rangle/\gamma}, \tag{10}$$

where $\mathcal{L}_{MSE}$ is the mean squared error, J_{reg} is the objective function of the total score regression, s_Q and $s_{\mathcal{E}}$ denote the action quality labels of the query $\mathcal{Q}$ and the exemplar $\mathcal{E}$, $\mathbb{N}$ denotes the total number of action types, P and $\hat{P}$ represent the true and predicted probability, $\gamma > 0$ is a smoothing parameter, $A_{\mathfrak{q},\mathfrak{e}} \subset \{0,1\}^{\mathfrak{q}\times\mathfrak{e}}$ is the set of alignment matrices of $\left\{f_{\mathcal{Q}_M}, f_{\mathcal{E}_M}\right\}$ feature sequences, D is the distance matrix of the two sequences, and the inner product $\langle A, D\rangle$ is the path cost sum. Following previous works [42,39,1], a multi-exemplar voting strategy is used in the contrast regression step during testing. For a test sample $\mathcal{V}_t$, K samples are randomly selected from the training set to form the video pair $\left\{\mathcal{V}_t, \mathcal{V}_{\mathcal{E}}^i\right\}_{i=1}^K$ and the exemplar action quality labels $\left\{s_{\mathcal{E}}^i\right\}_{i=1}^K$. The voting process can be written as:

$$\hat{s}_t^2 = \frac{1}{K}\sum_{i=1}^{K}\left(\mathcal{R}\left(\mathcal{C}\left(\mathcal{F}\left(\mathcal{V}_t|\Theta\right), \mathcal{F}\left(\mathcal{V}_{\mathcal{E}}^i|\Theta\right)\middle|\phi\right)\middle|\mathcal{W}\right) + s_{\mathcal{E}}^i\right). \tag{11}$$

3.2. Two-path contrastive regression

Problem formulation. As shown in Fig. 3, given a query and exemplar video pair $\{\mathcal{Q}, \mathcal{E}\}$, we sample them in two different ways: the base pathway segments the video into M clips, *i.e.*, $\left\{\mathcal{Q}_M, \mathcal{E}_M\right\}$, while the auxiliary pathway segments the video into N clips using sparse sampling, *i.e.*, $\left\{\mathcal{Q}_N, \mathcal{E}_N\right\}$, where $N < M$. Based on the target-aware contrastive regression, the process of obtaining the self-difference score $self\text{-}\Delta\hat{s}_Q$ and difference score $\Delta\hat{s}_Q$ from the two-path contrastive regression can be formulated as follows:

$$f_{\mathcal{Q}_M} = \mathcal{F}\left(\mathcal{Q}_M|\Theta\right), \tag{12}$$

$$f_{\mathcal{Q}_N} = \mathcal{F}\left(\mathcal{Q}_N|\Theta\right), \tag{13}$$

$$self\text{-}\Delta\hat{s}_Q = \mathcal{R}\left(\mathcal{C}\left(f_{\mathcal{Q}_M}, f_{\mathcal{Q}_N}\middle|\phi\right)\middle|\mathcal{W}\right), \tag{14}$$

$$\Delta\hat{s}_Q = \mathcal{R}\left(\mathcal{C}\left(\mathcal{M}\left(f_{\mathcal{Q}_M}, f_{\mathcal{Q}_N}\middle|\psi\right), \mathcal{M}\left(f_{\mathcal{E}_M}, f_{\mathcal{E}_N}\middle|\psi\right)\middle|\phi\right)\middle|\mathcal{W}\right), \tag{15}$$

where $\mathcal{M}$ is a model with parameters ψ to minimize the distance of similar samples in the two paths and to learn the commonality of information.

Two-pathway contrastive learning. Since the different number of clips on the base and auxiliary pathways, we use two convolutional modules to extend the N clips to M clips to meet the requirement that the dimensions of similar samples are the same in the pull operation. Each module consists of "3×3 convolution", "BatchNorm", and "ReLU". We then enhance the auxiliary pathway with the same target-aware attention module as the base, acquiring the features $\left\{f_{\mathcal{Q}_M}, f_{\mathcal{Q}_N}\right\}$ of the two pathways, which represent different visual field information of the same video and are regarded as similar samples. We use MSE and the following modified NT-Xent contrastive loss $\mathcal{L}_{cl}$ [4,33] to minimize the distance of similar samples:

$$h(\mu, \nu) = \exp\left(\frac{\mu^T\nu}{\|\mu\|_2\|\nu\|_2}/\tau\right), \tag{16}$$

$$\mathcal{L}_{cl}\left(f_{\mathcal{Q}_M}^j, f_{\mathcal{Q}_N}^j\right) = -\log\frac{h\left(f_{\mathcal{Q}_M}^j, f_{\mathcal{Q}_N}^j\right)}{h\left(f_{\mathcal{Q}_M}^j, f_{\mathcal{Q}_N}^j\right) + \sum_{k=1}^{B}\mathbb{1}_{\{k\neq j\}}h\left(f_{\mathcal{Q}_M}^j, f_q^k\right)}, \tag{17}$$

where h is the exponential of the cosine similarity measure, the indicator $\mathbb{1}_{\{k\neq j\}} \in \{0,1\}$ is 1 when $k \neq j$, τ is the temperature hyperparameter, B is a minibatch in training, and $q \in \{Q_M, Q_N\}$. The instances of the same video in two pathways have the same category label and similar semantic representations. However, the original NT-Xent loss applies similarity comparison directly to different instances without considering high-level semantic information, which may unintentionally encourage different representations for two similar pathways. Therefore, we use the modified NT-Xent contrastive loss that takes all path pairs $\left\{f^j_{Q_M}, f^j_{Q_N}\right\}$ from the same video as positive samples, and all $\left\{f^j_{Q_M}, f^k_q\right\}$ with $k \neq j$ and $q \in \left\{Q_M, Q_N\right\}$ as negative samples. This aims to focus the model on the highly consistent action types and difficulty levels between the two pathways, providing valuable information for understanding career rules.

In pulling closer similar samples, the model continuously explores the commonality of semantic information of the two visual fields. We use these commonalities and the score labels to build a relation model that guides the extraction of representative features that are more consistent with the career rules. We then fuse the commonalities in the two visual fields to remove the effects of irregular information distribution and subjective noise. Specifically, we use a lightweight self-attention [22] to mine similarities between commonalities and temporal context correlations. The resulting fused features are then used to represent the video for contrastive learning.

Self-difference contrastive learning. Considering the highly consistent semantic information between the two paths of the same video, they should ideally have the same quality score for the AQA task. This implies that the score difference between paths is theoretically 0, which serves as vital supervisory information and can be directly applied without additional annotation data. Therefore, we further perform self-difference contrastive learning on video features to compare the self-differences between the two paths of the same video. We input $\left\{f_{Q_M}, f_{Q_N}\right\}$ into the target-aware difference transformer, sharing parameters with the difference contrast between the two videos, to learn more subtle feature differences within the video itself. This is particularly important for AQA tasks that require capturing subtle action variations. The self-difference features of the video are also passed through the difference-score regressor to obtain the self-difference score $self\text{-}\Delta\hat{s}_Q$. By learning from richer information, we aim to establish a more accurate relationship model between features and scores. We impose the MSE constraint (J^{self}_{reg}) on $self\text{-}\Delta\hat{s}_Q$, which can be expressed as:

$$J^{self}_{reg} = \mathcal{L}_{MSE}\left(self\text{-}\Delta\hat{s}_Q, 0\right). \tag{18}$$

Optimization and inference. Based on the inference process of target-aware contrastive regression, two-path contrastive regression focuses more on the complementarity of information between different visual fields. In training, the complete objective function of our method is:

$$J = \lambda_1 J_{reg} + \lambda_2 J_{cls} + \lambda_3 J_{aln} + \lambda_4 J^{self}_{reg} + \lambda_5 \mathcal{L}_{cl}\left(f_{Q_M}, f_{Q_N}\right), \tag{19}$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 represent the balancing weights for each loss.

4. Experiments

4.1. Datasets

We conduct experiments on four mainstream AQA datasets including MTL-AQA [25], FineDiving [39], AQA-7 [24], and JIGSAWS [11]. We follow the criteria proposed by the datasets and previous works [35,42].

MTL-AQA [25] is a large-scale dataset commonly used in AQA. It consists of 16 different diving events and contains 1412 samples. MTL-AQA has rich data scenarios containing male and female athletes, single and double diving, 3 m springboard, and 10 m platform. In addition, MTL-AQA additionally annotates action categories and action commentary besides AQA scores. We follow the criteria proposed by the MTL-AQA dataset to use 1059 samples as the training set and 353 samples as the testing set.

FineDiving [39] is a recently proposed large-scale fine-grained diving dataset. It contains 3000 diving samples from the Olympics, World Cup, World Championships, and European Aquatics Championships, covering 52 action types, 29 sub-action types, and 23 difficulty degree types. In addition, FineDiving is annotated with fine-grained annotations such as action type, sub-action type, coarse-grained and fine-grained temporal boundaries, and action scores besides AQA scores. We follow the criteria proposed by the FineDiving dataset to use 75% of the samples as the training set and the remaining 25% as the testing set.

AQA-7 [24] is a multi-sport dataset widely used in AQA. It consists of 1189 samples from 7 scenarios, including 370 samples from diving - 10 m platform, 176 samples from gymnastic vault, 175 samples from big air skiing, 206 samples from big air snowboarding, 88 samples from synchronous diving - 3 m springboard, 91 samples from synchronous diving - 10 m platform, and 83 samples from trampoline. We follow the criteria proposed by the AQA-7 dataset and exclude the trampoline category where the videos are too long, using 803 samples as the training set and 303 as the testing set.

JIGSAWS [11] is a commonly used surgical action dataset. It consists of 3 surgical scenarios: “Suturing (S)”, “Needle Passing (NP)” and “Knot Tying (KT)”. The score label of each video in JIGSAWS consists of several scores under different assessment rules, and the final score is the sum of them. To facilitate comparison with previous work [23,35], we similarly only use the left view in the dataset and conduct experiments using a similar four-fold cross-validation strategy.

Table 1
Comparisons of performance with existing methods on AQA-7. We mark the best (bold) and second-best (underlined) results.

Sp. Corr.↑	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3 m	Sync. 10 m	Avg. Corr.
ST-GCN [41]	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-LSTM [26]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
C3D-SVR [26]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
JRG [23]	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL [35]	0.8099	0.7570	0.6538	**0.7109**	0.9166	0.8878	0.8102
CoRe [42]	0.8824	0.7746	0.7115	0.6624	0.9442	0.9078	0.8401
NL-Net [36]	0.8296	0.7938	0.6698	0.6856	0.9459	0.9294	0.8418
TSA-Net [36]	0.8379	0.8004	0.6657	0.6962	**0.9493**	0.9334	0.8476
TPT [1]	**0.8969**	0.8043	**0.7336**	0.6965	0.9456	0.9545	0.8715
HGCN [48]	0.8867	0.7917	0.7326	0.6447	0.9213	0.9424	0.8501
T^2CR(Ours)	0.8901	**0.8393**	0.7139	0.7052	0.9418	**0.9558**	**0.8726**
R-ℓ_2(×100)↓	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3 m	Sync. 10 m	Avg. R-ℓ_2
C3D-SVR [26]	1.53	3.12	6.79	7.03	17.84	4.83	6.86
USDL [35]	0.79	2.09	4.82	4.94	0.65	2.14	2.57
CoRe [42]	0.64	1.78	3.67	3.87	0.41	2.35	2.12
TPT [1]	**0.53**	1.69	**2.89**	**3.30**	**0.33**	1.33	1.68
HGCN [48]	0.59	1.85	3.59	3.61	0.82	1.40	1.98
T^2CR(Ours)	0.59	**1.37**	3.07	3.32	0.38	**0.98**	**1.62**

4.2. Metrics

Spearman's rank correlation. We first measure the performance of our method using Spearman's rank correlation [23], which is commonly used in AQA. Spearman's rank correlation (ρ) is defined as:

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}}, \tag{20}$$

where p and q denote the ranking of the two series, respectively, $\rho \in [-1, 1]$ and the larger the better. Fisher's z-value [24] was used when measuring the average Spearman's rank correlation across multiple action categories.

Relative ℓ_2-distance. Following [42,39,1], we also use the relative ℓ_2-distance (R-ℓ_2) to measure the performance of our method. R-ℓ_2 can be defined as:

$$\text{R-}\ell_2 = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{|s_l - \hat{s}_l|}{s_{max} - s_{min}}\right)^2, \tag{21}$$

where s_l and $\hat{s}_l$ denote the ground-truth score and predicted score for the l-th sample, respectively. s_{max} and s_{min} denote the highest and lowest scores of the action, respectively. The lower of R-ℓ_2, the better performance.

4.3. Implementation details

We adopt the I3D model pre-trained on the Kinetics [2] dataset with an initial learning rate of 1e-4 and the remaining two-path target-aware contrastive regression module with an initial learning rate of 1e-3. We use the Adam optimizer and set the weight decay to 0. As in previous works [35,42,39], we extract 103 frames for each video in MTL-AQA and AQA-7 experiments and segment them into 10 overlapping clips with 16 frames. In experiments on FineDiving, we extract 96 frames to form 9 overlapping 16-frame clips. In JIGSAWS, we evenly sample 160 frames for each video and segment it into 10 non-overlapping 16-frame clips. We select sample videos using the degree of difficulty (DD) and the dive numbers (DN) annotation information in MTL-AQA and FineDiving, and only according to the coarse category of the video in AQA-7 and JIGSAWS. For a prior knowledge of "action type," we incorporate category learning into the MTL-AQA and FineDiving datasets with 58 and 52 types, respectively, based on the existing annotations. In addition, we set M to the above number of segmented clips in the two-path contrastive regression, N to 8, and K to 10 in the multi-exemplar voting strategy. τ, λ_1, λ_2, λ_3, λ_4, and λ_5 values are taken to be 0.1, 1, 1, 0.1, 10, and 1 respectively.

In addition, we report ablation experiments based on the following baseline and different versions of our method:

- I3D+MLP(Baseline): The baseline uses the I3D backbone to extract features, proceeding through a simple two-layer MLP with ReLU non-linearity for action refinement classification and capturing video differences by directly subtracting features. A three-layer MLP is finally used to predict the difference scores.
- I3D+MLP+DR: We introduce the idea of direct regression (DR) into the baseline. DR establishes a relation mapping between the whole features and the scores and optimizes them using MSE loss.
- $\mathcal{F}+C$, $\mathcal{F}+C^{\star}$: Our target-aware contrastive regression framework (introduced in Sec. 3.1). We replace the two-layer MLP of I3D+MLP+DR with our target-aware attention and capture differences between videos using the target-aware difference trans-

Table 2
Comparisons of performance with existing methods on the MTL-AQA dataset. (w/o DD) and (w/ DD) denote random and use the degree of difficulty to select exemplars.

Method(w/o DD)	Sp. Corr.↑	R-ℓ_2(×100)↓
C3D-SVR [26]	0.7716	–
C3D-LSTM [26]	0.8489	–
MSCADC-STL [25]	0.8472	–
C3D-AVG-STL [25]	0.8960	–
MSCADC-MTL [25]	0.8612	–
C3D-AVG-MTL [25]	0.9044	–
USDL [35]	0.9066	0.654
MUSDL [35]	0.9158	0.609
CoRe [42]	0.9341	0.365
NL-Net [36]	0.9422	–
TSA-Net [36]	0.9393	–
TPT [1]	0.9451	0.322
HGCN [48]	0.9390	0.360
T^2CR(Ours)	**0.9464**	**0.308**
Method (w/ DD)	Sp. Corr.↑	R-ℓ_2(×100)↓
USDL [35]	0.9231	0.468
MUSDL [35]	0.9273	0.451
CoRe [42]	0.9512	0.260
TPT [1]	0.9607	0.238
HGCN [48]	0.9563	0.235
T^2CR(Ours)	**0.9638**	**0.222**

Table 3
Comparisons of performance with existing methods on the FineDiving dataset. (w/o DN) and (w/ DN) denote random and use the dive numbers to select exemplars.

Method (w/o DN)	Sp. Corr.↑	R-ℓ_2(×100)↓
USDL [35]	0.8302	0.5927
MUSDL [35]	0.8427	0.5733
CoRe [42]	0.8631	0.5565
TSA [39]	0.8925	0.4782
T^2CR(Ours)	**0.9234**	**0.3522**
Method (w/ DN)	Sp. Corr.↑	R-ℓ_2(×100)↓
USDL [35]	0.8913	0.3822
MUSDL [35]	0.8978	0.3704
CoRe [42]	0.9061	0.3615
TSA [39]	0.9203	0.3420
T^2CR(Ours)	**0.9275**	**0.3305**

former. ⋆ indicates the introduction of a prior knowledge of "action type," evaluated only on the MTL-AQA and FineDiving datasets with annotation information.

- T^2CR: The proposed method in Section 3.

4.4. Experiment results

Results on AQA-7 dataset. We evaluate our method on the AQA-7 dataset in multi-sport scenarios and report the experiment results in Table 1. The first contrastive regression method, CoRe [42], devised a group-aware regression tree to convert traditional AQA regression to classification and smaller interval regression but required the construction of the regression tree based on the score labels of the dataset. Our method is universal and achieves better performance without elaborate structural design. We achieve a new state-of-the-art Avg. Corr.(0.8726) and Avg. R-ℓ_2(1.62) on AQA-7 dataset. It can be noted that our T^2CR performs better on the R-ℓ_2 metric, indicating that our assessment results are more accurate, effectively modeling the relationship between features and quality scores.

Results on MTL-AQA dataset. The comparisons of our method with existing methods on MTL-AQA are reported in Table 2. The top half of the table indicates selecting samples without a degree of difficulty, and the bottom half denotes using a degree of difficulty. Our method achieves a new state-of-the-art Spearman's Rank Correlation (0.9638) and R-ℓ_2(0.222) with DD labels. Under 'w/o DD,' our approach also achieves the best performance of 0.9464(Sp. Corr.) and 0.308(R-ℓ_2), outperforming existing works. To intuitively demonstrate the results of our method, we visualize our prediction results compared with the state-of-the-art method

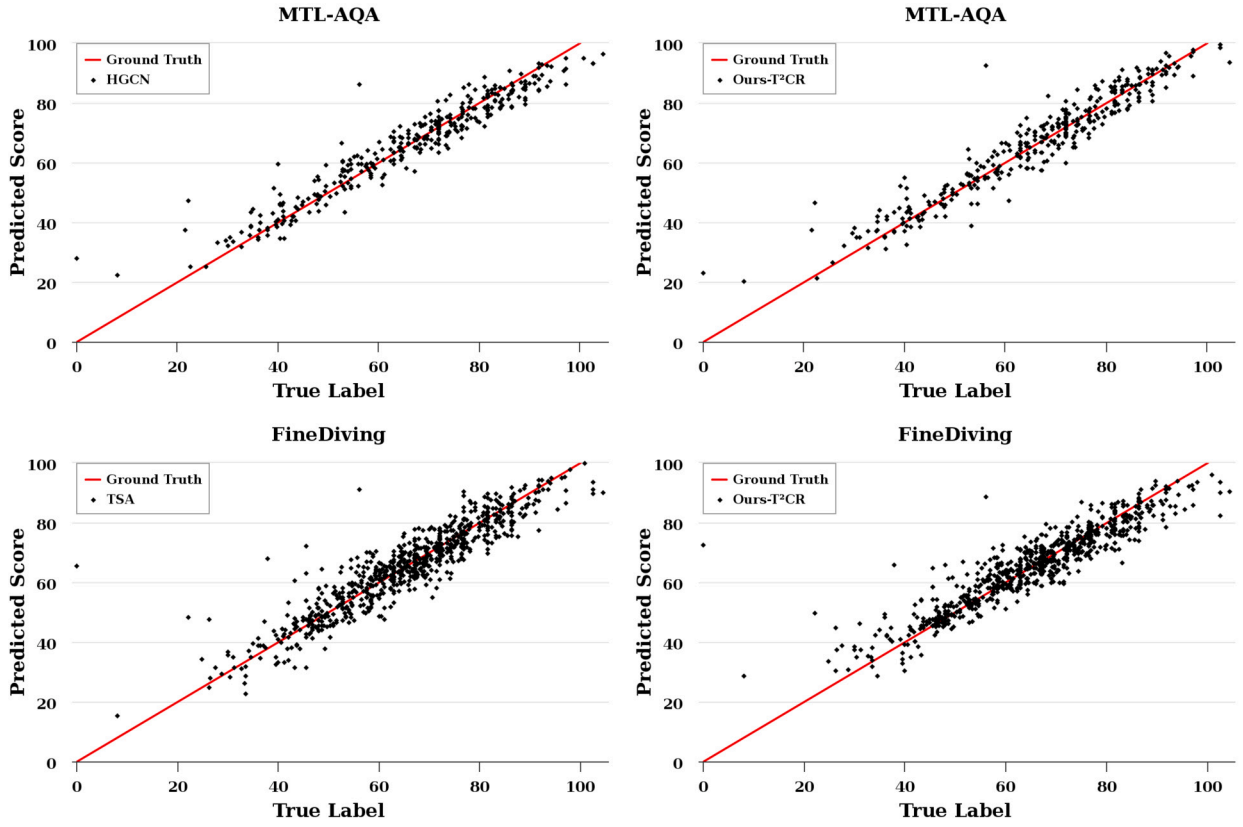

Fig. 5. Comparisons of scatter plots with state-of-the-art methods on MTL-AQA and FineDiving. Each point represents the predicted result for a video. The red line is the ground truth.

HGCN [48] on the R-ℓ_2 metric in the form of a scatter plot in Fig. 5 and show our training process compared with previous methods C3D-AVG [25], CoRe, TPT [1], HGCN, USDL and MUSDL [35] in Fig. 6. We can see that our predictions are much closer to the ground truth (red line) and that the training process is much smoother and more efficient in achieving the best results. The above analysis demonstrates our approach's powerful performance and robustness in various respects.

Results on FineDiving dataset. We report comparing our method with other AQA methods on FineDiving in Table 3. The top half of the table indicates selecting samples without dive numbers, and the bottom half denotes using dive numbers. The state-of-the-art method TSA [39] uses a temporal segmentation module to subdivide the action process into several sub-actions, learning the action process using fine-grained information that requires expert manual annotation. Our method performs better without fine-grained information. We see that T^2CR achieves 3.09% and 0.1260 significant improvements than TSA under Spearman's Correlation and R-ℓ_2 metric without DN labels, respectively. This demonstrates the effectiveness of introducing the prior knowledge of "action type" to guide the model in learning career rules. With the DN labels added, our method achieves 0.72% and 0.0115 improvements compared to TSA on two metrics. The fact that T^2CR does not rely on fine-grained annotation information further demonstrates the versatility and effectiveness of the proposed method in learning action spatial-temporal information and capturing subtle changes more comprehensively. We also show a visual comparison with the TSA in Fig. 5. Our predictions are much closer to the ground truth (red line).

Results on JIGSAWS dataset. We finally conduct experiments on the surgical action dataset JIGSAWS. Table 4 reports the final experiment results. Our method achieves new balanced and state-of-the-art results of 0.91 on Avg. Corr. and 3.279 on Avg. R-ℓ_2. The results in Tables 1 to 4 demonstrate that the proposed method outperforms the state-of-the-art methods and can be applied to various scenarios for action quality assessment.

4.5. Ablation studies

Effects of the proposed components. Table 5 shows the results of ablation experiments on the MTL-AQA dataset to verify the performance of each part of T^2CR. I3D+MLP+DR improves by 0.95% and 0.086 over I3D+MLP on two metrics, indicating the feasibility of combining direct regression with contrastive regression to establish a basis rule model for global feature-score mapping and effectively guide the learning of detailed differences. Incorporating our target-aware attention module further enhances the features to learn the underlying rules, resulting in an additional performance boost of 0.50% and 0.008. Introducing a prior knowledge of "action type" yields even better results, indicating that learning more accurate knowledge about career rules enhances

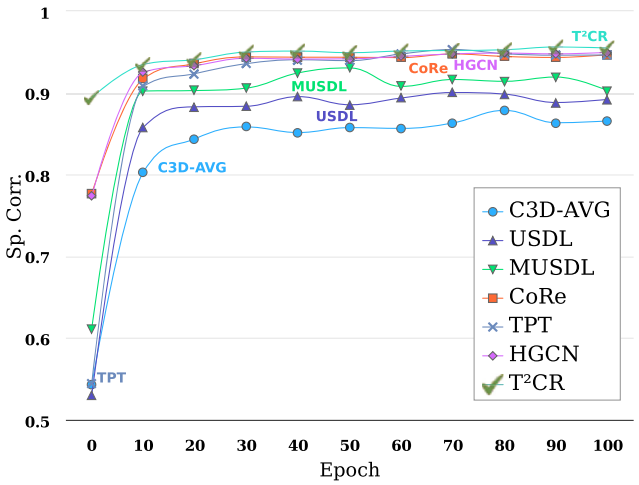

Fig. 6. Comparisons of the training process with existing methods on the MTL-AQA.

Table 4
Comparisons of performance with existing methods on the JIGSAWS dataset.

Sp. Corr.↑	S	NP	KT	Avg. Corr.
ST-GCN [41]	0.31	0.39	0.58	0.43
TSN [26]	0.34	0.23	0.72	0.46
JRG [23]	0.36	0.54	0.75	0.57
USDL [35]	0.64	0.63	0.61	0.63
MUSDL [35]	0.71	0.69	0.71	0.70
CoRe [42]	0.84	0.86	0.86	0.85
TPT [1]	0.88	0.88	**0.91**	0.89
HGCN [48]	0.89	**0.91**	0.90	0.90
T^2CR(Ours)	**0.93**	0.89	0.89	**0.91**
R-ℓ_2($\times$100)↓	S	NP	KT	Avg. R-ℓ_2
CoRe [42]	5.055	5.688	2.927	4.556
TPT [1]	2.722	5.259	3.022	3.668
HGCN [48]	4.784	3.927	3.380	4.031
T^2CR(Ours)	**2.203**	**5.119**	**2.516**	**3.279**

Table 5
Ablation experiments on the MTL-AQA dataset.

Method	Ablation	Sp. Corr.↑	R-ℓ_2($\times$100)↓
I3D+MLP	baseline	0.9408	0.339
I3D+MLP+DR	+DR	0.9503	0.253
$\mathcal{F}+C$	+Target-aware	0.9553	0.245
$\mathcal{F}+C^\star$	+Action Type	0.9572	0.239
T^2CR(Ours)	+Two-pathway	**0.9638**	**0.222**

the model's understanding of action semantics. The emphasis on profession-specific rules is crucial for action quality assessment in our approach. Moreover, the performance is further improved when incorporating our two-path contrastive regression, achieving new state-of-the-art results of 0.9638 and 0.222 under the two metrics. The above results demonstrate the effectiveness of the components of T^2CR.

Effects of each loss function. We report ablation experiments on the MTL-AQA dataset about each loss of two-path contrastive regression and target-aware contrastive regression in Table 6 and 7, respectively. Both MSE and NT-Xent* are used to maximize the consistency of two paths for the same video. NT-Xent* denotes the modified NT-Xent loss we use. Self-Reg indicates the J_{reg}^{self} in Eq. (18) used to fit the self-difference scores between the two paths. The best results are achieved when maximizing consistency using MSE+NT-Xent*, possibly due to NT-Xent* pulling similar samples closer together and MSE further exploring the inherently regular distribution of different visual fields. Combining Self-Reg with different losses further improves the performance, indicating that learning the subtle differences between the two pathways helps the model explore commonalities and capture the action differences across different videos.

Table 6
Ablation experiments on the MTL-AQA dataset about the loss of two-path contrastive regression.

MSE	NT-Xent*	Self-Reg	Sp. Corr.↑	R-$\ell_2(\times 100)$↓
√	×	×	0.9584	0.239
×	√	×	0.9603	0.233
×	×	√	0.9591	0.237
√	√	×	0.9624	0.230
×	√	√	0.9628	0.226
√	×	√	0.9619	0.228
√	√	√	**0.9638**	**0.222**

Table 7
Ablation experiments on the MTL-AQA dataset about the loss of target-aware contrastive regression.

MSE	Cross-Entropy	Soft-DTW	Sp. Corr.↑	R-$\ell_2(\times 100)$↓
√	×	×	0.9605	0.233
√	×	√	0.9618	0.227
√	√	×	0.9621	0.226
√	√	√	**0.9638**	**0.222**

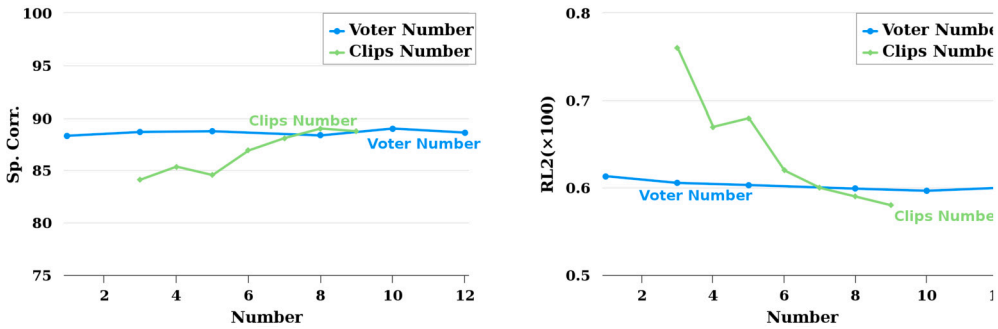

Fig. 7. Effects of the number of sparsely sampled clips of the auxiliary path (the green curves) and the number of exemplars for voting (the blue curves).

Table 8
Effects of the number of transformer decoder layers.

N_{layer}	1	3	5	7	10
Sp. Corr.↑	0.8226	**0.8901**	0.8734	0.8656	0.8648
R-$\ell_2(\times 100)$↓	0.68	**0.59**	0.66	0.64	0.64

Table 9
Effects of the number of multi-head cross-attention heads.

N_{head}	4	8	16	32
Sp. Corr.↑	0.8677	**0.8901**	0.8668	0.8579
R-$\ell_2(\times 100)$↓	0.65	**0.59**	0.67	0.71

On the other hand, in the target-aware regression, MSE serves as the foundational supervised loss used consistently in this framework. Meanwhile, Cross-Entropy loss represents the utilization of "action type" prior knowledge for classification learning, and Soft-DTW loss is employed for aligning the features of query and exemplar videos. It can be observed that incorporating Cross-Entropy loss and Soft-DTW loss can further improve the assessment's performance, demonstrating the importance of learning career rules, and aligning the action features of two videos enables more effective exploration of subtle differences between actions.

Effects of the number of sparsely sampled clips and exemplars for voting. Fig. 7 further reports our ablation experiments on the number of sparsely sampled clips of the auxiliary path (N) and the number of exemplars for voting (K) on the Diving category of AQA-7. We set the K to 10 for experiments on N and N to 8 for experiments on K. The best results are achieved when $N = 8$ and $K = 10$. The result of the voting number is consistent with the conclusion of previous works [42,39].

Effects of the number of transformer decoder layers. We conduct several experiments on the Diving category of the AQA-7 dataset. Table 8 summarizes the performance with different numbers of transformer decoder layers (denoted as N_{layer}), including

Table 10
Effects of the sampling strategies for sparse sampling.

Method	Sp. Corr.↑	R-ℓ_2(×100)↓
10-clip-2x	0.8839	0.62
8-clip-same	0.8749	0.65
8-clip-avg1	**0.8901**	**0.59**
8-clip-avg2	0.8799	0.64

Table 11
Effects of the number of target-aware attention layers.

$\alpha_t^{(1)}$	$\alpha_c^{(2)}$	$\alpha_t^{(3)}$	Sp. Corr.↑	R-ℓ_2(×100)↓
×	×	×	0.8755	0.68
√	×	×	0.8816	0.67
√	√	×	0.8891	0.63
√	√	√	**0.8901**	**0.59**

1, 3, 5, 7, and 10. We can see that when N_{layer} is raised from 1 to 3, Spearman's Correlation and R-ℓ_2 improve by 6.75% and 0.09, respectively. After that, the performance gradually decreases as N_{layer} rises. The possible reason is that too many transformer decoder layers lead to overfitting the training.

Effects of the number of multi-head cross-attention heads. To investigate the impact of the number of multi-head cross-attention heads (N_{head}), we conduct several ablation experiments on the Diving category of AQA-7 and present the results in Table 9. Our experiments demonstrate that increasing N_{head} from 4 to 8 results in improvements of 2.24% in Spearman's Correlation and 0.06 in R-ℓ_2. However, as the number of heads surpasses 8, the performance decreases consistently. This phenomenon may be attributed to the excessive convergence of multi-dimensional information caused by too many multi-head cross-attention heads, which results in noise interference and performance degradation.

Effects of the sampling strategies for sparse sampling. As shown in Table 10, we conduct several ablation experiments on the Diving category of AQA-7 to investigate the performance of different sampling strategies for sparse sampling. Using a 103-frame video clip as an example, the base path follows the previous works [35,42] using [0, 10, 20, 30, 40, 50, 60, 70, 80, 86] as the indices of beginning frames for the ten clips. We use "8-clip-same" to indicate that 8 clips are sparsely sampled in the same visual field as the base path, with the indices of beginning frames as [10, 20, 30, 40, 50, 60, 70, 80]. Both "8-clip-avg1" and "8-clip-avg2" split the 8 clips equally over different visual fields. Using 0 and 86 as the beginning and end of the indices, the 7 intervals in between would be 86/7 = 12.3. "8-clip-avg1" is rounded to adopt 12 as the interval, *i.e.*, [0, 12, 24, 36, 48, 60, 72, 84]. "8-clip-avg2" takes 12.3 as the interval, *i.e.*, [0, 12, 24, 36, 49, 61, 73, 86]. In addition, we also explore different rate processing strategies commonly used in video comparison learning, with "10-clip-2x" indicating a uniform sampling of 10 clips at 2x the rate. We can see that "8-clip-avg1" achieves the best results (namely 0.8901 and 0.59) on Spearman's rank Correlation and R-ℓ_2, and that both "8-clip-avg1" and "8-clip-avg2" are better than "8-clip-same", demonstrating the effectiveness of our method for fusing the information from multiple visual fields.

Effects of the number of target-aware attention layers. To investigate the effects of different layers of hierarchical attention mechanism on building the "feature-score" mapping relationship in our proposed target-aware attention module, we conduct ablation experiments on the Diving category and present the results in Table 11. The three attention weights $\alpha_t^{(1)}$, $\alpha_c^{(2)}$, and $\alpha_t^{(3)}$ correspond to the three components in Eq. (4) in our paper. The experimental results demonstrate that performance improves as our attention mechanism increases layer by layer. When using only the $\alpha_t^{(1)}$ component, which focuses on the importance of all actions in the temporal sequence and establishes the preliminary connection between "action-score", the performance is improved from 0.8755 Corr. to 0.8816 Corr. When $\alpha_c^{(2)}$ and $\alpha_t^{(3)}$ are further used, the performance improves from 0.8816 Corr. to 0.8901 Corr., indicating the effectiveness of our hierarchical attention mechanism. Learning the importance of actions, clips, and fine-grained rules can help understand professional rules and obtain more accurate assessment scores.

4.6. Visualization

To further demonstrate the effectiveness of our method in various action scenarios, we provide visualizations on all categories of the AQA-7 dataset using Intergated Gradients [17,34], as shown in Fig. 8. The odd rows represent input images, and the even rows are visualization results. Each model is trained solely on the data of the corresponding category. The results show that our method can effectively focus on action information, even in challenging scenarios such as low-light environments, small objects, and scenes with multiple people.

Moreover, we further visualize the comparison results of the Baseline and our T^{2}CR on the mainstream large-scale MTL-AQA dataset in Fig. 9. It can be observed that our method can focus more on the parts involved in the assessment (*e.g.*, body, splash, etc.) and less on irrelevant background information compared to the Baseline. The results indicate that the proposed method is more discriminative of action information and can effectively perceive important regions and mitigate the effects of background noise. This is in line with the motivation of our method.

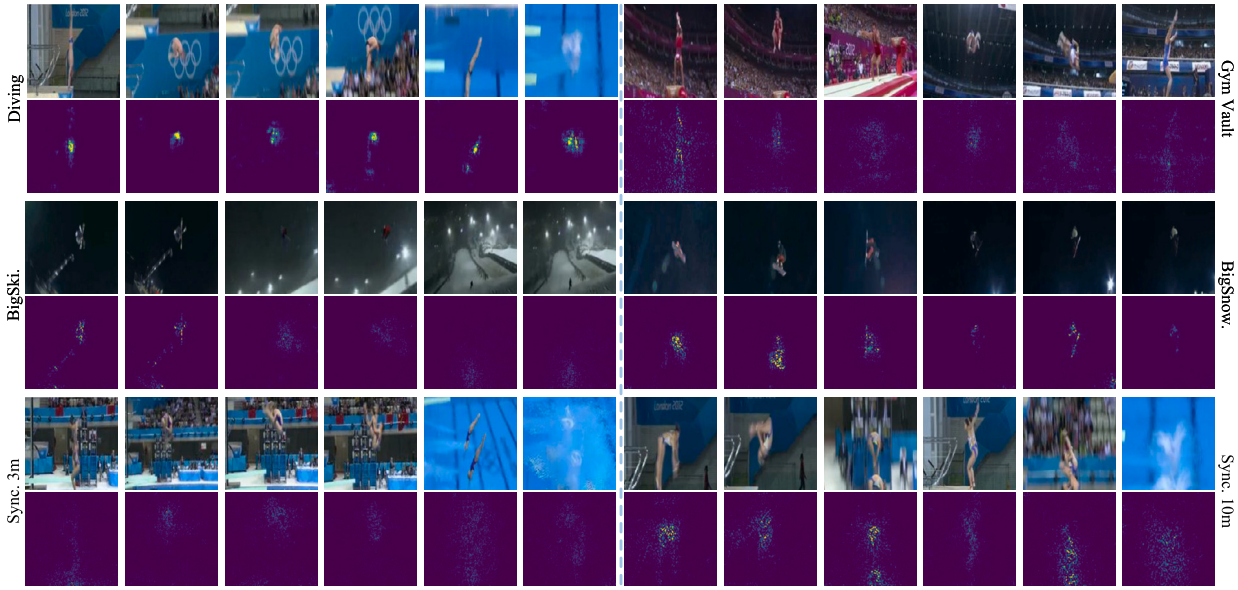

Fig. 8. Visualization of action scenarios for each category in the AQA-7 dataset. Our approach effectively focuses on motion information that is relevant to the assessment of all types of motion scenarios.

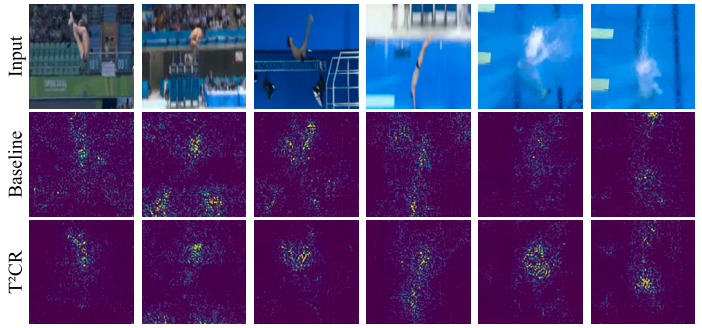

Fig. 9. Comparison of the visualization of the baseline and our T^2CR. Our method can focus on the important regions of the action quality assessment and mitigate the effects of background noise.

Case study. To better understand our model's implementation process, we present a case study with a visualization process on the MTL-AQA dataset in Fig. 10. For the given query and exemplar videos, we obtain both a full assessment score by directly regressing the whole video features and a relative score by comparing the differences in detail between the two videos. We then equalize the two scores to obtain a final score. Our results show that the full score obtained by directly regressing global features is larger, and the score obtained by contrastive learning is smaller. Our method combines the two scores to obtain a more accurate score.

5. Conclusion

In this paper, we have proposed the T^2CR framework for action quality assessment. To implement the framework, we have proposed target-aware contrastive regression, which fuses direct regression with contrastive regression, where direct regression learns occupational basis rules and contrastive regression learns difference scores based on rules. We have also proposed a two-path contrastive regression that fuses information from multiple visual fields to capture intrinsic feature representations and eliminate the effects of subjective noise from a single visual field. The extensive experiments on four AQA datasets have demonstrated that our approach can outperform state-of-the-art methods in various scenarios without fine-grained information and elaborate structural design.

Remark: Although our approach has shown effectiveness in AQA, there are some limitations and potential applications that can be further investigated in future research. Firstly, our method that utilizes the information from multiple visual fields imposes a high computational cost when assessing long-term action scenes. To explore long-term action quality assessment, we plan to explore a self-supervised approach to mitigate the subjective noise in the single visual field. Additionally, we plan to enhance the approaches to multi-view learning further, taking into full consideration the influence of various temporal and frequency domain information on quality assessment in action videos. Finally, an interesting potential application is to utilize multi-modal information, such as

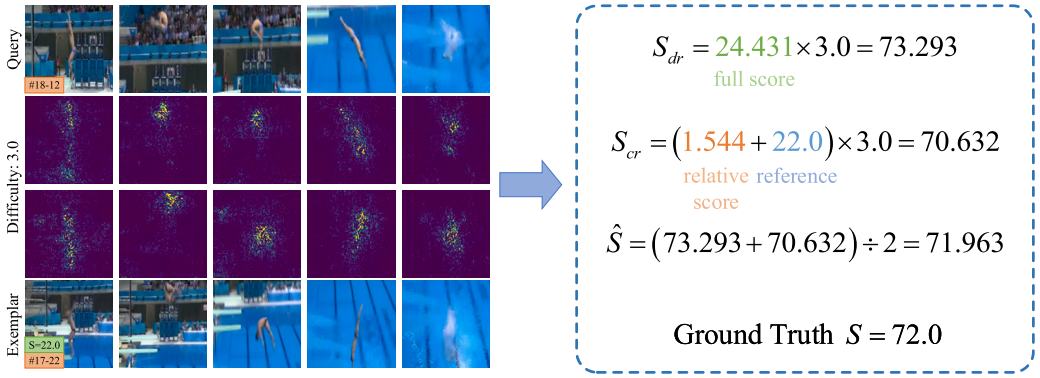

Fig. 10. Case study with visualization process. The query and exemplar videos have the same degree of difficulty (DD). S_{dr} and S_{cr} are the scores obtained from direct regression and contrastive regression respectively. Our method combines the two ideas to obtain a more accurate score $\hat{S}$. Moreover, our method can effectively focus on the relevant parts of the assessment.

combining visual and language information, which is consistent with human perception and can further enhance the model's ability to discriminate features. These potential applications could further improve the performance of AQA, and we will actively explore these avenues in the future.

CRediT authorship contribution statement

Xiao Ke: Conceptualization, Methodology, Software, Validation, Writing – review & editing. **Huangbiao Xu:** Data curation, Methodology, Software, Writing – original draft, Writing – review & editing. **Xiaofeng Lin:** Investigation, Visualization. **Wenzhong Guo:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant numbers 61972097, U21A20472]; the National Key Research and Development Plan of China [grant number 2021YFB3600503]; the Natural Science Foundation of Fujian Province [grant numbers 2021J01612, 2020J01494]; the Major Science and Technology Project of Fujian Province [grant number 2021HZ022007]; the Industry-Academy Cooperation Project of Fujian Province [grant number 2018H6010]; the Fujian Collaborative Innovation Center for Big Data Application in Governments; and the Fujian Engineering Research Center of Big Data Analysis and Processing.

References

[1] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, J. Wang, Action quality assessment with temporal parsing transformer, in: European Conference on Computer Vision, 2022, pp. 422–438.
[2] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
[3] A. Chandrasekar, T. Radhika, Q. Zhu, Further results on input-to-state stability of stochastic Cohen–Grossberg BAM neural networks with probabilistic time-varying delays, Neural Process. Lett. (2022) 1–23.
[4] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020, pp. 1597–1607.
[5] M. Cuturi, M. Blondel, Soft-dtw: a differentiable loss function for time-series, in: International Conference on Machine Learning, 2017, pp. 894–903.
[6] L.J. Dong, H.B. Zhang, Q. Shi, Q. Lei, J.X. Du, S. Gao, Learning and fusing multiple hidden substages for action quality assessment, Knowl.-Based Syst. 229 (2021) 107388.
[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[8] H. Doughty, W. Mayol-Cuevas, 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.

[9] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6202–6211.
[10] J. Gao, J. Pan, S. Zhang, W. Zheng, Automatic modelling for interactive action assessment, Int. J. Comput. Vis. 131 (2023) 659–679.
[11] Y. Gao, S.S. Vedula, C.E. Reiley, N. Ahmidi, 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, 2014.
[12] A.S. Gordon, Automated video assessment of human performance, in: Proceedings of AI-ED, 1995.
[13] S. Hu, X. Yan, Y. Ye, Joint specific and correlated information exploration for multi-view action clustering, Inf. Sci. 524 (2020) 148–164.
[14] H. Jain, G. Harit, A. Sharma, Action quality assessment using Siamese network-based deep metric learning, IEEE Trans. Circuits Syst. Video Technol. 31 (2020) 2260–2273.
[15] K. Keisham, A. Jalali, J. Kim, M. Lee, Multi-level alignment for few-shot temporal action localization, Inf. Sci. 119618 (2023).
[16] M. Li, H.B. Zhang, Q. Lei, Z. Fan, J. Liu, J.X. Du, Pairwise contrastive learning network for action quality assessment, in: European Conference on Computer Vision, 2022, pp. 457–473.
[17] Z. Li, W. Wang, Z. Li, Y. Huang, Y. Sato, Spatio-temporal perturbations for video attribution, IEEE Trans. Circuits Syst. Video Technol. 32 (2021) 2043–2056.
[18] Z. Liang, M. Yin, J. Gao, Y. He, W. Huang, View knowledge transfer network for multi-view action recognition, Image Vis. Comput. 118 (2022) 104357.
[19] T. Liu, Y. Ma, W. Yang, W. Ji, R. Wang, P. Jiang, Spatial-temporal interaction learning based two-stream network for action recognition, Inf. Sci. 606 (2022) 864–876.
[20] Y. Liu, N. Zhou, F. Zhang, W. Wang, Y. Wang, K. Liu, Z. Liu, APSL: action-positive separation learning for unsupervised temporal action localization, Inf. Sci. 630 (2023) 206–221.
[21] Z. Liu, Y. Wu, Z. Yin, Multi-layer representation for cross-view action recognition, Inf. Sci. 120088 (2024).
[22] S. Mehta, M. Rastegari, Separable self-attention for mobile vision transformers, Trans. Mach. Learn. Res. 2023 (2023).
[23] J.H. Pan, J. Gao, W.S. Zheng, Action assessment by joint relation graphs, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6331–6340.
[24] P. Parmar, B. Morris, Action quality assessment across multiple actions, in: IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1468–1476.
[25] P. Parmar, 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.
[26] P. Parmar, B. Tran Morris, Learning to score olympic events, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 20–28.
[27] H. Pirsiavash, C. Vondrick, A. Torralba, Assessing the quality of actions, in: European Conference on Computer Vision, Springer, 2014, pp. 556–571.
[28] S. Qiu, T. Fan, J. Jiang, Z. Wang, Y. Wang, J. Xu, T. Sun, N. Jiang, A novel two-level interactive action recognition model based on inertial data fusion, Inf. Sci. 633 (2023) 264–279.
[29] T. Radhika, A. Chandrasekar, V. Vijayakumar, Q. Zhu, Analysis of Markovian jump stochastic Cohen–Grossberg bam neural networks with time delays for exponential input-to-state stability, Neural Process. Lett. (2023) 1–18.
[30] A. Shahroudy, T.T. Ng, Y. Gong, G. Wang, Deep multimodal feature analysis for action recognition in RGB+ D videos, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 1045–1058.
[31] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 27 (2014).
[32] A. Singh, O. Chakraborty, A. Varshney, R. Panda, R. Feris, K. Saenko, A. Das, Semi-supervised action recognition with temporal contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10389–10399.
[33] A. Singh, O. Chakraborty, A. Varshney, R. Panda, R. Feris, K. Saenko, A. Das, Semi-supervised action recognition with temporal contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10389–10399.
[34] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, 2017, pp. 3319–3328.
[35] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, 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.
[36] S. Wang, D. Yang, P. Zhai, C. Chen, 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.
[37] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J.T. Zhou, X. Bai, Action recognition for depth video using multi-view dynamic images, Inf. Sci. 480 (2019) 287–304.
[38] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.G. Jiang, X. Xue, Learning to score figure skating sport videos, IEEE Trans. Circuits Syst. Video Technol. 30 (2019) 4578–4590.
[39] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, 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.
[40] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, C. Schmid, Multiview transformers for video recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3333–3343.
[41] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[42] X. Yu, Y. Rao, W. Zhao, J. Lu, J. Zhou, Group-aware contrastive regression for action quality assessment, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 7919–7928.
[43] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning (still) requires rethinking generalization, Commun. ACM 64 (2021) 107–115.
[44] Q. Zhang, B. Li, Relative hidden Markov models for video-based evaluation of motion skills in surgical training, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2014) 1206–1218.
[45] S.J. Zhang, J.H. Pan, J. Gao, W.S. Zheng, Semi-supervised action quality assessment with self-supervised segment feature recovery, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 6017–6028.
[46] S.J. Zhang, J.H. Pan, J. Gao, W.S. Zheng, Adaptive stage-aware assessment skill transfer for skill determination, IEEE Trans. Multimed. (2023).
[47] J. Zheng, Z. Jiang, R. Chellappa, Cross-view action recognition via transferable dictionary learning, IEEE Trans. Image Process. 25 (2016) 2542–2556.
[48] K. Zhou, Y. Ma, H.P. Shum, X. Liang, Hierarchical graph convolutional networks for action quality assessment, IEEE Trans. Circuits Syst. Video Technol. (2023).
[49] A. Zia, I. Essa, Automated surgical skill assessment in RMIS training, Int. J. Comput. Assisted Radiol. Surg. 13 (2018) 731–739.
[50] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, T. Ploetz, M.A. Clements, I. Essa, Automated video-based assessment of surgical skills for training and evaluation in medical schools, Int. J. Comput. Assisted Radiol. Surg. 11 (2016) 1623–1636.