FineParser: A Fine-grained Spatio-temporal Action Parser for Human-centric Action Quality Assessment

Jinglin Xu[1] Sibo Yin[2] Guohao Zhao[2] Zishuo Wang[2] Yuxin Peng[2*]
[1] School of Intelligence Science and Technology, University of Science and Technology Beijing
[2] Wangxuan Institute of Computer Technology, Peking University
xujinglinlove@gmail.com; 2000012982@stu.pku.edu.cn; ssee7235@gmail.com;
1900013093@pku.edu.cn; pengyuxin@pku.edu.cn

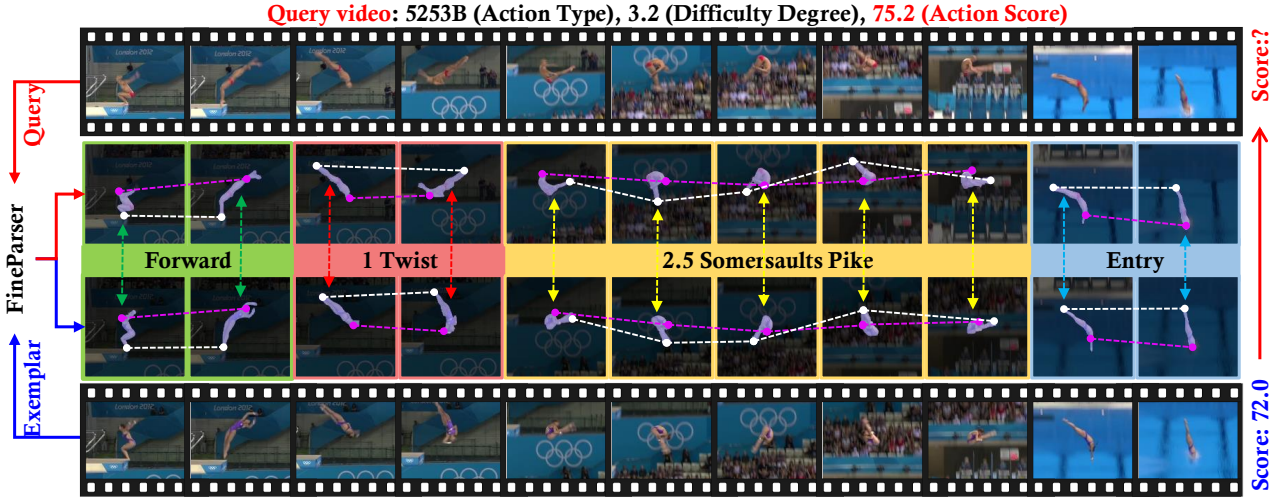

Figure 1. An overview of fine-grained spatial-temporal action parser (***FineParser***). It enhances human-centric foreground action representations by exploiting fine-grained semantic consistency and spatial-temporal correlation between video frames, improving the AQA performance. Green, red, yellow, and blue dashed lines represent the fine-grained alignment of target actions between query and exemplar videos in time and space within the same semantics.

Abstract

Existing action quality assessment (AQA) methods mainly learn deep representations at the video level for scoring diverse actions. Due to the lack of a fine-grained understanding of actions in videos, they harshly suffer from low credibility and interpretability, thus insufficient for stringent applications, such as Olympic diving events. We argue that a fine-grained understanding of actions requires the model to perceive and parse actions in both time and space, which is also the key to the credibility and interpretability of the AQA technique. Based on this insight, we propose a new fine-grained spatial-temporal action parser named ***FineParser****. It learns human-centric foreground action representations by focusing on target action regions within each frame and exploiting their fine-grained alignments in time and space to minimize the impact of invalid backgrounds during the assessment. In addition, we construct fine-grained annotations of human-centric foreground action masks for the FineDiving dataset, called* ***FineDiving-HM****. With refined annotations on diverse target action procedures, FineDiving-HM can promote the development of real-world AQA systems. Through extensive experiments, we demonstrate the effectiveness of FineParser, which outperforms state-of-the-art methods while supporting more tasks of fine-grained action understanding. Data and code are available at* https://github.com/PKU-ICST-MIPL/FineParser_CVPR2024.

1. Introduction

Video understanding is a crucial technique in computer vision that aims to analyze objects, actions, or events in videos automatically. It is essential for many real-world applications, e.g., human-computer interaction [9, 12, 21, 33], medical rehabilitation [11, 32], and sports analysis [6, 15, 30, 36]. Notably, a clear and accurate understanding of *actions* in videos provides critical and extensive technique

*Corresponding author.

support in action quality assessment (AQA). This considerably impacts sports analysis, helping evaluate athlete performance, designing targeted training programs, and preventing sports injuries.

Unlike general videos, sports videos are sequential processes with explicit procedural knowledge. Athletes have to complete a series of rapid and complex movements. Taking diving as an example, athletes will stretch, curl, and move their limbs and joints to finish different somersaults with three body positions, including straight, pike, and tuck, interspersed with varying twists. Then, the referee will assess the scores based on the athletes' take-off, somersault, twists, and entry. To achieve better competitive performance, athletes (1) take off decisively and forcefully at the right angle and with a proper height; (2) perform beautiful body positions, quick somersaults, and twists in the flight; (3) enter the water with a posture perpendicular to the surface, avoiding splashing water around. According to the diving rules, just a few degree differences in the take-off angle/height and the verticality of entry into the water can affect the number of points deducted. The difficulty lies in whether the human eye can accurately discern such subtle differences.

To address this issue, many video understanding-based AQA methods [24, 31, 35, 37] lack a fine-grained understanding of actions in videos. They cannot solve the problem of limitations of human eye judgment and lack credibility, which is inadmissible in real-world applications. There is an urgent need for a fine-grained understanding of actions, i.e., parsing the internal structures of actions in time and space with semantic consistency and spatial-temporal correlation, to obtain precise action representations and improve the usefulness of the AQA system.

To this end, we present a new framework for fine-grained action understanding, which learns human-centric foreground action representations with context information by developing a new fine-grained spatial-temporal action parser named ***FineParser***. FineParser consists of four components: (1) spatial action parser (SAP); (2) temporal action parser (TAE); (3) static visual encoder (SVE); (4) fine-grained contrastive regression (FineReg). Given query and exemplar videos, SAP first models the intra-frame feature distribution of each video by capturing multi-scale representations of human-centric foreground actions. The critical regions are concentrated around the athlete's body, springboard (or platform), and splash, guaranteeing the spatial parsing to be credible and visually interpretable. Then, TAP models semantic and temporal correspondences between videos by learning their spatial-temporal representations and parsing the actions into consecutive steps. Combining TAP and SAP, FineParser learns the target action representations at the fine-grained level, ensuring semantic consistency and spatial-temporal correspondence across videos. In addition, SVE enhances the above target action representations by capturing more contextual details. Finally, FineReg can quantify the quality differences in pairwise steps between query and exemplar videos and assess the action quality.

To promote the evaluation of credibility and visual interpretability of FineParser, we densely label human-centric foreground action regions of all videos in the FineDiving dataset and construct additional mask annotations, named ***FineDiving-HM***. Experimental results demonstrate that our fine-grained actions understanding framework accurately assesses diving actions while focusing on critical regions consistent with human visual understanding.

The contributions of this paper are summarized as follows: (1) We propose a new fine-grained spatial-temporal action parser, ***FineParser***, beneficial to the AQA task via human-centric fine-grained alignment. (2) FineParser captures the human-centric foreground action regions within each frame, minimizing the impact of invalid background in AQA. (3) We provide human-centric foreground action mask annotations for the FineDiving dataset, ***FineDiving-HM***, which we will release publicly to facilitate the evaluation of credibility and visual interpretability of the AQA system. (4) Extensive experiments demonstrate that our FineParser achieves state-of-the-art performance with significant improvements and better visual interpretability.

2. Related Work

Fine-grained Action Understanding. With ongoing advancements in action understanding, analyzing actions in finer granularity has become inevitable. Current endeavors in fine-grained action understanding mainly encompass tasks such as temporal action detection [10, 18, 28], action recognition [13, 19, 42], video question answering [5, 38, 39], and video-text retrieval [3, 7]. Recently, Shao *et al.* [30] constructed FineGym that provides coarse-to-fine annotations temporally and semantically for facilitating action recognition. Chen *et al.* [4] proposed SportsCap that estimates 3D joints and body meshes and predicts action labels. Li *et al.* [15] introduced MultiSports with spatio-temporal annotations of actions from four sports. Zhang *et al.* [39] constructed a temporal query network to answer fine-grained questions about event types and their attributes in untrimmed videos. Li *et al.* [16] presented a hierarchical atomic action network that models actions as combinations of reusable atomic ones to capture the commonality and individuality of actions. Zhang *et al.* [40] introduced a fine-grained video representation learning method to distinguish video processes and capture their temporal dynamics. These methods mainly concentrated on a fine-grained understanding of the temporal dimension. In contrast, our FineParser captures human-centric action representations by simultaneously building a fine-grained understanding in both time and space.

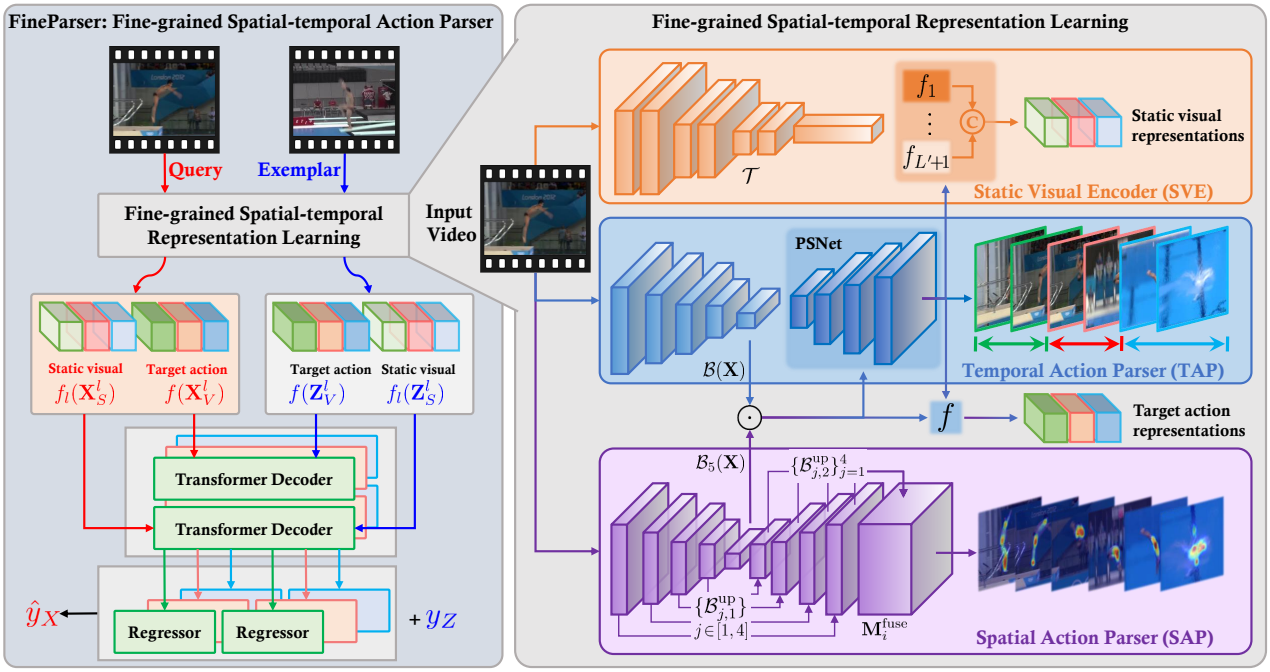

Figure 2. The architecture of the proposed ***FineParser***. Given a pair of query and exemplar videos, spatial action parser (SAP) and temporal action parser (TAP) extract spatial-temporal representations of human-centric foreground actions in pairwise videos, as well as predict both target action masks and step transitions. The static visual encoder (SVE) captures static visual representations combined with the target action representation to mine more contextual details. Finally, fine-grained contrastive regression (FineReg) utilizes the representations to predict the action score of the query video.

Action Quality Assessment. In early pioneering work, Pirsiavash *et al.* [29] formulated the AQA task as a regression problem from action representations to scores, and Parisi *et al.* [23] adopted the correctness of performed action matches to assess action quality. Parmar *et al.* [26] demonstrated the effectiveness of spatio-temporal features for estimating scores in various competitive sports. Recently, Tang *et al.* [31] introduced an uncertainty-aware score distribution learning method to alleviate the ambiguity of judges' scores. Yu *et al.* [37] developed a contrastive regression based on video-level features, enabling the ranking of videos and accurate score prediction. Wang *et al.* [34] introduced TSA-Net to generate action representations using the outputs of the VOT tracker, improving AQA performance. Xu *et al.* [36] contributed to a fine-grained sports video dataset for AQA and proposed a new action procedure-aware method to improve AQA performance. Zhang *et al.* [41] proposed a plug-and-play group-aware attention module to enrich clip-wise representations with contextual group information. In contrast, our FineParser parses action in space and time to focus on the human-centric foreground action, improving AQA's credibility and visual interpretability.

3. Approach

This section presents a fine-grained spatial-temporal action parser for human-centric action quality assessment, i.e., ***FineParser***. As illustrated in Fig. 2, FineParser consists of four components: spatial action parser (SAP), temporal action parser (TAP), static visual encoder (SVE), and fine-grained contrastive regression (FineReg).

3.1. Problem Formulation

Given a pair of query and exemplar videos with the same action type, denoted as $(\mathbf{X}, \mathbf{Z})$, our approach is formulated as a fine-grained understanding framework that predicts the action score of the query video $\mathbf{X}$. Inspired by fine-grained contrastive regression [36], our framework considers fine-grained quality differences between human-centric foreground actions in both time and space perspectives to model variations in their scores. The core is a new fine-grained action parser, FineParser $\mathcal{F}$, represented as

$$\hat{y}_X = \mathcal{F}(\mathbf{X}, \mathbf{Z}, y_Z; \mathbf{\Theta}), \tag{1}$$

where $\mathbf{\Theta}$ denotes all learnable parameters of $\mathcal{F}$, and $\hat{y}_X$ denotes the predicted action score of $\mathbf{X}$ referring to $\mathbf{Z}$ and its ground truth score y_Z.

3.2. Fine-grained Spatio-temporal Action Parser

FineParser is composed of four core components. In short, SAP, TAP, and SVE collaborate to learn fine-grained target action representations, and FineReg then uses these representations to predict the final score.

Spatial Action Parser (SAP). SAP parses the target action for each input video at a fine-grained spatial level. Inspired by I3D [2] and its fully convolutional version [20], transposed convolution layers are introduced before each max pooling layer to upsample the outputs of I3D submodules, and the rest after the last average pooling layer is discarded. These operations facilitate capturing multi-scale visual and semantic information that spans from short-term local features obtained from lower layers to long-term global semantic context derived from the last few layers.

Concretely, taking the query video $\mathbf{X}=\{\mathbf{X}_i\}_{i=1}^{N}$ as an example, the first I3D submodule $\mathcal{B}_1$ encodes each snippet $\mathbf{X}_i$ to capture short-term local features, as $\mathcal{B}_1(\mathbf{X}) = \{\mathcal{B}_1(\mathbf{X}_i)\}_{i=1}^{N}$. Similarly, other three submodules encode $\mathbf{X}_i$ to obtain middle representations, as $\mathcal{B}_j(\mathbf{X})=\mathcal{B}_j(\mathcal{B}_{j-1}(\mathbf{X}))$, with $j\in[2,4]$. For each $\mathcal{B}_j(\mathbf{X})$, two upsampling blocks are further inserted, denoted as $\mathcal{B}_{j,1}^{\text{up}}$ and $\mathcal{B}_{j,2}^{\text{up}}$. Both comprise convolution layers performed on the feature dimension and transpose convolution layers performed on both spatial and temporal dimensions. They can be presented as

$$\mathbf{M}_{j,i}^{\text{up}_1} = \mathcal{B}_{j,1}^{\text{up}}(\mathcal{B}_j(\mathbf{X}_i)),\ \mathbf{M}_{j,i}^{\text{up}_2} = \mathcal{B}_{j,2}^{\text{up}}(\mathcal{B}_j(\mathbf{X}_i)), \quad (2)$$

$$\mathbf{M}_{i}^{\text{fuse}} = \text{Conv3d}(\text{Concat}(\{\mathbf{M}_{j,i}^{\text{up}_1}\}_{j=1}^{4})), \quad (3)$$

where $\{\mathbf{M}_{j,i}^{\text{up}_2}\}_{j=1}^{4}$ are the predicted target action masks from different scales for optimizing SAP. These masks capture multi-scale human-centric foreground action information, from short-term local features obtained from lower layers (small scale) to long-term global semantic context derived from the last few layers (large scale). $\mathbf{M}_{i}^{\text{fuse}}$ is the final target action mask of $\mathbf{X}_i$ by fusing $\{\mathbf{M}_{j,i}^{\text{up}_1}\}_{j=1}^{4}$. SAP generates the above five target action masks and one target action mask embedding $\mathcal{B}_5(\mathbf{X})$, where the former are used to anticipate the human-centric foreground action mask, and the latter facilitates learning target action representations. With mask embedding $\mathcal{B}_5(\mathbf{X})$ and video embedding $\mathcal{B}(\mathbf{X})$, target action representations $\mathbf{X}_V$ are calculated by elements-wise multiplication, as $\mathbf{X}_V=\mathcal{B}(\mathbf{X}) \odot \text{sigmoid}(\mathcal{B}_5(\mathbf{X}))$. For the exemplar video $\mathbf{Z}$, the target action representations $\mathbf{Z}_V$ can be obtained similarly.

Temporal Action Parser (TAP). TAP parses each action procedure into consecutive steps with semantic and temporal correspondences. Specifically, PSNet [36] is adopted to parse $\mathbf{X}_V$ and $\mathbf{Z}_V$, which identifies the temporal transition when the step switches from one sub-action type to another. Supposed that L' step transitions are needed to be identified in the action, the submodule $\mathcal{S}$ predicts the probability of the k-th step transiting at the t-th frame, denoted as $\mathcal{S}(\mathbf{X}_V)[t,k]\in\mathbf{R}$. By

$$\hat{t}_k = \underset{\frac{T}{L'}(k-1)<t\leq\frac{T}{L'}k}{\arg\max}\ \mathcal{S}(\mathbf{X}_V)[t,k], \quad (4)$$

the timestamp $\hat{t}_k$ of the k-th step transition is predicted for each $k\in[1,L']$. Based on $\{\hat{t}_k\}_{k=1}^{L'}$, each action procedure is parsed into $L'+1$ consecutive steps, i.e., $\{\mathbf{X}_V^l\}_{l=1}^{L'+1}$ and $\{\mathbf{Z}_V^l\}_{l=1}^{L'+1}$, where l is the index of step. While the lengths of the above consecutive steps may differ in nature, they are fixed to the same size via downsampling or upsampling operations f along the temporal axis, ensuring that the dimensions of *query* and *key* are matched in the attention model. Therefore, the target action representations of query and exemplar videos become $\{f(\mathbf{X}_V^l)\}_{l=1}^{L'+1}$ and $\{f(\mathbf{Z}_V^l)\}_{l=1}^{L'+1}$.

Static Visual Encoder (SVE). SVE captures more contextual information to further enhance the action representations, especially for high-speed and complex actions like diving. It consists of two submodules: a ResNet model $\mathcal{T}$ and a set of projection functions $\{f_l\}_{l=1}^{L'+1}$. For the input video $\mathbf{X}$, the outputs of $\mathcal{T}$ can be obtained by

$$\begin{aligned}&\mathbf{X}_S^1 = \mathcal{T}(\mathbf{X})[:\hat{t}_1],\ \mathbf{X}_S^{L'+1} = \mathcal{T}(\mathbf{X})[\hat{t}_{L'}:],\\ &\mathbf{X}_S^l = \mathcal{T}(\mathbf{X})[\hat{t}_{l-1}:\hat{t}_l] \text{ s.t. } l\in[2,L'].\end{aligned} \quad (5)$$

Through post-projection, the static visual representations of $\mathbf{X}$ can be written as $\{f_l(\mathbf{X}_S^l)\}_{l=1}^{L'+1}$. Similarly, the static visual representation of $\mathbf{Z}$ are $\{f_l(\mathbf{Z}_S^l)\}_{l=1}^{L'+1}$.

Fine-grained Contrastive Regression (FineReg). It leverages the sequence-to-sequence representation ability of the transformer to learn powerful representations from pairwise steps and static visual representations via cross-attention. Specifically, the target action representations of pairwise steps $f(\mathbf{X}_V^l)$ and $f(\mathbf{Z}_V^l)$ interact with each other, helping the model focus on the consistent regions of motions in the cross-attention to generate the new features $\mathbf{D}_l^V$. Similarly, cross-attention between the static visual representations of pairwise steps $f_l(\mathbf{X}_S^l)$ and $f_s(\mathbf{Z}_S^l)$ generates the new features $\mathbf{D}_l^S$. Based on these two generated representations of the l-th step pairs, FineReg quantifies step quality differences between the query and exemplar by learning relative scores. This guides the framework to assess action quality at the fine-grained level with contrastive regression $\mathcal{R}$. The predicted score $\hat{y}_X$ of the query video $\mathbf{X}$ is calculated as

$$\hat{y}_X = \textstyle\sum_{l=1}^{L'+1} \lambda_l(\mathcal{R}_V(\mathbf{D}_l^V) + \mathcal{R}_S(\mathbf{D}_l^S)) + y_Z, \quad (6)$$

where $\mathcal{R}_\mathcal{V}$ and $\mathcal{R}_\mathcal{S}$ are two three-layer MLPs with ReLU non-linearity, y_Z is the ground truth score of the exemplar video $\mathbf{Z}$, and λ_l is the coefficient weighting the relative score of the l-th step pairs.

3.3. Training and Inference

Training. Given a pairwise query and exemplar videos $(\mathbf{X},\mathbf{Z})$ from the training set, FineParser is optimized by minimizing the following losses:

$$\mathcal{L} = \mathcal{L}_{\text{SAP}} + \mathcal{L}_{\text{TAP}} + \mathcal{L}_{\text{Reg}}. \quad (7)$$

Figure 3. Examples of human-centric action mask annotations for the FineDiving dataset. The right line indicates the action type.

$\mathcal{L}_{\text{SAP}}$ is used to optimize SAP, calculated by

$$\mathcal{L}_{\text{SAP}} = \sum \mathcal{L}_{\text{Focal}}(p(\mathrm{M}_{j,i})), \tag{8}$$

$$\mathcal{L}_{\text{Focal}}(p(\mathrm{M}_{j,i})) = -\alpha_t(1-p(\mathrm{M}_{j,i}))^{\gamma}\log(p(\mathrm{M}_{j,i})), \tag{9}$$

where $\mathrm{M}_{j,i}{=}\mathbf{M}^{\text{up}_2}_{j,i}[l,h,w]$ is the element of $\mathbf{M}^{\text{up}_2}_{j,i}$, $p(\mathrm{M}_{j,i}){=}\mathrm{M}_{j,i}$ if the ground-truth mask $\mathrm{M}^{\text{gt}}_i{=}1$, and $p(\mathrm{M}_{j,i}){=}1{-}\mathrm{M}_{j,i}$, otherwise. $\mathcal{L}_{\text{Focal}}$ is the focal loss [17] between predicted and ground truth masks. $\mathcal{L}_{\text{TAP}}$ is used to optimize TAP, calculated by

$$\mathcal{L}_{\text{TAP}}{=}{-}\textstyle\sum_t(\mathrm{p}_k(t)\log \mathrm{S}_{t,k}+(1{-}\mathrm{p}_k(t))\log(1{-}\mathrm{S}_{t,k})), \tag{10}$$

where $\mathrm{S}_{t,k}{=}\mathcal{S}(\mathbf{X}_V)[t,k]$ is the predicted probability of the k-th step transiting at the t-th frame, and $\mathbf{p}_k$ is a binary distribution encoded by the ground truth timestamp t_k of the k-th step transition, with $\mathrm{p}_k(t_k){=}1$ and $\mathrm{p}_k(t_m)|_{m\neq k}{=}0$. $\mathcal{L}_{\text{Reg}}$ is used to optimize $\mathcal{R}_V$ and $\mathcal{R}_S$ by minimizing the mean squared error between the ground truth y_X and prediction $\hat{y}_X$, which is written as

$$\mathcal{L}_{\text{Reg}} = \|\hat{y}_X - y_X\|^2. \tag{11}$$

Inference. For a query video $\mathbf{X}$ from the testing set, the multi-exemplar voting strategy [37] is adopted to select E exemplars $\{\mathbf{Z}_j)\}_{j=1}^{E}$ from the training set and construct pairwised $\{(\mathbf{X},\mathbf{Z}_j)\}_{j=1}^{E}$ with scores $\{y_{Z_j}\}_{j=1}^{E}$. The inference process can be written as

$$\hat{y}_X = \frac{1}{E}\textstyle\sum_{j=1}^{E}(\mathcal{F}(\mathbf{X},\mathbf{Z}_j;\mathbf{\Theta}) + y_{Z_j}). \tag{12}$$

4. Experiments

4.1. Datasets

FineDiving-HM. FineDiving [36] contains 3,000 videos covering 52 action types, 29 sub-action types, 23 difficulty degree types, fine-grained temporal boundaries, and official action scores. To evaluate the effectiveness of our FineParser and make the results more credible and interpretable visually, we provide additional human-centric action mask annotations for the FineDiving dataset in this work, called ***FineDiving-HM***. FineDiving-HM contains 312,256 mask frames covering 3,000 videos, in which each mask labels the target action region to distinguish the human-centric foreground and background. FineDiving-HM mitigates the problem of requiring frame-level annotations to understand human-centric actions from fine-grained spatial and temporal levels. We employ three workers with prior diving knowledge to double-check the annotations to control their quality. Fig. 3 shows some examples of human-centric action mask annotations, which precisely focus on foreground target actions. There are 312,256 foreground action masks in FineDiving-HM, where the number of action masks for individual diving is 248,713 and that for synchronized diving is 63,543. As shown in Fig. 4, the largest number of action masks is 35,287, belonging to the action type 107B; the second largest number of action masks is 34,054, belonging to the action type 407C; and

Methods	AQA Metrics	
	$\rho\uparrow$	$R\text{-}\ell_2\downarrow(\times100)$
C3D-LSTM [26]	0.6969	1.0767
C3D-AVG [25]	0.8371	0.6251
MSCADC [25]	0.7688	0.9327
I3D+MLP [31]	0.8776	0.4967
USDL [31]	0.8830	0.4800
MUSDL [31]	0.9241	0.3474
CoRe [37]	0.9308	0.3148
TSA [36]	0.9324	0.3022
FineParser	**0.9435**	**0.2602**

Methods	TAP Metrics	
	AIoU@0.5↑	AIoU@0.75↑
TSA [36]	0.9239	0.5007
FineParser	**0.9946**	**0.9467**

Methods	SAP Metrics		
	MAE↓	$F_\beta\uparrow$	$S_m\uparrow$
FineParser	**0.0408**	**0.1273**	**0.8357**

Table 1. Comparisons of performance with state-of-the-art AQA methods on the FineDiving-HM Dataset. Our result is highlighted in the **bold** format.

the smallest number of action masks is 101, corresponding to the action types 109B, 201A, 201C, and 303C. Coaches and athletes can use the above statistics to develop competition strategies, for instance, what led to the rise of 107B and 407C and how athletes gain a competitive edge.

MTL-AQA. It is a multi-task action quality assessment dataset [25] consisting of 1,412 samples collected from 16 different world events, with annotations containing the degree of difficulty, scores from each judge (7 judges), type of diving action, and the final score.

4.2. Evaluation Metrics

Action Quality Assessment. Following previous efforts [22, 25, 31, 36, 37], we utilize Spearman's rank correlation (ρ, the higher, the better) and Relative ℓ_2 distance (R_{ℓ_2}, the lower, the better) for evaluating the AQA task.

Temporal Action Parsing. Given the ground truth bounding boxes and a set of predicted temporal bounding boxes, we adopt the Average Intersection over Union (AIoU) [36] to evaluate the performance of TAP. The higher the value of AIoU@d, the better the performance of TAP.

Spatial Action Parsing. We adopt three evaluation metrics for comparison: MAE [27], F-measure F_β ($\beta = 0.3$) [1], and S-measure S_m [8]. MAE (the lower, the better) measures the average pixel-wise absolute error between the binary ground truth mask and normalized saliency prediction map. F-measure (the higher, the better) comprehensively considers precision and recall by computing the weighted harmonic mean. S-measure (the higher, the better) evaluates the structural similarity between the real-valued saliency map and the binary ground truth, considering object-aware (S_o) and region-aware (S_r) structure similarities ($\alpha=0.5$).

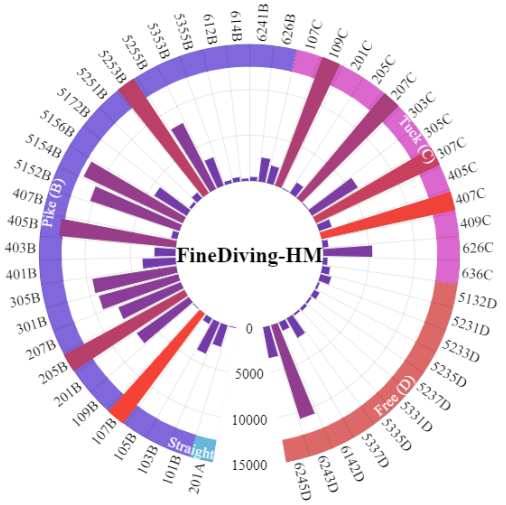

Figure 4. The distribution of human-centric foreground action masks. The largest number of mask instances is 35,287, belonging to the action type 107B. The smallest number of mask instances is 101, containing the action types 109B, 201A, 201C, and 303C.

4.3. Implementation Details

We adopted the I3D model pre-trained on the Kinetics [2] as the backbone of the SAP and TAP modules, where SAP is composed of $\{\mathcal{B}_j\}_{j=1}^{5}$ and $\{\mathcal{B}_{j,1}^{\text{up}}, \mathcal{B}_{j,2}^{\text{up}}\}_{j=1}^{4}$ with the initial learning rate 10^{-3} and TAP consists of $\mathcal{B}$ and $\mathcal{S}$ with the initial learning rate 10^{-4}. SAP and TAP did not share parameters. Besides, we set the initial learning rates of $\mathcal{T}$ (i.e., ResNet34) in SVE as 10^{-3}. We utilized Adam [14] optimizer and set weight decay as 0. In SAP and TAP, following previous works [31, 36, 37], we extracted 96 frames for each video and split them into 9 snippets, where each snippet contains 16 continuous frames with a stride of 10 frames. We set L' as 3 and the weights $\{\lambda_l\}_{l=1}^{L'}$ as $\{3, 5, 2\}$. Furthermore, we followed the exemplar selection criterion in [36] and [37] on the FineDiving-HM and MTL-AQA datasets, respectively. Following the experiment settings in [31, 36, 37], we selected 75 percent of samples for training and 25 percent for testing in all the experiments.

4.4. Comparison with the State-of-the-Arts

FineDiving-HM. Tab. 1 summarized the experimental results of state-of-the-art AQA methods on the FineDiving-HM dataset. Our FineParser significantly improved the performance of Spearman's rank correlation and Relative ℓ_2-distance compared to all methods. The advantages of FineParser stemmed from a fine-grained understanding of human-centric foreground actions, which requires the model to parse actions in time and space, making the model credible and interpretable visually. Compared to C3D-LSTM, C3D-AVG, MSCADC, I3D+MLP, USDL, and MUSDL, FineParser outperformed them significantly and achieved 24.66%, 10.64%, 17.47%, 6.59%, 6.05%, and

Methods	MTL-AQA	
	$\rho\uparrow$	$R\text{-}\ell_2\downarrow(\times100)$
Pose+DCT [29]	0.2682	/
C3D-SVR [26]	0.7716	/
C3D-LSTM [26]	0.8489	/
C3D-AVG-STL [25]	0.8960	/
C3D-AVG-MTL [25]	0.9044	/
USDL [31]	0.9231	0.4680
MUSDL [31]	0.9273	0.4510
TSA-Net [34]	0.9422	/
CoRe [37]	0.9512	0.2600
FineParser	**0.9585**	**0.2411**

Table 2. Comparisons of performance with representative AQA methods on the MTL-AQA dataset. Our result is highlighted in the **bold** format.

1.94% performance improvements in terms of Spearman's rank correlation as well as 0.8165, 0.3649, 0.6725, 0.2365, 0.2198, and 0.0872 in Relative ℓ_2-distance. Compared to CoRe, FineParser obtained 1.27% and 0.0546 performance improvements on Spearman's rank correlation and Relative ℓ_2-distance. FineParser further improved the performance of TSA on Spearman's rank correlation and Relative ℓ_2-distance, which also can be observed in the TAP metric.

MTL-AQA. Tab. 2 reported the experimental results of representative AQA methods on the MTL-AQA dataset. Our FineParser outperformed other methods on Spearman's rank correlation. For instance, FineParser achieved better AQA performance than CoRe and TSA-Net, demonstrating the effectiveness of additional human-centric foreground action masks and the meticulous design of a fine-grained action understanding of FineParser.

4.5. Ablation Study

We conducted an ablation study on the FineDiving-HM dataset to demonstrate the effectiveness of individual parts of FineParser by designing different modules, different backbones of SVE, and varied step durations of the projection function in SVE.

Different Modules in FineParser. We summarized the experimental results in Tab. 3. Under Spearman's rank correlation, the AQA performance of the model with SVE and TAP can be improved from 0.9334 to 0.9351. Significant improvements on AIoU@0.5 and AIoU@0.75 are directly proportional to the accuracy of action quality assessment, demonstrating that SVE can help the model perform more accurate temporal action parsing in the TAP module. Further introducing the SAP module into the model, the AQA performance can be further enhanced to 0.9435 in Spearman's rank correlation, demonstrating that incorporating SAP allows for capturing more characteristics of target action, achieving more accurate action quality assessment. If only SAP or SVE were introduced, Spearman's rank correlations would be 0.9313 or 0.9328, respectively, which cannot achieve the AQA performance of our final version.

Methods	Modules		
	SAP	SVE	TAP
A	✓		
B		✓	
C			✓
D		✓	✓
E	✓	✓	✓

Methods	$\rho\uparrow$	$R\text{-}\ell_2\downarrow(\times100)$
A	0.9313	0.3094
B	0.9328	0.3097
C	0.9334	0.3122
D	0.9351	0.2881
E	**0.9435**	**0.2602**

Methods	TAP Metrics	
	AIoU@0.5↑	AIoU@@0.75↑
C	0.9907	0.9039
D	0.9920	0.8932
E	**0.9946**	**0.9467**

Methods	SAP Metrics		
	MAE↓	$F_\beta\uparrow$	$S_m\uparrow$
E	**0.0408**	**0.1273**	**0.8357**

Table 3. Ablation study on different modules in FineParser on FineDiving-HM. The results of unavailable methods are omitted.

Different Step Durations in SVE. We studied the influence of different step durations used in the projection function of SVE on the AQA performance. As shown in Tab. 4, we set the step duration as 2, 4, and 8 and then observe that the AQA performance of FineParser is optimal when set to 4. It is attributed to proper step duration that can benefit mining more valuable information from human-centric foreground action and static visual representations.

Different Backbones of SVE. We conducted several experiments on the FineDiving-HM dataset to investigate the effects of different backbones of SVE on the performance of action quality assessment. In Tab. 5, ResNet34 outperforms other ResNet architectures while slightly inferior to ViT-S/16. For one thing, ResNet34 has a deeper network depth than ResNet18, allowing it to capture more global and high-level semantic information, whereas ResNet50 may lead to overfitting on the steps with relatively short durations (e.g., four frames). In addition, ViT allows the model to capture long-term dependencies among video frames rather than local relationships, which is beneficial to learning target action representations by capturing global features, further improving the AQA performance (i.e., $R\text{-}\ell_2$) of FineParser.

4.6. Visualization

To intuitively understand the benefits of our FineParser, we visualize the predicted masks obtained by SAP, as shown in Fig. 5. We see that the predictions can focus on target action regions in each frame, minimizing the impact of invalid backgrounds on action quality assessment.

Figure 5. Visualization of the predictions of target action masks produced by SAP. The predicted masks can focus on the target action regions in each frame, minimizing the impact of invalid backgrounds on action quality assessment.

Duration	AQA	
	$\rho \uparrow$	$R\text{-}\ell_2 \downarrow (\times 100)$
2	0.9320	0.2994
4	**0.9435**	**0.2602**
8	0.9337	0.2940
Duration	TAP	
	AIoU@0.5↑	AIoU@0.75↑
2	0.9987	0.9359
4	**0.9946**	**0.9467**
8	0.9973	0.9493
Duration	SAP		
	MAE↓	$F_\beta \uparrow$	$S_m \uparrow$
2	0.0532	0.1010	0.8643
4	**0.0408**	**0.1273**	**0.8357**
8	0.0535	0.1057	0.8616

Table 4. Ablation study on different step durations in the projection function in the SVE module.

Backbones	$\rho \uparrow$	$R\text{-}\ell_2 \downarrow (\times 100)$
ResNet18	0.9363	0.2829
ResNet34	**0.9435**	**0.2602**
ResNet50	0.9362	0.2859
ViT-S/16	0.9426	<u>0.2583</u>

Table 5. Ablation study on different backbones in SVE.

5. Conclusion and Discussion

We presented an end-to-end fine-grained spatial-temporal action parser named FineParser for the AQA task. It learned fine-grained representations for target actions via integrating spatial action parser, temporal action parser, static visual encoder, and fine-grained contrastive regression and achieved state-of-the-art. To understand human-centric actions from fine-grained spatial and temporal levels, we also provided human-centric foreground action mask annotations for the FineDiving dataset, named FineDiving-HM, to provide three quantitative metrics for the credibility and visual interpretability of the AQA model. We hope FineParser could be a baseline for fine-grained human-centric AQA and facilitate more tasks that require a fine-grained understanding of sports.

Limitations. The human-centric foreground action masks need to be manually adjusted and labeled. This work contributes new human-centric annotations for the dataset on diving events, while they are challenging to transfer to other competitive sports directly.

Acknowledgements. This work was supported by grants from the National Natural Science Foundation of China (61925201, 62132001, 62373043) and the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 6

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 4, 6

[3] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10638–10647, 2020. 2

[4] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *arXiv preprint arXiv:2104.11452*, 2021. 2

[5] Thilini Cooray, Ngai-Man Cheung, and Wei Lu. Attention-based context aware reasoning for situation recognition. In *CVPR*, pages 4736–4745, 2020. 2

[6] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, pages 9921–9931, 2023. 1

[7] Hazel Doughty and Cees GM Snoek. How do you do it? fine-grained action understanding with pseudo-adverbs. In *CVPR*, pages 13832–13842, 2022. 2

[8] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017. 6

[9] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 1

[10] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *CVPR*, pages 19999–20009, 2022. 2

[11] Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, 2023. 1

[12] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *CVPR*, pages 14708–14718, 2021. 1

[13] James Hong, Matthew Fisher, Michaël Gharbi, and Kayvon Fatahalian. Video pose distillation for few-shot, fine-grained sports action recognition. In *ICCV*, pages 9254–9263, 2021. 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *ICCV*, pages 13536–13545, 2021. 1, 2

[16] Zhi Li, Lu He, and Huijuan Xu. Weakly-supervised temporal action detection for fine-grained videos with hierarchical atomic actions. In *ECCV*, pages 567–584, 2022. 2

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017. 5

[18] Yi Liu, Limin Wang, Xiao Ma, Yali Wang, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *arXiv preprint arXiv:2105.11107*, 2021. 2

[19] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, pages 122–132, 2020. 2

[20] Purna Sowmya Munukutla and Siddhant Jain. One shot learning for video object segmentation using fully convolutional i3d network. 2018. 4

[21] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, pages 9890–9900, 2020. 1

[22] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *ICCV*, pages 6331–6340, 2019. 6

[23] German I Parisi, Sven Magg, and Stefan Wermter. Human motion assessment in real time using recurrent self-organization. In *RO-MAN*, pages 71–76, 2016. 3

[24] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, pages 1468–1476, 2019. 2

[25] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a ma multitask learningultitask learning approach to action quality assessment. In *CVPR*, pages 304–313, 2019. 6, 7

[26] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPRW*, pages 20–28, 2017. 3, 6, 7

[27] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 6

[28] AJ Piergiovanni and Michael S Ryoo. Fine-grained activity recognition in baseball videos. In *CVPRW*, pages 1740–1748, 2018. 2

[29] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *ECCV*, pages 556–571, 2014. 3, 7

[30] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 1, 2

[31] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, pages 9839–9848, 2020. 2, 3, 6, 7

[32] Chen Wang, Jingqi Kong, and Huiying Qi. Areas of research focus and trends in the research on the application of vr in rehabilitation medicine. In *Healthcare*, page 2056, 2023. 1

[33] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, pages 9401–9411, 2021. 1

[34] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action

quality assessment. In *ACM MM*, pages 4902–4910, 2021. 3, 7

[35] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *TCSVT*, 30(12):4578–4590, 2019. 2

[36] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *CVPR*, pages 2949–2958, 2022. 1, 3, 4, 5, 6

[37] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *ICCV*, pages 7919–7928, 2021. 2, 3, 5, 6, 7

[38] Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In *CVPR*, pages 23191–23200, 2023. 2

[39] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *CVPR*, pages 4486–4496, 2021. 2

[40] Heng Zhang, Daqing Liu, Qi Zheng, and Bing Su. Modeling video as stochastic processes for fine-grained video representation learning. In *CVPR*, pages 2225–2234, 2023. 2

[41] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *CVPR*, pages 2405–2414, 2023. 3

[42] Chen Zhu, Xiao Tan, Feng Zhou, Xiao Liu, Kaiyu Yue, Errui Ding, and Yi Ma. Fine-grained video categorization with redundancy reduction attention. In *ECCV*, pages 136–152, 2018. 2