IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

DOI:10.1587/transfun.2024SMP0003

Publicized:2024/08/20

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

A Hierarchical Joint Training based Replay-Guided Contrastive Transformer for Action Quality Assessment of Figure Skating

Yanchao LIU[†a)], **Xina CHENG**[††], ***Nonmembers***, ***and*** **Takeshi IKENAGA**[†], ***Member***

SUMMARY Action quality assessment (AQA) has gained prominence as it finds widespread applications in various scenarios. Most existing methods directly regress from single or pairwise videos, which leads to redundant temporal features and limited views affecting the scoring mechanism. Moreover, direct regression only applies supervision to the last layer, which leads to hardship in optimizing the intermediate layers such as gradient vanishing. To end this, we propose a Hierarchical Joint Training based Replay-Guided Contrastive Transformer, learned by a temporal concentration module. For network architecture, we design an extra contrastive module for the input and its replay, and the consistency of scores guides the model to learn the features of the same action under different views. A temporal concentration module is proposed to extract concentrated features such as errors or highlights, which are crucial factors affecting scoring. The proposed hierarchical joint training provides supervision on both shallow and deep layers, enhancing the performance of the scoring mechanism and speed of training convergence. Extensive experiments demonstrate that our method achieves Spearman's Rank Correlation of 0.9642 on the RFSJ dataset, which is the new state-of-the-art result.

key words: *Action Quality Assessment, Contrastive Learning, Multi-Supervision*

1. Introduction

Action Quality Assessment (AQA) aims to evaluate how well a specific action is performed and finds extensive applications in domains such as sports [1] and healthcare [2]. Diverging from video action recognition [3] or detection [4], the AQA task presents a greater challenge as it necessitates evaluating the nuanced visual distinctions among closely related actions.

In the past years, most existing works on AQA mainly regress the assessed scores from a single video [5], [6] or a pairwise exemplar and input videos [1], [7]. Despite their effectiveness, these methods are limited by the viewpoint and zoom scale of input videos. It impedes the ability to discern crucial nuances for accurate action quality assessment. For example, in figure skating competitions, the blade-ice contact angle during take-off and landing significantly influences the score. However, it is challenging to identify by only a single view. Relying solely on score regression from one perspective leads to inaccurate inferences, as the model struggles to differentiate changes in viewpoint from inherent action variations.

[†]The authors are with the Graduate School of Information, Production and Systems, Waseda University, Kitakyushu-shi, 808-0135 Japan.

[††]The authors are with Xidian University, Xi'an, 710126 China.

a) E-mail: liuyanchao@fuji.waseda.jp

Judges frequently review replay videos from various views to determine uncertain or disputed actions, ultimately establishing the final score in real competitive scenarios. Based on this fact, we contend that replay data from various angles holds great significance for AQA. Motivated by this, we introduce an innovative framework for action quality assessment, featuring a replay-guided triple-stream contrastive transformer. In line with recent conventional research efforts [1], [7]–[9], our framework discerns disparities between the pairwise exemplar and the input video. However, our novel triple-stream framework takes a step further by incorporating the input video and its corresponding replay, employing an additional contrastive branch guided by optimization consistency. In essence, the input video and its replay showcase the same athlete and action but offer different viewpoints and zoom scales, resulting in a relative score of zero. This concept draws inspiration from self-supervised learning, where the zero relative score acts as a constraint, steering the network's attention towards the athletes' actions rather than variations in viewpoint or scale.

We have observed that the occurrence of athlete errors or highlight moments significantly influences the scoring. These moments tend to cluster in specific sections of the video, instead of uniform distribution. Building upon this observation, we designed a Temporal Concentration Module. To elaborate, our initial step involves uniformly grouping the pairwise video features and devising a cross-attention concentration decoder for each group. This decoder extracts an attention heatmap that highlights the concentrated correlations between these features. Subsequently, we implement a dense resampling strategy guided by the hotspots on the heatmap, directing focused attention to errors or highlight moments. These resampled feature pairs are then fed into a contrastive decoder, and cross-attention mechanisms are employed to facilitate the network in learning from the errors or highlights within each group.

Note that this work is an extended version of our conference work that appeared in ACM Multimedia (MM2023) [10]. Compared with the conference version, this work includes a new Hierarchical Joint Training method for more reliable concentrated feature extraction and enhanced scoring performance and convergence speed.

The conference version primarily emphasizes the AQA through the direct regression of concentrated features. This causes inadvertently overlooking the intrinsic relationship between the generalized feature and the concentrated feature. Given that the concentrated feature stems from the

generalized feature, direct regression leads to gradient vanishing in the generation process, affecting reliability. our current work extends the conference version by introducing hierarchical joint training to systematically capture the interdependence between the generalized and concentrated features. Based on deep supervision, a shallow module is designed to supervise training directly in the shallow layer to ensure the reliability of extracting concentrated features.

Specifically, our novel strategy involves coordinated training to extract concentrated features while establishing a clear relationship with generalized features. This is achieved through a meticulous supervision mechanism, where the generalized feature undergoes score regression in the shallow layer. By incorporating this additional layer of supervision, we aim to enhance the reliability of the concentration ranking before the actual extraction of the concentrated feature. Through extensive experimentation, we provide empirical evidence to support the effectiveness of the proposed hierarchical joint training method.

In summary, the contributions of this work are listed as follows:

- We propose a replay-guided temporal concentration approach for action quality assessment, which inputs exemplar, input video, and replay simultaneously, learned by concentrated attention to quantify the quality difference of errors or highlight moments between videos.
- Different from the conference version, we propose a hierarchical joint training method to provide supervision on both shallow and deep layers, enhancing the performance of the scoring mechanism and speed of convergence.
- Extensive experiments demonstrate that our proposed method improves over state-of-the-art methods.

2. Related Work

In this section, we review existing AQA methods and multi-view learning architecture.

2.1 Action Quality Assessment

Existing methods for AQA can be broadly classified into two categories: single-stream regressive methods and double-stream contrastive methods. Single-stream regressive methods approach the AQA task as a regression problem optimized using labeled absolute scores. Pioneering this approach, Pirsiavash et al.[11] introduced a learning-based framework, training a regression model from spatiotemporal pose features to predict scores. Parmar et al.[12] employed 3D convolutional neural networks (C3D) to extract spatiotemporal features and utilized Long Short-Term Memory (LSTM) with Support Vector Regression (SVR) to regress the quality score. Their work [13] also introduced a multitask learning approach to AQA. Bertasius [14] utilized a convolutional LSTM network and Gaussian mixture to construct a non-linear spatiotemporal feature for assessing the superior player in a pair of videos. Xu et al.[15] integrated self-attentive LSTM and multi-scale convolutional skip LSTM in a single end-to-end framework. Tang et al.[5] proposed an Uncertainty-Aware Score Distribution Learning (USDL) framework, considering the subjectiveness of action scores from human judges. Wang et al. [6] introduced a tube self-attention network, generating representations with rich contextual information through a single-object tracker. In recent developments, double-stream contrastive methods have emerged, framing the AQA task as a ranking problem, offering more comprehensive supervision. Doughty et al.[16] evaluated pairwise actions by learning discriminative and shared features. Their subsequent work[17] presented a model for rank-aware attention, learning the most informative segments for assessing skill quality. Yu et al.[9] proposed Contrastive Regression (CoRe) to learn relative scores through pairwise comparison, guiding the network to discern differences between videos. Bai et al.[7] proposed a Temporal Parsing Transformer to decompose holistic features into temporal part-level representations. Li et al.[8] introduced a pairwise contrastive learning network to guide training. Xu et al.[1] proposed a procedure-aware approach to parse pairwise videos into consecutive steps with diverse semantics, supervised by temporal segmentation annotations. A notable departure from prior approaches, Liu et al. [10] introduced a replay-guided temporal concentration module (TCM) that concurrently analyzes differences between examples, input, and its replay. We further extend [10] in this paper, incorporating a hierarchical joint training method for more reliable concentrated feature extraction, aiming to explore an effective scoring mechanism.

2.2 Multi-View Learning Architecture

In the domain of action recognition and video prediction, the exploration of multi-view learning architectures has garnered attention. S. Vyas et al.[18] delve into learning a comprehensive internal representation of multi-view videos, enabling the prediction of a video clip from an unseen viewpoint and time for action recognition. On a similar note, S. Yan et al.[19] propose a model incorporating separate encoders to characterize distinct views of the input video, utilizing lateral connections to amalgamate information across views for enhanced video understanding. While the common objective is to cultivate viewpoint-invariant representations, existing methods often amalgamate features from multiple views of a single video to delineate discrete class clusters. However, the AQA task necessitates the model to meticulously attend to subtle differences between the exemplar and input video actions. Consequently, our approach refrains from directly incorporating multi-view information from a single video. We underscore the significance of contrastive learning, employing the relative score of 0 to constrict the input video and its replay. This strategic choice is intended to mitigate the impact of the network on viewpoint changes and uphold the integrity of the learning process.

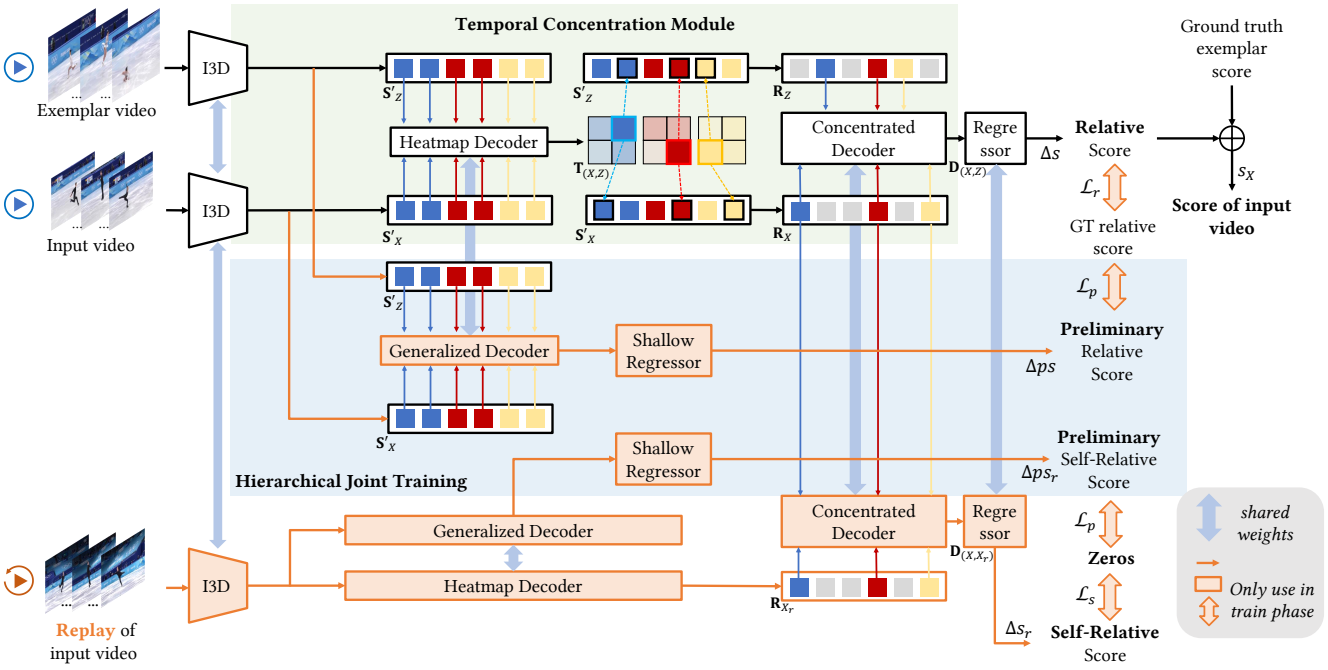

Fig. 1 The architecture of the proposed replay-guided triple-stream contrastive transformer with hierarchical joint training for action quality assessment. In the training phase, besides the loss between the relative score and ground truth, the other three losses are trained hierarchically. In the testing phase, the orange parts are not used. (Best viewed in color.)

2.3 Deeply Supervised Learning

To address the difficulty of optimization caused by a large number of layers, Lee et al. [20] first proposed deeply-supervised nets to directly supervise the intermediate layers of deep neural networks. In the past years, an increasing number of works applied deeply supervised learning for performance enhancement in various applications, such as object detection [21] and semantic segmentation [22]. Most of the existing work is directly using labels to regression features in the hidden layer. However, we do not have real labels of the heatmap to directly supervise the heatmap decoder. Therefore, we design a shallow module to perform indirect regression of heatmap generation by preliminary scores. Combined with deep supervision, the proposed hierarchical joint training benefits to the whole network to improve the accuracy.

3. Methodology

In this section, we first introduce the overview of the framework. Then, we revisit our temporal concentration module. Next, we describe our new hierarchical joint training method. Finally, we introduce the inference strategy.

3.1 Overview of Framework

The network architecture is shown in Figure 1. Given an input X, it corresponds to an exemplar Z and a replay of

Table 1 Notations and meanings.

Notation	Meaning
X	Input video
Z	Exemplar video
X_r	A replay video of the input X
T	Frame length of video
$\mathbf{F}$	Spatiotemporal features extracted by I3D
$\mathbf{S}$	Spatiotemporal features downsampled from $\mathbf{F}$
$\mathbf{S}'$	Generalized feature generated by group sampling
$\mathbf{T}$	Temporal heatmap
$\mathbf{P}$	The hottest point in $\mathbf{T}$
$\mathbf{R}$	Concentrated feature
$\mathbf{D}$	Fine-grained concentrated feature
$\mathbf{G}$	Generalized feature generated by generalized decoder
G	Group number of group resamling
M	Exemplars number of voting strategy
Δs	Relative score
Δs_r	Self-relative score
Δps	Preliminary relative score
Δps_r	Preliminary self-relative score
s_x	Final score of input X

the input X_r with the same frame length T. We extract spatiotemporal features $\mathbf{F}$ using the I3D backbone [23]. The I3D backbone extends 2D convolutional networks to 3D, allowing it to process both spatial dimensions (height and width) and the temporal dimension (time). This is crucial for capturing motion and dynamics in video data. Then we do the preprocess to features (down and grouping, omitted in the figure). In shallow layers, grouped generalized features are input to the generalized decoder with a shallow regressor, aiming to directly obtain the preliminary relative score to supervise the heatmap generation. In deep layers, we input grouped generalized features into the heatmap decoder to mine concentrated correlation, and then a dense resampling strategy is adopted according to the hot region in the heatmap to generate concentrated features. With the cooperation of shallow and deep layers, the proposed hierarchical joint training encourages the model to consider both generalized and concentrated features. Finally, the concentrated decoder and regressor quantify quality differences between concentrated features. Table 1 explains all notations used in subsequent sections.

3.2 Preprocessing

The utilization of high-dimensional features $\mathbf{F}$, integrating information across video clips, leads to significant redundancy and computational complexity. We employ the down module [1] to downsample the $\mathbf{F}$ to the $\mathbf{S} \in \mathbb{R}^{T \times D_S}$, where D_S is feature dimensions after downsampling. Aiming to reduce the noise information of long-term contrast, we evenly divide the feature $\mathbf{S}$ into G non-overlapping groups. Then we uniformly sample L frames in each group to keep the same length, aiming to meet the requirement that the dimensions of the query and key are the same in the transformer decoder [24]. The resulting generalized feature $\mathbf{S}^{'} \in \mathbb{R}^{G \times L \times D_S}$ is paired and inputted into a two-branched decoder module to facilitate the acquisition of a concentrated embedding through cross-attention.

3.3 Temporal Concentration Module

To extract the concentration of feature $\mathbf{S}^{'}$, we design a learnable temporal heatmap $\mathbf{T}_{(X,I)} \in \mathbb{R}^{G \times L \times L}$ to measure the concentrated correlation of pairwise features. Formally, the heatmap is represented as:

$$\mathbf{T}_{(X,I)} = \mathcal{P}\left(\frac{\exp^{(\delta_q(\mathbf{S}^{'}_X)\cdot\delta_k(\mathbf{S}^{'}_I)^T/\sqrt{s})}}{\sum_{k=1}^{n}\exp^{(\delta_q(\mathbf{S}^{'}_X)\cdot\delta_k(\mathbf{S}^{'}_I)^T/\sqrt{s})}}\right), I = Z, X_r, \tag{1}$$

where s is a scale factor in decoder [24], δ is a linear layer, $\mathcal{P}$ is an average pooling module to integrate multi-head attention features. The value of the heatmap $\mathbf{T}$ represents the spatiotemporal correspondence between the features of the pairwise videos, indicating the action difference. To ensure the correspondence reliability of the heatmap, a hierarchical joint training method is proposed, which is introduced in Section 3.4.

Utilizing the temporal heatmap as a basis, we employ a dense group resampling strategy on the feature vector $\mathbf{S}_I$. This approach aims to encourage the model to focus on nuanced differences by subsequently concentrating on the cross-attention decoder. To be more specific, we identify the highest-temperature point $\mathbf{P} \in \mathbb{R}^{G\times 2}$, representing the feature index in pairwise groups, within the heatmap of each group. The $\mathbf{P}$ is defined as:

$$\mathbf{P}_{(X,I)} = \left\{\text{argmax}\ \mathbf{T}^g_{(X,I)}(x,y)\right\}_{g=0}^{G}, \quad I = Z, X_r. \tag{2}$$

To extract concentrated features, we design a group resampling method. For each group, the feature $\mathbf{S}_X$ set an anchor on index $x_g \in \mathbf{P}^g_{(X,Z)}$ in the temporal dimension, and $\mathbf{S}_Z$ set an anchor on index $y_g \in \mathbf{P}^g_{(X,Z)}$. Then the feature $\mathbf{S}_X$ and $\mathbf{S}_Z$ are searched in forward and backward of the anchor with range μ to give a concentration range $[x_g-\mu, x_g+\mu]$ and $[y_g-\mu, y_g+\mu]$. We uniformly resample in each concentration range to keep the same length for the further decoder. Then the pairwise concentrated features $(\mathbf{R}_X, \mathbf{R}_I) \in \mathbb{R}^{G\times L\times D_S}$ are input to the concentrated cross-attention decoder to predict the score.

To mine the deep relation between the pairwise features, we design a concentrated decoder. The decoder leverages the seq-to-seq representation of the multi-head attention, intending to discern fine-grained differences in correspondence between concentrated features. For a pairwise concentrated feature $(\mathbf{R}_X, \mathbf{R}_I)$, the decoder identifies correspondence to produce a fine-grained concentrated feature $\mathbf{D} \in \mathbb{R}^{G\times L\times D_S}$.S The operation is represented as:

$$\mathbf{D}^{'}_{(X,I)} = \text{softmax}\left(\frac{\delta_q(\mathbf{R}_X)\cdot\delta_k(\mathbf{R}_I)^T}{\sqrt{s}}\right)\cdot\delta_v(\mathbf{R}_I), \tag{3}$$

$$\mathbf{D}_{(X,I)} = \text{MLP}(\mathbf{D}^{'}_{(X,I)}) + \mathbf{R}_X, \tag{4}$$

where s is a scale factor, and δ is a linear layer, the MLP block contains two layers with a GELU non-linearity.

To determine the relative score for pairwise features, a regressor denoted as $\mathcal{R}$ is employed to combine the contrasting regression components for each group. In the training process, regression is performed for both X, Z, and X, X_r, indicating that replay information is utilized to guide the training process. The regression of the training phase is represented as:

$$\Delta s = \frac{1}{G}\sum_{g=0}^{G}\mathcal{R}(\mathbf{D}^g_{(X,Z)}), \tag{5}$$

$$\Delta s_r = \frac{1}{G}\sum_{g=0}^{G}\mathcal{R}(\mathbf{D}^g_{(X,X_r)}). \tag{6}$$

3.4 Hierarchical Joint Training

In previous stages, we assessed the relative score by regressing the concentrated feature. However, direct supervision in

deep layers causes shallow gradient vanishing, affecting the reliability of the heatmap generated by the heatmap decoder, and ultimately affecting the stability of the entire scoring mechanism. Recent work [25] pointed out that shallow supervision is very effective in promoting model convergence. Motivated by this, we propose a hierarchical joint training method, consisting of a generalized decoder and joint losses. Specifically, as shown in Figure 2, we cooperate with the generalized feature $\mathbf{S}^{'}$ and the concentrated feature $\mathbf{R}$. Since $\mathbf{R}$ is generated from $\mathbf{S}^{'}$, supervising the process directly with a score at the deep level results in an unreliable and non-transparent generation process. To end this, we design a generalized decoder, which utilizes $\mathbf{S}^{'}$ to obtain the preliminary score. The resulting preliminary score is supervised in the optimization stage by directly updating feature concentration in $\mathbf{S}^{'}$ at shallow layers.

3.4.1 Generalized Decoder

Different from the heatmap decoder, the purpose of the generalized decoder is not to generate a heatmap, but to directly compare and learn the relationship between pairs of generalized features. The generalized decoder captures the spatial and temporal correspondences of small action differences in different aspects through a multi-head cross-attention mechanism, and generates new features among pairwise features. The generalized decoder is represented as:

$$\mathbf{G}_{(X,I)} = \mathrm{MLP}(\mathrm{MHCA}(\delta(\mathbf{S}^{'}_{X}), \delta(\mathbf{S}^{'}_{I}))) + \mathbf{S}^{'}_{X}, \tag{7}$$

where δ is a linear layer. The transformer decoder consists alternating layout of MHCA(Multi-Head Cross-Attention) and MLP(MultiLayer Perceptron).

3.4.2 Shallow Regressor

Based on the embedding features $\mathbf{G}$ learned in the previous step, we quantify the deviation between pairs of inputs by learning relative scores in advance. This guides the network to learn at a shallow level to evaluate action quality, improving overall reliability. To achieve this, a regression module is then connected to obtain the preliminary score. The preliminary score provides shallow supervision for the network during the optimization process, which is described in detail in the next subsection. Note that the preliminary score is not output as a final result.

The shallow regressor consists of three linear layers and two ReLU layers alternately, and outputs the preliminary score by parsing pairs of $\mathbf{G}$. The shallow regressor $\mathcal{SR}$ is formulated as:

$$\Delta ps = \frac{1}{G}\sum_{g=0}^{G} \mathcal{SR}(\mathbf{G}^{g}_{(X,Z)}), \tag{8}$$

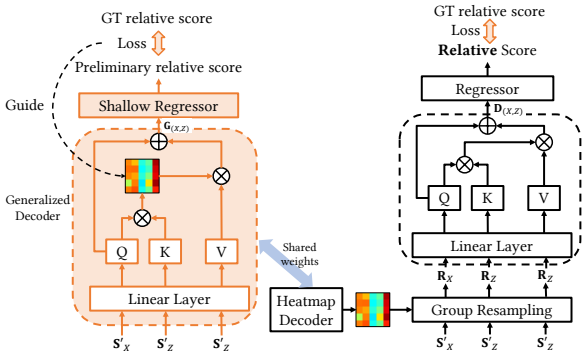

Fig. 2 Hierarchical Joint Training. We hierarchically regress the score from both shallow layers (generalized decoder) and deep layers (concentrated decoder). We use a loss to supervise the heatmap generation in the shallow layer. (We use input X and exemplar Z as an example. Best viewed in color)

$$\Delta ps_r = \frac{1}{G}\sum_{g=0}^{G} \mathcal{SR}(\mathbf{G}^{g}_{(X,X_r)}), \tag{9}$$

where the Δps is the preliminary score. To judge the intensity of each part in the generalized feature, the shallow branch is trained with TCM hierarchically. The attention weight of the generalized decoder is shared with Eq. (1).

3.4.3 Joint Loss and Optimization

We design three loss functions for joint training, which are self-relative loss, relative loss, and preliminary loss.

Given the availability of replay information, it becomes imperative to devise an appropriate loss function to steer the learning process. As both the input video and its corresponding replay inherently depict the same action performed by the same athlete, their associated scores should exhibit consistency. Taking inspiration from unsupervised learning principles, we introduce a self-replay loss function rooted in the notion of consistency. This loss function calculates the mean squared error between the self-relative score and zero, effectively emphasizing the alignment between these scores and enforcing consistency in the learning process. The self-relative loss function is represented as:

$$\mathcal{L}_s = ||\Delta s_r - 0||^2. \tag{10}$$

Besides, we also use the relative loss $\mathcal{L}_r$ to supervise the training process, which is represented as:

$$\mathcal{L}_r = ||\Delta s + \hat{s}_Z - \hat{s}_X||^2, \tag{11}$$

where $\hat{s}_X$ and $\hat{s}_Z$ is the groundtruth score of input and exemplar.

Additionally, a hierarchical joint training method contributes a shallow regression for more reliable extraction of the concentrated feature by supervision at shallow layers. We design a preliminary loss function, which is represented as:

$$\mathcal{L}_p = ||\Delta ps_r - 0||^2 + ||\Delta ps + \hat{s}_Z - \hat{s}_X||^2. \quad (12)$$

The final joint loss function is represented as:

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_s + \mathcal{L}_p. \quad (13)$$

Through the hierarchical joint training of the three loss functions, it not only improves the overall performance of the network, but also accelerates convergence.

3.5 Inference

In the inference phase, we only use the two-stream structure with input X and exemplar Z. We also adopt a multi-exemplar voting strategy. Given an input video X, we select M exemplars from training data to construct M pairs using these M different exemplars (X, Z_j) whose scores are s_{Z_j}. The inference progress is represented as:

$$s_X = \frac{1}{M}\sum_{j=1}^{M}(\mathcal{F}(X, Z_j) + s_{Z_j}), \quad (14)$$

where $\mathcal{F}$ means overall proposed framework.

4. Experiment

4.1 Dataset and Evaluation Metrics

We introduce the dataset and evaluation metrics in this subsection.

RFSJ (Replay Figure Skating Jumping) dataset. [10] Although there are many existing AQA datasets, most of them only provide a single-view video from the broadcasting. Since these datasets lack the additional video from other camera views, it is unfair and unsuitable to compare existing methods on these datasets. The conference version [10] proposed the RFSJ dataset including replay information with another camera view for action quality assessment. RFSJ focuses on various types of jumping actions, it consists of 768 live video sequences and 536 replay video sequences. These sequences are collected from the Olympic and European Championship figure skating competition videos. We randomly select 75 percent of live sequences for training and 25 percent of live sequences for testing.

Evaluation Metric. Following prior studies [1] [7] [10], we assess our experiments using two metrics.

- Spearman's rank correlation (ρ) aims to evaluate the ranks of the predicted scores. ρ is defined as:

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}}, \quad (15)$$

  where p and q denote the ranking of two series. The higher ρ means the result is better.
- Relative L2 distance (R-ℓ_2) aims to evaluate the numerical values of the predicted scores. R-ℓ_2 is defined as:

Table 2 Comparisons with state-of-the-art methods. w/ RP indicates using replay information; w/ AT indicates selecting exemplars by action type.

Method (w/ AT, w/ RP)	$\rho \uparrow$	R-ℓ_2 ($\times 100$) $\downarrow$
USDL [5]	0.8577	3.8001
CoRe [9]	0.9312	0.5551
TPT [7]	0.9317	0.5523
TSA [1]	0.8990	0.7449
TCM (conf.) [10]	0.9346	0.5500
Ours	**0.9642**	**0.2883**

Table 3 Ablation study of replay information and action type.

Method (w/ AT, w/o RP)	$\rho \uparrow$	R-ℓ_2 ($\times 100$) $\downarrow$
USDL [5]	0.6771	3.9765
CoRe [9]	0.9132	0.6837
TPT [7]	0.9154	0.6829
TSA [1]	0.8986	0.8168
TCM (conf.) [10]	0.9152	0.6784
Ours	**0.9562**	**0.4994**
Method (w/o AT, w/ RP)	$\rho \uparrow$	R-ℓ_2 ($\times 100$) $\downarrow$
USDL [5]	0.8577	3.8001
CoRe [9]	0.8606	1.3170
TPT [7]	**0.8620**	**1.3069**
TSA [1]	0.7905	2.6536
TCM (conf.) [10]	0.8005	1.7011
Ours	0.8090	1.6164

$$R\text{-}\ell_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|s_i - \hat{s}_i|}{s_{max} - s_{min}}\right)^2, \quad (16)$$

where s_i and $\hat{s}_i$ mean predicted score and groundtruth for the i-th sample. The lower R-ℓ_2 means the result is better.

In subsequent section 4.3, we assess our experimental results in comparison to state-of-the-art approaches on the RFSJ dataset. In section 4.4, we conduct ablation studies to assess the efficacy of the proposed model components and designs.

4.2 Implementation Detail

We implement our proposed method in PyTorch, and our experiments are conducted with an Nvidia RTX 3090 GPU on Ubuntu 20.04. We adopt the I3D pretrained on Kinetics [23] as the initialization of backbone $\mathcal{B}$ for spatiotemporal feature extraction. Following [1], [5], [9], [10], we uniform sample 96 frames for each video, which means $T = 96$. Then we split each video into 9 overlap clips with the same length, containing 16 continuous frames. Specifically, we took No. [0, 10, 20, 30, 40, 50, 60, 70, 80] frame as the start frame for 9 overlap clips. After the downsampling preprocess mentioned in section 3.2, the feature dimension D_S is 96. We input replay sequences to the third stream to guide the model. In some cases, the input sequence does

not have a corresponding replay sequence, so we use a data augmentation strategy. Specifically, we utilize zoom scale transformation and horizontal mirror flip of the input sequence to simulate the replay sequence. We group sample features $\mathbf{S}$ into $G = 3$ groups with $L = 5$ length and resample features in concentration range $\mu = 3$. The shallow regressor and deep regressor modules both contain three hidden linear layers, which are ReLU(FC(64,256)), ReLU(FC(256,64)), and FC(64,1), to generate the predicted score. The Adam optimizer with zero weight decay is utilized with a learning rate of 10^{-4} for backbone $\mathcal{B}$, and 10^{-3} for our proposed network. In the testing phase, we set the voting number M as 10 in the multi-exemplar voting strategy.

Note that, in previous methods [1], [5], [7], [9], the Difficulty Degree (DD) is used as prior information for training, because the DD is fixed before the competition. However, the base value (similar to the difficulty degree) in figure skating is changed according to the quality of the athlete's action. So the base value belongs to the posterior information and cannot be used to assist training. Therefore, we select exemplars from the same Action Type (AT) in the training set. Because all action types are confirmed in the program list before the competition, and athletes perform in sequence during the competition.

4.3 Comparison to State-of-the-art

We present quantitative experimental results comparing our method with recent state-of-the-art AQA approaches on the RFSJ dataset, as detailed in Table 2. To ensure a fair comparison, we incorporate the replay sequences into the training set and conduct training for these methods using the RFSJ dataset. The comparative analysis demonstrates the superiority of our method, establishing it as state-of-the-art in AQA performance. For a comprehensive evaluation, we consider two key factors in our experiments. The notation "w/ AT" signifies that both the training and test processes utilize action type labels, exclusively selecting exemplars from the same action type. Conversely, "w/ RP" indicates that the training process incorporates replay sequence data. As indicated in Table 2, under both conditions (w/ AT and w/ RP), our method outperforms all other approaches, including the prior version [10], achieving a Spearman's rank correlation of 0.9642 and an R-ℓ_2 value of 0.2883.

This notable performance improvement is primarily attributed to our proposed temporal concentration module and hierarchical joint training. The TCM enhances results by capturing the concentration of temporal features. Furthermore, the proposed hierarchical joint training method provides supervision at both shallow and deep layers. Consequently, our method achieves more accurate score predictions through regression, showcasing the efficacy of our proposed architecture in advancing the state-of-the-art in AQA on the RFSJ dataset.

As Figure 3 shows, we also present visualization results of some example sequences.

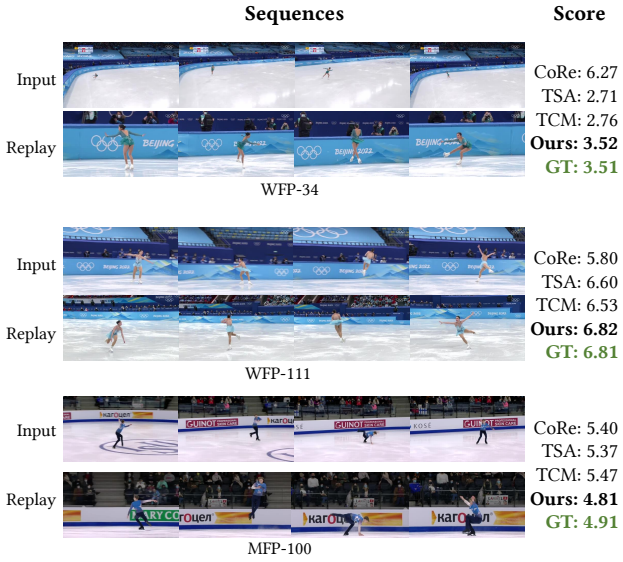

Fig. 3 Visualization results of state-of-the-art works and ours. GT means the ground truth

Table 4 Ablation study of different components

Method	TCM	$\mathcal{L}_s$	$\mathcal{L}_p$	$\rho\uparrow$	R-ℓ_2 ($\times100$)$\downarrow$
Baseline	×	×	×	0.8990	0.7449
	√	×	×	0.9209	0.6579
	×	√	×	0.9224	0.6305
TCM (conf.) [10]	√	√	×	0.9346	0.5500
Ours	√	√	√	**0.9642**	**0.2883**

4.4 Ablation Study

We conducted multiple experiments to present the outcomes of our ablation studies.

4.4.1 Ablation Study of Replay Information and Action Type

Table 3 illustrates the impact of replays and action types on the experimental outcomes. The configuration labeled "w/o RP" denotes training without the inclusion of replay sequences, while "w/o AT" signifies the selection of exemplars from random action types.

Comparing the results under the "w/o RP" setting reveals that the performance of all methods is inferior compared to the "w/ RP" configuration. This observation underscores the positive influence of replay information within our proposed RFSJ dataset [10] on enhancing the performance of existing state-of-the-art methods. Notably, our method achieves the highest Spearman's rank correlation and R-ℓ_2 result, showcasing its effectiveness in leveraging replay information. It is crucial to highlight that this superior performance is attributed to the innovative design of the TCM

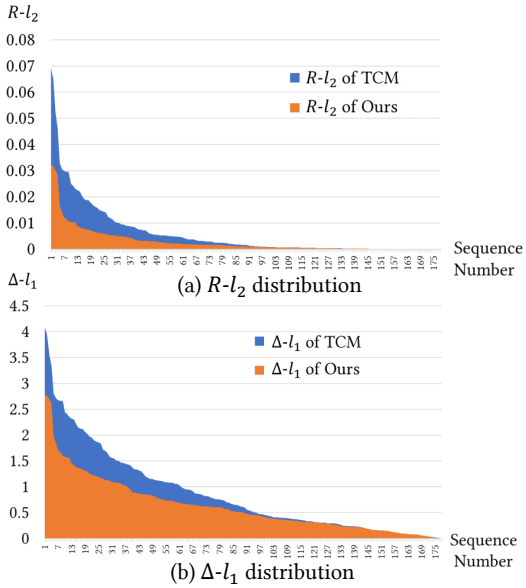

Fig. 4 R-ℓ_2 and Δ-ℓ_1 distribution, where Δ-ℓ_1 means the absolute difference between the predicted score and the ground truth. Our method has better results on both R-ℓ_2 and Δ-ℓ_1 distributions.

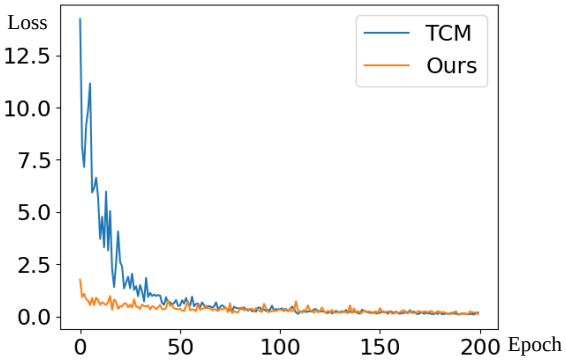

Fig. 5 Loss decrease curve of epochs.

coupled with a novel hierarchical joint training method. This combination enables the effective extraction of differences within the same action type under the supervision of action types (AT).

Conversely, under the "w/o AT" setting, our method lags behind state-of-the-art methods. This outcome is rooted in the fact that replay information, which guides network learning, is centered on the same action of the same athlete. In this scenario, our method focuses on discerning differences caused by action types rather than evaluating action quality, leading to a comparative decline in performance.

We compare the results in Table 3 vertically and find that selecting exemplars by action type (w/ AT) affects the results more than replay information (w/ RP). This is because all two-stream contrastive networks rely on AT, which was demonstrated in previous works [1], [7], [10], [25]. Replay and Input video must belong to the same AT, so RP relies on AT supervision and improves network performance under w/ AT conditions. In the absence of AT supervision, since the exemplar may choose other action types, the RP information limits the network's learning process, thereby damaging the network's performance.

4.4.2 Ablation Study of Different Components

Table 4 provides an analysis of the contribution of each component under the conditions of w/ AT and w/ RP. In comparison to the baseline, employing only TCM yields a notable 0.0219 enhancement in Spearman's rank correlation and a 0.087 improvement in R-ℓ_2. This outcome underscores the efficacy of our proposed TCM, which adeptly concentrates temporal features, and mitigates the interference of redundant features. Simultaneously, the triple-stream contrastive transformer is guided by the self-replay loss function $\mathcal{L}_s$. The introduction of the replay-guided method contributes a substantial 0.0234 improvement in correlation and a 0.114 enhancement in R-ℓ_2. Further augmenting the process, the model is supervised by $\mathcal{L}_p$ through the incorporation of a novel hierarchical joint training method. This integration results in a substantial 0.0652 improvement in correlation and a significant 0.4566 improvement in R-ℓ_2 compared to the baseline. The enhancement of results is due to the contribution of the joint training of deep and shallow layers.

Figure 4 shows the detail of result distribution, we observe that the proposed hierarchical joint training method has better R-ℓ_2 and absolute delta score results on both maximum and average values. It proves the effectiveness of the proposed hierarchical joint training.

Figure 5 shows the loss decrease curve during the training process, we observe that the loss value not only decreased rapidly at the initial stage, but also fluctuated smoothly in the subsequent stage. It proves that the proposed hierarchical joint training method also extremely speeds up the convergence of the network.

By combining each component, the final result attains a state-of-the-art performance.

4.4.3 Ablation Study of Different Concentration Range

The parameter μ serves as an indicator of the concentration degree of the TCM, prompting us to conduct a series of experiments. Figure 6 provides a summary of the performance with varying values of μ, including 1, 3, 5, 8, and 10. Two experiment setups were designed: one utilizing only the loss function $\mathcal{L}_s$, and the other implementing the hierarchical joint training method, incorporating both $\mathcal{L}_s$ and $\mathcal{L}_p$. Under the $\mathcal{L}_s$ setup. We observe distinct trends in performance as μ varies. Specifically, when μ ranges from 1 to 3, there is a noteworthy improvement of 1.02% and 0.0643 in Spearman's rank correlation and R-ℓ_2, respectively. However, as μ increases from 3 to 5, there is a deterioration, resulting in a decrease of 1.90% in correlation and 0.16 in R-ℓ_2. Subsequently, with further increases in μ, the performance stabilizes with slight fluctuations. In the $\mathcal{L}_s + \mathcal{L}_p$ setup, a similar distribution of performance is observed with

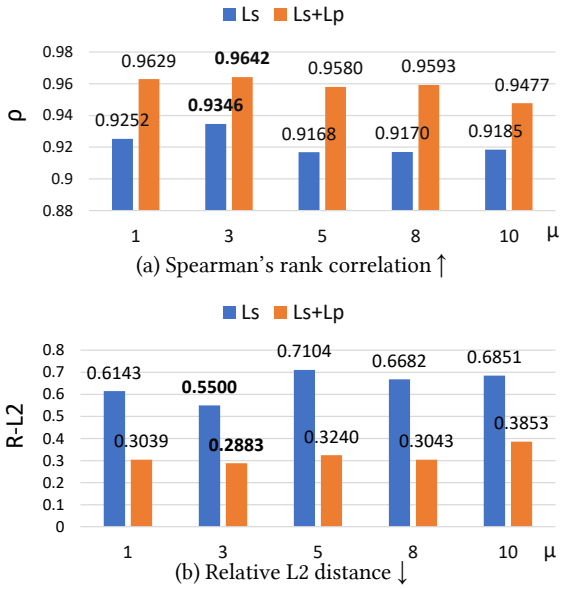

Fig. 6 Ablation study of concentration range μ. Note that $+\mathcal{L}_p$ means utilizing the hierarchical joint training method.

varying μ. Notably, when μ is set to 3, it yields the optimal performance in terms of both Spearman's rank correlation and R-ℓ_2. We attribute this observation to the nuanced nature of the concentration range: too narrow a range limits the model's perception of temporal actions, while too wide a range introduces redundant features, hindering the model from effectively discerning differences in actions. Furthermore, our analysis reveals that the introduction of $\mathcal{L}_p$ contributes to an enhancement in the generality of results, as evidenced by the comparison between the results of the two experiment setups. This underscores the beneficial impact of incorporating the additional loss function in refining the overall performance of the proposed model.

5. Conclusion

In this paper, we introduce a novel approach for action quality assessment of figure skating jumping, termed the replay-guided triple-stream contrastive transformer with hierarchical joint training. Our proposed method leverages replay sequences to facilitate training, enabling the learning of nuanced action quality quantization across various views and zoom scales. The proposed Temporal Concentration Module directs the model's focus toward discerning features related to athletes' errors or highlights critical elements influencing scoring outcomes. Additionally, the proposed Hierarchical Joint Training method is employed to provide supervision on both shallow and deep layers, enhancing the performance of the scoring mechanism and speed of training convergence. Extensive experiments on the RFSJ dataset prove the proposed method achieves an effective scoring mechanism and replay information is promoted to existing AQA methods. Our study delves into the impact of replays in the realm of figure skating for AQA. Looking ahead, we aspire to advocate for the collection of additional replay information from diverse camera views. This undertaking aims to foster a more thorough exploration of scoring mechanisms and a deeper understanding of actions, not only within figure skating but also in other domains.

6. Acknowledgments

This work was supported by KAKENHI (21K11816), and the National Natural Science Foundation of China under Grant 62006178.

References

[1] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2949–2958, 2022.

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

[3] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal Pyramid Network for Action Recognition," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp.588–597, IEEE, June 2020.

[4] R. Li, L. Yan, Y. Peng, and L. Qing, "Lighter Transformer for Online Action Detection," Proceedings of the 2023 6th International Conference on Image and Graphics Processing, Chongqing China, pp.161–167, ACM, Jan. 2023.

[5] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.9839–9848, 2020.

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

[7] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV, pp.422–438, Springer, 2022.

[8] M. Li, H.B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.X. Du, "Pairwise contrastive learning network for action quality assessment," Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV, pp.457–473, Springer, 2022.

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

[10] Y. Liu, X. Cheng, and T. Ikenaga, "A figure skating jumping dataset for replay-guided action quality assessment," ACM Multimedia (MM2023), 11 2023.

[11] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13, pp.556–571, Springer, 2014.

[12] P. Parmar and B. Tran Morris, "Learning to score olympic events," Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp.20–28, 2017.

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

Recognition, pp.304–313, 2019.
[14] G. Bertasius, H. Soo Park, S.X. Yu, and J. Shi, "Am i a baller? basketball performance assessment from first-person videos," Proceedings of the IEEE international conference on computer vision, pp.2177–2185, 2017.
[15] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.G. Jiang, and X. Xue, "Learning to score figure skating sport videos," IEEE transactions on circuits and systems for video technology, vol.30, no.12, pp.4578–4590, 2019.
[16] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.6057–6066, 2018.
[17] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.7862–7871, 2019.
[18] S. Vyas, Y.S. Rawat, and M. Shah, "Multi-view action recognition using cross-view video prediction," Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, pp.427–444, Springer, 2020.
[19] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, and C. Schmid, "Multiview transformers for video recognition," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.3333–3343, 2022.
[20] C.Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," Artificial intelligence and statistics, pp.562–570, Pmlr, 2015.
[21] C. Li, M. Zeeshan Zia, Q.H. Tran, X. Yu, G.D. Hager, and M. Chandraker, "Deep supervision with shape concepts for occlusion-aware 3d object parsing," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.5465–5474, 2017.
[22] Y. Zhang and A.C. Chung, "Deep supervision with additional labels for retinal vessel segmentation task," Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, pp.83–91, Springer, 2018.
[23] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6299–6308, 2017.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol.30, 2017.
[25] L. Zhang, X. Chen, J. Zhang, R. Dong, and K. Ma, "Contrastive deep supervision," European Conference on Computer Vision, pp.1–19, Springer, 2022.

Yanchao Liu received a B.E. degree in information engineering from the South China University of Technology, Guangzhou, China, in 2017, and an M.E. degree in engineering of information, production, and systems, in 2019, from Waseda University, Kitakyushu, Japan, where he is currently working towards the Ph.D. degree. His research focuses on sports analysis of computer vision algorithms.

Xina Cheng received her B.E. degree in the School of Optoelectronics from the Beijing Institute of Technology, China, in 2014. She received her M.E. degree and a Ph.D. degree from the Graduate School of Information, Production, and Systems of Waseda University, Japan, in 2015 and 2018 respectively. She is currently a lecturer at Xidian University. Her current research interests are sports analysis and computer vision.

Takeshi Ikenaga received his B.E. and M.E. degrees in electrical engineering and Ph.D degree in information computer science from Waseda University, Tokyo, Japan, in 1988, 1990, and 2002, respectively. He joined LSI Laboratories, Nippon Telegraph and Telephone Corporation (NTT) in 1990, where he had been undertaking research on the design and test methodologies for high performance ASICs, a real-time MPEG2 encoder chip set, and a highly parallel LSI system design for image understanding processing. He is presently a professor in the integrated system field of the Graduate School of Information, Production and Systems, Waseda University. His current interests are image and video processing systems, which covers video compression (e.g. VVC, SCC), video filter (e.g. super resolution, high-dynamic range imaging), and video recognition (e.g. sports analysis, ultra-high speed and ultra-low delay vision system). He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE), a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) and the Information Processing Society of Japan (IPSJ).