Multimodal Action Quality Assessment

Ling-An Zeng and Wei-Shi Zheng

***Abstract*— Action quality assessment (AQA) is to assess how well an action is performed. Previous works perform modelling by only the use of visual information, ignoring audio information. We argue that although AQA is highly dependent on visual information, the audio is useful complementary information for improving the score regression accuracy, especially for sports with background music, such as figure skating and rhythmic gymnastics. To leverage multimodal information for AQA, *i.e.*, RGB, optical flow and audio information, we propose a Progressive Adaptive Multimodal Fusion Network (PAMFN) that separately models modality-specific information and mixed-modality information. Our model consists of with three modality-specific branches that independently explore modality-specific information and a mixed-modality branch that progressively aggregates the modality-specific information from the modality-specific branches. To build the bridge between modality-specific branches and the mixed-modality branch, three novel modules are proposed. First, a Modality-specific Feature Decoder module is designed to selectively transfer modality-specific information to the mixed-modality branch. Second, when exploring the interaction between modality-specific information, we argue that using an invariant multimodal fusion policy may lead to suboptimal results, so as to take the potential diversity in different parts of an action into consideration. Therefore, an Adaptive Fusion Module is proposed to learn adaptive multimodal fusion policies in different parts of an action. This module consists of several FusionNets for exploring different multimodal fusion strategies and a PolicyNet for deciding which FusionNets are enabled. Third, a module called Cross-modal Feature Decoder is designed to transfer cross-modal features generated by Adaptive Fusion Module to the mixed-modality branch. Our extensive experiments validate the efficacy of the proposed method, and our method achieves state-of-the-art performance on two public datasets. Code is available at https://github.com/qinghuannn/PAMFN.**

***Index Terms*— Action quality assessment, multimodal learning, video understanding.**

I. Introduction

ACTION quality assessment (AQA) is the task of assessing how well an action is performed and is usually modeled as a score regression problem. Different from tasks such as action recognition and action localization, AQA is a fine-grained action understanding task, which not only needs to recognize actions but also to understand the subtle differences among actions. For instance, an insufficient leg lift angle in a leg-lifting action does not significantly affect action recognition but results in substandard actions and penalties [1]. AQA has many important real-world applications, including in sports [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], surgical training [7], [8], [13], [17] and other fields [18], [19], [20], [21].

Audio usually serves as complementary information in action recognition [22], [23], [24], [25], [26] and action localization [27], [28], [29], [30], [31], and has been shown to significantly improve performance. Similarly, although AQA is highly dependent on visual information, audio can be used as complementary information to improve the score regression accuracy, especially in sports with background music. In skating and rhythmic gymnastics, athletes need to perform actions to the rhythm of music, and a mismatch between an athlete's action and the rhythm of the music will result in a penalty. Therefore, exploring the consistency of the athlete's action and the rhythm of music is necessary to achieve accurate action quality assessment. Moreover, since the high technical action is usually accompanied by a drastic musical rhythm, music rhythms are natural signals for guiding viewers to focus on important parts of an action. However, previous AQA works [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] focus on only visual information and therefore cannot explore such a relation. Thus, we aim to design a multimodal action quality assessment model in this work.

It is a fact that the main advantage of multimodal methods is utilizing rich and diverse information from different modalities. However, existing multimodal works of other tasks [22], [23], [24], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42] model modality-specific and multimodal information simultaneously, causing that the information from different modalities inevitably influences each other. In such a way, it is almost impossible to extract pure modality-specific information, and hard to ensure that the network can extract modality-specific information instead of modality-general information. Thus, these methods cannot fully utilize rich and diverse information from different modalities.

To solve the above problem, we propose a multimodal AQA model called **Progressive Adaptive Multimodal Fusion Network** (PAMFN), which focuses on separately modeling modality-specific information and mixed-modality information. Specifically, our method separately models

Manuscript received 26 October 2022; revised 22 September 2023; accepted 19 January 2024. Date of publication 19 February 2024; date of current version 23 February 2024. This work was supported in part by NSFC under Grant U21A20471 and Grant U1911401 and in part by the Guangdong NSF Project under Grant 2023B1515040025 and Grant 2020B1515120085. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Charith Abhayaratne. *(Corresponding author: Wei-Shi Zheng.)*

Ling-An Zeng is with the School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, Guangdong 519082, China (e-mail: zenglan3@mail2.sysu.edu.cn).

Wei-Shi Zheng is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong 510275, China, also with the Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou, Guangdong 510275, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, Ministry of Education, Guangzhou, Guangdong 510275, China (e-mail: wszheng@ieee.org; zhwshi@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TIP.2024.3362135

modality-specific information and mixed-modality information to extract pure modality-specific information via three modality-specific branches and a mixed-modality branch. Thus, our PAMFN is designed as a pyramid architecture with N stages, and the information in the three modality-specific branches is progressively aggregated in the mixed-modality branch during each stage. In addition, to build the bridge between modality-specific branches and the mixed-modality branch, three novel modules are proposed to transfer information from modality-specific branches to the mixed-modality branch.

More specifically, to transfer modality-specific information to the mixed-modality branch, a **Modality-specific Feature Decoder** (MSFD) module is proposed. Since modality-specific features are aggregated in the mixed-modality branch during each stage, the three modality-specific branches and mixed-modality branch contain certain overlapping information after the first stage. Therefore, the MSFD components aims to decode the unseen or neglected information in the mixed-modality branch from modality-specific features.

Since different parts of an action can be always diverse, we argue that using an invariant multimodal fusion policy may lead to suboptimal results. To solve this issue, an **Adaptive Fusion Module** (AFM) composed of several fusion networks (FusionNets) and a policy network (PolicyNet) is proposed to learn an adaptive multimodal fusion policy. Different FusionNets are used to explore various fusion strategies and the PolicyNet is designed to adaptively select the optimal fusion strategies. In addition, by taking the similarity and diversity among video segments into consideration, we assume that some FusionNets are more general and some are more specific[1]; therefore, we propose a novel method called ranked FusionNets in which all FusionNets are ranked and the FusionNet with a higher rank indicates that it is more general. We use the Straight-Through Gumbel Estimator [43] to ensure that the decision process is differentiable.

After obtaining the cross-modal features generated by Adaptive Fusion Module, a module called **Cross-modal Feature Decoder** (CMFD) module is designed to transfer the cross-modal features to the modality-specific branch. Different from the MSFD module, CMFD aims to decode the unseen or neglected information in the mixed-modality branch and modality-specific branches from cross-modal features.

To demonstrate the effectiveness of our method, we conduct extensive experiments on two public action assessment datasets, *i.e.*, the Rhythmic Gymnastics dataset [1] and the Fis-V dataset [5]. Our method achieves state-of-the-art performance on both datasets, showing the advantages of our proposed model. In addition, we evaluate the proposed method on another task, *i.e.*, highlight detection, to demonstrate the generalizability of our method. The code for our approach will be released after publication.

Our main contributions can be summarized as follows:

- We propose a novel multimodal architecture, called Progressive Adaptive Multimodal Fusion Network (PAMFN), for action assessment that separately models modality-specific information and mixed-modality information, and progressively transfers information from modality-specific branches to the mixed-modality branch. To the best of our knowledge, it is the first work to perform action assessment with audio information.
- We propose an Adaptive Fusion Module with a novel ranked FusionNets strategy to learn an adaptive multimodal fusion policy, and use the ST Gumbel Estimator to efficiently train our model.
- We propose a Modality-specific Feature Decoder module and a Cross-modal Feature Decoder module to selectively transfer modality-specific information and cross-modal information to the mixed modality branch.

[1]If a fusion policy (FusionNet) is suitable for more parts of an action, it is more general. If a fusion policy is suitable for less parts of an action, it is more specific.

II. Related Works

A. Action Quality Assessment

Action quality assessment aims to assess the quality of an action and is a fine-grained action understanding task. Based on the length of processed videos, existing works can be divided into two categories: methods for short videos (averagely several seconds) and methods for long videos (averagely several minutes). Many works have focused on AQA for short videos and achieved remarkable progress. Wang et al. [44] propose a network to capture rich spatio-temporal contextual information in human motion. Bai et al. [15] propose a temporal parsing transformer to extract fine-grained temporal part-level representations. On the other hand, quite a few works focus on long videos. Zeng et al. [1] explore static information and dynamic information for AQA and propose a context-aware attention to learn context information of each segment. Xu et al. [3] design a Grade-decoupling Likert Transformer to explore the comprehensive effect of different grades exhibited in the video on the score.

Existing methods explore only visual information in videos ignoring audio information, which is an important cue for assessing the consistency of movement and the rhythm of the music and is a natural signal for guiding us to focus on the important parts of an action. Thus, in this work, we propose a multimodal network that leverages RGB, audio and optical flow information to explore modality-specific information and mixed-modality information for AQA. Since only long video datasets contain audio among existing AQA datasets, *i.e.*, Fis-V dataset [5] and Rhythmic Gymnastics dataset [1], our method focuses on long videos and all experiments are conducted on Fis-V dataset and Rhythmic Gymnastics dataset. However, as mentioned in [1], [5], works [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [17], [18], [19], [20], [21] designed for short videos are hard to work on long video datasets. Thus, these works will not be compared.

B. Multimodal Video Understanding

Multimodal video understanding refers to leveraging different representational modes to understand video, such as RGB frames, optical flows, audio and text. Since the action is usually accompanied with sounds, rich multimodal works [22], [23], [24], [27], [28], [29], [30], [31], [32], [33], [33], [34],

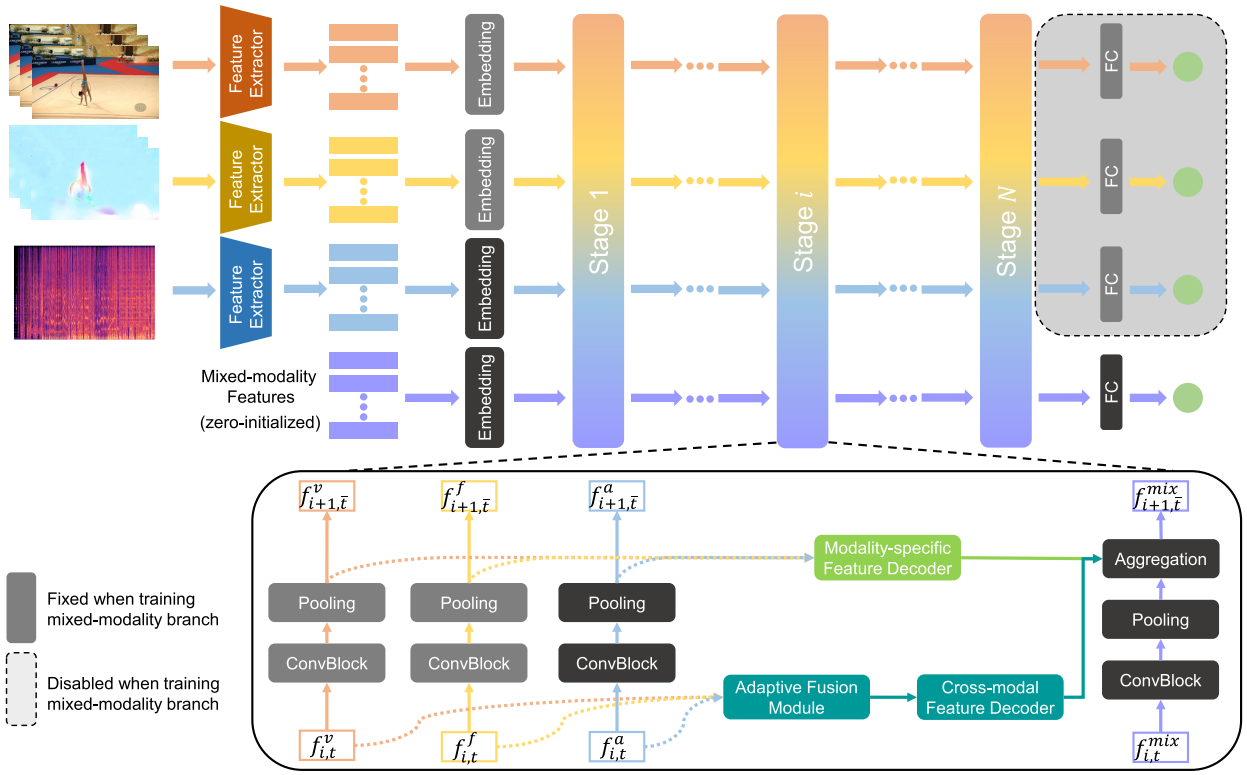


Fig. 1. The overall architecture of our proposed PAMFN. The RGB, optical flow and audio information are fed into three pretrained backbones to extract features respectively. Three modality-specific branches with the same structure are independently pretrained to explore the modality-specific information. Then, a mixed-modality branch progressively aggregates the modality-specific information via a Modality-specific Feature Decoder (MSFD) module and the cross-modal information via an Adaptive Fusion Module (AFM) and a Cross-modal Feature Decoder (CMFD) module. The Adaptive Fusion Module explores an adaptive cross-modal fusion policy. Our network is trained in two phases. We first separately train the modality-specific branches. Then we fix modality-specific branches except the audio branch since the quality of an action is almost impossible to assess using only audio information, and train the mixed-modality branch, MSFD, AFM and CMFD. Note that the initial mixed-modality features are initialized by zeros.

[35], [36], [37], [38] focus on the interaction of RGB, optical flow and audio information for video understanding. These works are mainly divided into four categories, *i.e.*, audio-visual separation and localization, audio-visual recognition, audio-visual representation and audio-visual corresponding learning. In general, the essence of these task are exploring the consistency between audio and video. For instance, Shi et al. [25] propose a novel relation model for exploring multimodal multi-action relations in videos, by leveraging both relational graph convolution networks and video multi-modality. Xia et al. [30] propose a cross-modal time-level and event-level background suppression to better solve the problem of inconsistent audio and visual information within an audiovisual event localization task.

Different from above multimodal works, our method solves the problem that the information from different modalities inevitably influences each other via separately modeling modality-specific information and mixed-modality information, and adaptively selects the optimal fusion policy for each segment conditioned on the input via our Adaptive Fusion Module.

III. Approach

In this section, we introduce the proposed Progressive Adaptive Multimodal Fusion Network (PAMFN). The overall framework is shown in Figure 1. First, we describe the preliminaries of our work in Section III-A. Then, we introduce the overview of PAMFN in Section III-B. Next, we detail the three main components of PAMFN, *i.e.*, Modality-specific Feature Decoder module, Adaptive Fusion Module and Cross-modal Feature Decoder module, in Sections III-C, III-D and III-E, respectively.

A. Preliminaries

1) Problem Formulation: Following previous AQA works [1], [2], [3], [4], [5], we formulate AQA as a regression problem, where the model observes a video containing a specific action and predicts a non-negative number as the action quality score. Similar to [1], [3], we normalize the labels to the range [0, 1] to ensure stable training:

$$y^i = \frac{\bar{y}^i}{C}, \tag{1}$$

where $\bar{y}^i$ is the ground truth and y^i is the normalized label for training. C is related to the maximum score of the dataset.

2) Feature Extraction: Similar to the common practice in AQA [1], [4], [5], we divide the input video into the non-overlapping video segments. We use the pretrained backbones to extract the features of each video segment, optical flow segment and audio segment. Then, the features of different modalities are fed into different embedding layers and are projected into the same dimension d. Formally, given an input

video with T segments, we denote the extracted video feature sequence, optical flow feature sequence and audio feature sequence as $\{f_t^v\}_{t=1}^T$, $\{f_t^f\}_{t=1}^T$ and $\{f_t^a\}_{t=1}^T$, respectively. For more details about the extraction process, please refer to Section IV-C.

B. Overview of Our Method

To leverage multimodal information for action quality assessment, we propose a novel multimodal architecture. In this architecture, the information of different modalities is independently learned and the mixed-modality information is progressively learned from modality-specific information. As shown in Figure 1, our PAMFN consists of four branches, *i.e.*, a RGB branch, an audio branch, an optical flow branch and a mixed-modality branch. These branches are for learning modality-specific assessment features and learning mixed-modality assessment features below.

1) Learning Modality-Specific Assessment Features: To explore modality-specific information, the modality-specific branches (i.e. the first three branches) except the audio branch are pretrained and fixed since the quality of an action is almost impossible to assess using only audio information (see Section IV-E for more details). All modality-specific branches have the same structure and consist of N convolution stages and a regression layer. Each stage contains a convolution block and a pooling layer. Each convolution block has the same structure as the residual block in ResNet [45] except that 1D convolutions are used instead of 2D convolutions. The final regression layer includes a fully-connected layer with a dropout layer and a sigmoid activation function.

2) Learning Mixed-Modality Assessment Features: After obtaining the modality-specific information, a mixed-modality branch (i.e. the last branch) is proposed to progressively aggregate the modality-specific information. The mixed-modality branch has same structure with modality-specific branches and the initial mixed-modality features are initialized by zeros. To transfer information from modality-specific branches to the mixed-modality branch, three novel modules are proposed. First, a Modality-specific Feature Decoder module is designed to selectively transfer modality-specific information to the mixed-modality branch. Second, an Adaptive Fusion Module is adopted to explore an adaptive multimodal fusion policy for different segments. Third, a Cross-modal Feature Decoder module is used to transfer cross-modal features, generated by Adaptive Fusion Module, to the mixed-modality branch.

Note that our PAMFN is a pyramid architecture with N stages, and the mixed-modality branch progressively extracts modality-specific information during each stage.

C. Decoding Modality-Specific Features

Since the modality-specific features are aggregated in the mixed features during every stage, the mixed features $f_{i,t}^m$ and the modality-specific features, *i.e.*, $f_{i,t}^v$, $f_{i,t}^f$ and $f_{i,t}^a$, contain certain the overlapping information in the i^{th} stage. Therefore, a Modality-specific Feature Decoder module is adopted to extract the unseen or neglected information in the mixed-modality branch. Inspired by [46], the Modality-specific Feature Decoder module is implemented by a cross-attention. The *query* is the mixed features $f_{i,t}^m$, and the *keys* and *values* are the concatenation of the modality-specific features:

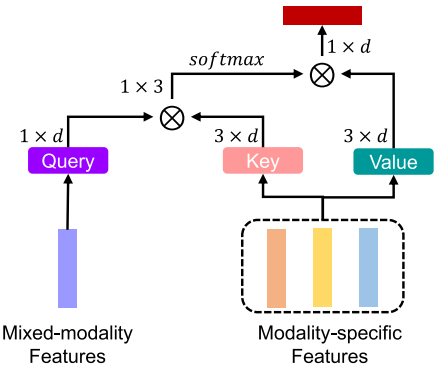


Fig. 2. Illustration of the Modality-specific Feature Decoder module. ⊗ denotes the matrix multiplication. The shapes of important tensors are shown in the figure. itQuery, *Key* and *Value* are three different linear projections.

$$Q_{i,t} = \mathbf{W}^q f_{i,t}^m, \quad K_{i,t} = \mathbf{W}^k f_{i,t}^{ms}, \quad V_{i,t} = \mathbf{W}^v f_{i,t}^{ms}, \tag{2}$$

$$f_{i,t}^{ms} = concatenate(f_{i,t}^v, f_{i,t}^f, f_{i,t}^a), \tag{3}$$

where $\{f_{i,t}^v\}_{t=1}^{T_i}$, $\{f_{i,t}^f\}_{t=1}^{T_i}$, $\{f_{i,t}^a\}_{t=1}^{T_i}$ and $\{f_{i,t}^m\}_{t=1}^{T_i}$ have the same dimension $\mathbb{R}^{T_i \times d}$, and T_i represents the length of feature sequences in the i^{th} stage. Then, the cross-attention is formulated as:

$$\bar{f}_{i,t}^{ms} = softmax(-\frac{Q_{i,t}K_{i,t}^T}{\sqrt{d}})V_{i,t}, \tag{4}$$

where $\sqrt{d}$ is a scaling factor and $\bar{f}_{i,t}^{ms}$ represents the final modality-specific features that are aggregated in the mixed-modality branch. Taking the negative of $Q_{i,t}K_{i,t}^T$ allows us to extract the least similar information between mixed-modality features and modality-specific features. Figure 2 shows the details of our Modality-specific Feature Decoder.

D. Learning an Adaptively Fusion Policy

Since interactions among different modalities are not explicitly modeled in the Modality-specific Feature Decoder module and mixed-modality branch, and using an invariant multimodal fusion policy during different parts of action may lead to sub-optimal results, we propose a novel Adaptive Fusion Module to explore an adaptive multimodal fusion policy. The Figure 3 illustrates an overview of the Adaptive Fusion Module. Our Adaptive Fusion Module consists of K fusion networks called FusionNets and a policy network called PolicyNet, the former for exploring different fusion strategies and the latter for deciding which fusion strategies will be enabled.

1) FusionNet: The three modality-specific features are first transformed via three different nonlinear projections. Each FusionNet takes the transformed modality-specific features as inputs and the outputs of the FusionNet, called cross-modal features, are then fed into a convolution block for refinement. The FusionNet is implemented by an attention network that consists of two fully-connected layers followed by a softmax function and the convolution block has the same structure as

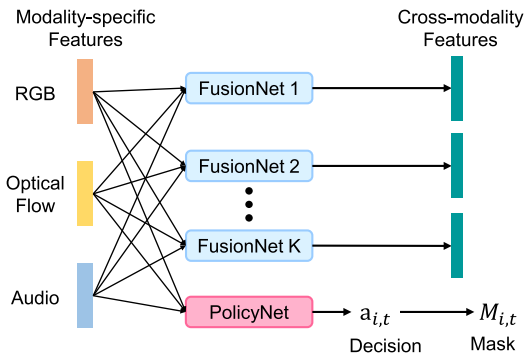


Fig. 3. Illustration of the Adaptive Fusion Module. K different FusionNets do not share parameters and explore different fusion policies. The PolicyNet generates an decision $a_{i,t}$ that determines which fusion strategies are enabled. $M_{i,t}$ is a binary mask vector that masks the not enabled cross-modal features.

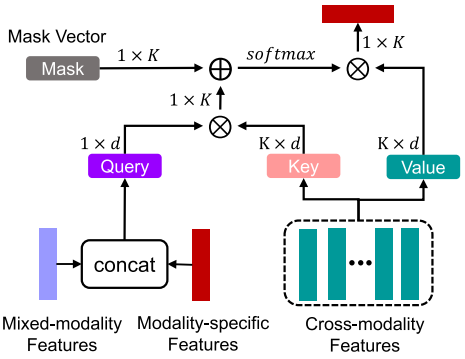


Fig. 4. Illustration of the Cross-modal Feature Decoder module. $\otimes$ denotes the matrix multiplication. $\oplus$ denotes the element-wise sum. The shapes of important tensors are shown in the figure. itQuery, *Key* and *Value* represent three different linear projections.

that in the modality-specific branch. Each attention network assigns weights for different modalities.

Here, we denote the weights generated by the k^{th} attention network as $\alpha^{v}_{i,t,k}$, $\alpha^{f}_{i,t,k}$ and $\alpha^{a}_{i,t,k}$, and cross-modal features as $f^{cs}_{i,t,k}$. Then, the FusionNet is formulated as:

$$\bar{f}^{cs}_{i,t,k} = \alpha^{v}_{i,t,k} f^{v}_{i-1,t} + \alpha^{f}_{i,t,k} f^{f}_{i-1,t} + \alpha^{a}_{i,t,k} f^{a}_{i-1,t}, \tag{5}$$

$$f^{cs}_{i,t,k} = pool(\Psi_f(\bar{f}^{cs}_{i,t,k})), \tag{6}$$

where $f^{v}_{i-1,t}$, $f^{f}_{i-1,t}$ and $f^{a}_{i-1,t}$ are modality-specific features in the $i-1^{th}$ stage, $\Psi_f(\cdot)$ is the convolution block and the pool function is used to ensure that the time dimension is the same as that of the other features in the i^{th} stage. Note that K FusionNets do not share weights, and we use the modality-specific features in the $i-1^{th}$ stage instead of those in the i^{th} stage to avoid the influence of the convolution block in the i^{th} stage. Therefore, we obtain K cross-modal features $\{f^{cs}_{i,t,k}\}^{K}_{k=1}$, with each focusing on a different fusion strategy.

2) Ranked FusionNets: By taking the similarity and diversity among video segments into consideration, we assume that some FusionNets are more general and some are more specific. Then, we proposed a rank mechanism of FusionNets, where the rank of a FusionNet corresponds to its generalizability and the FusionNet with a higher rank indicates it is more general. Specifically, we define that the subscript k ranging from 1 to K indicates the generality of FusionNet from most to least general. Then, the decision $a_{i,t} = c$ (the output of PolicyNet) represents only the first c FusionNets will be enabled. For example, suppose the decision $a_{i,t} = 3$; the top three FusionNets (ranked 1, 2, and 3, respectively) will be enabled, and the FusionNets with lower rank will be disabled. As you suggested, we have added the details to clarify the rank mechanism and improved the representation. Thus, the FusionNet with a higher rank will be used more frequently and focus on learning the more general fusion strategy. Other definitions of the decision $a_{i,t}$, such as setting $a_{i,t}$ as a one-hot vector (each element indicates the status of corresponding cross-modal features), are discussed in Section IV-F.3.

3) PolicyNet: The PolicyNet, implemented by two fully-connected layers, takes the modality-specific features as inputs and generates the logits $o_{i,t} \in \mathbb{R}^K$. As mentioned in Ranked FusionNets, we suppose that the logits $o_{i,t}$ represents the decision $a_{i,t} = c, 1 \leq c \leq K$, and the generated decision $a_{i,t} = c$ represents the first c cross-modal features $\{f^{cs}_{i,t,k}\}^{c}_{k=1}$ are enabled. To generate a discrete decision $a_{i,t}$, we first implement the PolicyNet to obtain the action probabilities $P_{i,t} \in \mathbb{R}^K$:

$$P_{i,t} = softmax(\Psi_p(f^{ms}_{i-1,t})), \tag{7}$$

$$f^{ms}_{i-1,t} = concatenate(f^{v}_{i-1,t}, f^{f}_{i-1,t}, f^{a}_{i-1,t}), \tag{8}$$

where $\Psi_p(\cdot)$ is PolicyNet. Then, the $arg\,max$ is used to obtain discrete decision $a_{i,t}$ from the decision probabilities $P_{i,t}$. Since the $arg\,max$ is not differentiable, we use the Straight-Through Gumbel Estimator [43] to solve this problem. Specifically, we first use the Gumbel-Max trick [47], [48] in the forward process to sample a decision from the decision probabilities $P_{i,t}$:

$$a_{i,t} = arg\,\max_{k}(log\, P_{i,t,k} + G_{i,t,k}), \tag{9}$$

where $\{G_{i,t,k}\}^{K}_{1}$ are i.i.d samples drawn from the Gumbel(0, 1) distribution $G_{i,t,k} = -log(-log\ U_{i,t,k})$ and $\{U_{i,t,k}\}^{K}_{1}$ are i.i.d samples drawn from the uniform distribution $U_{i,t,k} \sim Uniform(0, 1)$. Then, Gumbel-Softmax is used as a continuous, differentiable approximation to differentiate $arg\,max$ in the backward process:

$$\bar{a}_{i,t,k} = \frac{exp((log\ P_{i,t,k} + G_{i,t,k})/\tau)}{\sum_{k=1}^{K} exp((log\ P_{i,t,k} + G_{i,t,k})/\tau)}, \tag{10}$$

where τ is the softmax temperature, a non-negative number, and $\bar{a}_{i,t}$ is a continuous approximation of the one-hot encoding representation of $a_{i,t}$. τ controls the concentration level of the distribution. Inspired by CLIP [49] and [50], we set the softmax temperature τ as a learnable parameter. For more details of Straight-Through Gumbel Estimator, please refer to [43].

E. Decoding Cross-Modal Features

Similar to the Modality-specific Feature Decoder module, the Cross-modal Feature Decoder module is used to extract unseen or neglected information in the mixed-modality branch. As shown in Figure 4, the structure of Cross-modal Feature Decoder module is similar to that of the Modality-specific Feature Decoder module, except for the $query$ and an additional

mask vector $M_{i,t} = \{m_{i,t,k}\}_{k=1}^{K}$. The $queries$ are transformed as the concatenation of mixed-modality features and modality-specific features via linear projections. The additional mask vector $M_{i,t}$ masks the disabled cross-modal features and is determined by the decision $a_{i,t}$:

$$\bar{f}_{i,t}^{cs} = softmax(-\frac{\hat{Q}_{i,t}\hat{K}_{i,t}^{T}}{\sqrt{d}} + M_{i,t})\hat{V}_{i,t}, \tag{11}$$

where $\bar{f}_{i,t}^{cs}$ represents the refined cross-modal features. Taking the negative of $\hat{Q}_{i,t}\hat{K}_{i,t}^{T}$ allows us to extract the least similar information from cross-modal features. Since the mask vector $M_{i,t}$ is added before the softmax function, $m_{i,t,k}$ is set as a negative infinite number ξ when the i^{th} cross-features are disabled, or 0 when the i^{th} cross-features are enabled.

Since we assume that the decision $a_{i,t} = c, 1 \leq c \leq K$, indicates that the first c cross-modal features $\{f_{i,t,k}^{cs}\}_{k=1}^{c}$ are enabled, the mask vector $M_{i,t}$ is generated via a trick in PyTorch [51]:

$$M_{i,t} = \bar{M}_{i,t} + \bar{a}_{i,t} - stopgrad(\bar{a}_{i,t}), \tag{12}$$

where $\bar{M}_{i,t}$ is a preprocessed mask vector corresponding to the decision $a_{i,t}$. $stopgrad(\cdot)$ is a stop-gradient operation. Therefore, we obtain the mask vector $M_{i,t}$ through a differentiable approach.

F. Model Training

Since our method focuses on separately modeling modality-specific information and mixed-modality information, we train our network in two stages. In first stage, we train separately each modality-specific branch via a regression layer and the mean-squared error (MSE) loss. In second stage, we fix three modality-specific branches except the audio branch since the quality of an action is almost impossible to assess using only audio information (see Section IV-E for more details), and train the mixed-modality branch, MSFD, AFM and CMFD via the mean-squared error loss.

G. Discussion

As mentioned in Section I, different from existing multimodal works [22], [23], [24], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42] which models modality-specific and multimodal information simultaneously, our method separately models modality-specific information and mixed-modality information to extract pure modality-specific information via three modality-specific branches and a mixed-modality branch. Besides, these multimodal works use the same policy to fuse multimodal information for all parts of a video, ignoring diversity among different parts of a long video. By contrast, our method adaptively selects the optimal fusion policy for each segment conditioned on the input via a novel Adaptive Fusion Module. On the other hand, to avoid extracting redundant information from modality-specific branches into the mixed-modality branch in different stages (layers), both MSFD and CMFD aim to extract the unseen or neglected information in the mixed-modality branch, which has not been considered in existing modules.

Additionally, MINI-Net [42] also uses several networks named fusion submodules to learn multimodal information, but our FusionNets are constrained by the rank mechanism and PolicyNet to focus on different relations. Besides, with the help of the rank mechanism and PolicyNet, our Adaptive Fusion Module can adaptively decide which FusionNets to enable or disable.

IV. Experiments

In this section, we first introduce the datasets, evaluation metric and implementation details of our PAMFN. Then, we report the results with state-of-the-art AQA methods and multimodal methods implemented in AQA. Finally, we conduct extensive ablation study experiments and visualize qualitative results to demonstrate the effectiveness of our model.

A. Datasets

We conduct experiments on two long video AQA datasets, *i.e.*, the Fis-V dataset [5] and the Rhythmic Gymnastics dataset [1], since only long video datasets contain audio data among the existing AQA datasets.

1) Rhythmic Gymnastics (RG) Dataset: The RG dataset contains 1000 videos of four rhythmic gymnastics actions with different apparatuses, *i.e.*, ball, clubs, hoop and ribbon. Each video is about 1 minute and 35 seconds with 25 fps and only the duration from the moment of the beginning pose to the moment of the ending pose is preserved in each video. Each video is annotated with three scores, *i.e.*, a difficulty sore, an execution score and a total score, given by the referee on the spot. Following the evaluation protocol suggested in [1], we use 200 videos of each gymnastics routine type for training and 50 videos for testing.

2) Fis-V Dataset: The Fis-V dataset [5] contains 500 figure skating videos of the high standard international figure skating competition videos. Each video is about 2 minutes and 50 seconds with about 25 fps and shows the whole performance of only one skater. The irrelevant parts *i.e.*, warming up, bowing to the audience and so on, are pruned. Each video is labeled with two labels, namely, Total Element Score (TES) and Total program Component Score (PCS), which are provided by the referee on the spot. Following training-testing split in [5], 400 videos are used for training and the remaining 100 videos are used for testing.

B. Evaluation Metric

1) Spearman's Rank Correlation (Sp. Corr): To compare with previous works [1], [4], [5], Spearman's rank correlation is used to evaluate our method. Spearman's rank correlation represents the strength of the relation between the ground-truth labels and the predicted labels and is computed as follows:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}, \tag{13}$$

where x and y represent the rankings of two series. Its value ranges from -1 to 1 and a higher value indicates better results. Following previous works [3], [4], [6], [7], [8], [44], Fisher's z-value is used to compute the average Sp. Corr across actions.

TABLE I

THE SPEARMAN'S RANK CORRELATIONS OF OUR MODEL COMPARED WITH THE RESULTS OF STATE-OF-THE-ART AQA METHODS ON RG AND FIS-V DATASETS. FISHER'S Z-VALUE IS USED TO COMPUTE THE AVERAGE SP. CORR ACROSS ACTIONS AND THE HIGHER VALUES ARE BETTER. THE BEST RESULTS ARE INDICATED IN BOLD

	#Params	#Inference Time	Features	Rhythmic Gymnastics					Fis-V		
				Ball	Clubs	Hoop	Ribbon	**Avg.**	TES	PCS	**Avg.**
C3D+SVR [4]	-	-	C3D [55]	0.357	0.551	0.495	0.516	0.483	0.400	0.590	0.501
MS-LSTM [5]	1.08M	342ms	VST [17]	0.621	0.661	0.670	0.695	0.663	0.660	0.809	0.744
ACTION-NET [1]	3.54M	2ms	VST [17]+ResNet [45]	0.684	0.737	0.733	0.754	0.728	0.694	0.809	0.744
GDLT [3]	1.84M	5ms	VST [17]	0.746	0.802	0.765	0.741	0.765	0.685	0.820	0.761
PAMFN(Ours)	18.06M	33ms	VST [17]+I3D [52]+AST [53]	**0.757**	**0.825**	**0.836**	**0.846**	**0.819**	**0.754**	**0.872**	**0.822**

TABLE II

THE SPEARMAN'S RANK CORRELATIONS OF OUR MODEL COMPARED WITH THE RESULTS OF STATE-OF-THE-ART MULTIMODAL METHODS ON RG AND FIS-V DATASETS. FISHER'S Z-VALUE IS USED TO COMPUTE THE AVERAGE SP. CORR ACROSS ACTIONS AND THE HIGHER VALUES ARE BETTER. THE BEST RESULTS ARE INDICATED IN BOLD

	#Params	#Inference Time	Features	Rhythmic Gymnastics					Fis-V		
				Ball	Clubs	Hoop	Ribbo	**Avg.**	TES	PCS	**Avg.**
Joint-VA [39]	1.97M	4ms	VST [17]+AST [53]	0.719	0.674	0.749	0.820	0.746	0.751	0.844	0.802
MSAF [40]	5.56M	10ms	VST [17]+I3D [52]+AST [53]	0.743	0.795	0.734	0.836	0.781	0.751	0.843	0.802
UMT [41]	3.78M	5ms	VST [17]+AST [53]	0.725	0.588	0.678	0.823	0.714	0.716	0.822	0.774
PAMFN(Ours)	18.06M	33ms	VST [17]+I3D [52]+AST [53]	**0.757**	**0.825**	**0.836**	**0.846**	**0.819**	**0.754**	**0.872**	**0.822**

C. Implementation Details

1) Feature Extraction: As described in Section III-A, we divide RGB frames and optical flows into the non-overlapping segments and each segment contains 32 consecutive frames. To align with the video segments in time, audio is also divided into the same number of segments. Then we use three pretrained models to extract features from RGB frames, optical flows and audio. Following [3], we use the Video Swin Transformer (VST) [17] pretrained on Kinetics-600 dataset [52] to extract features from RGB frames. For optical flow, we use I3D [52] pretrained on Kinetics-400 dataset [52] to extract features. For audio, we use the Audio Spectrogram Transformer (AST) [53] pretrained on the large-scale AudioSet dataset [54] to extract audio features from the audio spectrogram. We randomly select 70 consecutive segments on RG and 130 consecutive segments on Fis-V for mini-batch training. If the number of segments is insufficient, we use zero-padding to maintain the same number of segments. All segments are used during testing.

2) Experimental Settings: We implement our PAMFN in three stages ($N = 3$) and set the number of FusionNets K to 10/6 to RG/Fis-V datasets. The feature dimension d is set to 256. The one-head cross-attention is used in MSFD and CMFD. Since our PAMFN is based on pretrained modality-specific branches, our model is trained in two phases. For the modality-specific branches, SGD [56] with a momentum of 0.9 and a weight decay of 10^{-4} and a cosine decay learning rate schedule are used to optimize our network. The batch size is 32 and the learning rate is 0.01. We train the modality-specific branches for 250 epochs to get the pretrained modality-specific branches. Then, we train our final model with AdamW [57] and cosine learning rate decay after fixing modality-specific branches except for the audio branch. The batch size is 32 and the learning rate is 5e-4/8e-4 for RG/Fis-V. Additionally, the learning rate of the regression layer is 0.1 times of previous layers when training the mixed-modality branch. Following [1], [3], we train different epochs on different datasets for better convergence: 400/500/300/500/500/500 for RG(Ball) / RG(Clubs) / RG(Hoop) / RG(Ribbon) / Fis-V(TES) / Fis-V(PCS). Following CLIP [49], the temperature τ is a learnable parameter and initialized as 10. Our method is implemented in PyTorch [58] and trained on a single RTX 3090Ti GPU.

D. Comparison With State-of-the-Art Methods

To demonstrate the effectiveness of our method, we compare our model with existing AQA methods on RG and Fis-V datasets. Besides, considering that existing AQA methods use only visual information, we also re-implement several multimodal methods used from other tasks for comparison.

1) Comparison With AQA Methods: As shown in Table I, our method achieves the best average correlation score on both datasets[2]. Compared with the previous state-of-the-art method GDLT [3], we achieve significant improvement of 0.054 and 0.061 on the average for RG and Fis-V datasets. Compared with all previous AQA works using only RGB information, our method fully utilizes multimodal information and explores the consistency of the athlete's action and the rhythm of the music, which helps our method achieve more accurate action quality assessment.

2) Comparison With Multimodal Methods From Other Tasks: We reimplement three multimodal methods from other tasks, *i.e.*, Joint-VA [39], MSAF [40] and UMT [41]. As shown in Table II, our method outperforms these multimodal methods on both RG and Fis-V datasets. Due to the benefits of separately modeling modality-specific information and mixed-modality information and adaptive fusion policy, our method achieves more accurate action quality assessment. Note that Joint-VA and UMT are not designed for AQA and UMT has some task-specific modules or losses, so it is not surprising that MSAF outperforms Joint-VA and UMT.

[2]The results of MS-LSTM [5] and ACTION-NET [1] in Table I are from the paper of GDLT [3]. MS-LSTM runs slowly since it uses custom LSTM [59].

TABLE III

THE SPEARMAN'S RANK CORRELATIONS WHEN USING STRONG FEATURES EXTRACTED BY UNMT [60] AND MAST [61] ON RG AND FIS-V DATASETS. FISHER'S Z-VALUE IS USED TO COMPUTE THE AVERAGE SP. CORR ACROSS ACTIONS AND THE HIGHER VALUES ARE BETTER. THE BEST RESULTS ARE INDICATED IN BOLD

	Features	Rhythmic Gymnastics					Fis-V		
		Ball	Clubs	Hoop	Ribbon	**Avg.**	TES	PCS	**Avg.**
GDLT [3]	UNMT [60]	0.785	0.776	0.768	0.776	0.776	0.710	0.823	0.773
Joint-VA [39]	UNMT [60]+MAST [61]	0.735	0.648	0.811	0.826	0.763	0.768	0.849	0.812
MSAF [40]	UNMT [60]+I3D [52]+MAST [61]	0.780	0.740	0.802	0.842	0.794	0.779	0.855	0.821
UMT [41]	UNMT [60]+MAST [61]	0.736	0.651	0.801	0.7998	0.753	0.732	0.810	0.774
PAMFN(Ours)	UNMT [60]+I3D [52]+MAST [61]	**0.822**	**0.813**	**0.853**	**0.848**	**0.835**	**0.791**	**0.865**	**0.832**

TABLE IV

THE SPEARMAN'S RANK CORRELATIONS OF OUR MODEL COMPARED WITH THE RESULTS OF THE UNIMODALITY METHODS AND MULTIMODAL METHODS USING WEIGHTED FUSION ON RG AND FIS-V DATASETS

	Modalities			Rhythmic Gymnastics					Fis-V		
	RGB	Flow	Audio	Ball	Clubs	Hoop	Ribbon	**Avg.**	TES	PCS	**Avg.**
Unimodality Methods	✓			0.636	0.720	0.769	0.708	0.711	0.665	0.823	0.755
		✓		0.536	0.674	0.684	0.716	0.657	0.606	0.772	0.698
			✓	0.286	0.297	0.423	0.254	0.317	0.517	0.628	0.575
Multimodal Methods using Weighted Fusion [62]	✓	✓		0.679	0.736	0.754	0.735	0.727	0.643	0.838	0.757
	✓		✓	0.601	0.623	0.746	0.749	0.686	0.735	0.825	0.784
		✓	✓	0.514	0.493	0.561	0.629	0.552	0.637	0.782	0.717
	✓	✓	✓	0.613	0.647	0.763	0.763	0.703	0.733	0.848	0.798
Ours	✓	✓		0.750	0.788	0.791	0.824	0.790	0.742	0.867	0.814
	✓		✓	0.749	0.757	**0.856**	0.769	0.787	0.736	0.863	0.809
		✓	✓	0.693	0.665	0.795	0.795	0.743	0.689	0.813	0.758
	✓	✓	✓	**0.757**	**0.825**	0.836	**0.846**	**0.819**	**0.754**	**0.872**	**0.822**

According to the results in Table I and Table II, we find that although these multimodal methods are not designed for AQA, these methods outperform all existing AQA methods with the help of audio information, demonstrating the benefits of audio information mentioned in the introduction. Besides, although our method takes more time in inference than existing AQA methods, it is still tolerable since it costs only 33ms which is fast enough, and the primary time cost is for the feature extraction but not the assessing network.

Additionally, since the used feature extractors may seem slightly suboptimal, we conduct experiments with stronger feature extractors to demonstrate the effectiveness of our method. Specifically, UNMT [60] pretrained on Kinectics-600 dataset and MAST [61] ptretrained on AudioSet are used as RGB and audio feature extractors. As shown in Table III, with the help of stronger features extracted by UNML and MAST, the performance of all methods boosts and our method achieves 0.835 and 0.832 on the average for RG and Fis-V datasets.

E. Comparison With Baselines

Since all existing AQA methods use only the RGB information, we compare our method with following multimodal baselines:

- Unimodality methods. We compare our method with modality-specific baselines in which we train three different models using RGB, optical flow and audio, respectively. The structure of each modality-specific model is the same to the modality-specific branch of our methods.
- Multimodal methods using Weighted Fusion. We compare our method with four joint learning multimodal baselines in which each model uses different combinations of the three modalities. A simple method, called Weighted Fusion [62], is used to fusion different modalities via late fusion with learnable weights in each multimodal baseline. And a softmax function is used to guarantee that the weight of each modality is positive. Note that these joint learning multimodal baselines are trained in one stage.
- We compare our method with different combinations of three modalities by removing one modality-specific branch from our model. Note that these baselines are trained in two stages, similar to our method.

Table IV shows the results of the above baseline comparisons on RG and Fis-V. Although the modality-specific branch of PAMFN is a simple network with three convolution blocks, the unimodality method using only RGB achieves a competitive performance (0.755) when compared with GDLT [3] (0.761) on Fis-V. Since AQA focuses on human actions, the unimodality method using only audio naturally obtains poor results, which is why we finetune the audio branch instead of fixing it. Additionally, it's surprise that the unimodality method using only audio achieves a performance of 0.575, which reveals that the action quality on skating has a strong correlation with the background music. Most multimodal methods using Weighted Fusion achieve worse results than unimodality methods on RG dataset but obtain conflicting results on Fis-V, which is most likely because the action quality on skating has a strong correlation with the background music. As shown in Table IV, our method clearly outperforms all multimodal methods using Weighted Fusion. In summary, multimodal information can improve the performance and our model effectively uses the multimodal information. Note that the mixed-modality branch is used in all multimodal

TABLE V
EVALUATION OF FUSION STRATEGIES

	Fusion	Rhythmic Gymnastics					Fis-V		
		Ball	Clubs	Hoop	Ribbon	**Avg.**	TES	PCS	**Avg.**
One-stage Training	AVG	0.596	0.644	0.746	0.778	0.698	0.715	0.840	0.785
	CAT	0.512	0.585	0.674	0.694	0.621	0.691	0.806	0.754
	Weighted [62]	0.613	0.647	0.763	0.763	0.703	0.733	0.848	0.798
	Attention	0.566	0.639	0.718	0.830	0.703	0.649	0.814	0.743
	Ours	0.741	0.728	0.819	**0.855**	0.792	0.735	0.853	0.802
Two-stage Training	AVG	0.637	0.747	0.781	0.771	0.739	0.720	0.857	0.799
	CAT	0.665	0.762	0.805	0.817	0.768	0.696	0.850	0.785
	Weighted [62]	0.638	0.752	0.781	0.776	0.742	0.721	0.860	0.801
	Attention	0.641	0.744	0.778	0.773	0.738	0.706	0.847	0.787
	Ours	**0.757**	**0.825**	**0.836**	0.846	**0.819**	**0.754**	**0.872**	**0.822**

TABLE VI
EVALUATION OF CROSS-MODAL FEATURE DECODER (CMFD) MODULE AND MODALITY-SPECIFIC FEATURE DECODER (MSFD) MODULE. → DENOTES REPLACING A MODULE WITH WEIGHTED FUSION

Fusion	Rhythmic Gymnastics					Fis-V		
	Ball	Clubs	Hoop	Ribbon	**Avg.**	TES	PCS	**Avg.**
w/o CMFD	0.672	0.789	0.769	0.775	0.755	0.733	0.858	0.804
w/o MSFD	0.700	0.814	0.833	0.839	0.802	0.734	0.871	0.813
Weighted Fusion [62] → CMFD	0.670	0.751	0.816	0.796	0.764	0.724	0.862	0.804
Weighted Fusion [62] → MSFD	**0.773**	0.782	0.810	0.836	0.802	0.726	0.862	0.804
Weighted Fusion [62] → MSDF and CMFD	0.629	0.804	0.819	0.816	0.777	0.712	0.858	0.796
Our Full Model	0.757	**0.825**	**0.836**	**0.846**	**0.819**	**0.754**	**0.872**	**0.822**

methods, and experiments without the mixed-modality branch are described in Section IV-F.1.

F. Ablation Study

1) Evaluation of Fusion Strategies: To demonstrate the effectiveness of our fusion strategy, we first compare with four common late fusion strategies:

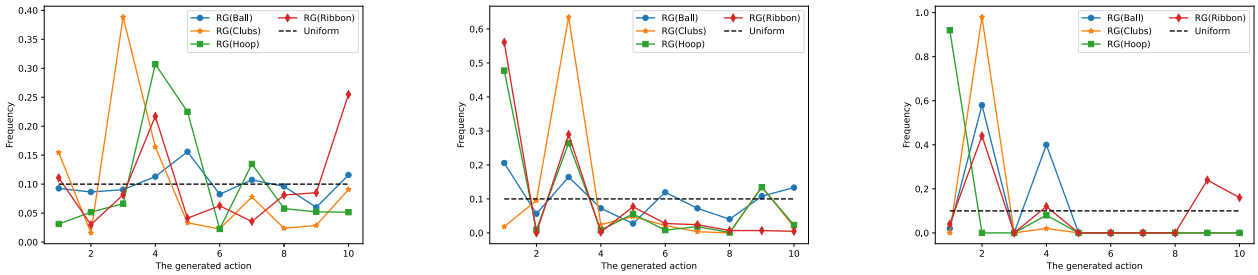
- AVG: We simply calculate the average of three modality features.
- CAT: We concatenate three modality features and use an additional fully connected (FC) layer to reduce the dimension.
- Weighted [62]: Learnable weights and a softmax function are used to fuse three modality features as described above.
- Attention: An FC layer takes the three modality features as inputs and generates weights for each modality. Then the generated weights are fed into a softmax function and the fused features is obtained by computing a weighted sum.

Since our method separately models modality-specific information and cross-modal information, we train our mixed-modality branch based on pretrained modality-specific branches. To further validate the superiority of our method and the effectiveness of two-stage training strategy, we train the above baselines with two-stage training and one-stage training. For one-stage training, fusion strategies are used to combine the outputs of the last convolution block, and two FC layers used as a regressor take the fused features as inputs.

The results are shown in Table V. Our method with two-stage training achieves the best average performance on RG and Fis-V, and outperforms other fusion strategies with a large margin on average. We observe that the models trained in two stages achieve better performance than the models trained in one stage. Remarkably, the models trained in one stage achieve worse average performance than the modality-specific model only using RGB on RG dataset, and we get an opposite conclusion on Fis-V. The weighted fusion performs better than the compared with other fusion strategies. Additionally, the average performance of our method decreases by 0.027 and 0.02 on RG and Fis-V when trained in one stage instead of in two stages, most likely because our method is designed based on the pretrained modality-specific branches and the modality-specific branches have difficulty extracting high-quality modality-specific information when trained in one stage.

2) Evaluation of the CMFD and MSFD: To demonstrate the effectiveness of our proposed Cross-modal Feature Decoder module and Modality-specific Feature Decoder module, we try to remove Cross-modal Feature Decoder and Modality-specific Feature Decoder from the full model. Additionally, we try to replace Cross-modal Feature Decoder and Modality-specific Feature Decoder with Weighted Fusion to show the superiority of our proposed components. Here, we choose Weighted Fusion because Weighted Fusion achieves the best performance among four common fusion strategies.

The results are shown in Table VI and → denotes replacing a module with Weighted Fusion. Removing the Cross-modal Feature Decoder or Modality-specific Feature Decoder causes an average performance drop of 0.064/0.018 and 0.017/0.009 on RG/Fis-V respectively, which shows that modality-specific information is more important than cross-modal information for action quality assessment. Replacing the Cross-modal Feature Decoder or Modality-specific Feature Decoder with Weighted Fusion alleviates the performance drop when compared with removing one of decoders. However, completely replacing two decoders with Weighted Fusion also leads to a performance drop. These results demonstrate the effectiveness



(a) Stage 1 (b) Stage 2 (c) Stage 3

Fig. 5. The frequency of the generated decisions at three stages on RG dataset. The line labeled with Uniform denotes the frequency of decisions under a uniform distribution. Best viewed in color.

TABLE VII
EVALUATION OF DIFFERENT FUSIONNETS ON RG AND FIS-V DATASETS

	Rhythmic Gymnastics					Fis-V		
	Ball	Clubs	Hoop	Ribbon	**Avg.**	TES	PCS	**Avg.**
Unranked FusionNets	0.689	0.822	0.789	0.832	0.789	0.730	0.868	0.810
Free FusionNets	0.755	0.793	0.835	0.843	0.809	0.733	0.870	0.812
Ranked FusionNets (Ours)	**0.757**	**0.825**	**0.836**	**0.846**	**0.819**	**0.754**	**0.872**	**0.822**

of our Cross-modal Feature Decoder and Modality-specific Feature Decoder.

3) Evaluation of Ranked FusionNets: To demonstrate the effectiveness of our proposed ranked FusionNets, we compare our ranked FusioNets with unranked FuionNets. To implement unranked FusionNets, we define that the decision $a_{i,t} \in R^K$, generated by the PolicyNet, is a one-hot vector and each element denotes the corresponding FusionNet is enabled or disabled. We also try to enable all FusionNets called Free FusionNets and remove the PolicyNet since all FusionNets is enabled. Note that the numbers of FusionNets K remains the same in each model.

As shown in Table VII, replacing ranked FusionNets with unranked FusionNets or Free FusionNets causes an average performance drop of 0.03/0.012 and 0.01/0.009 on RG/Fis-V, respectively. Although unranked FusionNets has more freedom to select the optimal combination of FusionNets than ranked FusionNets, the model has difficulty converging to the optimal solution and achieves poor results. Similarly, most information, *i.e.*, all FusionNets, is used in Free FusionNets, but this method also obtains poor results, especially on RG. Different from unranked FusionNets and free FusionNets, since we assume that the FusionNet with a higher rank indicates that it is more general and the first c FusionNets are enabled when given an decision $a_{i,t} = c$, the model converges more easily than the other two models.

To verify that our ranked FusionNets does not converge to a trivial solution, *i.e.*, the generated decisions are almost the same and only a few FusionNets are used, we visualize the generated decisions on the RG test set. The frequencies of the generated decisions in three stages are shown in Figure 5 and the line labeled Uniform represents the frequency of decisions according to a uniform distribution. As shown in Figure 5, our ranked FusionNets does not converge to a trivial solution and our method really adopts different fusion policies for different parts of an action. We notice that although we adopt the same number of FusionNets across all actions for simplicity, our method can learn to disable redundant FusionNets, as shown in Figure 5-c. Remarkably, the frequency curves of different actions have similar trends, most likely because these actions are all rhythmic gymnastics actions, with only the apparatus differing.

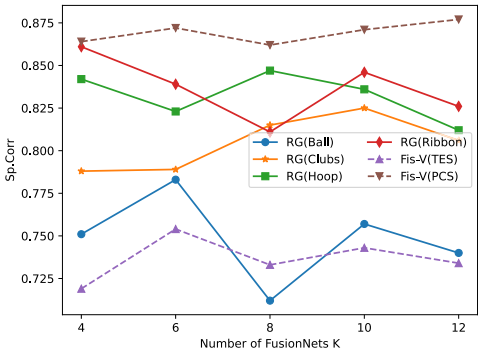


Fig. 6. Evaluation of the number of FusionNets.

4) Evaluation of the Progressive Fusion Strategy: Our method is based on a progressive fusion strategy which means that the mixed-modality branch extracts information from the modality-specific branches in different stages. To demonstrate the effectiveness of our progressive fusion strategy, we try to only fuse modality information at specific layers. Specifically, we evaluate our method with fusion in only the first stage or in the first two stages. The results are shown in Table VIII. We can observe that fusion during the first stage achieves the worst performance of 0.786 on RG and the performance improves as the number of fusion stages increases. In addition, fusion during the first stage or first two stages leads to a slight performance drop on Fis-V.

5) Evaluation of the Number of FusionNets K: We implement our method with different numbers of FusionNets K to evaluate the impact of the number of FusionNets on model performance. As shown in Figure 6, different action types has different optimal K values, and $K = 10/6$ achieves the best average performance on RG/Fis-V dataset. The performance

TABLE VIII
EVALUATION OF FUSING MODALITY INFORMATION AT DIFFERENT STAGES ON RG AND FIS-V DATASETS

Fusion Stage	Rhythmic Gymnastics					Fis-V		
	Ball	Clubs	Hoop	Ribbon	**Avg.**	TES	PCS	**Avg.**
First Stage	0.733	0.785	0.795	0.824	0.786	0.753	0.863	0.815
First Two Stages	0.730	0.786	**0.836**	**0.864**	0.810	**0.755**	**0.873**	**0.823**
All Stages (Ours)	**0.757**	**0.825**	**0.836**	0.846	**0.819**	0.754	0.872	0.822

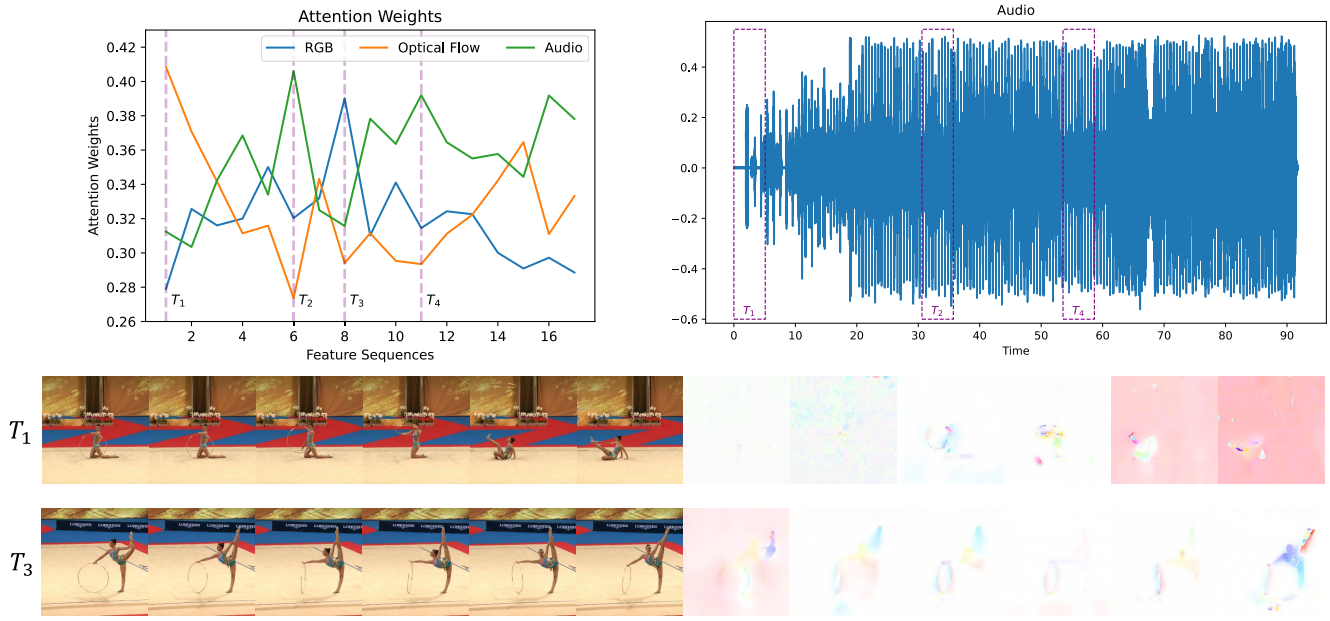


Fig. 7. Visualization of attention weights of the Modality-specific Feature Decoder in the second stage on RG (Hoop). The top left figure shows the attention weights of three modalities on feature sequences and four purple vertical dashed lines indicate the time windows corresponding to specific features. The top right figure shows the audio waveform of the video and three rectangles with dashed lines represent the time windows, *i.e.*, T_1, T_2 and T_4, in the audio waveform. The next two rows show the RGB frames and optical flows corresponding to the time windows, *i.e.*, T_1 and T_3. Note that each feature in the second stage corresponds to a video segment about 2 seconds and all frames are cropped to keep the gymnast in the center of frames for visualization. This figure is best viewed in color.

on RG/Fis-V dataset decreases slightly when the number of FusionNets is larger than 10. In addition, we observe clear changes in the performance as the number of FusionNets ranges from 4 to 12 on RG and the number of FusionNets is more important for RG than Fis-V.

G. Extension to Highlight Detection

Although our method is designed for action quality assessment, the architecture of our method is general and can be easily used for other tasks. To demonstrate the generalizability of our method, we evaluate our method on the highlight detection task, which is a task of detecting interesting moments called "highlights", within videos. To apply our method, we remove all pooling layers and replace the regression layer with a classification head. We conduct experiments on YouTube Highlights dataset [65] which is a commonly used dataset for highlight detection.

1) Dataset: YouTube Highlights dataset [65] contains six topic categories, *i.e.*, dog, gymnastics, parkour, skating, skiing, and surfing, and each topic contains about 100 videos. The annotations are segment-level and indicate whether a segment is a highlight segment. We follow the training-test split [41] to evaluate our model. Moreover, we train a highlight detector for each topic, following existing works.

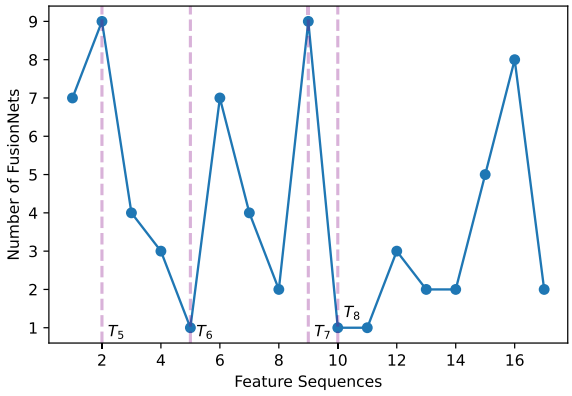


Fig. 8. Visualization of decisions generated by PolicyNet in second stage on RG(Hoop). Three purple vertical dashed lines indicate the time windows corresponding to specific features and the corresponding RGB frames, and optical flows are shown in Figure 9.

2) Implementation Details: We use RGB and audio features released by UMT [41]. For the optical flows, we use I3D [52] pretrained on Kinetics-400 dataset [52] to extract features. The number of FusionNets K is set to 10. We use AdamW [57] to optimize our network and train the modality-specific branches and mixed-modality branch for 300 epochs. The batch size is 1 and the learning rate is 5e-4. Others are the same as those used in AQA task.

TABLE IX

THE MEAN AVERAGE PRECISION OF OUR MODEL COMPARED WITH THE RESULTS OF STATE-OF-THE-ART METHODS ON YOUTUBE HIGHLIGHTS DATASET. THE BEST RESULTS ARE INDICATED IN BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED

Method	Modality	Dog	Gymnastics	Parkour	Skating	Skiing	Surfing	**Average**
RRAE [63]	RGB	0.490	0.350	0.500	0.250	0.220	0.490	0.383
GIFs [64]	RGB	0.308	0.335	0.540	0.554	0.328	0.541	0.464
LSVM [65]	RGB	0.600	0.410	0.610	0.620	0.360	0.610	0.536
CLA [66]	RGB	0.502	0.217	0.309	0.505	0.379	0.584	0.416
LM [66]	RGB	0.579	0.417	0.670	0.578	0.486	0.651	0.564
Mini-Net [42]	RGB+Audio	0.582	0.617	0.702	0.722	0.587	0.651	0.644
Trail. [67]	RGB	0.633	**0.825**	0.623	0.529	**0.745**	0.793	0.691
DL-VHD [39]	RGB	0.708	0.532	0.772	0.725	0.661	0.762	0.693
Joint-VA [39]	RGB+Audio	0.645	0.719	0.808	0.620	0.732	0.783	0.718
PLD [68]	RGB	**0.749**	0.702	0.779	0.575	0.707	0.790	0.730
CO-AV [69]	RGB+Audio	0.609	0.660	**0.890**	**0.741**	0.690	0.811	0.747
UMT [41]	RGB+Audio	0.659	0.752	0.816	0.718	0.723	**0.827**	**0.749**
PAMFN(Ours)	RGB	0.652	0.677	0.591	0.629	0.710	0.825	0.681
	Flow	0.626	0.657	0.662	0.420	0.672	0.684	0.620
	Audio	0.502	0.639	0.720	0.467	0.655	0.737	0.620
	RGB+Audio+Flow	0.720	0.690	0.822	0.722	0.720	0.810	0.747

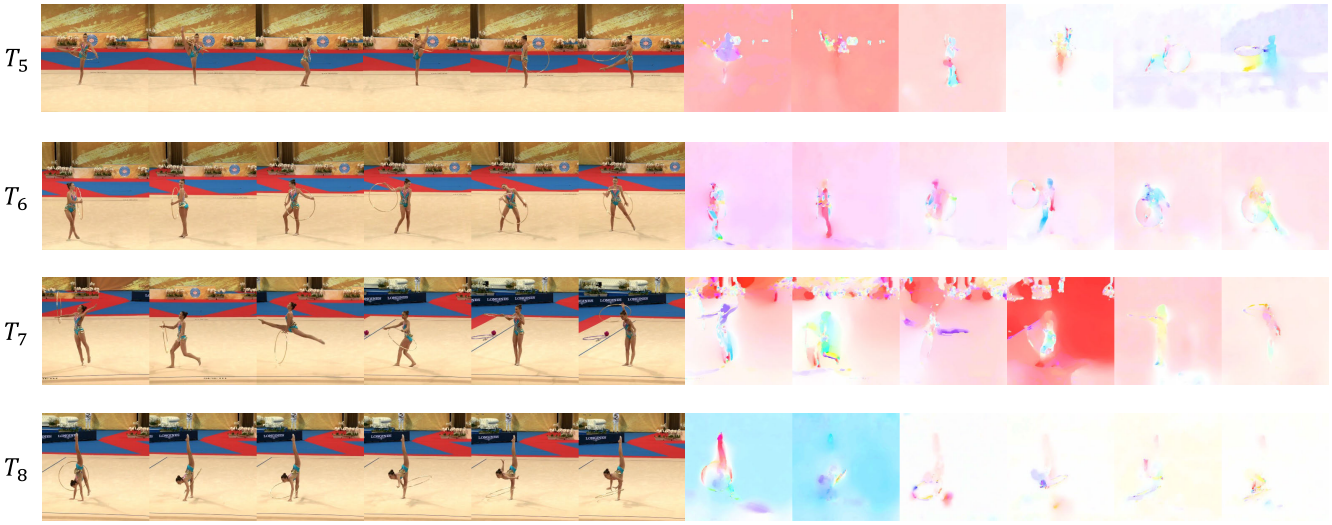


Fig. 9. Visualization of RGB frames and optical flows corresponding to the time windows marked in Figure 8, *i.e.*, T_5, T_6, T_7 and T_8. This figure is best viewed in color. Note that each feature in the second stage corresponds to a video segment about 2 seconds and all frames are cropped to keep the gymnast in the center of frames for visualization.

The experiment results are shown in Figure IX and the mean average precision (mAP) is used as the evaluation metric. Our method achieve comparable results to the state-of-the-art method UMT [41] (0.749 vs. 0.747). Moreover, our method achieves performance improvements of 0.066, 0.127 and 0.127 over the modality-specific branches. Remarkably, the modality-specific branch using only audio achieves not bad performance (0.620). This result suggests the audio contains rich information for inferring whether a segment is a highlight moment. These results demonstrate the generalizability and effectiveness of our method.

H. Qualitative Results

1) Visualization of Attention Weights: To show the importance of different modalities learned by our method, we visualize the attention weights of modality-specific feature decoder on a video feature sequence during the second stage, as shown in Figure 7. Additionally, we visualize the audio, video frames and optical flows corresponding to these specific features. As shown in Figure 7, since the gymnast always strikes a pose and begins to move only when the music starts at the beginning of the video, our model pays more attention to optical flows to assess actions at T_1 and we observe that our method usually focuses on optical flows at the beginning of a video on RG dataset. In addition, the audio is assigned to high attention weight at T_2 and T_4 because T_2 includes the chorus of the background music, which has a simple and quick rhythm, and T_4 includes a transition from a verse to the chorus. Additionally, when the gymnast rotates the hoop while maintaining body posture at T_3, our method focuses on the RGB information to observe the body posture of the gymnast. In summary, we observe that our method uses the background music to assist in action assessment and focuses on the modality that provides the most useful information for action assessment.

2) Visualization of Decisions: We visualize the decisions generated by the PolicyNet in the second stage on RG (Hoop) in Figure 8, and Figure 9 visualizes the RGB frames and

optical flows at the four time windows marked in Figure 8. The PolicyNet chooses to use all FusionNets at T_5 and T_7, and one FusionNet at T_6 and T_8. As shown in Figure 9, the gymnast turns her body around with leg raised at T_5 and performs a split leap at T_7, while the gymnast rotates the hoop while twisting or posing at T_6 and T_8. By comparing the RGB frames and optical flows at T_5 and T_7 with those at T_6 and T_8, we find that the PolicyNet tends to use more FusionNets when the gymnast is performing an action with large movements or changing in posture.

V. Conclusion

In this work, we propose a novel multimodal method for action quality assessment, called Progressive Adaptive Multimodal Fusion Network (PAMFN), that separately models modality-specific information and cross-modal information. Our PAMFN consists of three modality-specific branches for independently exploring modality-specific information and a mixed-modality for progressively aggregating the modality-specific information. Then, to build the bridge between modality-specific branches and the mixed-modality branch, we propose three novel modules. We first propose a Modality-specific Feature Decoder module to selectively transfer modality-specific information to the mixed-modality branch. Then, by taking the potential diversity during different parts of action into consideration, we design an Adaptive Fusion Module to learn adaptive multimodal fusion policies in different parts of an action. Third, we propose a Cross-modal Feature Decoder module to transfer cross-modal features generated by Adaptive Fusion Module to the mixed-modality branch. The state-of-the-art results on two public action quality assessment datasets demonstrate the effectiveness of the proposed model.

References

[1] L.-A. Zeng et al., "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2526–2534.

[2] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Proc. Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, Sep. 2014, pp. 556–571.

[3] A. Xu, L.-A. Zeng, and W.-S. Zheng, "Likert scoring with grade decoupling for long-term action assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3232–3241.

[4] P. Parmar and B. T. Morris, "Learning to score Olympic events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–28.

[5] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4578–4590, Dec. 2020.

[6] Y. Tang et al., "Uncertainty-aware score distribution learning for action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9839–9848.

[7] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6331–6340.

[8] J. Gao et al., "An asymmetric modeling for action assessment," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 222–238.

[9] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7919–7928.

[10] D. Liu et al., "Towards unified surgical skill assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9517–9526.

[11] T. Nagai, S. Takeda, M. Matsumura, S. Shimizu, and S. Yamamoto, "Action quality assessment with ignoring scene context," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1189–1193.

[12] M. Nekoui, F. O. Tito Cruz, and L. Cheng, "EAGLE-eye: Extreme-pose action grader using detail bird's-eye view," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 394–402.

[13] J.-H. Pan, J. Gao, and W.-S. Zheng, "Adaptive action assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8779–8795, Dec. 2022.

[14] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, "Semi-supervised action quality assessment with self-supervised segment feature recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6017–6028, Sep. 2022.

[15] Y. Bai et al., "Action quality assessment with temporal parsing transformer," 2022, *arXiv:2207.09270*.

[16] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2949–2958.

[17] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 3, pp. 443–455, Mar. 2018.

[18] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? Who's best? Pairwise deep ranking for skill determination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6057–6066.

[19] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7854–7863.

[20] Z. Li, Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4385–4395.

[21] P. Parmar, A. Gharat, and H. Rhodin, "Domain knowledge-informed self-supervised representations for workout form assessment," 2022, *arXiv:2202.14019*.

[22] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 631–648.

[23] F. Xiao, Y. Jae Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual SlowFast networks for video recognition," 2020, *arXiv:2001.08740*.

[24] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10454–10464.

[25] Z. Shi et al., "Multi-modal multi-action video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13678–13687.

[26] S. Alfasly, J. Lu, C. Xu, and Y. Zou, "Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20176–20185.

[27] D. Liu, T. Jiang, and Y. Wang, "Completeness modeling and context separation for weakly supervised temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1298–1307.

[28] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6291–6299.

[29] J.-T. Lee, M. Jain, H. Park, and S. Yun, "Cross-attentional audio-visual fusion for weakly-supervised action localization," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–17.

[30] Y. Xia and Z. Zhao, "Cross-modal background suppression for audio-visual event localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19957–19966.

[31] H. Jiang, C. Murdock, and V. K. Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10534–10542.

[32] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020.

[33] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 247–263.
[34] D. Hu, Y. Wei, R. Qian, W. Lin, R. Song, and J.-R. Wen, "Class-aware sounding objects localization via audiovisual correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9844–9859, Dec. 2022.
[35] P. Wang, J. Li, M. Ma, and X. Fan, "Distributed audio-visual parsing based on multimodal transformer and deep joint source channel coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 4623–4627.
[36] R. Tao, R. K. Das, and H. Li, "Audio-visual speaker recognition with a cross-modal discriminative network," 2020, *arXiv:2008.03894*.
[37] S. Ma, Z. Zeng, D. McDuff, and Y. Song, "Contrastive learning of global-local video representations," 2021, *arXiv:2104.05418*.
[38] J. F. Montesinos, V. S. Kadandale, and G. Haro, "VoViT: Low latency graph-based audio-visual voice separation transformer," 2022, *arXiv:2203.04099*.
[39] T. Badamdorj, M. Rochan, Y. Wang, and L. Cheng, "Joint visual and audio learning for video highlight detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8107–8117.
[40] L. Su, C. Hu, G. Li, and D. Cao, "MSAF: Multimodal split attention fusion," 2020, *arXiv:2012.07175*.
[41] Y. Liu, S. Li, Y. Wu, C. W. Chen, Y. Shan, and X. Qie, "UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 3032–3041, doi: 10.1109/cvpr52688.2022.00305.
[42] F.-T. Hong, X. Huang, W.-H. Li, and W.-S. Zheng, "Mini-net: Multiple instance ranking network for video highlight detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 345–360.
[43] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2016, *arXiv:1611.01144*.
[44] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "TSA-Net: Tube self-attention network for action quality assessment," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4902–4910.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.
[46] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[47] E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical*, vol. 33. Washington, DC, USA: US Government Printing Office, 1954.
[48] C. J. Maddison, D. Tarlow, and T. Minka, "A* sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3086–3094.
[49] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.
[50] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.
[51] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[52] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2017, pp. 6299–6308.
[53] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575.
[54] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 776–780.
[55] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
[56] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[58] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.
[59] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[60] K. Li et al., "Unmasked teacher: Towards training-efficient video foundation models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 19891–19903.
[61] W. Zhu and M. Omar, "Multiscale audio spectrogram transformer for efficient audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[62] R. Panda et al., "AdaMML: Adaptive multi-modal learning for efficient video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7556–7565.
[63] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo, "Unsupervised extraction of video highlights via robust recurrent auto-encoders," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4633–4641.
[64] M. Gygli, Y. Song, and L. Cao, "Video2GIF: Automatic generation of animated GIFs from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1001–1009.
[65] M. Sun, A. Farhadi, and S. Seitz, "Ranking domain-specific highlights by analyzing edited videos," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 787–802.
[66] B. Xiong, Y. Kalantidis, D. Ghadiyaram, and K. Grauman, "Less is more: Learning highlight detection from video duration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1258–1267.
[67] L. Wang, D. Liu, R. Puri, and D. N. Metaxas, "Learning trailer moments in full-length movies with co-contrastive attention," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 300–316.
[68] F. Wei, B. Wang, T. Ge, Y. Jiang, W. Li, and L. Duan, "Learning pixel-level distinctions for video highlight detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3063–3072.
[69] S. Li et al., "Probing visual-audio representation for video highlight detection via hard-pairs guided contrastive learning," in *Proc. 33rd Brit. Mach. Vis. Conf. (BMVC)*, London, U.K., Nov. 2022, p. 709. [Online]. Available: https://bmvc2022.mpi-inf.mpg.de/709/

Ling-An Zeng received the B.S. degree in software engineering from the University of Electronic Science and Technology of China in 2019 and the M.S. degree in computer technology from Sun Yat-sen University, Guangzhou, China, in 2021, where he is currently pursuing the Ph.D. degree in computer science and technology with the School of Artificial Intelligence. His research interests include computer vision and machine learning. He is currently focusing on the topic of action understanding.

Wei-Shi Zheng is now a full Professor with Sun Yat-sen University. His research interests include person/object association and activity understanding, and the related weakly supervised/unsupervised and continuous learning machine learning algorithms. He has now published more than 200 papers, including more than 150 publications in main journals (TPAMI, IJCV, TIP) and top conferences (ICCV, CVPR, SIGGRAPH, ECCV, NeurIPS). He has ever served as an area chairs of ICCV, CVPR, ECCV, BMVC, NeurIPS and etc. He is an Associate Editors/on the Editorial Board of IEEE-TPAMI, *Artificial Intelligence Journal*, Pattern Recognition. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He is a Cheung Kong Scholar Distinguished Professor, a recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and a recipient of the Royal Society-Newton Advanced Fellowship of the U.K.