IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 April 2024, accepted 26 June 2024, date of publication 4 July 2024, date of current version 12 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3423462

RESEARCH ARTICLE

MMW-AQA: Multimodal In-the-Wild Dataset for Action Quality Assessment

TAKASUKE NAGAI, SHOICHIRO TAKEDA, SATOSHI SUZUKI, AND HITOSHI SESHIMO
Nippon Telegraph and Telephone Corporation, Yokosuka 239-0847, Japan

Corresponding author: Takasuke Nagai (takasuke.nagai@ntt.com)

Prior to the test, the athletes were informed of the study's purpose, possibility of datasets being made publicly available and the right to withdraw at any time. All athletes agreed to participate, and the study was conducted in accordance with the Declaration of Helsinki.

ABSTRACT Action quality assessment (AQA) is a task for assessing a specific action quality in videos. Since existing AQA datasets provide only two-dimensional (2D) video data captured from fewer viewpoints, existing AQA methods based on deep neural networks (DNNs) often struggle to assess complex three-dimensional (3D) actions accurately, and their robustness against diversified viewpoints remains unknown. We created a dataset called multimodal in-the-wild (MMW)-AQA in freestyle windsurfing that addresses these concerns. In addition to video data, MMW-AQA provides inertial measurement unit (IMU) and global positioning system (GPS) data. The 3D information of IMU data helps DNNs accurately assess complex 3D actions. Moreover, MMW-AQA provides wild video data captured by a single unmanned aerial vehicle (UAV). These wild video data enable us to evaluate whether AQA methods can work well on diversified viewpoints. Furthermore, we also present the baseline multimodalization framework with a transformer-based fusion module. These frameworks multimodalize existing unimodal DNN models easily to assess action quality using multimodal data. Our experimental results demonstrate that multimodal data improves the AQA accuracy compared with unimodal video data.

INDEX TERMS Action quality assessment, deep learning, multimodal dataset, multimodal learning.

I. INTRODUCTION

Action quality assessment (AQA) is a computer vision task for assessing the action quality in videos. AQA has been applied to various domains, such as competitive sports [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], surgical training [18], [19], [20], [21], daily skills [22], [23], health care [24], [25], [26], and musical performances [27]. In competitive sports, AQA has attracted significant attention because it can replace human judges and provide feedback to athletes, helping them improve their action quality during practice. In fact, an AQA system was introduced at the 2019 Artistic Gymnastics World Championships in Germany for scoring [28].

Most AQA methods, including state-of-the-art ones, are based on deep neural networks (DNNs) to learn/predict action quality and its corresponding score in a video. These methods use existing AQA datasets that provide not only video data and action-quality scores but also various additional information; the MTL-AQA dataset [4] provides action classes and action-described texts, FineDiving dataset [9] provides temporal action segmentations in each video, and LOGO dataset [29] provides formation graphs of group competitive sports, such as artistic swimming.

While various AQA datasets have been proposed as described above, two major concerns remain regarding the AQA dataset. First, these datasets provide only unimodal video data. It is widely known that DNNs often struggle to extract three-dimensional (3D) features from two-dimensional (2D) videos [30], [31], [32], [33]. Thus, DNNs trained on unimodal video data cannot accurately assess complex 3D actions. Second, the video data in the existing AQA datasets are captured from fewer viewpoints. Thus, it is still unknown whether the existing AQA methods can work on *wild* videos, where actions are captured from diversified viewpoints, e.g., an unmanned aerial vehicle (UAV) camera [34], [35].

In this paper, we propose a new AQA dataset in freestyle windsurfing, called multimodal in-the-wild (MMW)-AQA, that addresses the aforementioned two

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Heng Foh.

concerns. MMW-AQA is novel in two aspects: (I. **Multi-modality**) it newly provides multimodal data to help DNNs learn complex 3D actions, and (II. **In-the-wild**) it consists of wild video data with diversified viewpoints to evaluate the robustness of AQA methods against the wild setting. In more detail, (I) MMW-AQA contains not only video data but also inertial measurement unit (IMU) and global positioning system (GPS) data, as illustrated in Figure 1. The IMU data can provide 3D rotation angles during actions, which helps DNNs accurately assess complex 3D freestyle actions. Also, the GPS data provides how far an athlete is from a shore; in freestyle windsurfing, it is difficult to perform actions if the athlete is close to the shore because the wind weakens, and this data thus assists DNNs in understanding the difficulty of actions. (II) The video data in MMW-AQA are captured from diversified viewpoints by a single UAV flying around an athlete at sea, as shown in Figure 2. These wild video data enable us to evaluate whether existing AQA methods can work well on diversified viewpoints. Furthermore, we evaluate the effectiveness of multimodal data for this wild setting. Note that freestyle windsurfing naturally involves two aspects (I, II); acrobatic 3D actions are performed in various locations on the sea. In summary, MMW-AQA addresses the aforementioned two concerns because it has 3D information of the IMU data and the wild video setting.

This paper also presents the baseline multimodalization framework with a transformer-based fusion module to multimodalize existing unimodal AQA models [7], [8], [11]. Using this framework, we easily obtain multimodal AQA models. The transformer-based fusion module is inspired by the transformer architecture [36]; it is plausible for handling multimodal datasets because its multi-head self-attention mechanism can efficiently learn relationships across multiple modalities [37], [38], [39], [40], [41], [42]. Therefore, we introduced this transformer-based fusion module into state-of-the-art unimodal AQA models to fuse the multimodal inputs in MMW-AQA. Our experimental results on three baseline multimodal AQA models indicate that multimodal data in MMW-AQA improves the AQA accuracy compared with unimodal video data.

The contributions of this paper are as follows.

- We propose MMW-AQA that introduces multimodal data into the AQA research field for the first time. This dataset also introduces the wild video data with diversified viewpoints.
- We present the baseline multimodalization framework for evaluating the effectiveness of multimodal data in MMW-AQA.
- Extensive comparisons and ablation studies demonstrate that using multimodal data improves the AQA accuracy compared with unimodal video data.

Our dataset will be publicly available at https://github.com/ntthilab-cyb/mmwaqa_dataset.

II. RELATED WORK

A. DATASETS FOR ACTION QUALITY ASSESSMENT

The MIT [1] and UNLV [2] datasets were the earliest AQA datasets that provide video data inputs and action-quality scores as annotation labels in diving, gymnastic vault, and figure skating sports. Subsequent AQA datasets have built upon these earliest datasets from various perspectives. For example, AQA-7 dataset [3] increases the total number of competitive sports to seven, including the new skiing, snowboarding, trampoline, and synchronized diving to generalize DNNs across multiple sports. MTL-AQA dataset [4] introduces the concept of multi-task learning into AQA by providing action class and action-described text other than action-quality score. Moreover, FineDiving dataset [9] provides temporal action segmentation that divides the action into multiple steps to help DNNs understand the action procedures. Recently, LOGO dataset [29] extends the application scope of AQA from small-group to large-group scenarios by providing artistic swimming sports with formation graphs representing the position relationships between multiple athletes.

While various AQA datasets have been proposed as summarized above, all these datasets still provide only video data captured from fewer viewpoints to assess action quality. This characteristic hinders the AQA accuracy for complex 3D actions and the robustness evaluation against diversified viewpoints. In contrast, our MMW-AQA provides multimodal data, including the IMU and GPS data, and the video data captured from diversified viewpoints, as shown in Table 1.

B. METHODS FOR ACTION QUALITY ASSESSMENT

In competitive sports, such as diving, gymnastics vault, and figure skating, most AQA methods have tried to extract spatiotemporal action features from videos using DNNs to assess action quality accurately. These methods can be categorized into two main approaches: train DNNs with single/pair-video inputs.

Parmar and Morris proposed the C3D-SVR and C3D-LSTM [2] models that use 3D convolution neural networks to extract spatiotemporal action features from single-video inputs. Parmar and Morris [4] also proposed the C3D-AVG-MTL model that learns spatiotemporal action features across multiple tasks (e.g., the action classification and video captioning tasks) and showed higher AQA accuracy than DNNs trained on only a single AQA task. Moreover, Tang et al. [7] introduced the uncertainty-aware score distribution learning (USDL) model that estimates not score values but score distributions, representing the ambiguity of human judges, from single-video inputs.

In contrast to the approach of single-video inputs, recent AQA methods use pair-video inputs to help DNNs accurately distinguish the difference between two actions. This concept, known as contrastive regression (CoRe), was initially

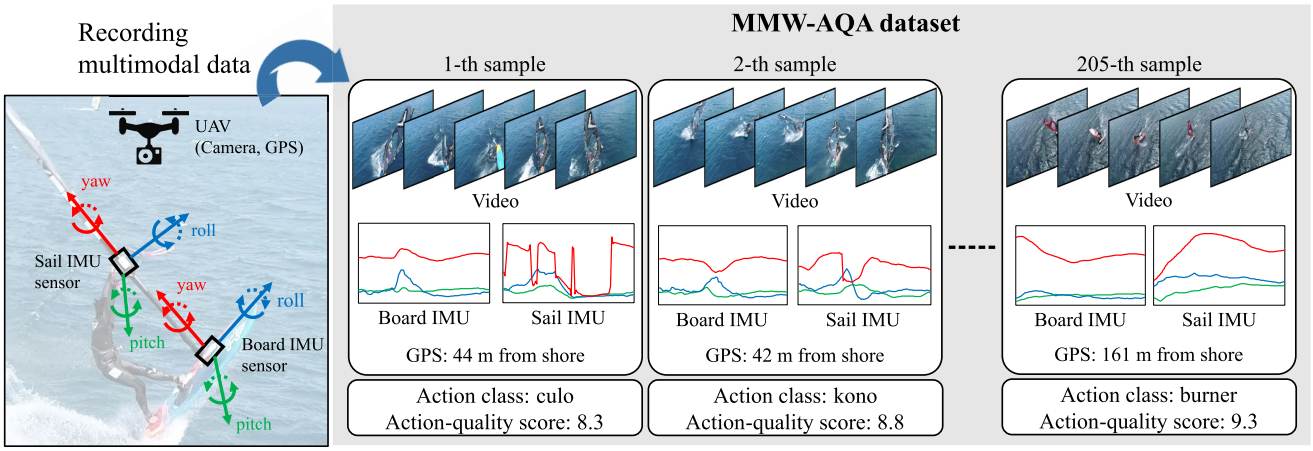


FIGURE 1. Illustration of MMW-AQA. This is a novel multimodal in-the-wild dataset for action quality assessment (AQA), focusing on freestyle windsurfing. This dataset contains multiple modalities, including video, inertial measurement unit (IMU), and global positioning system (GPS) data. It also includes wild video data captured from diversified viewpoints by a single unmanned aerial vehicle (UAV).

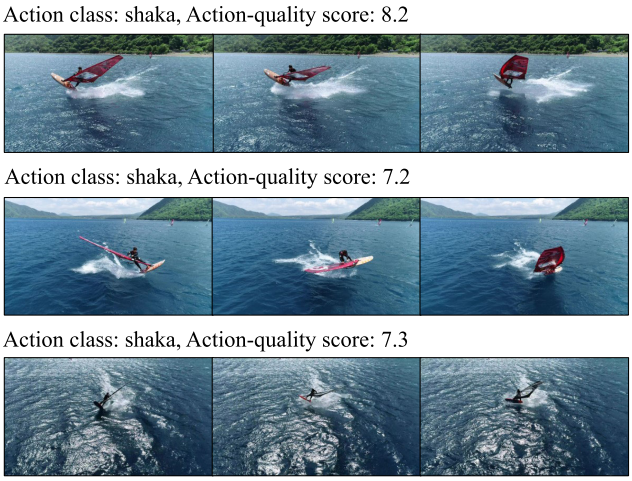


FIGURE 2. Three videos captured when performing the action, shaka. Even for the same action, viewpoints changed significantly due to UAV movements.

proposed by Yu et al. [8]. Later, Xu et al. [9] and Bai et al. [10] proposed procedure-aware CoRe models that segment an action into multiple steps and distinguish the pair-video inputs at each step. Moreover, Li et al. [11] introduced the pairwise contrastive learning network (PCLN) model that distinguishes the pair-video inputs and estimates the action-quality score of each video input simultaneously.

While these methods showed that the good use of single/pair-video inputs is critical for improving the AQA accuracy, the input type they can handle is limited to unimodal video data. Thus, the existing unimodal AQA models cannot directly apply to our MMW-AQA, which consists of multimodal inputs. In this work, we present the baseline multimodalization framework with a transformer-based fusion module. These frameworks multimodalize existing unimodal AQA models for evaluating the effectiveness of multimodal data in MMW-AQA.

III. PROPOSED DATASET: MMW-AQA

We explain our multimodal in-the-wild dataset MMW-AQA. MMW-AQA contains various freestyle windsurfing actions by professional athletes. In this sport, athletes perform various actions at sea, and human judges assess the action quality on the basis of overall finish and impression. We give details of the data modalities, annotations, and statistics in MMW-AQA below.

A. DATA MODALITIES

1) VIDEO DATA

The video data show actions of freestyle windsurfing by professional athletes. To capture these data, we used a camera of a single UAV Phantom 4 Pro. A skilled UAV pilot operated the UAV alongside the athletes at sea, capturing actions from diversified viewpoints, as illustrated in Figure 2. These data were recorded with 3840 × 2160 pixels and 60 fps. For privacy protection, all individuals with identifiable faces who did not consent were blurred.

2) IMU DATA

As shown in Figure 3, IMU sensors were attached to both the board and sail–the tools of freestyle windsurfing maneuvered by the athletes. These motions are closely related to action quality. The IMU data provide 3D rotation angles of the board and sail motions, namely pitch, roll, and yaw, as illustrated in Figure 1. The IMU data were recorded fluctuating at about 15–25 fps and later up-sampled and standardized to 30 fps. These data were synchronized with the video data using timestamps.

3) GPS DATA

The GPS data indicates the distance between the athlete and the shore during the action performance. We used the UAV's GPS as the reference for the athlete's position because it flew close to them. These data are used to determine the action difficulty and annotate the action-quality score, as explained in the following section. This data was recorded as one scalar value for each action; the timing of this record is when an action was performed.

TABLE 1. Comparison of existing AQA datasets [1], [2], [3], [4], [9], [29] and MMW-AQA. Our dataset is novel in two aspects: multimodal data and diversified viewpoints.

Dataset	Modality			Viewpoint	Avg.Dur.	Anno.Type	Samples	Act.Clas.	Sports
	Video	IMU	GPS						
MIT [1]	✓	—	—	small	6.0s(diving)	Score	309	—	Diving,Skating
UNLV [2]	✓	—	—	small	3.8s(diving)	Score	716	—	Diving,Vault,Skating
AQA-7 [3]	✓	—	—	mid	4.1s(diving)	Score	1189	—	Seven sports
MTL-AQA [4]	✓	—	—	mid	4.2s	Action,Score	1412	—	Diving
FineDiving [9]	✓	—	—	mid	4.2s	Action,Score	3000	52	Diving
LOGO [29]	✓	—	—	mid	3m,24s	Action,Formation,Score	200	12	Artistic swimming
MMW-AQA (ours)	✓	✓	✓	diversified	6.0s	Action,Score	205	14	Freestyle windsurfing

B. ANNOTATION

To create this dataset, we employed three active freestyle windsurfing judges as annotators. These annotators labeled the action class and action-quality score under the official rules of freestyle windsurfing competitions and their expert knowledge. The action-quality score is determined by considering two main factors: (1) the level of perfection and overall impression of the action, which encompasses aspects such as jump height and speed, and (2) the athlete's position from the shore when the action was performed. In this sport, action-quality scores generally tend to be higher when athletes perform actions closer to the shore because the wind conditions are usually weaker in that area, making successful performance more challenging. Please note that if the athlete falls into the water after performing the action, the score is zero. We provided the annotators with both video and GPS data to annotate the action class and action-quality score. The action-quality scores are annotated in 0.5-point increments, ranging from the lowest score of 0.0 to the highest score of 10.5. In this paper, the average of the three action-quality scores is used as the final annotation, ranging from the lowest score of 0.0 to the highest score of 9.3, as illustrated in Figure 4 (b).

C. DATASET STATISTICS

MMW-AQA consists of 205 samples from 4 athletes, 3 locations, and 14 action classes, as illustrated in Figure 4 (a). A comparison between MMW-AQA with other AQA datasets is listed in Table 1. MMW-AQA differs from the other datasets regarding modality type and viewpoints. Specifically, MMW-AQA includes multimodal data, including video, IMU, and GPS data. Furthermore, this dataset provides the wild setting, where video data were captured from diversified viewpoints by a single UAV.

IV. MULTIMODALIZATION FRAMEWORK

We present the baseline multimodalization framework that extends existing unimodal AQA models into multimodal AQA models for MMW-AQA, as illustrated in Figure 5. The objective of our framework is to fuse multimodal inputs and predict action-quality scores on MMW-AQA. In our framework, we introduced a transformer-based fusion module, which is used widely and commonly in multimodal learning [37], [38], [39], [40], [41], [42], into the existing unimodal AQA models. In this section, we explain the transformer-based fusion module and two types of multimodalization framework for the existing unimodal AQA models in detail.

A. TRANSFORMER-BASED FUSION MODULE

As the evidence from many studies [36], [43], [44], [45], [46], [47], the multi-head self-attention mechanism of the transformer encoder can effectively learn the relationships between input features. Thus, the transformer encoder is often introduced to models of multimodal learning [37], [38], [39], [40], [41], [42]. Inspired by these studies, we use the transformer encoder to fuse action features of multimodal inputs in MMW-AQA.

As shown in Figure 5, we assume that the transformer-based fusion module receives action features of i-th sample in MMW-AQA. The i-th sample consists of four multimodal inputs: video data $x_{i,0} \in \mathbb{R}^{L\times H\times W\times 3}$, board IMU data $x_{i,1} \in \mathbb{R}^{L\times 3}$, sail IMU data $x_{i,2} \in \mathbb{R}^{L\times 3}$, and GPS data $x_{i,3} \in \mathbb{R}$, where H is the video height, W is the video width, and L is the number of frames. Before the transformer-based fusion module, two processes are performed for the i-th sample: the feature extraction process and the embedding process.

In the feature extraction process, the feature of m-th modal input in the i-th sample is extracted as follows:

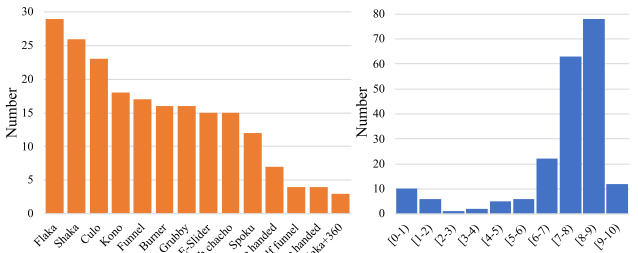
$$f_{i,m} = \begin{cases} F_m(x_{i,m};\theta_m) & (m=0,1,2), \\ x_{i,m}\cdot \mathbf{1}_D & (m=3), \end{cases} \tag{1}$$

where $f_{i,0}, f_{i,1}, f_{i,2} \in \mathbb{R}^{L'\times D}$ are the features of $x_{i,0}, x_{i,1}, x_{i,2}$, respectively, through the feature extraction module $F_m(\cdot;\theta_m)$ parameterized by θ_m, $f_{i,3} \in \mathbb{R}^D$ is the feature of $x_{i,3}$, and $\mathbf{1}_D$ is the D-dimensional all ones vector. Here, L' indicates the temporal dimension of the features. Then, the features, $f_{i,m}$ $(m=0,1,2,3)$, are embedded by modality-type information in the embedding process. This process is necessary because the transformer encoder operates on all input features in parallel and cannot distinguish them. Specifically, we embed these features as follows:

$$f'_{i,m} = f_{i,m} + e_m \tag{2}$$

(a) Board IMU sensor (b) Sail IMU sensor

FIGURE 3. Illustration of IMU sensors attached to (a) board and (b) sail.



(a) Action class (b) Action-quality score

FIGURE 4. Statistics of MMW-AQA. (a) Distribution of action class. (b) Distribution of action-quality score.

where

$$e_m = \begin{cases} m \cdot \mathbf{1}_{L' \times D} & (m = 0, 1, 2), \\ m \cdot \mathbf{1}_{D} & (m = 3). \end{cases}$$

$\mathbf{1}_{L' \times D}$ is the all ones $L' \times D$ matrix. The e_m represents the m-th modality-type information. After the above two processes, the extracted features are fed into the transformer-based fusion module as follows:

$$h_i = \mathcal{T}(f'_{i,0}, f'_{i,1}, f'_{i,2}, f'_{i,3}; \phi), \tag{3}$$

where $h_i \in \mathbb{R}^{(3L'+1) \times D}$ is the fused feature of the all modality-embedded features $f'_{i,0}, f'_{i,1}, f'_{i,2}, f'_{i,3}$ and $\mathcal{T}(\cdot; \phi)$ is the transformer-based fusion module parameterized by ϕ. We used the original transformer encoder [36] as $\mathcal{T}(\cdot; \phi)$.

We introduce the embedding process and transformer-based fusion module described above into existing unimodal AQA models based on two approaches, single/pair-video inputs as described in Section II, to multimodalize them for MMW-AQA. We explain this multimodalization framework for each AQA approach in the following section.

B. MULTIMODALIZATION FRAMEWORK FOR AQA MODELS WITH SINGLE-VIDEO INPUTS

In the existing unimodal AQA models with single-video inputs, the action-quality score s_i with respect to the video data $x_{i,0}$ is predicted as follows:

$$s_i = \mathcal{A}(f_{i,0}; \psi), \tag{4}$$

where $\mathcal{A}(\cdot; \psi)$ is the assessment module parameterized by ψ. To multimodalize these unimodal AQA models for MMW-AQA, we fuse multimodal features before Equation (4), namely, we insert the transformer-based fusion module before the assessment module. Consequently, we can describe our framework as simply replacing the feature $f_{i,0}$ in Equation (4) with the fused feature h_i in Equation (3):

$$s_i = \mathcal{A}(h_i; \psi). \tag{5}$$

Therefore, the existing unimodal AQA models turn to the multimodal AQA models by Equation (5) as shown in Figure 5 (a), and these models will predict action-quality scores accurately because of considering multimodal inputs.

C. MULTIMODALIZATION FRAMEWORK FOR AQA MODELS WITH PAIR-VIDEO INPUTS

In contrast to the AQA models with single-video inputs, the existing unimodal AQA models for AQA with pair-video inputs try to predict a relative score $r_{i,j} = s_i - s_j$ directly. Specifically, these models predict $r_{i,j}$ as

$$r_{i,j} = \mathcal{A}(f_{i,0}, f_{j,0}; \psi). \tag{6}$$

Note that one of the state-of-the-art unimodal AQA model with pair-video inputs, PCLN [11], proposed to predict both the action-quality scores s_i, s_j and their relative score $r_{i,j}$ as follows:

$$[r_{i,j}, s_i, s_j] = \mathcal{A}(f_{i,0}, f_{j,0}; \psi). \tag{7}$$

Similar to the previous Section IV-B, to multimodalize these unimodal AQA models for MMW-AQA, we fuse multimodal features before Equations (6) and (7), namely, we insert the transformer-based fusion module before the assessment module. Consequently, we can describe our framework as simply replacing the features $f_{i,0}$ and $f_{j,0}$ in Equations (6) and (7) with the fused features h_i and h_j in Equation (3):

$$r_{i,j} = \mathcal{A}(h_i, h_j; \psi), \tag{8}$$

$$[r_{i,j}, s_i, s_j] = \mathcal{A}(h_i, h_j; \psi). \tag{9}$$

Therefore, the existing unimodal AQA models turn to the multimodal AQA models by Equations (8) and (9) as shown in Figure 5 (b), and these models will predict relative score accurately because of considering multimodal inputs.

As described above, the heart of our multimodalization frameworks is newly inserting the transformer-based fusion module (and the embedding process) between the feature extraction and assessment modules. Thanks to this simple procedure, our frameworks can be introduced easily into various AQA models. Moreover, our frameworks can adopt the optimization schemes that are used to learn the existing AQA models [7], [8], [11] because our frameworks simply insert the learnable transformer-based fusion module parameterized by ϕ as in Equation (3).

V. EXPERIMENTS

To evaluate the effectiveness of multimodal data for AQA, we conducted experiments to compare the AQA accuracy on multimodal data with unimodal video data. We first describe the evaluation metric and implementation details used in the experiments. We then evaluate the effectiveness of multimodal data and an ablation study.

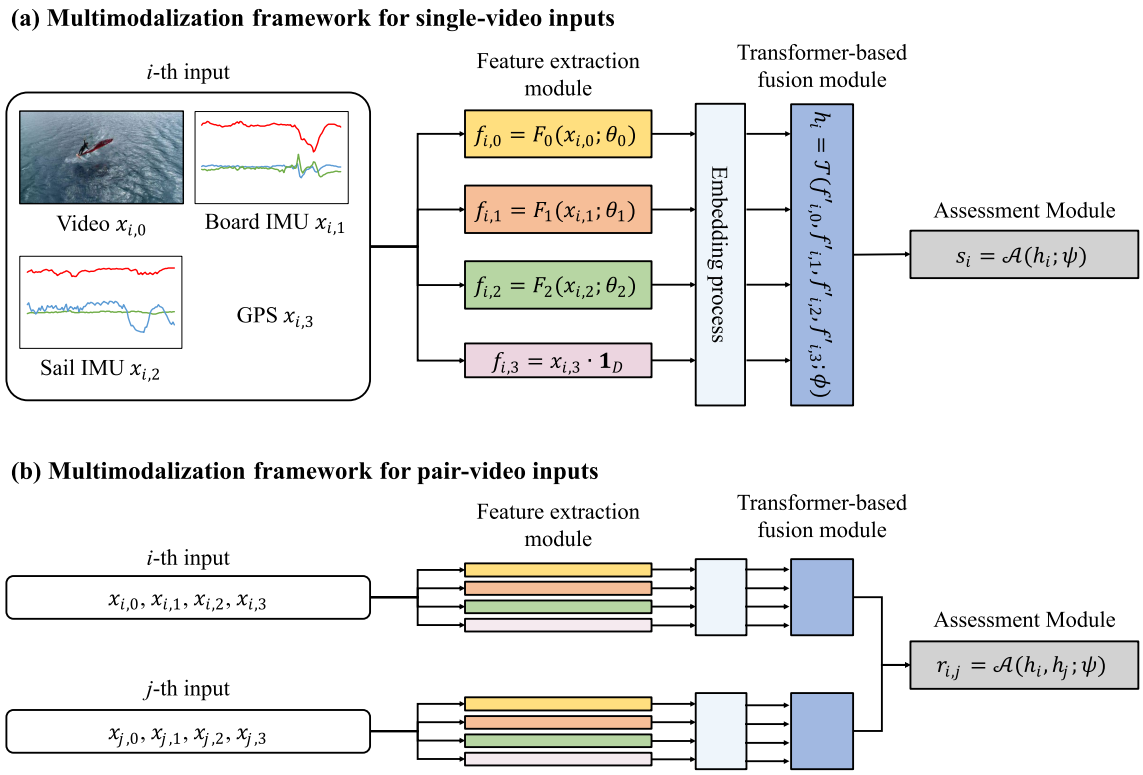


FIGURE 5. Overview of multimodalization frameworks with a transformer-based fusion module for two AQA approaches: (a) single-video and (b) pair-video inputs. These frameworks use multimodal inputs: video, IMU, and GPS data. Transformer-based fusion module effectively fuses features from multiple modalities through its multi-head self-attention mechanism. These frameworks can be introduced easily into various unimodal AQA models.

A. EVALUATION METRIC

To compare the AQA accuracy of our multimodalization frameworks with that of the unimodal versions, we used Spearman's rank correlation (Sp.Corr.) to measure the rank correlation between the outputted action-quality scores $s_1, \ldots, s_K$ and the ground-truth scores $y_1, \ldots, y_K$ where K is the number of samples in a dataset, following existing AQA methods [7], [8], [11]. This metric is defined as

$$\frac{\sum_{k=1}^{K}(p_k - \bar{p})(q_k - \bar{q})}{\sqrt{\sum_{k=1}^{K}(p_k - \bar{p})^2 \sum_{k=1}^{K}(q_k - \bar{q})^2}},$$

where p_k and q_k denote the ranks of s_k and y_k, respectively, and $\bar{p}$ and $\bar{q}$ denote the average of $p_1, \ldots, p_K$ and $q_1, \ldots, q_K$, respectively. Note that since the AQA models with pair-video inputs output the relative score $r_{i,j}$, the action-quality score s_i is calculated by $s_i = r_{i,j} + y_j$.

B. IMPLEMENTATION DETAILS

We implemented all models using PyTorch version 1.12.0. Since the original frame lengths of multimodal data in MMW-AQA differ, we sampled these data along the temporal axis. We set their frame lengths to the same length L for inputting them into the AQA models described in Section IV-A. Specifically, we sampled the m-th original multimodal time-series data ($m = 0, 1, 2$) in the i-th sample of MMW-AQA at the following frame indices n:

$$n = \left\lfloor N_{i,m} \frac{l}{L-1} \right\rfloor \quad (l = 0, 1, \ldots, L-1),$$

where $N_{i,m}$ is the number of frames of the m-th original multimodal data in the i-th sample of MMW-AQA. After this sampling process, the multimodal time-series inputs $x_{i,0}, x_{i,1}, x_{i,2}$ have the same frame length L even if $N_{i,m}$ differ among each sample. We set $L = 103$ for all models following the previous studies [7], [8] for a fair comparison. The resolution of the original video data was resized to 456×256 pixels, and a center cropping of 224×224 pixels was applied, following [7], [8], [11]. Random horizontal flipping for the input video data $x_{i,0}$ was also carried out during training following [7], [8], [11]. We used the Adam optimizer [48] to train all models following [7], [8], [11]. We set the learning rate to 1e-4 for the feature extraction and assessment modules, and 1e-5 for the transformer-based fusion module. All accuracy results are based on 100 epochs. We split 205 samples in MMW-AQA into 80% for training data and 20% for test data. We present only the test data results in this section.

By applying our multimodalization framework to unimodal AQA models, we prepared three multimodal ones: MM-USDL, MM-Core+GART, and MM-PCLN as follows.

- **MM-USDL**: This model is based on USDL [7], the state-of-the-art unimodal AQA model with single-video inputs. We followed [7] for the implementation of this model. For the feature extraction module of the video inputs, we used the pre-trained I3D with the Kinetics dataset [49]. Both video input $x_{i,0}$ and IMU inputs $x_{i,1}, x_{i,2}$ were divided into 10 overlapping clips with 16 frames, and we extracted 1024-dimensional features from each clip (denoted as $f_{i,0}, f_{i,1}, f_{i,2} \in \mathbb{R}^{10\times1024}$). All input clips started from $\{0, 10, \ldots, 80, 86\}$ -th frame. The assessment module and optimization schemes followed the original model [7]. This model estimates not score value but score distributions and is optimized on these distributions. The score distributions were generated using a Gaussian distribution with a standard deviation of 5.0 and a mean equal to the ground-truth action-quality score.
- **MM-Core+GART**: This model is based on Core+GART [8], one of the state-of-the-art unimodal AQA models with pair-video inputs. We followed [8] for the implementation of this model. For the feature extraction module of the video inputs, we used the pre-trained I3D with the Kinetics dataset [49]. Both video input $x_{i,0}$ and IMU inputs $x_{i,1}, x_{i,2}$ were divided into 10 clips, and we extracted 1024-dimensional features like the implementation of the above MM-USDL. The assessment module and optimization schemes followed the original model [8]. We set the depth of the assessment module to three. For pair-video inputs, we randomly selected two samples from the same action class.
- **MM-PCLN**: This model is based on PCLN [11], one of the state-of-the-art unimodal AQA models with pair-video inputs. We followed [11] for the implementation of this model. For the feature extraction module of the video inputs, we used the pre-trained ResNet-50 [50] with the ImageNet dataset [51] and the temporal encoder network [52]. The temporal encoder network consists of two stacked encoding blocks, and each block includes 1×1 temporal convolution, activation function, and max-pooling layer. In this model, the weights of ResNet-50 were fixed, and only the temporal encoder network was trained to extract features from the video inputs (denoted as $f_{i,0} \in \mathbb{R}^{25\times1024}$). For feature extraction from IMU inputs, these inputs $x_{i,1}, x_{i,2}$ were divided into 25 overlapping clips with 6 frames, and we extracted 1024-dimensional features $f_{i,1}, f_{i,2} \in \mathbb{R}^{25\times1024}$ from each clip. Clips of IMU inputs started from $\{0, 4, \ldots, 92, 96\}$-th frame. The assessment module and optimization schemes followed the original model [11]. For pair-video inputs, we randomly selected two samples from the total samples.

Moreover, for our multimodal AQA models, we use ConvGRU [53] as the feature extraction module of IMU inputs. Moreover, we use the transformer encoder [36] as the transformer-based fusion module with two layers. Each layer contains a self-attention block with four heads. In our module, unlike the original transformer [36], we did not add position information because we confirmed that it decreases the AQA accuracy in our setting. However, it is worth noting that our proposed multimodalization frameworks are not limited to this type of transformer.

TABLE 2. Comparing the AQA accuracy across different modality combinations in three multimodal AQA models. * and ♯ indicate the models with single/pair-video inputs, respectively.

Model	Modality			Sp.Corr.
	Video	IMU	GPS	
USDL* [7]	✓	—	—	0.5438
MM-USDL*	✓	✓	–	0.5714
	✓	—	✓	0.6015
	✓	✓	✓	**0.6165**
Core+GART♯ [8]	✓	—	—	0.5635
MM-Core+GART♯	✓	✓	—	0.5664
	✓	—	✓	0.5771
	✓	✓	✓	**0.6487**
PCLN♯ [11]	✓	—	—	0.2824
MM-PCLN♯	✓	✓	—	0.3985
	✓	—	✓	0.2991
	✓	✓	✓	**0.4250**

TABLE 3. L_1 score error of each model for two samples in Figure 6. * and ♯ indicate the models with single/pair-video inputs, respectively. *Full* indicate the models trained with all modalities of the video, IMU, and GPS data.

Model	Modality	L_1 Score Error	
		(a)	(b)
USDL* [7]	video	0.242	1.388
Core+GART♯ [8]	video	0.586	2.568
PCLN♯ [11]	video	1.303	2.964
MM-USDL*	Full	0.316	1.202
MM-Core+GART♯	Full	0.276	1.503
MM-PCLN♯	Full	0.151	1.310

C. EFFECTIVENESS OF MULTIMODAL DATA

In this experiment, we tested whether the utilization of multimodal data could improve the AQA accuracy compared with unimodal video data. Table 2 presents the AQA accuracy when training models with different modalities as inputs. We first discuss the AQA accuracy of three unimodal AQA models that rely solely on video data: USDL [7], Core+GART [8], and PCLN [11]. As shown in Table 2, all unimodal AQA models showed lower AQA accuracy compared with results reported in previous research [7], [8], [11], e.g., those for snowboarding on the AQA-7 [3] dataset, which includes a similar data volume to MMW-AQA. This result suggests that accurately assessing action quality using wild video data with diversified viewpoints is more challenging compared with fewer viewpoints as observed in existing AQA datasets [1], [2], [3], [4], [9], [29]. We then compared the AQA accuracy between using unimodal video

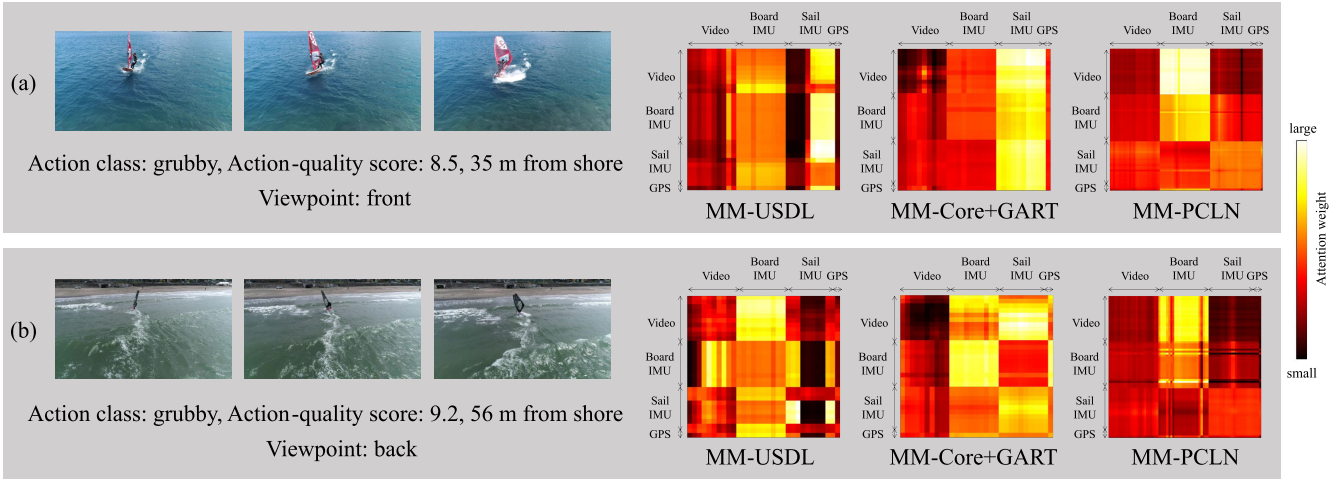


FIGURE 6. Two samples performing the same action, grubby, with different viewpoints and heatmaps of attention weights in the transformer-based fusion module. The scales of each heatmap are tuned for the best view.

data and multimodal data. The table clearly shows that, across all multimodal AQA models, the highest AQA accuracy was achieved when all modalities were used. This result indicates that the utilization of multimodal data helps DNNs learn complex 3D actions and improves the AQA accuracy compared with unimodal video data.

Next, to verify the effectiveness of multimodal data for complex 3D action and diversified viewpoints, we compared the AQA accuracy of the two action samples. Figure 6 shows the same action samples, whose action class is named by *grubby*. As we can see, these samples are captured from diversified viewpoints: sample (a) from the front of the athlete and sample (b) from the back of the athlete. We emphasize that the back viewpoint of the grubby class is included in only test data; for the trained DNNs, sample (b) can be regarded as a complex 3D action sample captured from an unexpected viewpoint in the grubby class. We here evaluated the L_1 score error between the predicted action-quality score of each model and the ground-truth one for these two samples. We first compared the results between samples (a) and (b) in Table 3. In all unimodal and multimodal AQA models, we can see that the score errors of sample (b) are larger than those of sample (a). This is because sample (b) is more complex than sample (a) due to its unexpected viewpoint. These results imply that assessing complex 3D action quality, such as grubby action in sample (b), is difficult in AQA. We then compared the results in terms of using data modality in Table 3, namely the unimodal (only video) AQA models or our multimodal AQA ones. In both samples (a) and (b), the table shows that our multimodal AQA models achieved totally lower score errors than the unimodal AQA ones. In particular, a significant improvement was shown in MM-Core+GART and MM-PCLN for sample (b). These results indicate that multimodal data improves robustness against complex 3D action and diversified viewpoints.

Finally, to verify whether the multimodal AQA models focus on multimodal data effectively, we visualize the attention weights in the transformer-based fusion module. Figure 6 presents heatmaps of attention weights, denoted as $(3L' + 1) \times (3L' + 1)$ matrices, generated from the transformer-based fusion module of each multimodal AQA model. The value at coordinate (x, y) in the heatmap indicates the attention weight from a x-th input to a y-th input. In MM-USDL and MM-Core+GART, the range $0 \leq x, y < 10$ represents video data, $10 \leq x, y < 20$ represents board IMU data, $20 \leq x, y < 30$ represents sail IMU data, and $x, y = 30$ represents GPS data, respectively. In MM-PCLN, the range $0 \leq x, y < 25$ represents video data, $25 \leq x, y < 50$ represents board IMU data, $50 \leq x, y < 75$ represents sail IMU data, and $x, y = 75$ represents GPS data, respectively. In MM-USDL and MM-Core+GART, the attention weights for sample (b) show larger values across various modalities compared with those of sample (a), which is biased to sail IMU data. In MM-PCLN, the attention weights for both samples (a) and (b) show strong responses to board IMU data. These different attention weights indicate that each multimodal AQA model can select effective modality data for each sample to improve the AQA accuracy.

D. ABLATION STUDY

1) EFFECTIVENESS OF FUSION APPROACH

In the above experiments, we used a specific type of fusion approach for multimodal data, i.e., the transformer-based fusion approach. To validate the effectiveness of our transformer-based approach, we evaluated the effects of various fusion approaches for multimodal data. Table 4 shows the AQA accuracy of the three multimodal AQA models with different fusion modules. In this table, *Concat* module simply concatenates features outputted from the feature extraction module, and *MLP* module uses two fully connected layers for fusing these features. We compared the results across Concat, MLP, and transformer-base fusion modules. All multimodal AQA models showed the highest AQA accuracy when using the transformer-based fusion module. These results indicate

TABLE 4. Effectiveness of fusion approach. * and ♯ indicate the models with single/pair-video inputs, respectively.

Model	Fusion module	Sp.Corr.
MM-USDL*	Concat	0.5323
	MLP	0.6050
	Transformer	**0.6165**
MM-Core+GART♯	Concat	0.4014
	MLP	0.4570
	Transformer	**0.6487**
MM-PCLN♯	Concat	0.1745
	MLP	-0.0524
	Transformer	**0.4250**

TABLE 5. Effectiveness of embedding process. * and ♯ indicate the models with single/pair-video inputs, respectively.

Model	e_m	Sp.Corr.
MM-USDL*	—	0.4664
	✓	**0.6165**
MM-Core+GART♯	—	0.5895
	✓	**0.6487**
MM-PCLN♯	—	0.3767
	✓	**0.4250**

TABLE 6. Comparison on feature extraction module $F_{i,1}, F_{i,2}$ for IMU inputs. * and ♯ indicate the models with single/pair-video inputs, respectively.

Model	$F_{i,1}, F_{i,2}$	Sp.Corr.
MM-USDL*	BiLSTM [54]	0.3305
	GRU [55]	0.5365
	ConvGRU [53]	**0.6165**
MM-Core+GART♯	BiLSTM [54]	0.3572
	GRU [55]	0.5379
	ConvGRU [53]	**0.6487**
MM-PCLN♯	BiLSTM [54]	0.4003
	GRU [55]	**0.4847**
	ConvGRU [53]	0.4250

the effectiveness of the transformer-based fusion module for fusing multimodal data in MMW-AQA.

2) EFFECTIVENESS OF EMBEDDING PROCESS FOR DISTINGUISHING INPUT MODALITY TYPE

We also evaluated the effect of the embedding process, which aims to distinguish the input modality type in the transformer-based fusion module. Table 5 shows the AQA accuracy of multimodal AQA models that were trained with or without the embedding process. In all multimodal AQA models, the AQA accuracy is improved by incorporating the modality-type information with the embedding process. These results suggest that considering modality-type information is effective in the multimodal AQA models for MMW-AQA.

3) EFFECTIVENESS OF FEATURE EXTRACTION MODULE $F_{I,1}, F_{I,2}$

We also evaluated the feature extraction module $F_{i,1}, F_{i,2}$ for IMU inputs. Table 6 shows the AQA accuracy of multimodal AQA models trained with different feature extraction modules for IMU inputs: Bidirectional LSTM (BiLSTM) [54], GRU [55], and ConvGRU [53]. ConvGRU achieved the highest AQA accuracy in MM-USDL and MM-Core+GART, and ranked second place in MM-PCLN. Therefore, we chose ConvGRU, which demonstrated consistently high average accuracy, as the feature extraction module $F_{i,1}, F_{i,2}$ for IMU inputs in the experimental section.

VI. CONCLUSION

In this paper, we presented a novel AQA dataset, MMW-AQA, in freestyle windsurfing. In contrast to existing AQA datasets that only provide unimodal video data, MMW-AQA newly provides multimodal data, including video, IMU, and GPS data. The IMU data helps DNNs accurately assess complex 3D actions, and the GPS data is important to determine the action difficulty. With extensive experimental results with MMW-AQA, we showed that using multimodal data is plausible to improve the AQA accuracy. As another characteristic of MMW-AQA, its video data were captured by a single UAV flying around an athlete at sea with diversified viewpoints. Through the evaluation of such *wild* video data, we observed that existing AQA methods do not have robustness against viewpoints that models have never seen during training (see Table 3). In contrast, we observed that using multimodal data in MMW-AQA improves the robustness against such viewpoints.

This paper also presented the baseline multimodalization frameworks with a transformer-based fusion module, as shown in Figure 5. This framework extends the existing AQA models, which assume unimodal video data as input, into models that can receive multimodal data. By applying this framework to the existing state-of-the-art unimodal AQA models, we showed the effectiveness of using multimodal data compared with unimodal video data.

A. FUTURE WORK

While we confirmed that the multimodal AQA models with MMW-AQA performed well in our experiments, MMW-AQA is relatively smaller in data volume compared with other recent datasets [4], [9], [29]. Constructing a larger multimodal dataset across various scenarios is a crucial future task for further exploring the effectiveness of multimodal data in the AQA research field. Furthermore, discussing the dependence of AQA accuracy with respect to the action class is difficult due to the small data volume. In future work, we will discuss this point by constructing a larger multimodal dataset.

REFERENCES

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

[2] P. Parmar and B. T. Morris, "Learning to score Olympic events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 76–84.

[3] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *Proc. WACV*, Jan. 2019, pp. 1468–1476.

[4] P. Parmar and B. T. Morris, "What and how well you performed? A multitask learning approach to action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 304–313.

[5] J. Pan, J. Gao, and W. Zheng, "Action assessment by joint relation graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6330–6339.

[6] S. Zahan, G. M. Hassan, and A. Mian, "Learning sparse temporal video mapping for action quality assessment in floor gymnastics," 2023, *arXiv:2301.06103*.

[7] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9836–9845.

[8] 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. (ICCV)*, Oct. 2021, pp. 7899–7908.

[9] 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. (CVPR)*, Jun. 2022, pp. 2939–2948.

[10] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 422–438.

[11] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 457–473.

[12] 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.

[13] 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.

[14] 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. 3222–3231.

[15] C. Zhou, Y. Huang, and H. Ling, "Uncertainty-driven action quality assessment," 2022, *arXiv:2207.14513*.

[16] B. Zhang, J. Chen, Y. Xu, H. Zhang, X. Yang, and X. Geng, "Auto-encoding score distribution regression for action quality assessment," 2021, *arXiv:2111.11029*.

[17] M.-Z. Li, H.-B. Zhang, L.-J. Dong, Q. Lei, and J.-X. Du, "Gaussian guided frame sequence encoder network for action quality assessment," *Complex Intell. Syst.*, vol. 9, no. 2, pp. 1963–1974, Apr. 2023.

[18] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, T. Ploetz, M. A. Clements, and I. Essa, "Automated video-based assessment of surgical skills for training and evaluation in medical schools," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 9, pp. 1623–1636, Sep. 2016.

[19] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Towards unified surgical skill assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9517–9526.

[20] J. Gao, J.-H. Pan, S.-J. Zhang, and W.-S. Zheng, "Automatic modelling for interactive action assessment," *Int. J. Comput. Vis.*, vol. 131, no. 3, pp. 659–679, Mar. 2023.

[21] D. Anastasiou, Y. Jin, D. Stoyanov, and E. Mazomenos, "Keep your eye on the best: Contrastive regression transformer for skill assessment in robotic surgery," *IEEE Robot. Autom. Lett.*, vol. 8, no. 3, pp. 1755–1762, Mar. 2023.

[22] 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.

[23] 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.

[24] T. Elgamal and K. Nahrstedt, "Multicamera summarization of rehabilitation sessions in home environment," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 585–602.

[25] P. Parmar and B. T. Morris, "Measuring the quality of exercises," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 2241–2244.

[26] K. Zhou, R. Cai, Y. Ma, Q. Tan, X. Wang, J. Li, H. P. H. Shum, F. W. B. Li, S. Jin, and X. Liang, "A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 5, pp. 2456–2466, May 2023.

[27] P. Parmar, J. Reddy, and B. Morris, "Piano skills assessment," in *Proc. IEEE 23rd Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2021, pp. 1–5.

[28] A. Ejiri, K. Iida, H. Tomimori, Y. Ikai, S. Yamao, K. Teduka, K. Yanai, and M. Nishikawa, "3D sensing of gymnastics competition using MEMS mirror laser sensor," in *Proc. 60th Annu. Conf. Soc. Instrum. Control Engineers Jpn. (SICE)*, Sep. 2021, pp. 1175–1180.

[29] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "LOGO: A long-form video dataset for group action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2405–2414.

[30] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13222–13232.

[31] J. Zhan, X. Nie, and J. Feng, "Inference stage optimization for cross-scenario 3D human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst. (NeuIPS)*, 2020, pp. 2408–2419.

[32] M. Gholami, B. Wandt, H. Rhodin, R. Ward, and Z. J. Wang, "AdaptPose: Cross-dataset adaptation for 3D human pose estimation by learnable motion generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13065–13075.

[33] B. Wandt, J. J. Little, and H. Rhodin, "ElePose: Unsupervised 3D human pose estimation by predicting camera elevation and learning normalizing flows on 2D poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6625–6635.

[34] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16261–16270.

[35] G. Shi, X. Fu, C. Cao, and Z.-J. Zha, "Alleviating spatial misalignment and motion interference for UAV-based video recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 193–202.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[37] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7463–7472.

[38] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7073–7083.

[39] M. Yasuda, Y. Ohishi, S. Saito, and N. Harada, "Multi-view and multi-modal event detection utilizing transformer-based multi-sensor fusion," in *Proc. ICASSP*, 2022, pp. 4638–4642.

[40] J. Chen and C. M. Ho, "MM-ViT: Multi-modal video transformer for compressed video action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 786–797.

[41] M. M. Islam and T. Iqbal, "MuMu: Cooperative multitask learning-based guided multimodal fusion," in *Proc. AAAI*, 2022, pp. 1043–1051.

[42] X. Gong, S. Mohan, N. Dhingra, J.-C. Bazin, Y. Li, Z. Wang, and R. Ranjan, "MMG-Ego4D: Multi-modal generalization in egocentric action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6481–6491.

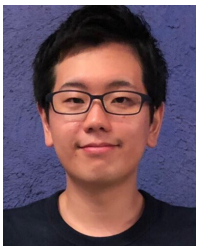
[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[45] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[47] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[49] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[52] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1003–1012.
[53] S. Moon, A. Madotto, Z. Lin, A. Dirafzoon, A. Saraf, A. Bearman, and B. Damavandi, "IMU2CLIP: Multimodal contrastive learning for IMU motion sensors from egocentric videos and text," 2022, *arXiv:2210.14395*.
[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[55] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Deep Learn. Represent. Learn. Workshop*, 2014.

TAKASUKE NAGAI received the B.E. and M.E. degrees in engineering from Tsukuba University, Japan, in 2017 and 2019, respectively. In 2019, he joined Nippon Telegraph and Telephone Corporation (NTT). His research interests include development of computer vision and machine learning.



SHOICHIRO TAKEDA received the B.E. and M.E. degrees in biosciences and informatics from Keio University, Japan, in 2014 and 2016, respectively, and the Ph.D. degree in engineering from the University of Tsukuba, Japan, in 2021. In 2016, he joined Nippon Telegraph and Telephone Corporation (NTT). His research interests include development of image and signal processing, computer vision, and machine learning. He received the Third Science Intercollegiate Incentive Award from the Ministry of Education, the IPSJ AVM Award, in 2019, and the IPSJ CGVI Research Award, in 2019.

SATOSHI SUZUKI received the B.E., M.E., and Ph.D. degrees from the University of Electro-Communications, in 2015, 2017, and 2022, respectively. In 2017, he joined Nippon Telegraph and Telephone (NTT). He is currently a Researcher with NTT Human Informatics Laboratories. His current research interests include neural networks, computer vision, surveillance systems, and machine learning. He received the IEEE CIS Japan Chapter Young Researcher Award, in 2015. He is a member of the Information Processing Society of Japan (IPSJ).

HITOSHI SESHIMO received the B.E. and M.E. degrees in mechanical engineering from Waseda University, Tokyo, Japan, in 1995 and 1997, respectively. In 1997, he joined Nippon Telegraph and Telephone Corporation (NTT). His research interests include computer-aided instruction, web-based learning, content distribution and navigation systems, geographical information services, and cybernetics.