Sci

MDPI

Article

Creating a Novel Attention-Enhanced Framework for Video-Based Action Quality Assessment

Wenhui Gong [1], Wei Li [1,*], Huosheng Hu [2,*], Zhijun Song [3], Zhiqiang Zeng [1], Jinhua Sun [1] and Yuping Song [4]

[1] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China; 2322081015@stu.xmut.edu.cn (W.G.), zqzeng@xmut.edu.cn (Z.Z.), jhsun@xmut.edu.cn (J.S.)
[2] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK
[3] School of Physics and Information Engineering, Jiangsu Second Normal University, Nanjing 210013, China; wlp2022@jssnu.edu.cn
[4] School of Mathematical Sciences, Xiamen University, Xiamen 361005, China; ypsong@xmu.edu.cn
* Correspondence: weili@xmut.edu.cn (W.L.), hhu@essex.ac.uk (H.H.)

Abstract: Action Quality Assessment (AQA)—the task of evaluating how well human actions are performed—is essential in domains such as sports and medicine. Existing AQA methods typically rely on score regression following feature extraction but often neglect the ambiguity inherent in extracted features. In this work, we introduce a novel AQA framework that incorporates a modified attention module to better capture relevant information. Our approach segments video data into clips, extracts features using the I3D network, and applies attention mechanisms to highlight salient features while suppressing irrelevant ones. To assess feature quality, we employ score distribution regression and propose an uncertainty-aware score distribution learning strategy that models features as Gaussian distributions. We further leverage Variational Autoencoders (VAEs) to capture complex latent representations and quantify uncertainty. Extensive experiments on the MTL-AQA and JIGSAWS datasets demonstrate the effectiveness and robustness of our proposed method.

Keywords: attention mechanism; I3D network; feature extraction; video action quality assessment

Academic Editor: Anastasios Doulamis

Received: 12 February 2025
Revised: 25 April 2025
Accepted: 28 April 2025
Published: 6 May 2025

Citation: Gong, W.; Li, W.; Hu, H.; Song, Z.; Zeng, Z.; Sun, J.; Song, Y. Creating a Novel Attention-Enhanced Framework for Video-Based Action Quality Assessment. *Sci* **2025**, *7*, 54. https://doi.org/10.3390/sci7020054

1. Introduction

Action Quality Assessment (AQA) extends the scope of human action recognition by not only identifying actions within videos but also evaluating and scoring their execution, aiming to mitigate biases inherent in manual judgment. Due to its relevance in practical applications, AQA has emerged as a significant research focus in the field of computer vision. In recent years, a variety of methods have been developed to address AQA, with much of the research is centered on sports video analysis [1–3] and medical care [4–6], driven by the availability of domain-specific datasets and the complexity of assessment tasks. Typical vision-based AQA frameworks begin by extracting action features from videos, followed by quality evaluation and score prediction. Currently, convolutional 3D (C3D) [7] and inflatable 3D (I3D) [8] networks are among the most widely used feature extractors, due to their successful pre-training in large-scale action recognition datasets.

However, despite their success, most existing AQA methods rely heavily on the direct regression of features, often overlooking the need to emphasize features that are most relevant to accurate quality assessment. To address this limitation, we propose a novel

approach that integrates channel attention into the feature evaluation process. Drawing inspiration from the squeeze-and-excitation (SE) framework [9], we adapt and enhance its compression mechanism to develop a more effective attention module. This module selectively amplifies discriminative features while suppressing irrelevant or redundant information. By incorporating channel attention, our method refines feature representations and significantly improves the model's ability to capture subtle variations in action quality, leading to more accurate and robust assessments.

Traditional approaches often fail to account for uncertainties in data, resulting in suboptimal performance. Since datasets are constructed subjectively, observational noise can corrupt target values, introducing inconsistencies. To mitigate this, we adopted an uncertainty-aware score distribution learning strategy [10] combined with Variational Autoencoders (VAE) [11]. By leveraging the VAE's capability to model complex latent distributions, we enhance feature evaluation robustness while capturing the underlying uncertainties associated with each feature. This not only boosts the discriminative power of the extracted features but also provides valuable insights into their reliability, leading to a more refined understanding of feature significance. An overview of our work's primary contributions is given below.

- Attention-Enhanced Feature Learning: To enhance the model's capacity to concentrate on important details while eliminating extraneous features, we included an attention module. By giving distinct channels different weights, the module prioritizes features that significantly influence predicted scores, dynamically refining the model's feature learning capacity.
- Uncertainty-Aware Feature Modeling: We applied a VAE to the extracted features, encoding latent variables as Gaussian distributions. The model captures uncertainty in the generated samples while learning the latent structure of the data through variational inference and reparameterization.
- Comprehensive Evaluation and Analysis: Numerous experiments were carried out using datasets that are accessible to the public, showing that our approach attains state-of-the-art results with respect to Spearman rank correlation. Additionally, ablation studies systematically assess the contributions of each model component, validating the effectiveness of our approach.

The remainder of this paper is organized as follows: Section 2 outlines some related research works conducted in the field of Action Quality Assessment, attention mechanism, and uncertainty learning. Section 3 presents our proposed framework, which contains three components: feature extraction, attention module, and score distribution regression. Section 4 describes the conducted experiments to show the viability and effectiveness of the suggested strategy in contrast to a few previous works. Finally, Section 5 presents a concise conclusion along with suggestions for future work.

2. Related Work

This section presents a review of three research topics: Action Quality Assessment, attention mechanism, and uncertainty learning, which are closely related to our research work conducted in this article.

2.1. Action Quality Assessment

The goal of Action Quality Assessment (AQA) is to develop systems capable of automatically and objectively evaluating the execution of specific human actions from video input. AQA has a wide range of practical applications, including analyzing athlete performance to support coaching, assessing surgical skills in medical training, and evaluating motor function in rehabilitation scenarios.

Early AQA approaches relied on hand-crafted features and traditional machine learning classifiers. Gordon [12] was among the first to explore video-based AQA, using gymnastics scoring as a case study to demonstrate its feasibility. Ilg et al. [13] introduced a spatio-temporal deformation model that established correspondences at both the global action sequence level and the individual motion element level for comparative analysis. Further advancing the field, Pirsiavash et al. [14] proposed a method based on frequency domain analysis, combining high-level pose features with low-level visual cues for quality assessment.

The emergence of general-purpose deep neural networks (DNNs) has significantly advanced AQA research. The strong representational capacity of DNNs [15] has driven a shift from hand-crafted features to end-to-end learning models, resulting in substantial performance improvements. Xu et al. [16] proposed a dual-branch architecture incorporating Long Short-Term Memory (LSTM) and Skip-LSTM modules to capture both global and local temporal dependencies in video sequences. Li et al. [17] introduced a new framework that uses C3D for feature extraction and enhances learning through a combination of ranking loss and mean squared error (MSE) loss. Building on this, Doughty et al. [18] leveraged I3D and further improved score prediction by designing a more sophisticated ranking loss function to better capture the nuances of action quality.

2.2. Attention Mechanism

Attention mechanisms are designed to focus on the most relevant regions of an image while filtering out less important areas, emulating the human visual system's ability to efficiently interpret complex visual scenes. In computer vision, attention mechanisms dynamically select and weight features based on their importance, thereby enhancing a model's ability to process and understand visual input.

Among various types, channel attention specifically improves feature representations by adaptively adjusting the weights of individual channels. This enables the model to emphasize informative features while suppressing less relevant ones, ultimately boosting overall performance. Building upon this idea, Hu et al. [9] introduced the Squeeze-and-Excitation (SE) module, which explicitly models inter-channel relationships to strengthen feature representation. Similarly, Wang et al. [19] proposed the Non-Local Network, leveraging self-attention to capture long-range dependencies within visual data.

Given the demonstrated success of attention mechanisms in various computer vision tasks, recent research has increasingly explored their application in AQA. For example, Wang et al. [20] integrated a single-object tracker with AQA and introduced the Tube Self-Attention (TSA) module to enhance performance. Lei et al. [21] proposed an end-to-end temporal attention framework that improves Action Quality Assessment in sports videos by mimicking human perceptual and evaluative behavior during temporal modeling.

2.3. Uncertainty Learning

Uncertainty learning focuses on modeling uncertainty to enhance model performance, particularly for complex tasks or incomplete data. The main objective is to improve the model's reliability by quantifying confidence in its predictions. Key approaches include Bayesian neural networks [22] and variational inference [23]. Gal et al. [24] interpreted dropout as a form of Bayesian inference, making deep learning models capable of estimating uncertainty. Ovadia et al. [25] highlighted the importance of uncertainty in deep learning, especially in areas such as image classification and object detection. As deep learning evolved, numerous AE variants emerged, including Denoising Auto-encoders for image denoising [26] and Convolutional Auto-encoders for feature extraction [27]. Building on variational Bayesian inference, Kingma et al. [11] proposed the Variational Auto-encoder

(VAE), which models the latent space distribution to generate samples and quantify the uncertainty involved in the generation process.

Owing to its strong performance in deep learning, uncertainty learning has been increasingly adopted in Action Quality Assessment (AQA) tasks. Among them, Tang et al. [10] proposed a Score Distribution Learning approach with an emphasis on uncertainty awareness, which models each action as a score distribution rather than a single value. This approach provides a more reliable representation of action quality. To reflect inconsistencies across different judges' assessments, Zhou et al. [28] designed a probabilistic method known as Uncertainty-Driven AQA (UD-AQA), which focuses on modeling score variation.

Both attention mechanisms and uncertainty learning have shown promising performance in AQA tasks. However, few studies have effectively integrated these two approaches, we propose an innovative method that combines the strengths of both techniques for a more robust Action Quality Assessment.

3. Approach

This section describes our proposed framework with three components: feature extraction, attention module, and regression of score distribution.

3.1. Feature Extraction

First, an n-frame action video is split into clips, after which the feature extraction is performed. As shown in Figure 1, we split the entire video sequence into n clips, each consisting of 16 frames, denoted as $c_1, c_2, \ldots, c_n$, in accordance with most of the previous work [10,29]. Using a weight-sharing Inflated 3D backbone network [8], these clips $c_1, c_2, \ldots, c_n$ are individually encoded into features $f_1, f_2, \ldots, f_n$. For clip c_i, the process of extracting the corresponding feature f_i is described by the following equation

$$f_i = F_{I3D}(c_i). \tag{1}$$

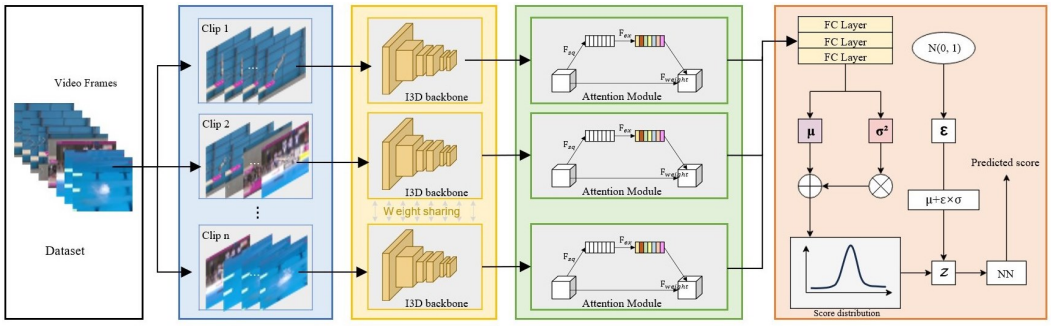


Figure 1. The proposed framework in which input video frames are segmented into n clips and processed through an I3D backbone for feature extraction. An attention module is then applied to improve the model's focus on important features. A trio of fully connected layers synthesize the final features, which are encoded as a Gaussian distribution. The VAE and re-parameterization trick are applied to obtain the predicted scores.

The input to the I3D network consists of images resized to 224 × 224, and the output feature dimension is 1024. The full architecture of the I3D network is illustrated in Figure 2. Specifically, Figure 2a depicts the 3D convolutional layers, pooling layers, and Inception modules, while Figure 2b provides a detailed view of the Inception module design.

The Inception module is designed to extract features at multiple spatial and temporal scales within a single layer while maintaining computational efficiency. It achieves this by processing the input through four parallel branches, each with different kernel sizes and operations. This architecture enables the module to capture a diverse range of features

across various receptive fields and to integrate them effectively, thereby enhancing the model's representational power and flexibility.

In contrast to the standard I3D implementation, we removed the step where predictions are made after pooling the features, and instead output the obtained features. This allows us to further process the features in the subsequent attention module.

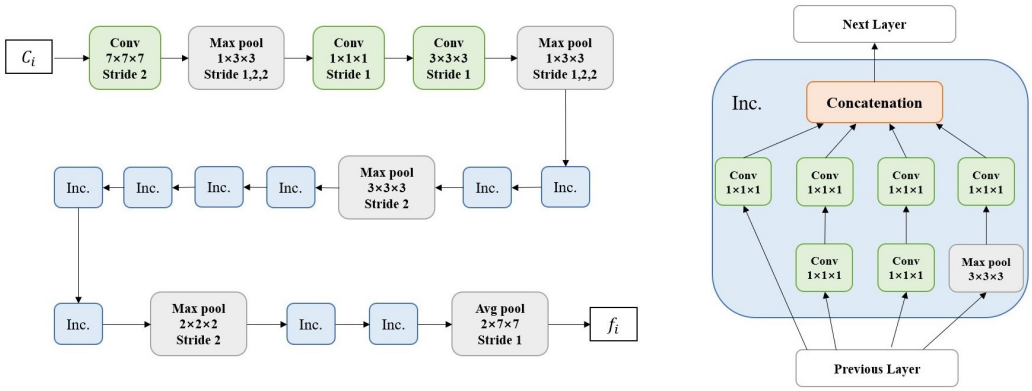


Figure 2. The complete I3D network which is built upon an existing 2D convolutional network by Inflated its convolutional kernels into 3D. (**a**) shows the transformation through multiple convolutional layers, pooling layers, and Inception modules. In our work, we modify the final part of the traditional I3D model. (**b**) shows the detailed structure of the Inception module involves processing the input data through four different paths: one $1 \times 1 \times 1$ convolution path, two paths that transition from $1 \times 1 \times 1$ convolutions to $3 \times 3 \times 3$ convolutions, and one path that applies max pooling followed by a $1 \times 1 \times 1$ convolution. The outputs of these paths are then aggregated.

3.2. Attention Module

We introduced an attention module to enhance the network's sensitivity to informative features, enabling the model to better utilize these features during subsequent transformations while reducing the influence of less relevant ones. This was achieved by explicitly modeling inter-channel dependencies, allowing the network to recalibrate filter responses through a two-step process before passing them to the next layer. Specifically, in the output of the I3D network, we applied global average pooling to aggregate spatial information and generate channel-wise statistical descriptors. This operation compresses the global spatial information of the feature map into a compact channel descriptor.

Let the input feature map be $X \in R^{H \times W \times C}$, a statistic $z \in R^C$ is generated by compressing x across spatial dimensions $H \times W$, where the c-th dimension of z is computed as follows:

$$z_c = F_{squ}(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j). \tag{2}$$

To enable each feature channel to generate an adaptive weight, we model inter-channel dependencies using two fully connected layers. This design allows the network to learn meaningful correlations between channels, with the output dimension matching the number of input channels to preserve alignment. To capture nonlinear interactions among channels, we incorporated a lightweight gating mechanism consisting of a ReLU activation followed by a sigmoid function. This setup enables the module to learn complex, nonlinear relationships while producing normalized attention weights that recalibrate the original feature channels accordingly.

$$w = F_{exp}(z, \mathrm{W}) = \sigma(\mathrm{W}_2 \, \mathrm{ReLU}(\mathrm{W}_1 z)). \tag{3}$$

where $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$.

To parameterize the gating mechanism effectively, we introduced a bottleneck architecture composed of two fully connected (FC) layers surrounding a nonlinearity. This design reduces model complexity and enhances generalization by limiting the number of parameters. Specifically, this includes a dimensionality expansion layer specified by parameters W_2, a ReLU activation, and a layer for dimensionality reduction with parameters W_1 and a reduction factor r. Finally, to perform the channel recalibration, we applied the obtained normalized weights to the features of each channel by multiplying the weights by the original input, element by element. The final features obtained are shown as follows:

$$\tilde{X} = F_{weight}(x_c, w_c) = x_c \cdot w_c. \tag{4}$$

3.3. Score Distribution Regression

Due to the inherently subjective nature of scoring in AQA datasets, the mapping between the input data and the corresponding labels often contains noise, which can degrade the prediction performance. To address this aleatoric uncertainty, rather than directly regressing the final score as in previous approaches, we treat the score as a random variable and aim to learn its probability distribution. The predicted score is then sampled from this learned distribution. Specifically, we modeled the video features as heteroskedasticity Gaussian distributions using a probabilistic encoder, where the output variance captures the data-dependent noise present in each video. This allows the model to explicitly represent uncertainty in the scoring process and improves robustness to label noise.

For a given feature x, a probabilistic encoder is employed to map x into a random scoring variable s. Assume that the scoring variable s follows a Gaussian distribution, as presented in the following formula:

$$g(s \mid x) = \frac{1}{\sqrt{2\pi\sigma^2(x)}} \exp(-\frac{(s-\mu(x))^2}{2\sigma^2(x)}). \tag{5}$$

where the quality and uncertainty of the action score are measured using the parameters $\mu(x)$ and variance $\sigma^2(x)$.

Our approach is inspired by the Variational Autoencoder (VAE) framework [11], which aims to reconstruct input data x by learning a compact, low-dimensional latent representation that effectively captures the underlying data distribution. This ensures that the generated samples $\tilde{x}$ closely approximate the original input x. The overall reconstruction and sampling process is illustrated in the right part of Figure 1.

To obtain the final predicted score, we sampled from the learned score distribution, which is modeled as a Gaussian. Although the distribution is parameterized by the model's predicted mean and variance, directly optimizing these parameters through sampling is non-differentiable. To address this, we applied the Reparameterization Trick, which enables gradient-based optimization by expressing the sampling operation as a deterministic function of a noise variable. This reformulation allows the model to be trained end-to-end while preserving the stochastic nature of score sampling.

An external vector $\epsilon \sim N(0,1)$ was introduced to compute z, with a uni-variate Gaussian distribution $z \sim p(z \mid x) = N(\mu, \sigma^2)$. Sampling z from this distribution can be achieved by first sampling ϵ from $N(0,1)$, and then defined as follows:

$$z = \mu(x) + \epsilon \odot \sigma(x). \tag{6}$$

Thus, we transformed the sampling process from $N(\mu, \sigma^2)$ into sampling from $N(0,1)$, followed by a parameterized transformation to recover a sample from $N(\mu, \sigma^2)$. This

reparameterization ensures that the entire model remains differentiable and trainable through backpropagation.

3.4. Loss Function

To evaluate the effectiveness of the model, we designed a loss function with two mean squared error (MSE) components, which is expressed as follows:

$$\begin{aligned} L &= \mathrm{MSE}(\tilde{x}, x) + \mathrm{MSE}(x, \mu) \\ &= \frac{1}{N}\sum_{i=1}^{N}(\tilde{x}_i - x_i)^2 + \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_i)^2. \end{aligned} \tag{7}$$

where $\tilde{x}$ denotes the predicted score from the model, x denotes the true score of the i-th sample.

During training, the optimizer adjusts the model parameters through the backpropagation method to minimize the loss function mentioned above.

4. Experiment

4.1. Datasets and Evaluation Metrics

MTL-AQA: It is a dataset [30], standing as the largest publicly available resource for Action Quality Assessment (AQA), featuring 1412 detailed diving event samples. Covering individual and synchronized dives from male and female athletes across 10 m platform and 3 m springboard disciplines, it offers diverse annotations for action quality evaluation, commentary generation, and action recognition. Raw scores, including judges' marks and Difficulty Degree (DD), are included. Following the protocol in [31], the dataset is split into 1059 training and 353 test samples.

JIGSAWS: It is a dataset [32] focused on surgical actions, featuring three types of surgical tasks: Suture (S), Needle Passing (NP), and Knotted (KT). The final score is the sum of the sub-scores assigned to each video sample for these tasks, which are used to evaluate different facets of the surgical actions. We adopted a strategy of using left-side video for experiments and four-fold cross-validation similar to the previous work [10].

Evaluation Metrics: As described in previous studies [33–35], we evaluated the performance of AQA methods using Spearman's rank correlation coefficient. It is a measure of the performance of our method between a real scoring sequence and a predicted scoring sequence. Spearman's correlation is defined as follows:

$$\rho = \frac{\sum_i (p_i - p)(q_i - q)}{\sqrt{\sum_i (p_i - p)^2 (q_i - q)^2}}. \tag{8}$$

4.2. Implementation Details

Our method was developed on a remote server with an NVIDIA RTX 3080 GPU using the PyTorch framework (https://pytorch.org/). As a feature extractor, we used the I3D model that had already been trained on the Kinetics dataset; it generated a feature vector of 1024 dimensions after receiving 16-frame action sequences as input.

For the MTL-AQA datasets, we extracted 103 frames from each clip, following the approach in previous works [10,29,30], and then divided them into 10 segments, each consisting of 16 consecutive frames. For the JIGSAWS dataset, we adhered to the method in [10] by sampling 160 frames and splitting them into 10 non-overlapping 16-frame segments.

We adopted the Adam optimizer for its combination of momentum and adaptive learning rate strategies, which facilitates efficient and stable convergence. Given the differentiability enabled by the reparameterization trick, the first-order optimization scheme of Adam, coupled with its adaptive learning rate mechanism, facilitates stable and efficient

training of the encoder parameters. The learning rate was set to the weight decay rate of 10^{-4}. For the attention module, the input dimension of the first fully connected layer (FC1) is 1024, with a reduction ratio of 16, input dimension of 64 for the second fully connected layer (FC2). There was a cap of 100 training epochs.

4.3. Results and Analysis

MTL-AQA: the performance of both existing methods and our approach on the MTL-AQA is shown in Table 1. The results without DD information are displayed in the upper part of the table, where our method achieved a predicted correlation coefficient of 0.9269, a notable enhancement compared to the previous method. The bottom half of the table shows that our method still performed well with the use of DD information, with a final correlation coefficient of 0.9478, outperforming the benchmark model MUSDL [10].

Table 1. Comparison of correlation coefficients on MTL-AQA.

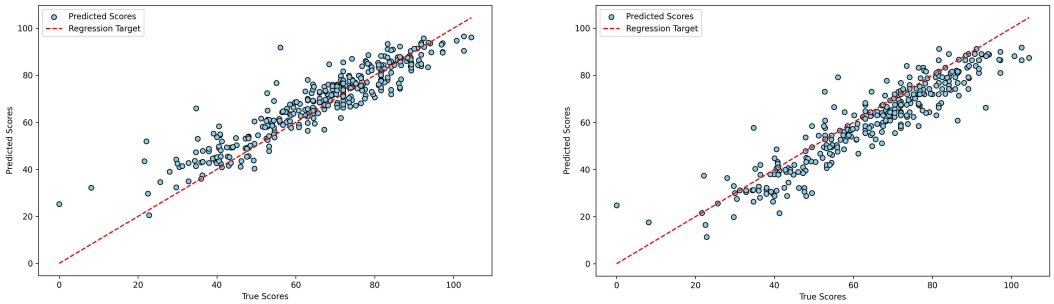
DD	Methods	Sp. Corr.
w/o	C3D-SVR [36]	0.7716
	C3D-LSTM [36]	0.8489
	MSCADC-STL [31]	0.8472
	MSCADC-MTL [31]	0.8612
	C3D-AVG-STL [31]	0.8960
	C3D-AVG-MTL [31]	0.9044
	USDL [10]	0.9066
	MUSDL [10]	0.9158
	I3D + MLP [29]	0.9196
	Ours	0.9269
w/o	USDL [10]	0.9231
	MUSDL [10]	0.9273
	I3D + MLP [29]	0.9381
	Ours	0.9478

JIGSAWS: We conducted experiments on the JIGSAWS surgical activity dataset. We divided each video's 160 frames into ten segments and uniformly sampled them to use as model inputs. The tests on JIGSAWS are shown in Table 2. Our approach performed better than MUSDL in all three surgical videos, with scores of 0.79 (S), 0.75 (NP), 0.79 (KT), and 0.78 (Ave).

Table 2. Comparison of correlation coefficients on JIGSAWS.

Methods	S	NP	KT	Avg. Corr.
ST-GCN [33]	0.31	0.39	0.58	0.43
TSN [36]	0.34	0.23	0.72	0.46
JRG [33]	0.36	0.54	0.75	0.57
USDL [10]	0.64	0.63	0.61	0.63
MUSDL [10]	0.71	0.69	0.71	0.70
I3D + MLP [29]	0.61	0.68	0.66	0.65
Ours	0.79	0.75	0.79	0.78

Visualization: We conducted a detailed comparison between our methods and regression baselines using scatter plots. Figure 3 shows the results of our method compared to MUSDL on the MTL-AQA. The regression target represents the ideal outcome, with scatter points closer to this line indicating superior regression performance. As shown in the figure, the scatter points produced by our methods are more tightly clustered around the target line compared to the regression baseline, reflecting more accurate and consistent predictions.



(**a**) (**b**)

Figure 3. A scatterplot comparing several techniques. Each point on the graph represents a video from the test set. The ideal predictions are indicated by the red line. (**a**) The results of our method on MTL-AQA dataset. (**b**) The results of MUSDL on MTL-AQA dataset.

The training process of the model on the MTL-AQA dataset is illustrated in Figure 4. The model starts to converge after 35 epochs and reaches its optimal performance at epoch 73.

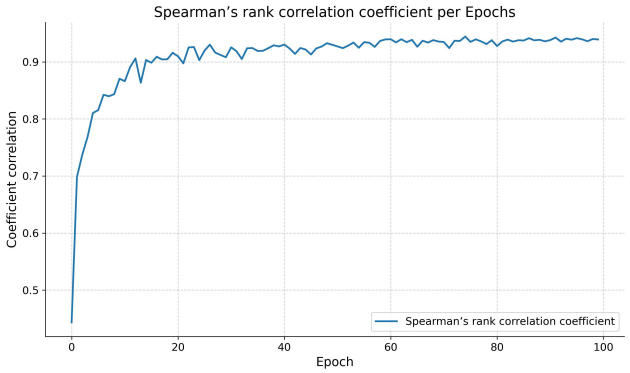


Figure 4. Evolution of Spearman's rank correlation coefficient in the training process.

Ablation experiment: Using the MTL-AQA dataset, we conducted an ablation study to investigate the performance of each module in our model. Both the attention module and the VAE-based score distribution regression method contributed to the model's performance improvement. The effective combination of these two components further enhanced the results, as shown in Table 3. Additionally, Table 4 shows that the inclusion of the attention module can only slightly increase the inference time.

Table 3. Explore the impact of each part on the overall model.

Attention Module	Score Regression	Sp. Corr.
No Attention Module	Direct regression of scores	0.9381
No Attention Module	Score Distribution Regression	0.9423
With Attention Module	Direct regression of scores	0.9441
With Attention Module	Score Distribution Regression	0.9478

Table 4. Comparison results of inference time.

Method	Inference Time (ms)
No Attention Module	1.42
With Attention Module	1.51

5. Conclusions

This paper presents a novel approach to Action Quality Assessment (AQA), incorporating an attention module to improve the model's focus on critical features while suppressing irrelevant information. The proposed attention mechanism assigns adaptive weights to different feature channels, prioritizing those most influential to the predicted scores. In addition, we integrate a Variational Autoencoder (VAE) to model the extracted features, representing latent variables as Gaussian distributions. Through reparameterization and variational inference, the model effectively captures the underlying data structure and accounts for uncertainty in predictions.

Extensive experiments on publicly available benchmarks demonstrate that our method achieves state-of-the-art performance, particularly in terms of Spearman rank correlation. Ablation studies further validate the contribution of each component, underscoring the overall effectiveness of the proposed framework.

In future work, our aim is to extend this approach to the assessment and correction of robot-assisted surgical actions, with potential applications in medical training and performance evaluation.

Author Contributions: Conceptualization, W.G.; methodology, W.L. and H.H.; software, Z.S.; experiment and validation, Z.Z. and J.S.; writing—original draft preparation, W.G.; writing—review and editing, W.L.; supervision, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported in part by Ministry of Education of China (23YJAZH067), in part by Xiamen Municipal Science and Technology Bureau of China (2023CXY0409).

Data Availability Statement: Data in this paper are available from the corresponding authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zeng, L.-A.; Zheng, W.-S. Multimodal Action Quality Assessment. *IEEE Trans. Image Process.* **2024**, *33*, 1600–1613. [CrossRef] [PubMed]
2. Dong, L.-J.; Zhang, H.-B.; Shi, Q.; Lei, Q.; Du, J.; Gao, S.; Learning and fusing multiple hidden sub-stages for action quality assessment. *Knowl.-Based Syst.* **2021**, *229*, 107388. [CrossRef]
3. Zhou, K.; Ma, Y.; Shum, H.P.H.; Liang, X. Hierarchical graph convolutional networks for action quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7749–7763. [CrossRef]
4. Ismail, Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Evaluating surgical skills from kinematic data using convolutional neural networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018, Granada, Spain, 16–20 September 2018; pp. 214–221.
5. Benmansour, M.; Malti, A.; Jannin, P. Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 929–937. [CrossRef] [PubMed]
6. Zia, A.; Sharma, Y.; Bettadapura, V.; Sarin, E.L.; Essa, I. Video and accelerometer-based motion analysis for automated surgical skills assessment. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 443–455. [CrossRef] [PubMed]
7. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
8. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
9. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.

10. Tang, Y.; Ni, Z.; Zhou, J.; Zhang, D.; Lu, J.; Wu, Y.; Zhou, J. Uncertainty-aware score distribution learning for action quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 9839–9848.
11. Kingma, D.P. Auto-encoding variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
12. Gordon, A.S. Automated video assessment of human performance. In Proceedings of the AI-ED, Washington, DC, USA, 16–19 August 1995; Volume 2, pp. 10–20.
13. Ilg, W.; Mezger, J.; Giese, M. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, 10–12 September 2003*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 523–531.
14. Pirsiavash, H.; Vondrick, C.; Torralba, A. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part VI; Springer: Berlin/Heidelberg, Germany, 2014; pp. 556–571.
15. Hinton, G.E.; Osindero, S.; Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
16. Xu, C.; Fu, Y.; Zhang, B.; Chen, Z.; Jiang, Y.-G.; Xue, X. Learning to score figure skating sport videos. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4578–4590. [CrossRef]
17. Li, Y.; Chai, X.; Chen, X. End-to-end learning for action quality assessment. In Proceedings of the Pacific Rim Conference on Multimedia, Hefei, China, 21–22 September 2018; pp. 125–134.
18. Doughty, H.; Mayol-Cuevas, W.; Damen, D. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7862–7871.
19. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
20. Wang, S.; Yang, D.; Zhai, P.; Chen, C.; Zhang, L. Tsa-net: Tube self-attention network for action quality assessment. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 4902–4910.
21. Lei, Q.; Zhang, H.; Du, J. Temporal attention learning for action quality assessment in sports video. *Signal Image Video Process.* **2021**, *15*, 1575–1583. [CrossRef]
22. Neal, R.M. *Bayesian Learning for Neural Networks*; Springer Science & Business Media: New York, NY, USA, 2012; Volume 118.
23. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
24. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33nd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
25. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 13969–13980.
26. Masci, J.; Meier, U.; Cireşan, D.; Schmidhuber, J. Stacked convolutional auto-encoders for hierarchical feature extraction. In Proceedings of the International Conference on Artificial Neural Networks (ICANN), Espoo, Finland, 14–17 June 2011; pp. 52–59.
27. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning (ICML), Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
28. Zhou, C.; Huang, Y.; Ling, H. Uncertainty-driven action quality assessment. *arXiv* **2022**, arXiv:2207.14513.
29. Yu, X.; Rao, Y.; Zhao, W.; Lu, J.; Zhou, J. Group-aware contrastive regression for action quality assessment. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 7919–7928.
30. Parmar, P.; Morris, B. Action quality assessment across multiple actions. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1468–1476.
31. Parmar, P.; Morris, B.T. What and how well you performed? A multitask learning approach to action quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 304–313.
32. Gao, Y.; Vedula, S.S.; Reiley, C.E.; Ahmidi, N.; Varadarajan, B.; Lin, H.C.; Tao, L.; Zappella, L.; Yuh, D.D.; Chen, C.C.G.; et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A surgical activity dataset for human motion modeling. In Proceedings of the MICCAI Workshop: M2CAI, Boston, MA, USA, 14 September 2014; Volume 3, pp. 3–10.
33. Pan, J.H.; Gao, J.; Zheng, W.S. Action assessment by joint relation graphs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6331–6340.
34. Jain, H.; Harit, G.; Sharma, A. Action quality assessment using Siamese network-based deep metric learning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2260–2273. [CrossRef]

35. Zhang, B.; Chen, J.; Xu, Y.; Zhang, H.; Yang, X.; Geng, X. Auto-encoding score distribution regression for action quality assessment. *Neural Comput. Appl.* **2024**, *36*, 929–942. [CrossRef]
36. Parmar, P.; Morris, B.T. Learning to score Olympic events. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 20–28.