systems

MDPI

Article

MLA-LSTM: A Local and Global Location Attention LSTM Learning Model for Scoring Figure Skating

Chaoyu Han [1], Fangyao Shen [2], Lina Chen [1,*], Xiaoyi Lian [1], Hongjie Gou [3] and Hong Gao [3]

[1] Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua 321004, China
[2] Mathematics and Computer Science, Zhejiang Normal University, Jinhua 321004, China
[3] Computer Science and Technology, Harbin Institute of Technology, Harbin 150006, China
* Correspondence: chenlina@zjnu.cn

Abstract: Video-based scoring using neural networks is a very important means for evaluating many sports, especially figure skating. Although many methods for evaluating action quality have been proposed, there is no uniform conclusion on the best feature extractor and clip length for the existing methods. Furthermore, during the feature aggregation stage, these methods cannot accurately locate the target information. To address these tasks, firstly, we systematically compare the effects of the figure skating model with three different feature extractors (C3D, I3D, R3D) and four different segment lengths (5, 8, 16, 32). Secondly, we propose a Multi-Scale Location Attention Module (MS-LAM) to capture the location information of athletes in different video frames. Finally, we present a novel Multi-scale Location Attentive Long Short-Term Memory (MLA-LSTM), which can efficiently learn local and global sequence information in each video. In addition, our proposed model has been validated on the Fis-V and MIT-Skate datasets. The experimental results show that I3D and 32 frames per second are the best feature extractor and clip length for video scoring tasks. In addition, our model outperforms the current state-of-the-art method hybrid dynAmic-statiC conText-aware attentION NETwork (ACTION-NET), especially on MIT-Skate (by 0.069 on Spearman's rank correlation). In addition, it achieves average improvements of 0.059 on Fis-V compared with Multi-scale convolutional skip Self-attentive LSTM Module (MS-LSTM). It demonstrates the effectiveness of our models in learning to score figure skating videos.

Keywords: scoring figure skating; feature extraction; feature attention fusion; multi-scale location attention module

Citation: Han, C.; Shen, F.; Chen, L.; Lian, X.; Gou, H.; Gao, H. MLA-LSTM: A Local and Global Location Attention LSTM Learning Model for Scoring Figure Skating. *Systems* **2023**, *11*, 21. https://doi.org/10.3390/systems11010021

Academic Editors: Shixuan Fu, Bo Yang and Alex Zarifis

Received: 22 November 2022
Revised: 28 December 2022
Accepted: 28 December 2022
Published: 2 January 2023

1. Introduction

Nowadays, how to objectively assess and regularize the action of athletes has attracted more and more attention from worldwide sports committees. As a supplementary tool, action quality assessment (AQA) is developed to assess the performance of an athlete and provide detailed feedback to improve his/her action quality. It is widely used in various sport events, such as diving [1–3], rhythmic gymnastics [4,5], and basketball [6]. Among these sports, figure skating usually consists of long-term actions (average 2 min and 50 s) which contain richer and more complex information than others. As a result, assessing the action quality of figure skating is difficult, and few works have been proposed to score figure skating.

Figure 1 depicts a popular figure skating scoring system that consists primarily of four stages. First, figure skating videos are gathered from sports events. Second, each video is split into several clips, and clip-level features are extracted. Third, an aggregation module is proposed to fuse all clip-level features into video-level features. Finally, a regression model will predict figure skating scores based on video-level features. The second and third stages are the primary focus of this paper.

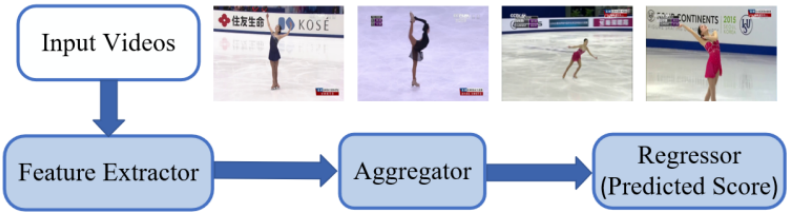

Figure 1. The flow chart of the figure skating scoring system.

Designing a video feature extraction method is important because the quality of features largely determines the performance of models. To learn Spatio-temporal features for videos, Tran et al. [7] designed a convolutional 3D (C3D) model to extract features from 16-frame video clips. C3D could model appearance and motion information simultaneously, outperforming 2D convolutional features on video analysis tasks. Carreira et al. [8] introduced a Two-Stream Inflated 3D ConvNet (I3D) to learn video features. Their study proved that I3D outperformed previous methods in action classification. Similarly, Hara et al. [9] proposed a ResNets-based 3D convolutional neural network (R3D) for improved action representation and discovered that it outperformed relatively shallow networks. Although many feature extractors have been proposed in recent years, a problem still exists. On the one hand, all these methods are compared on the video classification task, while it is unknown whether the effect is good or bad on the video scoring task. On the other hand, people often use different feature extraction methods and clip lengths to compare model effects, which is very unfair. Therefore, it is critical to select the best feature extraction method and clip length for scoring figure skating videos consistently.

In the aggregation stage, many previous studies [1,3,10–12] obtained video-level features by averaging clip-level features, which did not reflect the different importance of the clip. To solve this problem, Xu et al. [13] proposed a deep model that includes two parallel LSTM modules, to efficiently learn the local and global sequential information in each video. To extract more robust stream features, Zeng et al. [4] proposed a context-aware attention module. In this module, the temporal instance-wise graph convolutional network unit is used to explore the relations between instances, and the attention unit is designed to assign a proper weight to each instance. However, these works only focused on dynamic information, which mainly reflects the category of action, while ignoring the location information, which represents the coordinates of specific behavior at a given time. Location information is critical for figure skating scoring, especially in video frames with dark backgrounds. For example, as shown in Figure 2a, video frames with bright backgrounds show the full body of the skaters. Therefore, models can successfully recognize action categories and assess action qualities. However, as shown in Figure 2b, for video frames with dark backgrounds, it is difficult for models to distinguish the body parts apart from the background. The model will be unable to recognize action categories, resulting in the loss of some key clip-level features. Thus, locating the coordinates of athletes is important during the aggregation stage.

(a) Success case (b) Failed case

Figure 2. Examples of video frames in the dataset.

To address the above issues, we propose a novel model, named Multi-scale Location Attentive Long Short-Term Memory (MLA-LSTM), to effectively assess the action qualities of figure skaters from videos. First, to determine the best feature extractor and the best

video clip length, we compare the performance of three pre-trained feature extractors (C3D, I3D, and R3D) and four clip lengths (5, 8, 16, and 32) on the two public figure skating datasets (MIT-skate [10] and Fis-v [13]). Second, to identify and locate athletes with similar backgrounds, we propose a Multi-Scale Location Attention Module (MS-LAM), which can capture the location information of athletes in different video frames. Third, a Multi-scale convolutional skip Self-attentive LSTM Module (MS-LSTM [13]) is used to fuse features from different clips. It learns spatial sequence information at multiple scales and selects important clips by a spatial attention strategy to fuse features. Finally, linear regression is used to predict the action quality scores from fused features. Extensive experimental results show that our model achieves Spearman's rank correlation of 0.684 (0.069 higher than the current state-of-the-art ACTION-NET [4]) on the MIT-Skate for feature extraction using an I3D feature extractor, and achieves average Spearman's rank correlation of 0.765 (0.05 improvement over state-of-the-art MS-LSTM [13]) on the Fis-V for feature extraction using an I3D feature extractor.

The contributions of our work can be concluded as two-fold:

(1) Since the most suitable feature extraction method and sampling clip lengths for the video scoring task are not yet conclusive, in this paper, a systematical comparison between different feature extraction and different clip lengths for the performance of figure skating scoring has been conducted. It will provide an effective reference for the research of feature extraction methods of long videos in the future;
(2) To accurately identify the frames, we proposed a novel position aggregation network MS-LAM to capture the position information of athletes, as much as possible without losing athlete clip-level features. The location information of figure skaters is calculated automatically, and it cannot be affected by the size of skaters in the background.

2. Related Work

2.1. Video Understanding

With the explosion of online video data, video understanding plays an important role in the field of computer vision. Although traditional 2D convolutional neural networks (CNNs) based feature extractors are computationally inexpensive, they cannot capture the long-term temporal relationships in video. To solve this problem, most studies extracted temporal and spatial features by two-stream CNNs (RGB and Optical Flow) or 3D CNNs. However, the number of network model parameters of these methods is larger than 2D CNNs, and their computational efficiency is lower than 2D CNNs. They usually require large-scale datasets for training.

To improve computational efficiency, Feichtenhofer et al. [14] proposed a two-path SlowFast network model for action recognition, where the slow path captures spatial semantic information at a lower frame and slower refresh rate, and the fast path learns temporal semantic information at a fast refresh rate and high temporal resolution; Wang et al. [15] proposed Temporal Difference Networks to efficiently recognize action. They explicitly captured short-term and long-term action information. Specifically, for short-term information, temporal difference over consecutive frames is utilized to supply 2D CNNs with finer action patterns. For long-term information, temporal difference across segments is incorporated to capture long-term structure for action feature excitation.

C3D [7] stacked multiple 3D convolutional kernels with the size of $3 \times 3 \times 3$ to form a model structure, which was similar to the VGG-16 [16] network. It showed strong generalization capability and better feature representation than 2D CNNs for feature extraction in a variety of scenes. Carreira et al. [8] proposed an I3D model, which used RGB images and stacked optical streams as input and finally obtained the fusion of the two-stream output results. Tran et al. [17] deconstructed the Spatio-temporal 3D convolution kernels into two 3D convolutional kernels with the size of $1 \times 3 \times 3$ and $3 \times 1 \times 1$, which further reduced the number of parameters and improved the computational efficiency. Hara et al. [9] proposed R3D networks of different depths and used residual connections

for experimental verification. The results proved that their networks have strong fitting degrees and generalization ability.

Although there are many feature extractors that have been proposed, the performance of them only been validated and analyzed in action recognition tasks. As for the action scoring the task, there is no unified conclusion as to which kind of feature extractors is most suitable for long-term spatial-temporal feature learning. In this paper, we mainly solve this issue. We first extract features from videos by C3D, I3D, and R3D, respectively. Then, we feed them into MLA-LSTM and ACTION-NET [4] models and analyze the performance of these feature extractors, which attempts to obtain the most applicable feature extraction method for long-term videos. In addition, we analyze the results in sampling steps of 5, 8, 16, and 32 frame clips, and find that the applicability of the feature extraction methods varies in different networks. It is more beneficial to handle longer frame clips.

2.2. Feature Fusion

In the long video action recognition task, the attention mechanism is an important tool to improve detection performance. It enables networks to recognize important parts of the video and locate the region of the interested object. It is widely used in object localization [18] and small object detection [19]. It enables networks to recognize important parts of the video and locate the region of object of interest. However, in current research [1,2,10,11], most of them fused video-level features with a simple concatenate or averaging pooling operation, which treated actions in different clips as equally important, without considering further processing and optimization of the features.

With the rapid development of CNNs, different attention mechanism modules have been proposed. Hu et al. [20] proposed the Squeeze-and-Excitation Networks (SENet) to integrate global spatial information into channel information to capture channel-related information across the network space. Doughty et al. [21] proposed a learnable rank-aware temporal attention module applicable to a mass task, which was used to extract the parts of the video that were more relevant to the skill, and the attention module focused on the skills with higher (pros) and lower (cons) representation of the video part, respectively. Nakano et al. [22] detected highlight moment clips by recording the blinking frequency of people when watching the video. After that, based on these, they determined the importance of athletes' actions in the whole sports video. Lei et al. [23] improved the relevance of prediction scores by weighted fusion of spatio-temporal features of sports videos.

Since these works have not considered scale variation and background similarity problems, the Multi-Scale Location Attention Module (MS-LAM) incorporated in the early video feature fusion stage can effectively fuse local and global sequence information in features and can simply and effectively solve the feature scale inconsistency and background similarity problem. By aggregating the multi-scale context information along the channel dimension, MS-LAM can simultaneously emphasize large objects that distribute more globally and highlight small objects that distribute more locally, facilitating the network to recognize and detect video features of the athlete's scale changes when the camera is constantly switched.

3. Approach

In this section, we introduce our proposed MLA-LSTM in detail. We first describe the video feature representation in Section 3.1. In Section 3.2, we introduce multi-scale location attention module. Finally, we present the overall framework of the MLA-LSTM in Section 3.3. The overall framework of the MLA-LSTM is shown in Figure 3.

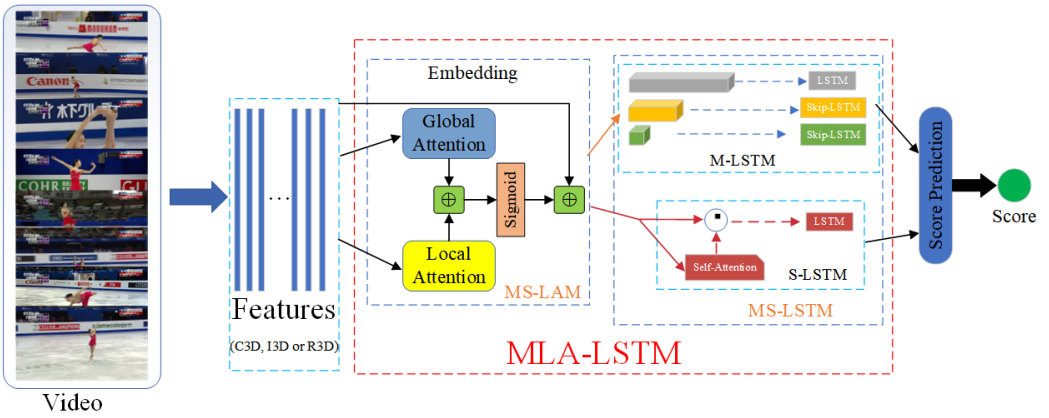

Figure 3. Our Approach: First, use pretrained different feature extractors to extract feature sequences from the video. The feature extractor can be C3D, I3D, or R3D, and the sampling clip lengths can be 5 (5 frames per clip), 8 (8 frames per clip), 16 (16 frames per clip), or 32 (32 frames per clip). Then, through the Multi-scale Location Attention Module (MS-LAM) embedding the feature, and then enter two independent modules, Multi-scale Convolutional Skip LSTM (M-LSTM) and Self-attentive LSTM (S-LSTM) finally aggregate the feature sequence and used to predict score.

3.1. Video Features

We first divide the video into non-overlapping clips, and then extract video features from each clip. In different literature, the clip length is different. For example, in the literature [4,13,24], authors set the clip length 5, 16, and 32 frames, respectively. However, the optimal clip length has not been determined. Therefore, in this paper, we set the clip length as 5, 8, 16, and 32 frames, and study the effect of different clip lengths on the scoring of figure skating videos, so as to obtain the best clip length.

Next, we extract features from each clip video. There are three main feature extraction methods (C3D [7], I3D [8], R3D [9]). However, there is no unified conclusion on which feature extraction method is the most effective for video scoring tasks, so we used these three methods for video feature extraction separately and compared their performance on two publicly available datasets (MIT-Skate and Fis-V), and finally determined the best feature extraction method based on the performance.

3.2. Multi-Scale Location Attention Module (MS-LAM)

A long video is approximately 4500 frames (2 min and 50 s), and it is a challenge to predict the score for such a long video. Meticulous processing and calculation of the features of each frame can lead to high computational costs and inefficiencies. Since not all moments in a long video are useful for predicting scoring, we propose MS-LAM to selectively learn the video feature representation. Figure 4 shows more details of it. Suppose we have a video feature $X \in R^{C \times L}$ with C channels and feature sequence of size L, the channels' attention weights in SENET [20] can be computed as

$$w = \sigma(\beta(W_2(\delta(\beta(W_1(g(X))))))) \quad (1)$$

where $g(X) \in R^C$ denotes the global feature context and $g(X) = \frac{1}{L}\sum_{i=1}^{L} X_{[:,i]}$ is the global average pooling. σ is the Sigmoid function. δ denotes the Rectified Linear Unit (ReLU), and β denotes the Batch Normalization (BN) [25], where $W_1 \in R^{\frac{C}{r} \times C}$ is a dimension reduction layer, and $W_2 \in R^{C \times \frac{C}{r}}$ is dimension increasing layer. r is the channel reduction ratio.

After the operation, we can only obtain the attention of the channel with a global feature. However, when capturing the action of figure skaters, the camera is usually not fixed; instead, it is switched between different cameras to select the viewpoint that best characterizes the athlete's action, so small targets will inevitably appear in the screen.

Therefore, we try to aggregate multi-scale contextual features in the attention module to alleviate the size variation and small-target problems.

The realization of channel attention at multiple scales is mainly achieved by changing the size of the spatial pool. We choose the 1D point-wise convolution (PWConv) as the context aggregator for local channels and only perform point-by-point interactions for each spatial location of the channel. Eventually, the local channel context is computed by a bottleneck structure $L(X) \in R^{C \times L}$ as follows:

$$L(X) = \beta(PWConv_2(\delta(\beta(PWConv_1(X))))) \tag{2}$$

The kernel size of $PWConv_1$ and $PWConv_2$ are $\frac{C}{r} \times C \times 1$ and $C \times \frac{C}{r} \times 1$, respectively. It is inspiring that the global context g(X) and the local context L(X) have the same shape as the input feature, the refined feature $X^{'} \in R^{C \times L}$ can be obtained from the existing and, namely MS-LAM calculated as follows:

$$X^{'} = X \otimes M(X) = X \otimes \delta(L(X) \oplus g(X)) \tag{3}$$

where $M(X) \in R^{C \times L}$ denotes the attentional weights generated by MS-LAM. $\oplus$ denotes the broadcasting addition and $\otimes$ denotes element-wise multiplication.

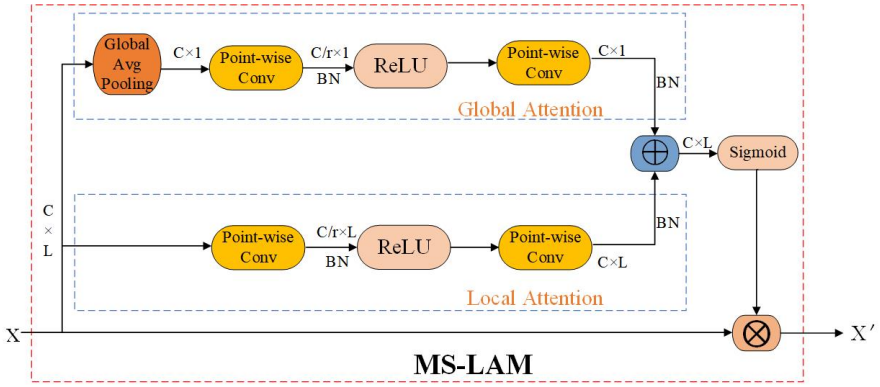

Figure 4. Multi-Scale Location Attention Module (MS-LAM) architecture. The upper branch represents Global Attention, and the lower branch represents Local Attention.

3.3. MLA-LSTM

The overall structure diagram of our proposed MLA-LSTM is shown in Figure 3. MS-LSTM mainly includes Self-Attentive LSTM (S-LSTM) and Multi-scale Convolutional Skip LSTM (M-LSTM). The S-LSTM employs a simple self-attentive strategy to select important clip features which are directly used for regression tasks. The M-LSTM models the local and global sequential information at multi-scale, and utilizes the skip LSTM to efficiently save total computational cost. We add MS-LAM to the early feature fusion stage, taking the I3D feature extractor as an example. First, we extract 1024-dimensional clip-level features of the avg-pool layer of I3D pretrain on Kinetics 400 from the whole video. Second, we input the feature sequences into two independent local channel attention and global channel attention. The local channel attention first passes through a 1D point-wise convolution with a convolution kernel size of 1×1024 output channel of 128, a batch normalization with 128 features and a ReLU activation function, and then a 1D point-wise convolution with a convolution kernel size of 1×128 output channels of 1024 and a batch normalization with a feature number of 1024. The global channel attention is based on the local channel attention, an adaptive average pooling operation with an output size of 1 is added to the front, so that all subsequent input sizes are 1. The global attention and the local attention do broadcasting addition operations and element-wise multiplication with the original feature sequence by sigmoid activation function. After that, the early fusion features are input to the MS-LSTM model for regression prediction score.

4. Experiments

4.1. Settings and Evaluation

4.1.1. Datasets

We evaluate our method on two public datasets, MIT-Skate [10] and Fis-V [13]. MIT-Skate dataset is collected from international figure skating events, which contains 171 competition videos of both male and female athletes. Each video has a frame rate of 24 and lasts about 2 min and 50 s. The final score of each video is labeled by professional judges, which ranges from 0 (worst) to 100 (best). There are 120 videos for training and the remaining 51 videos for testing. We repeat the experiment 500 times using different random data splits and cycling through five epochs to count as a complete experiment.

The FIS-V dataset has 500 competition videos of women's singles short program of figure skating; each video lasts about 2 min and 50 s with a frame rate of 25. Each video is annotated with three scores, Total Element Score (TES), Total Program Component Score (PCS), and final score. The TES score is used to determine the difficulty and execution of all technical actions. The PCS score assesses the skater's performance and interpretation of the music. The final score is the sum of the TES and PCS and subtracts the penalty score. Following [13], 400 videos are used for training and the remaining 100 videos for testing.

4.1.2. Evaluation Metric

Following previous works [3,4,10,11,13], we utilized Spearman's rank correlation to evaluate the method. Spearman's rank correlation is used to measure the degree of correlation between predicted series and ground-truth series. It ranges from -1 to $+1$, and the higher the value is, the higher the rank correlation between these two series. It is calculated as follows:

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - x)^2 + \sum_i (y_i - y)^2}} \tag{4}$$

where ρ is Spearman's rank correlation, x and y represent the ranking of two series, respectively. Additionally, we evaluated methods with Mean Square Error (MSE).

4.1.3. Experimental Settings

To better compare with the data from previous works, we used the same data settings for training and testing. More specifically, for the feature extraction stage, we sampled 16 frames per second on average, that is, sampling a continuous clip containing 16 frames per second with a sliding window. All frames are scaled to the shortest side length of 256, then the center is cropped to 224 $\times$ 224, and we augment the video by flipping the frames horizontally. We then extracted 4096-dimensional clip-level features from the fc6 layer of C3D [7] pretrained on Sports-1M [26]; 1024-dimensional clip-level features of the avg-pool layer of I3D [8] pretrained on Kinetics 400 [27] and 2048-dimensional clip-level features of the avg-pool layer of R3D [9] pretrained on Kinetics 400 [27].

In the multi-scale channel attention module, we set different channels and r for different feature extraction networks; the channels and r of I3D, R3D and C3D are set to 1024, 8; 2048, 16 and 4096, 32 respectively. We build the model by Pytorch and the entire framework is trained on a 1 NVIDIA 1080Ti GPU card, which can achieve 100 epochs of convergence. The specific Loss and Spearman's rank correlation curves are shown in Figure 5. It takes about 30 min to train a model. Our model is an end-to-end network, so we can directly input the extracted feature sequence into the model for training and predicting scores.

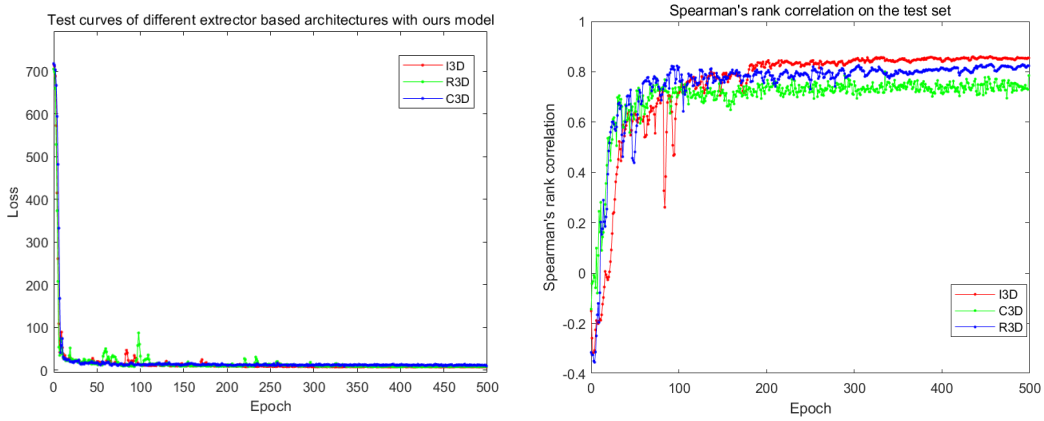

Figure 5. Test curves of PCS scores obtained by training on the Fis-V dataset using C3D, I3D, and R3D as feature extractors. The plot on the left shows the test loss obtained using our model as aggregated various feature extractors. The plot on the right contains Spearman's rank correlation on the test set.

4.1.4. Competitors

We demonstrate and compare previous works [3,4,10,11,13] to validate the effectiveness of our proposed method. Since few methods have been evaluated on the Fis-V dataset, we replicated the algorithm of [4] and evaluated it on this dataset.

- Input features: We use clip-level features as input. To verify the effectiveness of different feature extraction methods, we extract them as described in Section 3.1.
- We use max and average pooling for video features embedding representation and using linear kernel SVR to regress the prediction scores.
- We emulate the C3D-LSTM architecture used in Parmar et al. [11] to directly generate video descriptions, and we also use the bi-directional LSTM (Bi-LSTM) system to maintain fairness issues that may arise due to the long video sequences. In addition, the hidden layer size in LSTM/Bi-LSTM is set to 256/128, using the same regression as our model. In addition, the method is extended to I3D and R3D features.

4.2. Results

4.2.1. Results of Different Variants

As the regression tasks, we further explore different variants in Table 1, we compare different pooling and regression methods on C3D, R3D and I3D features, respectively.

(1) C3D Vs. I3D Vs. R3D. After comparing all methods, we can conclude that I3D performs well on datasets of different sizes and resolutions, and its experimental results are significantly better than the other two feature methods. The reason for this result is that I3D can extract spatial information while selecting the correct kernel size to extract local and global distribution information, which makes I3D display richer feature information than other features.

(2) Max vs. Avg pooling. Because linear SVR has been shown to be better in [13], we use linear regression for both pooling methods. Actually, we cannot give a uniform result on which pooling method is better. On both datasets, the maximum pooling is better on C3D and R3D features, while the average pooling is better on I3D features. This demonstrates the inherent difficulty of the video feature regression tasks.

(3) Different variants of LSTM. In general, we found that the Bi-LSTM has better performance than the LSTM on various features and datasets. However, results of all these LSTM methods are lower than our framework.

Table 1. Results of the Spearman's rank correlation (the higher the better) on the MIT-Skate and Fis-V. Avg. is the average Spearman's rank correlation across all classes computed using Fisher's z-value.

		MIT-Skate	Fis-V		
			TES	PCS	Avg.
Pose + DCT [10]		0.350 **	-	-	-
C3D + LSTM [11]		0.530 **	-	-	-
ConvISA [12]		0.450 **	-	-	-
MSE + Ranking Loss [3]		0.575 **	-	-	-
MS-LSTM [13]		0.590 **	0.650 **	0.780 **	0.715 **
ACTION-NET [4]		0.615 **	0.580 *	0.794 *	0.697
C3D	Max + SVR	0.480	0.470	0.610	0.540
	Avg + SVR	0.420	0.400	0.590	0.495
	LSTM	0.370	0.590	0.770	0.680
	Bi-LSTM	0.580	0.560	0.730	0.645
	MLA-LSTM (ours)	0.616	0.604	0.808	0.706
I3D	Max + SVR	0.442	0.547	0.695	0.621
	Avg + SVR	0.531	0.558	0.703	0.631
	LSTM	0.472	0.520	0.742	0.631
	Bi-LSTM	0.587	0.629	0.705	0.667
	MS-LSTM	0.628	0.656	0.809	0.733
	MLA-LSTM (ours)	**0.684**	**0.673**	**0.857**	**0.765**
R3D	Max + SVR	0.560	0.557	0.669	0.613
	Avg + SVR	0.526	0.482	0.606	0.544
	LSTM	0.552	0.537	0.765	0.651
	Bi-LSTM	0.593	0.554	0.745	0.650
	MS-LSTM	0.621	0.648	0.808	0.728
	MLA-LSTM (ours)	0.652	0.640	0.830	0.735

The symbol "*" indicates that only the ACTION-NET dynamic stream evaluation is used; The symbol "**" indicates statistic significance (paired t-test, $p < 0.01$) of performance improvement of the proposed method (MLA-LSTM) in comparison with other methods.

4.2.2. Results of the Spearman's Rank Correlation

In Table 1, we report the results on the MIT-Skate and Fis-V datasets, and it is evident from the results that our method obtains the best performance on both datasets and significantly outperforms the baseline of the previous study (including [3,4,10–13]). Our model achieves an average Spearman's rank correlation of 0.765 (0.050 improvements over state-of-the-art MS-LSTM) on the Fis-V for feature extraction using the I3D feature extractor. This demonstrates the effectiveness of our proposed method. TES and PCS measure the performance of figure skaters from different aspects. Since PCS scoring is highly subjective in judgment, it considers how well the participants' movements fit the background music, yet the music cannot be measured as good or bad. Our model improves the prediction ability by extracting more useful location information (the best result is 7.7% better than MS-LSTM). In contrast, in TES scoring, deep 3D feature extraction adds redundant information and therefore has a negative effect on the scoring results, so our model shows the best prediction results using I3D extracted features (the best result is 2.3% better than MS-LSTM). On the MIT-Skate dataset, we can find that I3D extracting feature information for scoring is the best result, where our MS-LAM can optimize the channel information and add important location information well (6.9% improvements compared to state-of-the-art ACTION-NET).

To illustrate the statistical significance between our proposed MLA-LSTM method and the other methods (ACTION-NET, MS-LSTM, MSE + Ranking Loss, ConvISA, C3D + LSTM and Pose + DCT), we perform the paired t-test on their classification results. The hypothesis is that "the classification performance of MLA-LSTM is greater than that of the other

methods". Each test is run on the two sequences of the classification results obtained by MLA-LSTM and the given method. The statistical test results are represented by the symbol "**", which means that the hypothesis is correct with probability 0.99. For example, on the MIT-Skate dataset, the Spearman's rank correlation of MLA-LSTM is 0.684 and that of ACTION-NET is 0.615. The appended "**" means the hypothesis of MLA-LSTM is superior to ACTION-NET is true based on the statistical test. In summary, from Table 1, we conclude that MLA-LSTM achieves better performance than the other compared methods on both datasets.

4.2.3. Results of the Mean Squared Error

We compared the results for the MSE metric on the Fis-V dataset, which we also compared in Table 2. In particular, we found that the improved model tested on I3D far outperforms all other baselines. In addition, we also observe that our model has greatly improved the scoring of PCS performance for all features but does not produce significant improvements in the scoring results for TES performance. This is conceivable because PCS is easier to understand and interpret relative to TES scoring. Since the TES task is for technical action scoring, this requires a more fine-grained labeling study to remove redundant information and thus produce more accurate determinations.

Table 2. Results of the MSE (the lower the better) on Fis-V.

		TES	**PCS**
MS-LSTM [13]		19.91	8.35
ACTION-NET [4]		26.35 *	7.62 *
C3D	Max + SVR	27.42	13.98
	Avg + SVR	30.25	15.96
	LSTM	22.96	8.70
	Bi-LSTM	23.80	10.36
	MLA-LSTM	24.09	9.17
I3D	Max + SVR	24.55	13.02
	Avg + SVR	25.46	12.78
	LSTM	30.626	14.62
	Bi-LSTM	24.99	13.54
	MS-LSTM	23.87	10.70
	MLA-LSTM	**19.07**	**6.63**
R3D	Max + SVR	28.06	14.25
	Avg + SVR	31.22	16.82
	LSTM	30.01	12.18
	Bi-LSTM	27.51	12.55
	MS-LSTM	24.05	9.32
	MLA-LSTM	21.46	8.62

The symbol "*" indicates that only ACTION-NET dynamic stream evaluation is used.

4.2.4. Impact of Clip Length

We experimentally verify the impact of clip length (the number of frames) on model performance. Due to the lack of resources and open-source pretraining weights, we only use I3D to extract features with different sampling steps on the MIT-Skate dataset and perform comparison experiments in two different models. Table 3 shows the results, and it is clear that the performance increases as the number of frames in each clip increases. This indicates that, for long action videos longer clips, using I3D can extract a wider range of information in the temporal dimension and more fully characterize the continuity of the action.

Table 3. The Spearman's rank correlation (the higher the better) of different sampling steps using I3D on the MIT-Skate.

Clip Length\Model	ACTION-NET *	MLA-LSTM
5	0.603	0.612
8	0.606	0.625
16	0.612	0.654
32	0.630	0.671

The symbol "*" indicates that only ACTION-NET dynamic stream evaluation is used.

4.2.5. Ablation Study on Multi-Scale Location Attention Strategy

We apply GradCAM [28] to ResNet-50, SENet-50, and MS-LAM-ResNet-50 for the visualization results of images from the Fis-V dataset. Figure 6 shows that the regions recognized by the MS-LAM-ResNet-50 are highly overlapping with the skaters, which indicates that it learns object locations and utilizes features in the regions well. On the contrary, ResNet-50 has poor localization ability and is unable to target in many cases. Although SENet-50 was able to localize some of the location information, the locked region was too large and contained too much background information. This is because SENet-50 only utilizes global channel attention, which is biased toward the global range of contextual information. However, our proposed MS-LAM also aggregates local channel contextual information, which helps the network to identify small targets and character locations in contextually similar situations.

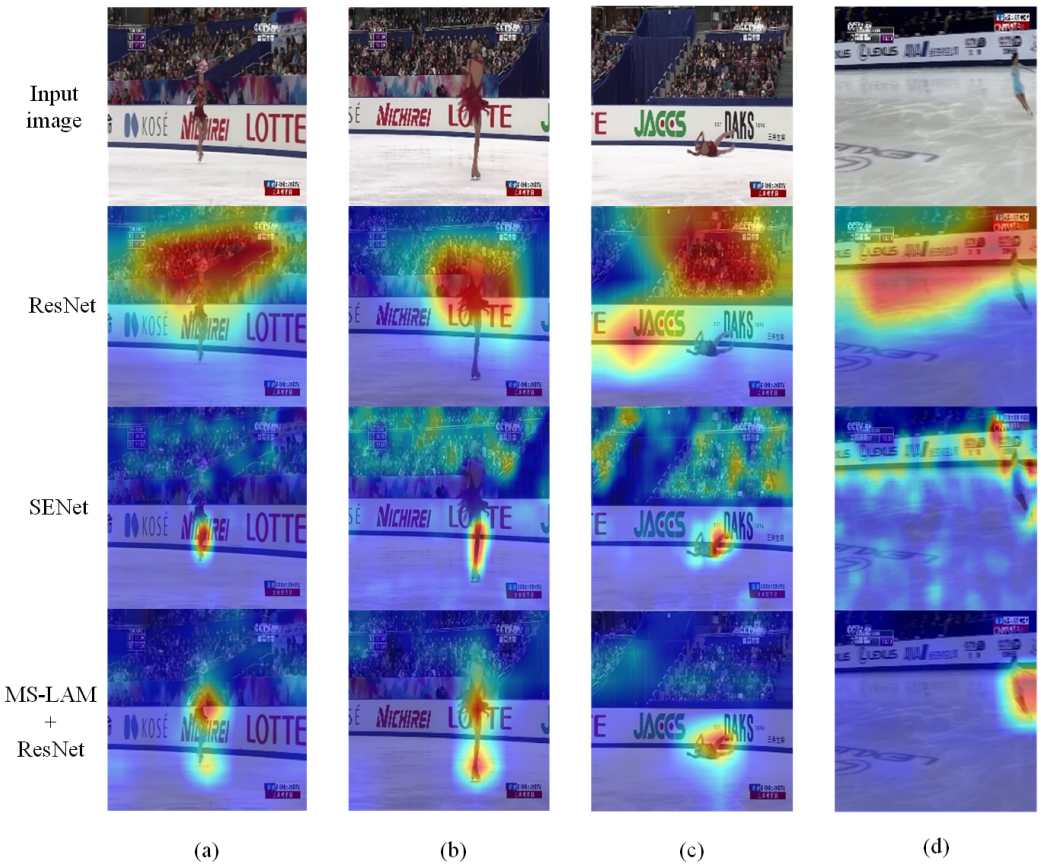

Figure 6. Network visualization with Grad-CAM. Our method can accurately locate the target; specifically, (**a**,**b**) show mainly the localization ability of the network in the case of similarity between the person and the background. (**c**,**d**) show that for small target detection. The comparison results show that the proposed MS-LAM is beneficial to blurred background recognition and object localization.

4.2.6. Ablation Study on Attention Strategy

We also conducted further ablation studies to explore the contribution of the respective attention components of MS-LAM and S-LSTM. As shown by the results in Table 4, we use I3D as an extractor to extract features, and it works best on two datasets. Furthermore, we designed to directly add MS-LAM or double-layer MS-LAM Iterative Attentional Feature Fusion (IAFF) in the early fusion stage, as well as remove the S-LSTM module, named LA-M-LSTM and IAFF-M-LSTM, and show the results in Tables 4 and 5. We find that the strategy of directly adding MS-LAM in the early fusion stage can better help the model capture useful location information from the rich frames, which in turn improves the model performance for more accurate prediction capability. However, the poor performance of IAFF results from the loss of feature information due to excessive convolution.

Table 4. Ablation study of the MS-LAM model. We show the results of the Spearman's rank correlation (the higher the better) and MSE (the lower the better) on Fis-V. We use I3D features.

Model	Correlation		MSE	
	TES	**PCS**	**TES**	**PCS**
MS-LSTM	0.626	0.809	23.87	10.70
MLA-LSTM	**0.673**	**0.857**	**22.07**	**6.63**
LA-M-LSTM	0.646	0.824	24.86	9.39
IAFF-MS-LSTM	0.586	0.807	27.13	12.62
IAFF-M-LSTM	0.601	0.803	26.35	12.10

Table 5. Ablation study of the MS-LAM model. We show the results of the Spearman's rank correlation (the higher the better) and MSE (the lower the better) on MIT-Skate. We use I3D features.

Model	Correlation	MSE
MS-LSTM	0.618	127.42
MLA-LSTM	**0.684**	**112.46**
LA-M-LSTM	0.625	124.84
IAFF-MS-LSTM	0.660	118.66
IAFF-M-LSTM	0.642	120.81

5. Discussion

In this paper, we argue that evaluating the quality of action requires a combination of the most fitting feature extractor and sampling clip length based on locking the athlete's position information. To this end, we propose a modular attentional network that adequately aggregates local and global semantic information of position features, and evaluates them over dynamic features extracted with sampling length of 32 and I3D feature extractor. Our experimental evaluation demonstrates that our method achieves state-of-the-art results on long-term action assessment in comparison to prior works. In addition, compared to MS-LSTM, our method can precisely locate the athlete position, which allows for better localization scoring in certain background similar situations. Compared to ACTION-NET, our method extracts information and captures athlete positions in a continuous dynamic stream, and the richness of information is significantly better than the feature information extracted on some specific frames.

6. Conclusions

Summarizing, in the present work, we have proposed a Multi-scale Location Attentive Long Short-Term Memory (MLA-LSTM) learning model for scoring figure skating. The Multi-Scale Location Attentive Module (MS-LAM), which can capture the location information of athletes in different video frames, is able to provide useful spatial information for subsequent tasks. We also summarize the best feature extraction method and clip length

for the video scoring task, which are feature extractor I3D and sampling 32 frames per second, respectively. Experiments on the MIT-Skate and Fis-V datasets demonstrate that the proposed MLA-LSTM model can effectively learn video scoring, and it can achieve state-of-the-art performance on two datasets in terms of Spearman's rank correlation.

Future research can be focused on the real-time design of the network [19,29]. Since in figure skating competitions, the referees will incrementally add scores with the progress of the whole competition on-the-fly, ideally, we want the scores for each technical movement. Therefore, we plan a more specialized and fine-grained annotation of the figure skating video [30], to allow the model to supervised learning of the scoring criteria of the skater's technical movements, to achieve the real-time of the network. This real-time design can provide timely feedback data for athlete training, enable artificial intelligence for refereeing earlier, and eliminate the subjectivity of scoring penalties in sports events.

Author Contributions: Conceptualization, C.H. and L.C.; methodology, C.H. and F.S.; software, H.G. (Hongjie Gou); validation, C.H. and X.L.; formal analysis, L.C.; writing—original draft preparation, C.H.; writing—review and editing, F.S.; funding acquisition, H.G. (Hong Gao). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the contribution of the anonymous reviewers and the editors.

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

References

1. Parmar, P.; Morris, B.T. What and how well you performed? A multitask learning approach to action quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 304–313.
2. 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/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9839–9848.
3. Li, Y.; Chai, X.; Chen, X. End-to-end learning for action quality assessment. In *Pacific Rim Conference on Multimedia*; Springer: Cham, Switzerland, 2018; pp. 125–134.
4. Zeng, L.A.; Hong, F.T.; Zheng, W.S.; Yu, Q.Z.; Zeng, W.; Wang, Y.W.; Lai, J.H. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12 October 2020; pp. 2526–2534.
5. Chen, X.; Pang, A.; Yang, W.; Ma, Y.; Xu, L.; Yu, J. SportsCap: Monocular 3D human motion capture and fine-grained understanding in challenging sports videos. *Int. J. Comput. Vis.* **2021**, *129*, 2846–2864. [CrossRef]
6. Zuo, K.; Su, X. Three-Dimensional Action Recognition for Basketball Teaching Coupled with Deep Neural Network. *Electronics* **2022**, *11*, 3797. [CrossRef]
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, 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, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
9. Hara, K.; Kataoka, H.; Satoh, Y. Learning spatio-temporal features with 3d residual networks for action recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 3154–3160.
10. Pirsiavash, H.; Vondrick, C.; Torralba, A. Assessing the quality of actions. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 556–571.
11. Parmar, P.; Tran Morris, B. Learning to score olympic events. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 20–28.
12. Le Q.V.; Zou, W.Y.; Yeung, S.Y.; Ng, A.Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3361–3368.
13. 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]

14. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
15. Wang, L.; Tong, Z.; Ji, B.; Wu, G. Tdn: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1895–1904.
16. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
17. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
18. Roy, A.M.; Bhaduri, J.; Kumar, T.; Raj, K. WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecol. Inform.* **2022**, *2022*, 101919. [CrossRef]
19. Sun, W.; Dai, L.; Zhang, X.; Chang, P.; He, X. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. *Appl. Intell.* **2022**, *52*, 8448–8463. [CrossRef]
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
21. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7862–7871.
22. Nakano, T.; Sakata, A.; Kishimoto, A. Estimating blink probability for highlight detection in figure skating videos. *arXiv* **2020**, arXiv:2007.01089.
23. 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]
24. Xu, A.; Zeng, L.A.; Zheng, W.S. Likert Scoring With Grade Decoupling for Long-Term Action Assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 3232–3241.
25. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 448–456.
26. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
27. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
28. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
29. Sahoo, J.P.; Prakash, A.J.; Pławiak, P.; Samantray, S. Real-Time Hand Gesture Recognition Using Fine-Tuned Convolutional Neural Network. *Sensors* **2022**, *22*, 706. [CrossRef] [PubMed]
30. Shao, D.; Zhao, Y.; Dai, B.; Lin, D. Finegym: A hierarchical video dataset for fine-grained action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2616–2625.