Skating-Mixer: Long-Term Sport Audio-Visual Modeling with MLPs

Jingfei Xia[*1,2], **Mingchen Zhuge**[*1,3],
Tiantian Geng[1], **Shun Fan**[1], **Yuantai Wei**[1], **Zhenyu He**[4], **Feng Zheng**[†1]
[1]Southern University of Science and Technology
[2]The Chinese University of Hong Kong
[3]AI Initiative, King Abdullah University of Science and Technology (KAUST)
[4]Harbin Institute of Technology (Shenzhen)
xj022@ie.cuhk.edu.hk, mingchen.zhuge@kaust.edu.sa, zhenyuhe@hit.edu.cn,
{gengtiantian97, 27957322s, santyelegy, zfeng02}@gmail.com

Abstract

Figure skating scoring is challenging because it requires judging the technical moves of the players as well as their coordination with the background music. Most learning-based methods cannot solve it well for two reasons: 1) each move in figure skating changes quickly, hence simply applying traditional frame sampling will lose a lot of valuable information, especially in 3 to 5 minutes long videos; 2) prior methods rarely considered the critical audio-visual relationship in their models. Due to these reasons, we introduce a novel architecture, named **Skating-Mixer**. It extends the MLP framework in a multimodal fashion and effectively learns long-term representations through our designed memory recurrent unit (MRU). Aside from the model, we collected a high-quality audio-visual **FS1000** dataset, which contains over 1000 videos on 8 types of programs with 7 different rating metrics, overtaking other datasets in both quantity and diversity. Experiments show the proposed method achieves SOTAs over all major metrics on the public Fis-V and our FS1000 dataset. In addition, we include an analysis applying our method to the recent competitions in Beijing 2022 Winter Olympic Games, proving our method has strong applicability.

Introduction

Due to the importance of fair competition, many worldwide sports committees devote themselves to regularizing the behaviors of both athletes and referees. As a supplementary tool, crowd workers seek to employ objective machine intelligence to judge performances in competitions. Therefore, many learning-based assessment models (Pirsiavash, Vondrick, and Torralba 2014; Parmar and Tran Morris 2017; Pan, Gao, and Zheng 2019; Jain, Harit, and Sharma 2020; Xiang et al. 2018; Li, Chai, and Chen 2018) have been introduced in recent years. However, few works proposed to score figure skating videos due to several key challenges:

- **Requiring strong video representation learning.** (1) Figure skating videos are 3−5 minutes long and contain manifold technical movements, requiring effective representation learning on the long videos with large fps (frame per second) which results in large size of inputs. (2) Both audio and video should be considered when calculating the scores in figure skating.
- **Missing high-quality dataset.** Unlike common videos, figure skating videos are sourced from live sporting tournaments that require extensive manual process effort. This could be the reason that existing datasets (Xu et al. 2019; Liu et al. 2020) are not comprehensive enough (in scale or diversity) to cover figure skating.

Earlier work (Liu et al. 2020) utilized a hierarchical LSTM model (Hochreiter and Schmidhuber 1997) to capture the local and global information in figure skating videos. The recent Eagle-eye method (Nekoui, Cruz, and Cheng 2021) considers posing heatmaps and appearance features jointly. Although these two methods achieve comparably good results, they remain obvious deficiencies. Firstly, their methods can hardly generalize to different categories of competition in the real scene. More importantly, they merely consider visual modality. While technical action (visual modality) is important in figure skating, we argue that background music (audio modality) should not be ignored as well. An effective model which considers both visual and audio cues is urgently needed in this field.

Convolutional Neural Networks (Fukushima 1979) have long been used for various computer vision tasks (Zhuge et al. 2022). Later, Transformers (Vaswani et al. 2017) have been introduced into this area and show great power in multimodal learning compared with CNNs (Zhuge et al. 2021). However, the complexity of Vision Transformer (Dosovitskiy et al. 2020) is quadratic in the number of input patches and it usually requires a large amount of data to train (Liu et al. 2021b), which is hardly satisfied in our task. Recently, a pure MLP-based structure MLP-Mixer has been proposed and shows promising results on the image classification task (Tolstikhin et al. 2021) with linear complexity. Yet, this simple structure has not been deeply explored in the audio-visual area. Therefore, we introduce Skating-Mixer, which is the pioneer of MLP-based multimodal architecture and also the first to score figure skating with auditory and visual features. Skating-Mixer has the following properties: (1) It simultaneously models audio and visual features and

[*]The first two authors contributed equally: work is done when they worked as visiting scholars in SUSTech.
[†]Corresponding author
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

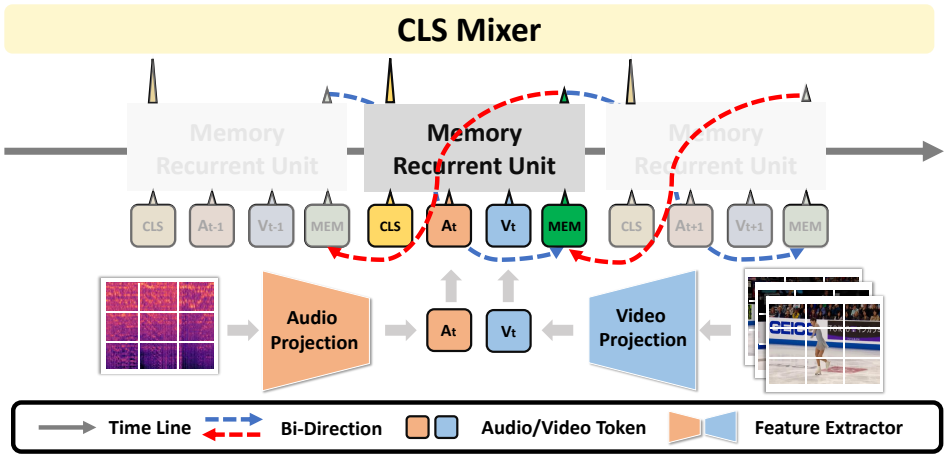


Figure 1: Pipeline of Skating-Mixer. The memory recurrent unit (MRU) of Skating-Mixer works for learning sequential temporal multimodal information. After integral learning in both spatial and temporal information, Skating-Mixer obtains a representation of long-range video.

learns the long-term joint representation in an effective way; (2) With the memory recurrent unit (MRU), this approach accurately predicts the results using extremely long-range cues; (3) With its simple design, it could avoid the gradient vanishing and exploding problems in the vanilla recurrent neural network (Schuster and Paliwal 1997).

Moreover, we observe that few high-quality datasets are set up for this task. Fis-V (Xu et al. 2019) only contains ladies short program videos. Since the data distribution is monotonous, this dataset is less challenging. In this case, we build a new dataset FS1000 with more than 1000 videos, 8 categories of figure skating, and 7 detailed scores to increase the diversity and quantity. The dataset requires the model to learn the underlying features across different figure skating videos. Experiments conducted on both datasets demonstrate that Skating-Mixer not only achieves state-of-the-art results but also has the ability to generalize to all sorts of figure skating, making a good example to tackle multimodal representation in sports. In summary, the primary contributions of this paper are listed as follows:

- We present the pioneer MLP-based multimodal framework that can model extremely long-range videos and score figure skating with auditory and visual information.
- We collect a comprehensive FS1000 dataset, including more long-range videos that contain all types of figure skating with more detailed score records.
- We benchmark recent methods in this field on both Fis-V and FS1000 datasets. The proposed framework outperforms other CNN-based (Parmar and Tran Morris 2017; Parmar and Morris 2019), LSTM-based (Xu et al. 2019), and Transformer-based (Lee et al. 2020) methods.

Dataset

To further facilitate the research of learning-based figure skating scoring, we present the largest figure skating dataset FS1000 with high-quality videos. The dataset is designed for predicting scores in figure skating competitions, with rich annotations like player ID and program categories, which may facilitate this field even more.

Data Collection

Data Source. With an aim to obtain high-quality figure skating videos, we carefully selected and downloaded videos only from top-tier international skating competitions. Besides, we also collected and checked scores from referee reports to annotate our dataset in authoritative.[1] Specifically, ISU World Figure Skating Championships, ISU Grand Prix of Figure Skating, *etc.*, are selected as our main data sources. Normally, figure skating consists of 4 primary categories: mens singles, ladies singles, pair skating and ice dance. Each primary category contains short programs and free skating (in ice dance, they are called rhythm dance and free dance). So, there are 8 subdivided categories, as shown in Figure 2.
Pre-processing. The raw videos collected from figure skating competitions are usually untrimmed and record the whole procedure ranging from 1 to 5 hours. It consists of the performances of all players as well as highlight replay, warming up parts, and waiting for score parts. These redundant contents are usually not helpful for score judgment. We initially downloaded over 400-hour videos and then manually processed all videos, only reserving pure competition performance clips of players from the exact beginning to the ending moment of background music.

Annotation and Statistics

Annotation. As mentioned above, there are totally eight categories of figure skating competitions, namely, men/ladies/-pairs short program (**MS**, **LS**, **PS**), men/ladies/pairs free skating (**MF**, **LF**, **PF**) and ice dance rhythm dance/free dance (**IR**, **IF**). We carefully labeled each video with its official scores according to the referee reports, and also player

[1] https://www.isu.org/figure-skating/entries-results/fsk-results

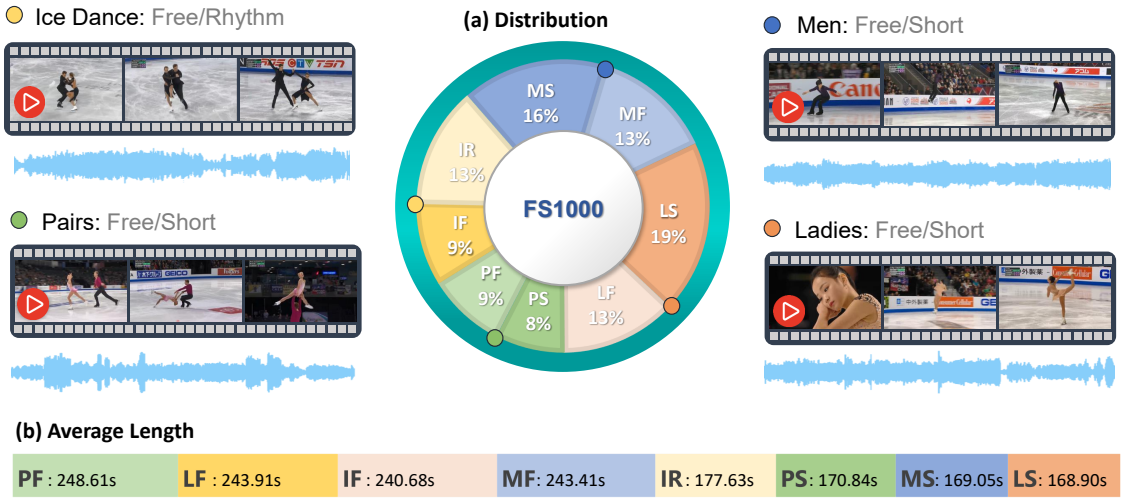


Figure 2: The distribution, average length, and video samples for each category in FS1000. MS: men short program *(16%)*, MF: men free skating *(13%)*, LS: ladies short program *(19%)*, LF: ladies free skating *(13%)*, PS: pairs short program *(8%)*, PF: pairs free skating *(9%)*, IF: ice dance free skating *(9%)*, and IR: ice dance rhythm dance *(13%)*.

ID and corresponding category. The scoring result can be divided into two parts: Technical Element Score (***TES***) and Program Component Score (***PCS***). ***TES*** evaluates the difficulty and execution of all technical movements, and ***PCS*** describes the overall performance, considering five aspects: the Skating Skills (***SS***), Transitions (***TR***), Performance (***PE***), Composition (***CO***), and Interpretation of music (***IN***). Skating Skills assess the skater's command of the blade over the ice; Transitions evaluate skaters' ability to transit between technical elements naturally; Performance shows the appeal and personality of the program; Composition reflects the choreography and the purpose of the way the program is constructed, and Interpretation is more concerned with the consistency between each movement and a corresponding beat in music. Besides, there is a factor that indicates different weights of ***PCS*** scores in different competitions.

Statistics. There are 1604 figure skating videos in our FS1000 dataset: 1247 videos are for training and validation while 357 videos from contests in 2022 are for testing, including the Beijing 2022 Olympics. Each video has ∼5000 frames with a frame rate of 25 and is annotated with detailed ground-truth scores. Some example frames and the percentage of the number of each category in the FS1000 dataset are shown in Figure 2(a). As the videos contain complete snippets of each performance, they are relatively long with a duration ranging from 2.5 minutes to 4.3 minutes, and the average duration of all videos is about 3 minutes and 20 seconds. The details of each category's average length are given in Figure 2(b). We can see that compared with the duration of the short program and rhythm dance, free skating and free dance generally have a longer duration.

Comparison with Other Datasets

Here, we show the comparison between our proposed FS1000 dataset and other existing figure skating video datasets: MCFS (Liu et al. 2021a), FSD-10 (Liu et al. 2020), FisV (Xu et al. 2019) and MIT-Skate dataset (Pirsiavash, Vondrick, and Torralba 2014; Parmar and Tran Morris 2017). FSD-10 and MCFS focus on the technical actions in figure skating and they only contain short clips of each action instead of complete videos, which is not suitable for our multimodal scoring task. MIT-Skate, Fis-V, and our proposed FS1000 are all set up for long-video scoring tasks. MIT-Skate consists of videos that happened before 2012, which is outdated because of the regulation change. Also, it only contains 171 ladies short program videos, which is too limited in scale. Therefore, we focus on the last two datasets in our work. We can see that our dataset is the first one to utilize audio features in this area and overtakes the other datasets in quantity and diversity.

Dataset	Task	# Video	Length	# Score	# Type	Feature
FSD-10	AR	1484	10h	1	-	V
MCFS	AS	271	5h	-	-	V
MIT-Skate	LS	171	8h	1	1	V
FisV	LS	500	24h	2	1	V
FS1000	LS	1604	91h	7	8	A+V

Table 1: Dataset comparison in the figure skating area. The length refers to the total length of all videos. (AR: Action Recognition, AS: Action Segmentation, LS: Long-video Scoring.)

Method

In this section, we comprehensively introduce our MLP-based multimodal model, the **Skating-Mixer**. In our task, the model should be capable of handling extremely long videos with large fps, which means gigantic input audio-visual feature. Models like ViT (Dosovitskiy et al. 2020) can hardly handle these data due to the memory limit. Unlike

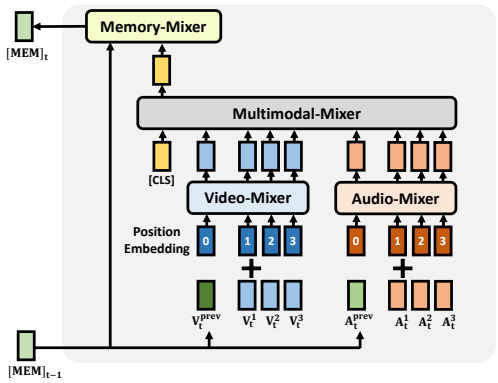


Figure 3: The structure of memory recurrent unit.

previous methods (Bain et al. 2021; Lei et al. 2021) that sample several frames to represent a whole video, we segment the video into multiple 5-second clips and input them to projection models (Gong, Chung, and Glass 2021; Bertasius, Wang, and Torresani 2021) to obtain audio and video features respectively. This is because our model needs to deal with both audio and visual features and sampling frames will cause these two features unaligned. Also, sampling frames will lose the information of fast motions. The whole framework is shown in Figure 1. Suppose there are T clips in the video. For the t-th clip in the video, $\mathbf{A}_t$ and $\mathbf{V}_t$ denote the input audio feature and video feature, respectively.

Memory Recurrent Unit. The structure of the memory recurrent unit (MRU) is shown in Figure 3. This unit tends to capture the long-term relationships in the video and obtain a comprehensive overview of the whole video. The main idea of MRU is to split the video into several clips and process clip by clip. However, it is also necessary to let each clip acquire the information from the previous clips. Therefore, a memory token `[MEM]` is applied with audio and video input. The memory token is a learnable randomly-initialized vector. At the t-th clip, the input memory $\texttt{[MEM]}_{t-1}$ is first passed into two bottleneck structures to extract audio and video-related features from the previous clips denoted as $\mathbf{A}_t^{prev}$ and $\mathbf{V}_t^{prev}$. $\mathbf{A}_t^{prev}$ and $\mathbf{V}_t^{prev}$ are concatenated with the feature input $\mathbf{A}_t$ and $\mathbf{V}_t$. Position embeddings are added with the concatenated feature to identify the time order of the input tokens:

$$\begin{aligned} \tilde{\mathbf{A}}_t &= [\mathbf{A}_t^{prev}\ \mathbf{A}_t] + PE_a \\ \tilde{\mathbf{V}}_t &= [\mathbf{V}_t^{prev}\ \mathbf{V}_t] + PE_v \end{aligned} \tag{1}$$

The audio and video features are passed through two separate MLP-Mixer blocks, Audio-Mixer and Video-Mixer. The two MLP-Mixer blocks enable each modality to fuse information along the time dimension. The result audio and video features, $\hat{\mathbf{A}}_t$ and $\hat{\mathbf{V}}_t$, could capture the long-term feature from previous clips. Then, we use the same strategy as ViT (Dosovitskiy et al. 2020) by using a `[CLS]` token. `[CLS]` token is also a learnable vector concatenated with $\hat{\mathbf{A}}_t$ and $\hat{\mathbf{V}}_t$. The concatenated feature is input to another MLP-Mixer block, Multimodal Mixer to mix information across different modalities. The output of `[CLS]` token, $\texttt{[CLS]}_t$ is used to represent the whole video clip with multimodal information. Finally, $\texttt{[CLS]}_t$ is concatenated with the input memory token $\texttt{[MEM]}_{t-1}$ and pass through another MLP-Mixer block, Memory-Mixer. In the Memory-Mixer, the current clip representation will interact with the memory token and update the memory content. The output at the position of $\texttt{[MEM]}_{t-1}$ is $\texttt{[MEM]}_t$, which is the input memory token for the next clip. The memory output of the last clip $\texttt{[MEM]}_T$ will be used for scoring. Besides, we collect all the `[CLS]` tokens output and add with position embeddings:

$$\tilde{\mathbf{C}} = [\texttt{CLS}_1\ \texttt{CLS}_2\ ...\ \texttt{CLS}_T] + PE_c \tag{2}$$

$\tilde{\mathbf{C}}$ is then input to another Mixer block, CLS Mixer. This block helps the model further learn the local information from each clip. The outputs of the CLS Mixer are averaged and concatenated with $\texttt{[MEM]}_T$. The final score is generated by these two features with a linear layer.

Skating Mixer *vs* LSTM. The recurrent mechanism is similar to RNN (Schuster and Paliwal 1997) and LSTM (Hochreiter and Schmidhuber 1997), but the design of MLP-Mixer is simpler and is capable of dealing with multimodal data. In vanilla RNN architecture, when the input sequence becomes longer, gradient vanishing and gradient exploding happen. In LSTM, this problem is solved by adding a gate strategy. In our architecture, it is not necessary to have delicately designed gates. Gradient vanishing and exploding issues could be mitigated since there is skip-connection within MLP-Mixer blocks and no extra projection is implemented for the memory token. Additionally, the proposed structure is capable of dealing with multimodal information, which is not considered in LSTM. Mixing the memory token will enable the current clip to see the previous information and thus generate a comprehensive view of the whole video.

Bi-Direction Mixer. In our model, we use a similar strategy as in bidirectional recurrent neural network (Schuster and Paliwal 1997), that is to add a backward direction in the model to grasp a comprehensive view of the whole video. For the backward direction, the last clip of the video will be processed first, then the memory flows to the first clip. The averaged value of `[CLS]` outputs from forward and backward directions at each clip are input to CLS Mixer; the averaged `[MEM]` output of the last clip in forward and backward directions will be used for scoring.

Experiment

In this section, we show the experimental results on Fis-V dataset and FS1000 of different methods. Fis-V dataset (Xu et al. 2019) contains 400 ladies short program videos for training and 100 videos for validation. Our proposed FS1000 dataset contains a training set of 1000 videos and a validation set of 247 videos. AST (Gong, Chung, and Glass 2021) and TimeSformer (Bertasius, Wang, and Torresani 2021) are used as our feature extractors. The Mean Square Error (MSE) and Spearman Correlation are used as our evaluation metrics. We compare our method with several related methods (Xu et al. 2019; Parmar and Tran Morris 2017; Parmar

Datasets	Methods	*Mean Square Error*(↓)							*Spearman Correlation*(↑)						
		TES	***PCS***	***SS***	***TR***	***PE***	***CO***	***IN***	***TES***	***PCS***	***SS***	***TR***	***PE***	***CO***	***IN***
Fis-V	C3D-LSTM	39.25	21.97	†	†	†	†	†	0.29	0.51	†	†	†	†	†
	MSCADC	25.93	11.94	†	†	†	†	†	0.50	0.61	†	†	†	†	†
	M-LSTM	25.70	12.48	†	†	†	†	†	0.53	0.68	†	†	†	†	†
	S-LSTM	22.31	10.21	†	†	†	†	†	0.57	0.74	†	†	†	†	†
	MS-LSTM	22.64	9.84	†	†	†	†	†	0.59	0.73	†	†	†	†	†
	M-BERT (Early)	28.04	13.31	†	†	†	†	†	0.54	0.69	†	†	†	†	†
	M-BERT (Mid)	33.32	17.79	†	†	†	†	†	0.54	0.71	†	†	†	†	†
	M-BERT (Late)	27.73	12.38	†	†	†	†	†	0.53	0.72	†	†	†	†	†
	Ours	**19.57**	**7.96**	†	†	†	†	†	**0.68**	**0.82**	†	†	†	†	†
FS1000 (Ours)	C3D-LSTM	308.30	25.85	0.92	0.99	1.21	0.97	1.01	0.78	0.53	0.50	0.52	0.52	0.57	0.47
	MSCADC	148.02	15.47	0.51	0.57	0.78	0.55	0.60	0.77	0.70	0.69	0.69	0.71	0.68	0.71
	M-LSTM	104.62	15.57	0.49	0.72	0.89	0.46	0.56	0.84	0.69	0.74	0.59	0.64	0.78	0.71
	S-LSTM	83.79	10.90	0.40	0.43	0.70	0.40	0.44	0.87	0.79	0.79	0.80	0.78	0.80	0.80
	MS-LSTM	94.55	11.03	0.45	0.49	0.76	0.43	0.47	0.86	0.80	0.77	0.78	0.76	0.79	0.78
	M-BERT (Early)	139.09	14.49	0.44	0.44	0.71	0.47	0.50	0.78	0.77	0.80	0.80	0.76	0.79	0.79
	M-BERT (Mid)	170.57	21.28	0.57	0.54	0.69	0.56	0.56	0.77	0.75	0.79	0.79	0.80	0.79	0.80
	M-BERT (Late)	131.28	15.28	0.44	0.43	0.67	0.47	0.55	0.79	0.75	0.80	0.81	0.80	0.80	0.76
	Ours	**81.24**	**9.47**	**0.35**	**0.35**	**0.62**	**0.37**	**0.39**	**0.88**	**0.82**	**0.80**	**0.81**	**0.80**	**0.81**	**0.81**

Table 2: Experiment Results on Fis-V (Xu et al. 2019) and FS1000. CNN-based: (Parmar and Tran Morris 2017; Parmar and Morris 2019), LSTM-based: (Xu et al. 2019; Parmar and Tran Morris 2017), Transformer-based: (Lee et al. 2020), MLP-based: Ours. MSE and Spearman Correlation are used for evaluation. For MSE, the lower the better; for Spearman Correlation, the higher the better. † denotes the dataset does not include the ground truth.

and Morris 2019; Lee et al. 2020). M-Bert (Lee et al. 2020) uses both audio and visual features while others only consider visual features. It should be noted that we find around 40 out of 500 videos in Fis-V (Xu et al. 2019) dataset contains redundant information (such as the interview and replay), so we cut the videos and re-extract the features.

Results on Fis-V

From Table 2, it can be observed that our proposed Skating-Mixer outperforms other models. Although the Transformer model performs better than MLP-Mixer on general tasks like image classification (Tolstikhin et al. 2021), it does not have an obvious advantage in this specific task. C3D-LSTM (Parmar and Tran Morris 2017) has the worst results since the model is too simple to learn such complex data. 3D CNN-based method, MSCADC (Parmar and Morris 2019) is also struggled to understand such long videos. MS-LSTM (Xu et al. 2019) and our proposed Skating Mixer performs better than the Transformer model, indicating that a strong memory mechanism plays an essential role in long video learning; besides, the attention mechanism is less effective to capture extremely long-term dependencies across clips over several minutes. Additionally, by comparing our model and M-Bert (Lee et al. 2020), it could be found that our model could better learn multimodal features from long videos.

Results on FS1000

Scoring on the FS1000 dataset is much more challenging since Fis-V (Xu et al. 2019) dataset only contains ladies short program videos while the FS1000 dataset consists of different types of figure skating videos, which highly tests the robustness of the model. As shown in Table 2, CNN-based (Parmar and Tran Morris 2017; Parmar and Morris 2019) and Transformer-based (Lee et al. 2020) models still cannot capture long-term dependencies across extremely long videos. LSTM-based (Xu et al. 2019) models achieve better results but fail to further improve the results with only visual features. Our proposed model could handle multimodal information in long figure skating videos and shows great flexibility to fit different types of competitions and obtain the best result among all the models. Another interesting observation is that performance score ***PE*** is harder to learn than the other four sub-scores for all the models, indicating that future work could focus on improving the learning of performance score for a better overall result.

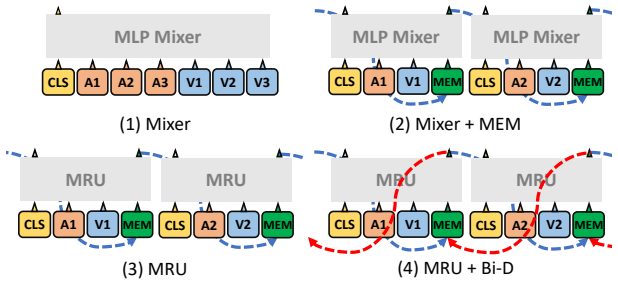


Figure 4: Four of our designed fusion structures.

Ablation Studies

We have conducted ablation experiments on both datasets. Here we analyze two major scores, ***TES*** and ***PCS***.

Component. We first examine the effectiveness of our designed memory recurrent unit (MRU) and bi-direction propagation through four different fusion model structures, as shown in Figure 4. The first one (**Mixer**) is to simply input all the audio and video features into a large MLP-Mixer model. We then optimize this process with the memory recurrent flow with MLP-Mixer (**Mixer+MEM**). Then,

Factor		Baseline	Component			Modality		Scoring token	
			Mixer	Mixer+MEM	MRU	A	V	[CLS]	[MEM]
Fis-V	*TES*	**19.57 (0.68)**	23.25 (0.55)	20.34 (0.66)	20.07 (0.66)	33.04 (0.49)	20.56 (0.67)	21.53 (0.66)	20.10 (0.67)
	PCS	**7.96 (0.82)**	10.87 (0.73)	8.75 (0.79)	8.23 (0.81)	14.68 (0.67)	9.97 (0.76)	8.55 (0.81)	8.10 (0.82)
FS1000 (Ours)	*TES*	**81.24 (0.88)**	93.65 (0.85)	90.80 (0.87)	85.71 (0.88)	94.94 (0.82)	82.59 (0.88)	83.16 (0.86)	88.78 (0.86)
	PCS	**9.47 (0.82)**	13.81 (0.75)	10.96 (0.80)	10.44 (0.81)	20.40 (0.55)	10.92 (0.79)	10.37 (0.81)	10.71 (0.79)

Table 3: Ablation studies on different components, modalities, and scoring tokens. The value outside the brackets is MSE while inside is Spearman correlation. Baseline is our proposed framework, which means MRU+Bi-D, A+V, and [CLS]+[MEM] in component, modality, and scoring token, respectively.

Dataset	Backbone	*TES*	*PCS*	# Params	MACs
Fis-V	Transformer	25.60	13.18	20.48M	35.55G
	MLP-Mixer	**19.57**	**7.96**	**14.32M**	**24.95G**
FS1000 (Ours)	Transformer	108.68	13.40	20.48M	48.92G
	MLP-Mixer	**81.24**	**9.47**	**14.32M**	**34.36G**

Table 4: Comparison of Transformer and MLP-Mixer as the backbone in our Skating Mixer. MACs mean Multiply–Accumulate Operations.

the vanilla MLP-Mixer is replaced by our proposed MRU (**MRU**). Finally, the bi-directional mechanism is applied with our proposed MRU (**MRU+Bi-D**). The result is shown in Table 3. The model fails to learn the long-term relationship between audio and video modalities when directly inputting all the features into a single model. Introducing the memory mechanism helps the model capture long-term information and our proposed MRU could further better fuse multimodal features within extremely long videos. Moreover, a bi-direction flow could effectively enhance the understanding of long videos and improve the performance.

Modality. Here we analyze the importance of using multimodal features in figure skating tasks. We have run experiments on our proposed model with only audio and only visual features. Results in Table 3 show that Skating Mixer with only audio features generates the worst result. This shows that music plays an auxiliary part in figure skating and the visual technique moves always play a key role in scoring. Using both audio and video clues in figure skating scoring performs better than using only visual features, indicating that audio-visual learning is important in this field.

Scoring token. In this section, we are going to discuss the effectiveness of [CLS] and [MEM] tokens in our proposed architecture. In our model, both [CLS] and [MEM] token output are used for final scoring. We first generate scores with only the output from CLS Mixer, which means only [CLS] token outputs are used for scoring. Then we utilize only the [MEM] output from the last clip for scoring. The result shows that using both [CLS] and [MEM] token generates the best result. This is because [MEM] token passes across the whole video and contains global information while [CLS] token contains local information of each clip. Combining these two types of features could help the model improve its performance.

Backbones. In this part, we show the difference between MLP-Mixer and Transformer. Transformer has been widely adopted in the multimodal area, but as mentioned in (Tolstikhin et al. 2021), the computation complexity of Transformer is $\mathcal{O}(n^2)$ while for MLP-Mixer it is $\mathcal{O}(n)$, where n is the number of input tokens. Also, (Liu et al. 2021b) mentions that Transformer usually requires a large amount of data for training. We conduct experiments by replacing the MLP-Mixer with the same number of Transformer blocks. The results in Table 4 demonstrate that in our task with a relatively small dataset (compared to Imagenet (Deng et al. 2009)), MLP-Mixer yields better performance than Transformer with less computation resource.

Visualization

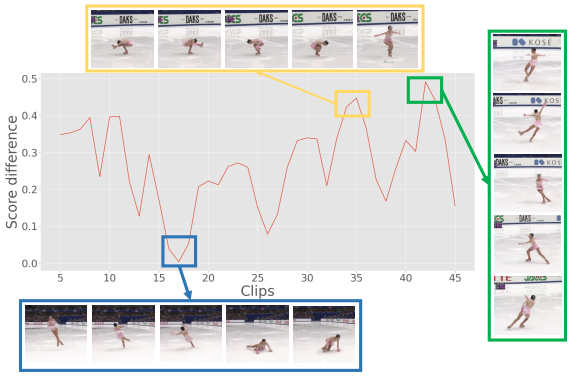
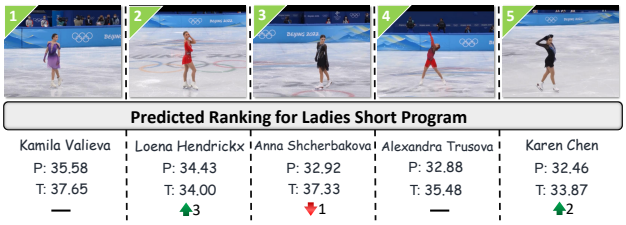


Figure 5: Score difference when sequentially adding clips. The score is low when the action fails (blue box); the score is relatively high when finishing a complex move (yellow box) or skating fluently with the music (green box).

In this section, we will demonstrate how our model scores in a certain video. We produce scores starting with the first clip and add one following clip at each time. When adding one clip, the score difference is computed to show the score of this clip. An example is shown in Figure 5. It can be seen that when an action is failed, the score for this clip is much lower (blue box). When the athlete successfully performs a complex technical move, the score is much higher (yellow box). Also, we found that when the athletes skate with the beat of the music, they also get higher scores (green box). This result clearly demonstrates that our proposed method could learn some basic patterns in figure skating.

Performances on Beijing 2022 Olympic Games

To verify the robustness and effectiveness of the figure skating model, we apply our trained model to competitions that



(a) Predicted results on Ladies Short Program

(b) Predicted results on Pairs Free Program

Figure 6: Predicted TOP-5 Ranking in Beijing 2022 Winter Olympic Games. P stands for predicted score and T stands for true score. The last row is the ranking difference compared to the real ranking.

are not included in the training data to see its actual effect. Therefore, we take the figure skating videos from the Beijing 2022 Winter Olympic Games and obtain the rankings of athletes. In practice, predicting correct ranking is more meaningful than predicting the exact scores because the evaluation scales are inconsistent in different competitions. Here we use our model to predict the ***PCS*** for samples. Results in Figure 6 show that although the score may not be accurate, the top-5 ranking does not change too much compared to the real one. This is because top-tier athletes share similar technique moves and maintain high-quality performances. Such results could satisfy the practical need for auxiliary judgment. This demonstrates that our model could actually learn some of the scoring standards and identify better performance from a group of athletes.

Related Work

Audio-Visual Learning

There exists a rich exploration in audio-visual multimodal learning, especially in the deep learning era (Zhu et al. 2021). Datasets (Chen et al. 2020; Lee et al. 2021; Gemmeke et al. 2017; Abu-El-Haija et al. 2016) involving representation learning in this field have been also developed. M-BERT (Lee et al. 2020) focuses on reducing the complexity when introducing Transformer into audio-visual learning. AVAS (Morgado, Li, and Vasconcelos 2020) attempts to learn the spatial alignment between audio and vision by introducing a new self-supervised proxy task. Besides, another work (Nagrani et al. 2021) also proposes to design a middle bottleneck to integrate audio and video modalities. Although many existing methods explore audio-visual learning by various techniques, our proposed Skating-Mixer is the first attempt to apply MLP architecture to tackle this problem.

MLP-Based Architecture

Attention-based network Transformer (Vaswani et al. 2017; Ji et al. 2022; Dosovitskiy et al. 2020) achieve unparalleled success in the computer vision field. Recently, MLP-Mixer (Tolstikhin et al. 2021) argues that MLP can be an alternative solution for visual representation learning. Subsequently, many MLP-based architectures emerged in the computer vision area. S^2-MLP (Yu et al. 2021) only contains channel-mixing MLP and utilizes a spatial-shift operation for communication among patches. In addition, there are several specific-function models being proposed, such as CycleMLP (Chen et al. 2021) used in dense prediction, MixerGAN (Cazenavette and De Guevara 2021) and CrossMLP (Ren, Tang, and Sebe 2021) in image translation, and Mixer-TTS (Tatanov, Beliaev, and Ginsburg 2021) in the text-to-speech task. Distinguished from the aforementioned works, our proposed model is the first MLP model to solve audio-visual problems.

Figure Skating

In computer vision, the earliest work about figure skating can be traced back to MIT-Skating (Pirsiavash, Vondrick, and Torralba 2014), which gathered Olympic videos and assessed the actions. Similar research (Xu et al. 2019) also focuses on scoring figure skating videos and collects a Fis-V dataset with 500 videos. However, it only considers ladies single programs, making it hard to generalize in this field. (Liu et al. 2020) introduces action recognition in figure skating and meanwhile designs an FSD-10 dataset. Another fine-grained, motion-centered MCFS dataset (Liu et al. 2021a) is proposed for the temporal action segmentation task. Additionally, several dedicated models have been proposed. 1D CNN (Nakano, Sakata, and Kishimoto 2020) detects the highlight in figure skating programs with people's reactions. EAGLE-Eye (Nekoui, Cruz, and Cheng 2021) creates a two-stream pipeline to learn the long-term representation of figure skating actions. Compared with the others, our model is the first attempt to solve figure skating scenarios using both audio and video cues.

Conclusion and Future Work

This paper introduces the MLP-based multimodal architecture for scoring figure skating. The proposed model solves learning multimodal information in long videos, which is essential for this task. Besides, an elaborated-designed dataset has been collected. We set the benchmarks that compare our model with CNN-Based, LSTM-based, and Transformer-Based methods on the Fis-V and our proposed FS1000 datasets. The experiments show the effectiveness of our method, indicating that MLP-based architecture is capable of the multimodal task. Furthermore, we apply our model to the 2022 Winter Olympics to verify the model's effectiveness and applicability. However, this work mainly focuses on figure skating. Extending the dataset and method to other sports would be interesting. Also, the pose information can be great supplementary to action-based tasks, which can be used to improve the method in the future.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant NO. 2022YFF1202903) and the National Natural Science Foundation of China (Grant NO. 61972188 and 62122035).

References

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1728–1738.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095.

Cazenavette, G.; and De Guevara, M. L. 2021. MixerGAN: An MLP-Based Architecture for Unpaired Image-to-Image Translation. *arXiv preprint arXiv:2105.14110*.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, S.; Xie, E.; Ge, C.; Liang, D.; and Luo, P. 2021. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fukushima, K. 1979. Neural network model for a mechanism of pattern recognition unaffected by shift in position-Neocognitron. *IEICE Technical Report, A*, 62(10): 658–665.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. AST: Audio Spectrogram Transformer. arXiv:2104.01778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jain, H.; Harit, G.; and Sharma, A. 2020. Action quality assessment using siamese network-based deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6): 2260–2273.

Ji, G.-P.; Zhuge, M.; Gao, D.; Fan, D.-P.; Sakaridis, C.; and Van Gool, L. 2022. Masked Vision-Language Transformer in Fashion. *arXiv preprint arXiv:2210.15110*.

Lee, S.; Chung, J.; Yu, Y.; Kim, G.; Breuel, T.; Chechik, G.; and Song, Y. 2021. ACAV100M: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10274–10284.

Lee, S.; Yu, Y.; Kim, G.; Breuel, T.; Kautz, J.; and Song, Y. 2020. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124*.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Li, Y.; Chai, X.; and Chen, X. 2018. End-to-end learning for action quality assessment. In *Pacific Rim Conference on Multimedia*, 125–134. Springer.

Liu, S.; Liu, X.; Huang, G.; Qiao, H.; Hu, L.; Jiang, D.; Zhang, A.; Liu, Y.; and Guo, G. 2020. FSD-10: A fine-grained classification dataset for figure skating. *Neurocomputing*, 413: 360–367.

Liu, S.; Zhang, A.; Li, Y.; Zhou, J.; Xu, L.; Dong, Z.; and Zhang, R. 2021a. Temporal Segmentation of Fine-gained Semantic Action: A Motion-Centered Figure Skating Dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 2163–2171. AAAI Press.

Liu, Y.; Sangineto, E.; Bi, W.; Sebe, N.; Lepri, B.; and Nadai, M. D. 2021b. Efficient Training of Visual Transformers with Small Datasets. In *NeurIPS*.

Morgado, P.; Li, Y.; and Vasconcelos, N. 2020. Learning representations from audio-visual spatial alignment. *arXiv preprint arXiv:2011.01819*.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *arXiv preprint arXiv:2107.00135*.

Nakano, T.; Sakata, A.; and Kishimoto, A. 2020. Estimating Blink Probability for Highlight Detection in Figure Skating Videos. *arXiv preprint arXiv:2007.01089*.

Nekoui, M.; Cruz, F. O. T.; and Cheng, L. 2021. EAGLE-Eye: Extreme-Pose Action Grader Using Detail Bird's-Eye View. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 394–402.

Pan, J.-H.; Gao, J.; and Zheng, W.-S. 2019. Action assessment by joint relation graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6331–6340.

Parmar, P.; and Morris, B. T. 2019. Action Quality Assessment Across Multiple Actions. arXiv:1812.06367.

Parmar, P.; and Tran Morris, B. 2017. Learning to score olympic events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 20–28.

Pirsiavash, H.; Vondrick, C.; and Torralba, A. 2014. Assessing the quality of actions. In *European Conference on Computer Vision*, 556–571. Springer.

Ren, B.; Tang, H.; and Sebe, N. 2021. Cascaded Cross MLP-Mixer GANs for Cross-View Image Translation. *arXiv preprint arXiv:2110.10183*.

Schuster, M.; and Paliwal, K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11): 2673–2681.

Tatanov, O.; Beliaev, S.; and Ginsburg, B. 2021. Mixer-TTS: non-autoregressive, fast and compact text-to-speech model conditioned on language model embeddings. *arXiv preprint arXiv:2110.03584*.

Tolstikhin, I.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Xiang, X.; Tian, Y.; Reiter, A.; Hager, G. D.; and Tran, T. D. 2018. S3d: Stacking segmental p3d for action quality assessment. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 928–932. IEEE.

Xu, C.; Fu, Y.; Zhang, B.; Chen, Z.; Jiang, Y.-G.; and Xue, X. 2019. Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30(12): 4578–4590.

Yu, T.; Li, X.; Cai, Y.; Sun, M.; and Li, P. 2021. S^2-MLPv2: Improved Spatial-Shift MLP Architecture for Vision. *arXiv preprint arXiv:2108.01072*.

Zhu, H.; Luo, M.-D.; Wang, R.; Zheng, A.-H.; and He, R. 2021. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 1–26.

Zhuge, M.; Gao, D.; Fan, D.-P.; Jin, L.; Chen, B.; Zhou, H.; Qiu, M.; and Shao, L. 2021. Kaleido-BERT: Vision-Language Pre-training on Fashion Domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12647–12657.

Zhuge, M.; Lu, X.; Guo, Y.; Cai, Z.; and Chen, S. 2022. CubeNet: X-shape connection for camouflaged object detection. *Pattern Recognition*, 127: 108644.