Audio-Visual MLP for Scoring Sport

1、Motivation

这应该是一篇2022刚放出来的要投TCSVT的文章

在过往的研究中，只有很少的工作关注到对figure skating videos进行assessment，作者认为主要原因是：

- **Requiring strong representation learning.** (1) Figure skating videos are 3–5 minutes long and contain manifold technical movements, requiring extremely long-term representation learning, and cannot be sampled. (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 demand extensive manual effort to process. 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.

虽然过往的研究提出了一些不错的方法，但他们都没有关注到听觉信息，但事实上，对于滑冰比赛来说，视听信息的融合是非常重要的。基于这些问题，作者提出了一个新的数据集FS1000，并设计了一个基于MLP-Mixer的模型。

> Skating-Mixer has the following properties: (1) It simultaneously model audio and visual features, and learns in an effective way; (2) By adopting 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).

2、Dataset

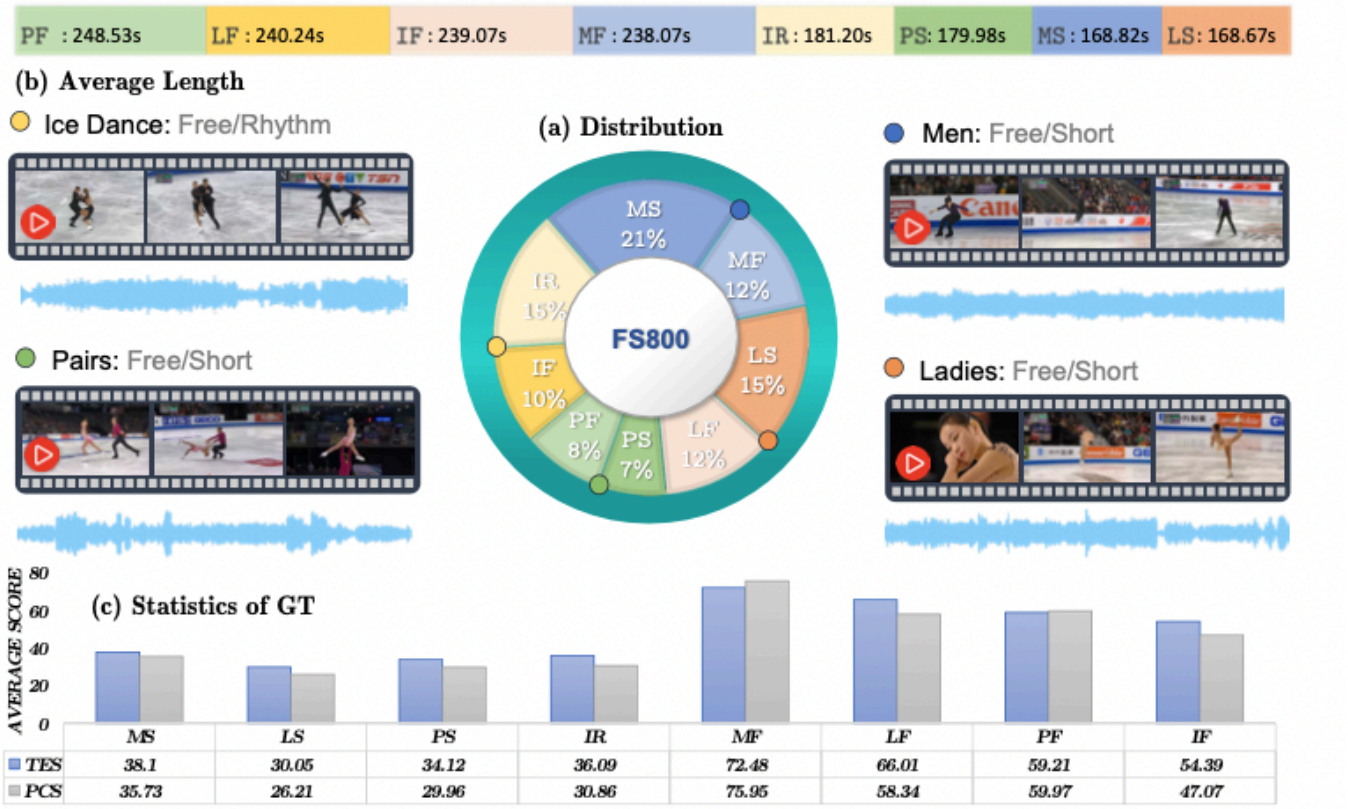


Figure 2: The average length (a), distribution (b), and statistics of GT (c) for each category in FS1000. **MS:** men's short program *(21%)*, **MF:** men's free skating *(12%)*, **LS:** ladies' short program *(15%)*, **LF:** ladiess free skating *(12%)*, **PS:** pairs short program *(7%)*, **PF:** pairs free skating *(8%)*, **IF:** ice dance free skating *(10%)*, and **IR:** ice dance rhythm dance *(15%)*.

TES——Technical Element Score：表示难度和表现分

PCS——Program Component Score：总体表现，包含 the Skating Skills (SS), Transitions (TR), Performance (PE), Composition (CO), and Interpretation of music (IN)五个方面

3、Method

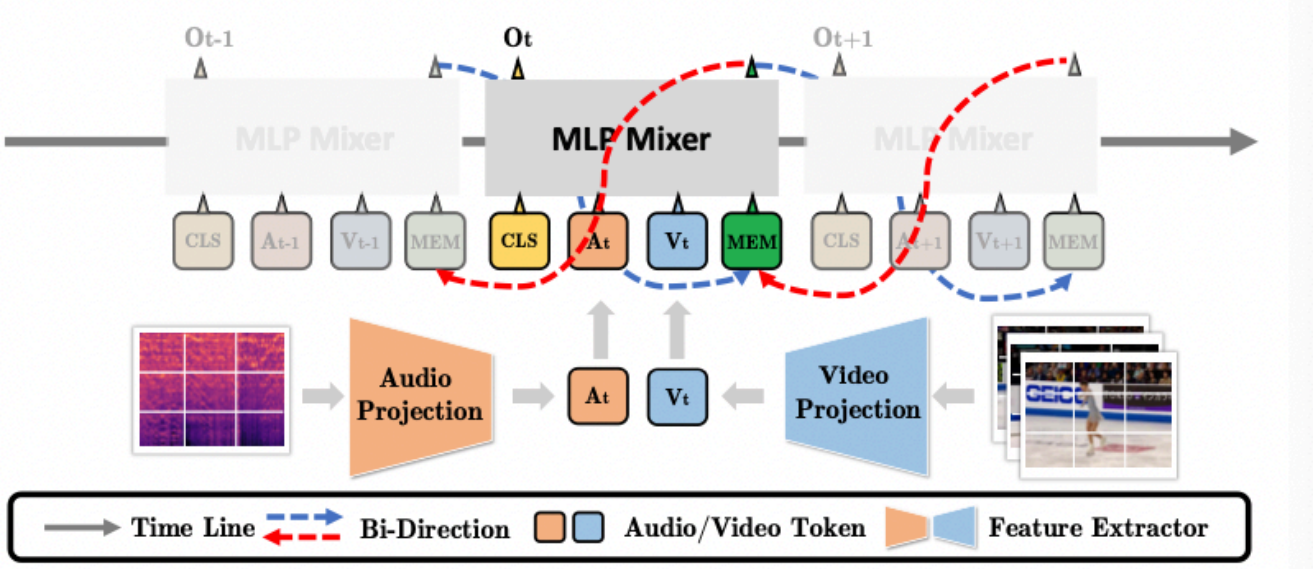


Figure 1: **Pipeline of Skating-Mixer.** We adopt patch modeling methods just like (Dosovitskiy et al., 2020; Tolstikhin et al., 2021), and use TimeSformer (Bertasius et al., 2021) and AST (Gong et al., 2021) as our projection backbones. The memory recurrent unit (MRU) of Skating-Mixer works for learning sequential temporal information. After integral learning in both spatial and temporal information, Skating-Mixer obtains a representation of long-range video. We integrate the outputs from each Skating-Mixer block into a head module as mentioned in Section 4.2 and finish the figure skating scoring. `[MEM]` denotes the memory token and `[CLS]` denotes the class token.

Algorithm 1 Skating Mixer

Input: $\mathbf{A}$: projected acoustic feature; $\mathbf{V}$: projected visual feature;
Output: S: predicted score;
Parameter: `CLS`: class token; $\texttt{MEM}_0/\texttt{MEM}_{T-1}$: the same initial memory token

for t $\leftarrow 0$ **to** $T-1$ **do**
  $\mathbf{Y}_t = \text{MLP-Mixer}([\texttt{CLS}\ \mathbf{A}_t\ \mathbf{V}_t\ \texttt{MEM}_t])$
  $\mathbf{O}_t^{forward} = \mathbf{W}_{out}\mathbf{Y}_t[0]$
  $\texttt{MEM}_{t+1} = \mathbf{Y}_t[-1]$
end for
for t $\leftarrow T-1$ **to** 0 **do**
  $\mathbf{Y}_t = \text{MLP-Mixer}([\texttt{CLS}\ \mathbf{A}_t\ \mathbf{V}_t\ \texttt{MEM}_t])$
  $\mathbf{O}_t^{backward} = \mathbf{W}_{out}\mathbf{Y}_t[0]$
  $\texttt{MEM}_{t-1} = \mathbf{Y}_t[-1]$
end for
$\hat{\mathbf{O}} = \text{Average}(\frac{\mathbf{O}_0^{forward}+\mathbf{O}_0^{backward}}{2}, \ldots, \frac{\mathbf{O}_{T-1}^{forward}+\mathbf{O}_{T-1}^{backward}}{2})$
$S = \text{ScoreHead}(\hat{\mathbf{O}})$
return S

作者的这个设计很像Bi-LSTM的设计，但是相比于LSTM，MLP-Mixer更具有优势：

LSTM makes the model focus on important parts of input and thus reduces the effective input length to avoid the gradient problem.

Gradient vanishing and exploding issues could be mitigated since there is skip-connection within channel-mixing and token-mixing MLP block and no extra projection is implemented for memory token.

4、Experiment

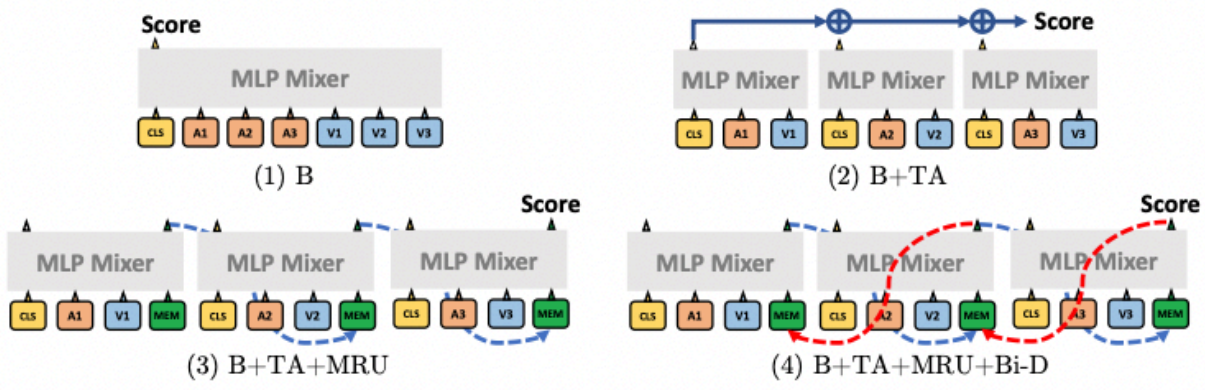
4.1 comparison results

Datasets	Methods	*TES*	*PCS*	*SS*	*TR*	*PE*	*CO*	*IN*
Fis-V (Xu et al., 2019)	C3D-LSTM (Parmar and Tran Morris, 2017)	42.79	25.81	†	†	†	†	†
	MSCADC (Parmar and Morris, 2019)	27.27	16.00	†	†	†	†	†
	M-LSTM (Xu et al., 2019)	20.80	9.09	†	†	†	†	†
	S-LSTM (Xu et al., 2019)	21.95	9.60	†	†	†	†	†
	MS-LSTM (Xu et al., 2019)	22.54	9.09	†	†	†	†	†
	M-BERT (Early) (Lee et al., 2020)	28.50	14.85	†	†	†	†	†
	M-BERT (Mid) (Lee et al., 2020)	35.51	20.31	†	†	†	†	†
	M-BERT (Late) (Lee et al., 2020)	30.25	14.93	†	†	†	†	†
	Ours (A)	36.66	19.46	†	†	†	†	†
	Ours (V)	20.45	9.03	†	†	†	†	†
	Ours (A+V)	**20.39**	**8.93**	†	†	†	†	†
FS1000 (Ours)	C3D-LSTM (Parmar and Tran Morris, 2017)	291.61	33.88	1.24	1.32	1.66	1.30	1.34
	MSCADC (Parmar and Morris, 2019)	84.99	18.96	0.69	0.76	1.12	0.71	0.82
	M-LSTM (Xu et al., 2019)	189.35	14.82	0.59	0.63	1.00	0.55	0.66
	S-LSTM (Xu et al., 2019)	210.99	17.83	0.53	0.57	0.98	0.55	0.61
	MS-LSTM (Xu et al., 2019)	204.17	16.56	0.58	0.62	0.95	0.53	0.61
	M-BERT (Early) (Lee et al., 2020)	90.48	17.79	0.52	0.49	0.91	0.60	0.58
	M-BERT (Mid) (Lee et al., 2020)	127.73	25.91	0.63	0.60	1.06	0.70	0.73
	M-BERT (Late) (Lee et al., 2020)	95.27	18.27	0.54	0.57	0.92	0.55	0.56
	Ours (A)	97.17	21.47	0.81	0.84	1.18	0.82	0.92
	Ours (V)	68.80	11.56	0.42	0.48	0.84	0.49	0.55
	Ours (A+V)	**65.08**	**10.44**	**0.42**	**0.45**	**0.83**	**0.46**	**0.49**

Table 1: Experiment Results on Fis-V (Xu et al., 2019) and ours 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. The scores are computed by MSE. A: only audio, V: only video, A+V: Audio+Video. † denotes the dataset does not include the GT.

在setting上，M-BERT使用了A+V其他的过往的方法只有V。

4.2 Ablation study



Factor		Component				Layer			Dropout			
		B	B+TA	B+MRU	B+MRU+Bi-D	1	2	3	0%	10%	20%	30%
					✓		✓		✓			
Fis-V	*TES*	23.25	22.34	20.80	20.39	21.22	20.39	21.79	20.39	21.31	20.80	21.16
(Xu et al., 2019)	*PCS*	11.32	9.98	9.37	8.93	9.73	8.93	9.21	8.93	9.61	10.49	10.33
FS1000	*TES*	70.93	71.15	68.80	65.08	65.20	65.08	66.21	65.08	64.24	68.02	69.35
(Ours)	*PCS*	13.70	11.90	11.17	10.44	11.60	10.44	12.26	10.44	11.33	11.52	11.98

performance on test set

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

The result shows that although the score may not be accurate, the top-5 ranking does not change too much compare to real ranking because top-tier athletes share some similar technique moves and maintain high-quality performances.