ACM multimedia
Chengdu, China OCT 20-24 2021
ACM Multimedia 2021

TSA-Net: Tube Self-Attention Network for Action Quality Assessment

Shunli Wang, *Dingkang Yang*, *Peng Zhai*, *Chixiao Chen*, *Lihua Zhang*∗

Mail: slwang19@fudan.edu.cn, lihuazhang@fudan.edu.cn

Fudan University, Shanghai, China

https://github.com/Shunli-Wang/TSA-Net

DOI: 10.1145/3474085.3475438

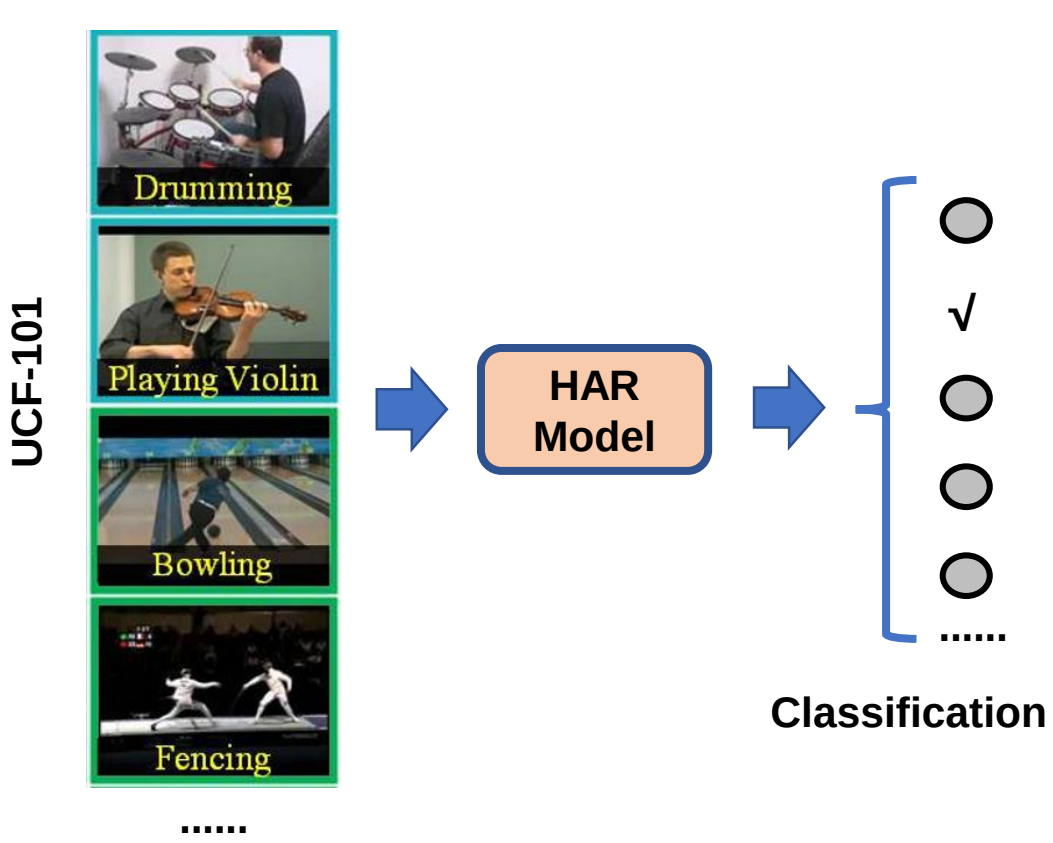
Outline

- 1. Background & Motivations
- 2. Proposed TSA-Net
- 3. Experimental Results
- 4. Conclusion

1. Background & Motivations

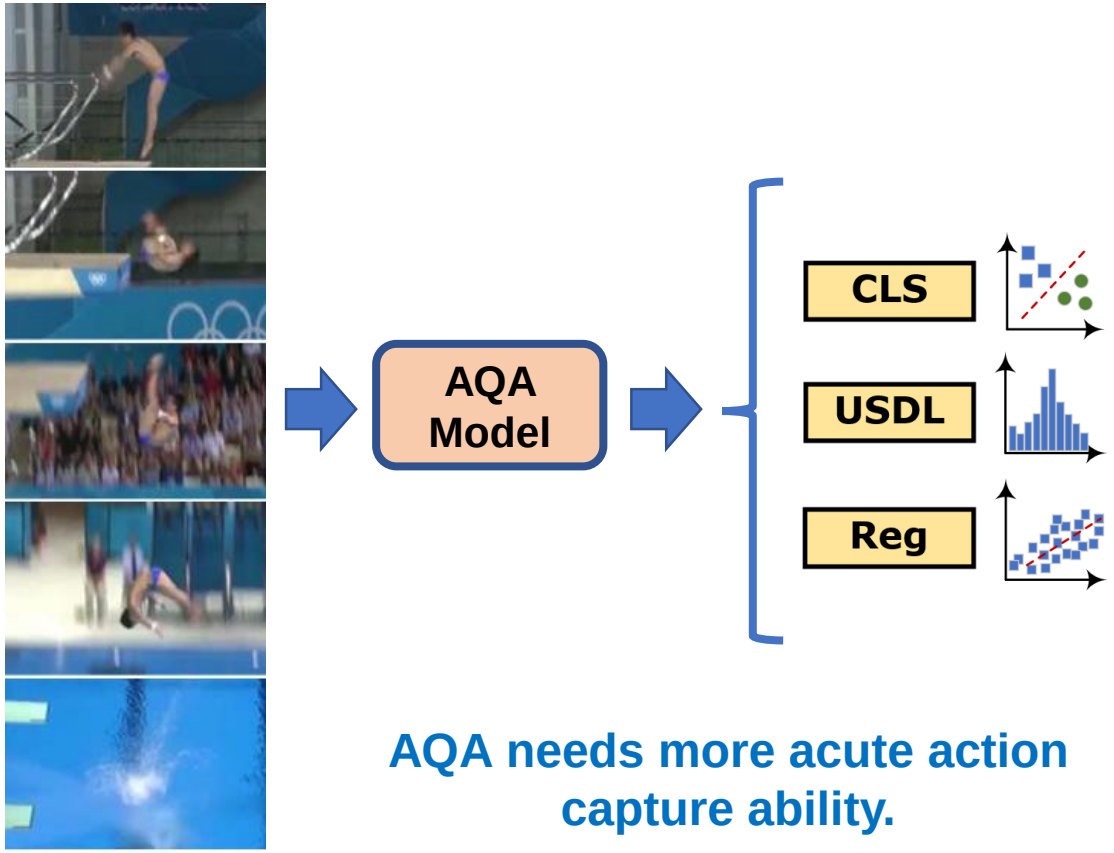
Human Action Recognition (HAR)

Models in HAR require distinguishing subtle differences between different actions.



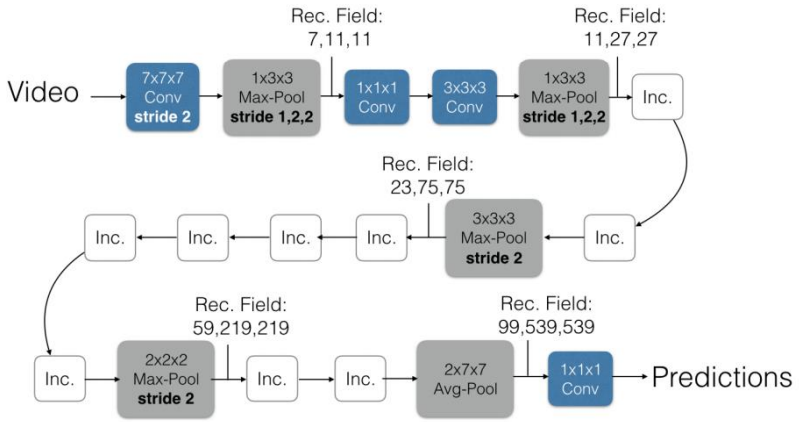
Action Quality Assessment (AQA)

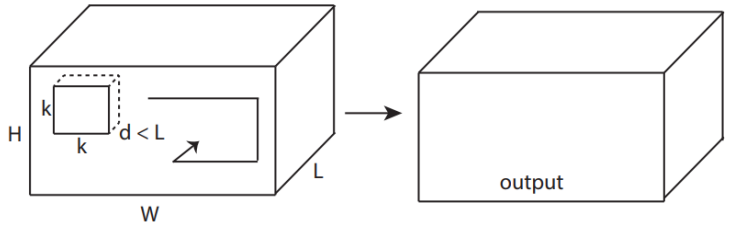
Models in AQA require evaluating a specific action's advantages and disadvantages.



AQA needs more acute action capture ability.

1. Background & Motivations





Current AQA Methods

Challenge 1: There is a huge GAP between HAR and AQA.

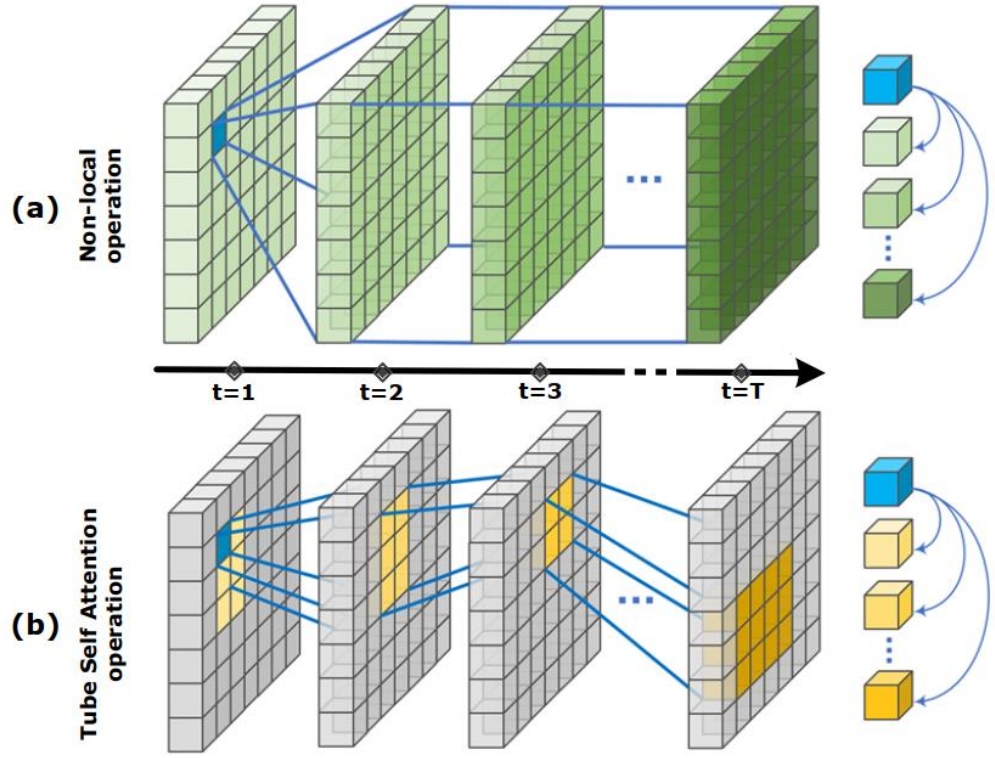
Challenge 2: Existing methods cannot perform feature aggregation efficiently.

The performances and efficiency of AQA methods are indeed limited.

How can we design a network structure suitable for AQA to complete effective and efficient feature aggregation?

1. Background & Motivations

AQA models require rich temporal contextual information and do not require irrelevant spatial contextual information.



& { • Tube Mechanism • Self-attention Mechanism }

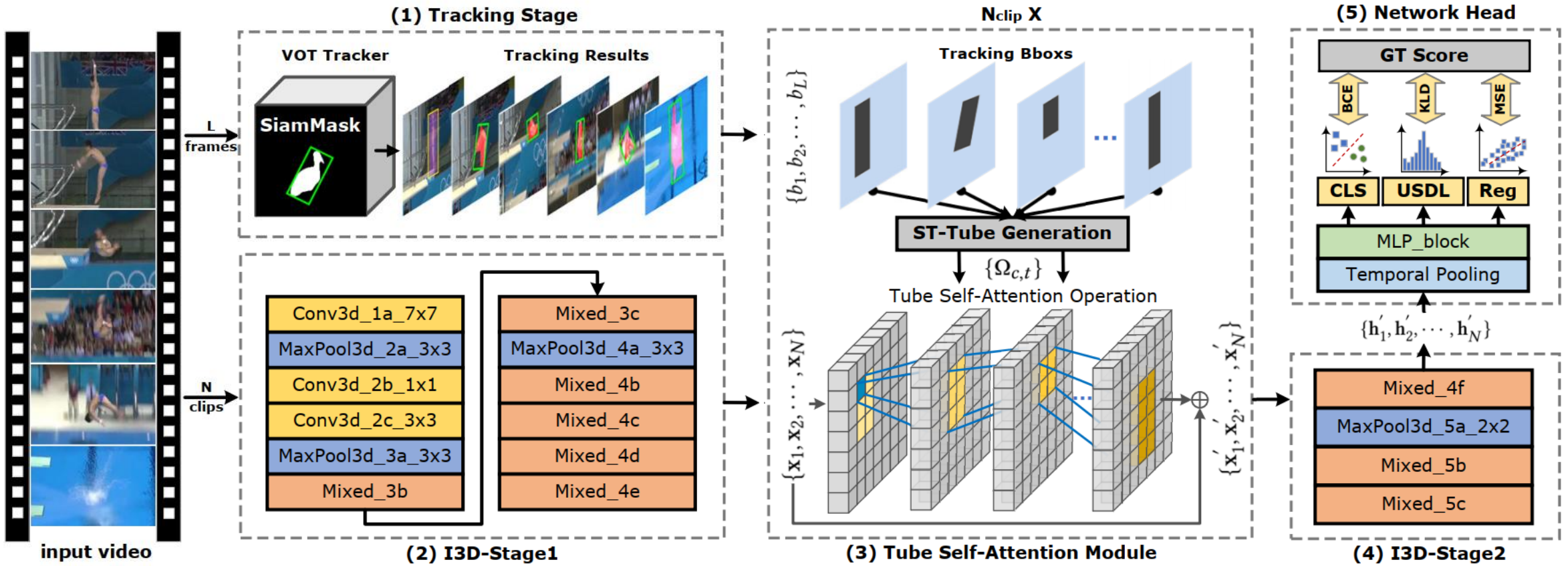
↓

Tube Self-Attention Module

Merits of the TSA module

- ✓ High efficiency
- ✓ Effectiveness
- ✓ Flexibility

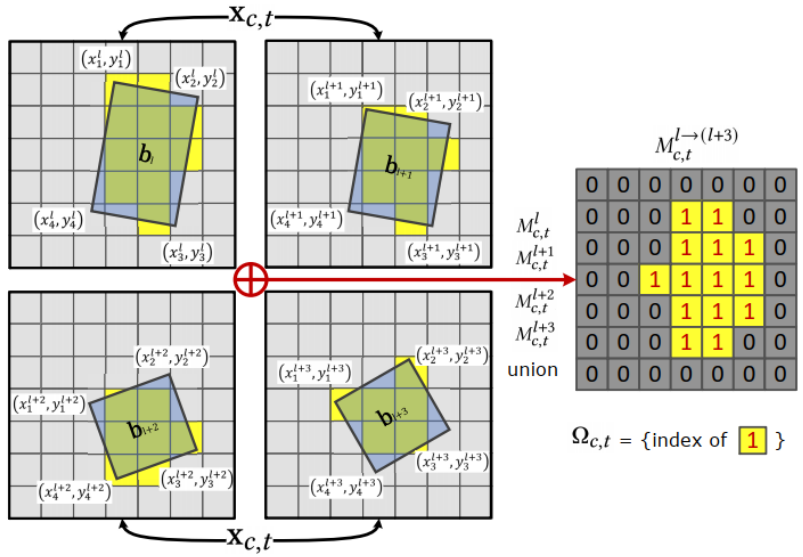
2. Proposed TSA-Net



$V = \{F_l\}_{l=1}^{L}$	**tracking results** $B = \{b_l\}_{l=1}^{L} \quad b_l = \{(x_p^l, y_p^l)\}_{p=1}^{4}$	**TSA Module** $\mathbf{X}' = \{\mathbf{x}'_n\}_{n=1}^{N} \quad \mathbf{x}'_n \in \mathbb{R}^{T\times H\times W\times C}$	**average pooling** $\overline{\mathbf{h}} = \frac{1}{N}\sum_{n=1}^{N} h_n$
input video	$\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathbb{R}^{T\times H\times W\times C}$ **I3D Stage-1**		$\mathbf{H} = \{\mathbf{h}_n\}_{n=1}^{N}$ **I3D Stage-2**

2. Proposed TSA-Net

Step 1: Spatio-temporal Tube Generation



Mask Generation

$$M_{c,t}^{l}(i,j) = \begin{cases} 1, S(b_l,(i,j)) \geqslant \tau \\ 0, S(b_l,(i,j)) < \tau \end{cases}$$

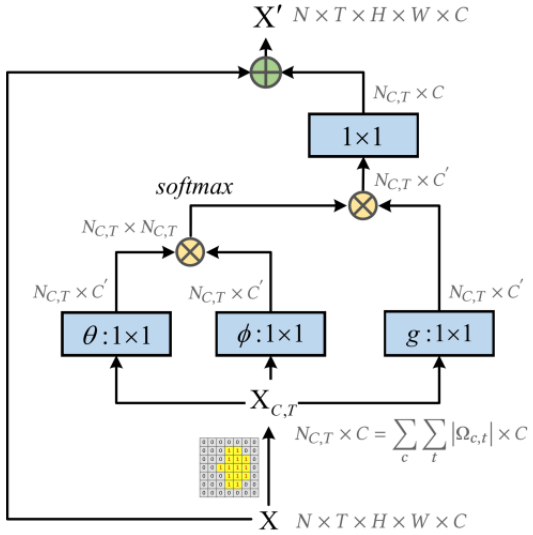
Element-wise OR

$$M_{c,t}^{l\to(l+3)} = Union(M_{c,t}^{l}, M_{c,t}^{l+1}, M_{c,t}^{l+2}, M_{c,t}^{l+3})$$

Indicator Set

$$\Omega_{c,t} = \left\{(i,j) | M_{c,t}^{l\to(l+3)}(i,j) = 1\right\}$$

Step 2: Tube Self-attention Operation



TSA mechanism

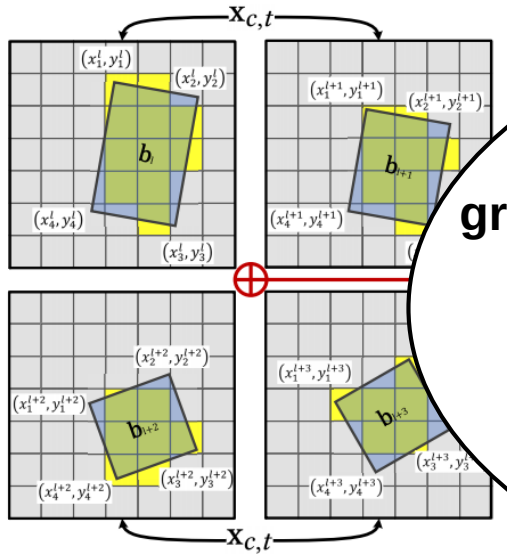
$$\mathbf{y}_p = \frac{1}{C(\mathbf{x})} \sum_{\forall c} \sum_{\forall t} \sum_{\forall (i,j)\in\Omega_{c,t}} f\left(\mathbf{x}_p, \mathbf{x}_{c,t}(i,j)\right) g(\mathbf{x}_{c,t}(i,j))$$

Normalized Constant $C(\mathbf{x}) = \sum_c \sum_t |\Omega_{c,t}|$

Similarity Func. $f\left(\mathbf{x}_p, \mathbf{x}_{c,t}(i,j)\right) = \theta(\mathbf{x}_p)^T \phi(\mathbf{x}_{c,t}(i,j))$

2. Proposed TSA-Net

Step 1:
Spatio-temporal Tube Generation



Mask Generation

$$M_{c,t}^{l}(i,j) = \begin{cases} 1, S(b_l,(i,j)) \geqslant \tau \\ 0, S(b_l,(i,j)) < \tau \end{cases}$$

Element-wise OR

$$M_{c,t}^{l\to(l+3)} = Union(M_{c,t}^{l}, M_{c,t}^{l+1}, M_{c,t}^{l+2}, M_{c,t}^{l+3})$$

Indicator Set

$$\Omega_{c,t} = \left\{(i,j) | M_{c,t}^{l\to(l+3)}(i,j) = 1\right\}$$

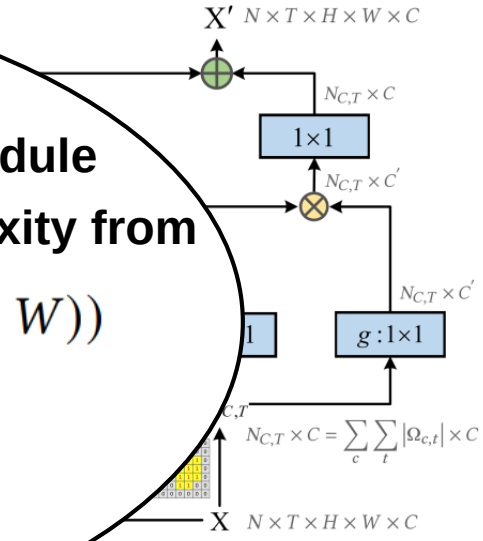
Compared with Non-local, the TSA module greatly reduces the computational complexity from

$$O\left((N\times T\times H\times W)\times(N\times T\times H\times W)\right)$$

to

$$O\left(\left(\sum_c\sum_t|\Omega_{c,t}|\right)\times\left(\sum_c\sum_t|\Omega_{c,t}|\right)\right)$$

Step 2:
Tube Self-attention Operation



TSA mechanism

$$\mathbf{y}_p = \frac{1}{C(\mathbf{x})}\sum_{\forall c}\sum_{\forall t}\sum_{\forall(i,j)\in\Omega_{c,t}} f\left(\mathbf{x}_p, \mathbf{x}_{c,t}(i,j)\right) g(\mathbf{x}_{c,t}(i,j))$$

Normalized Constant $C(\mathbf{x}) = \sum_c\sum_t|\Omega_{c,t}|$

Similarity Func. $f\left(\mathbf{x}_p, \mathbf{x}_{c,t}(i,j)\right) = \theta(\mathbf{x}_p)^T\phi(\mathbf{x}_{c,t}(i,j))$

3. Experimental Results: Validation of Effectiveness

Table 1: Comparison with state-of-the-arts on AQA-7 Dataset.

Method	Diving	Gym Vault	Skiing	Snowboard	Sync. 3m	Sync. 10m	Avg. Corr.
Pose+DCT [27]	0.5300	-	-	-	-	-	-
ST-GCN [41]	0.3286	0.577	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-LSTM [23]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
C3D-SVR [23]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
JRG [22]	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL [33]	0.8099	0.757	0.6538	**0.7109**	0.9166	0.8878	0.8102
NL-Net	0.8296	0.7938	**0.6698**	0.6856	0.9459	0.9294	0.8418
TSA-Net (Ours)	**0.8379**	**0.8004**	0.6657	0.6962	**0.9493**	**0.9334**	**0.8476**

Table 4: Comparison with state-of-the-arts on MTL-AQA.

Method	Avg. Corr.
Pose+DCT [27]	0.2682
C3D-SVR [23]	0.7716
C3D-LSTM [23]	0.8489
C3D-AVG-STL [25]	0.8960
C3D-AVG-MTL [25]	0.9044
MUSDL [33]	0.9273
NL-Net	**0.9422**
TSA-Net	0.9393

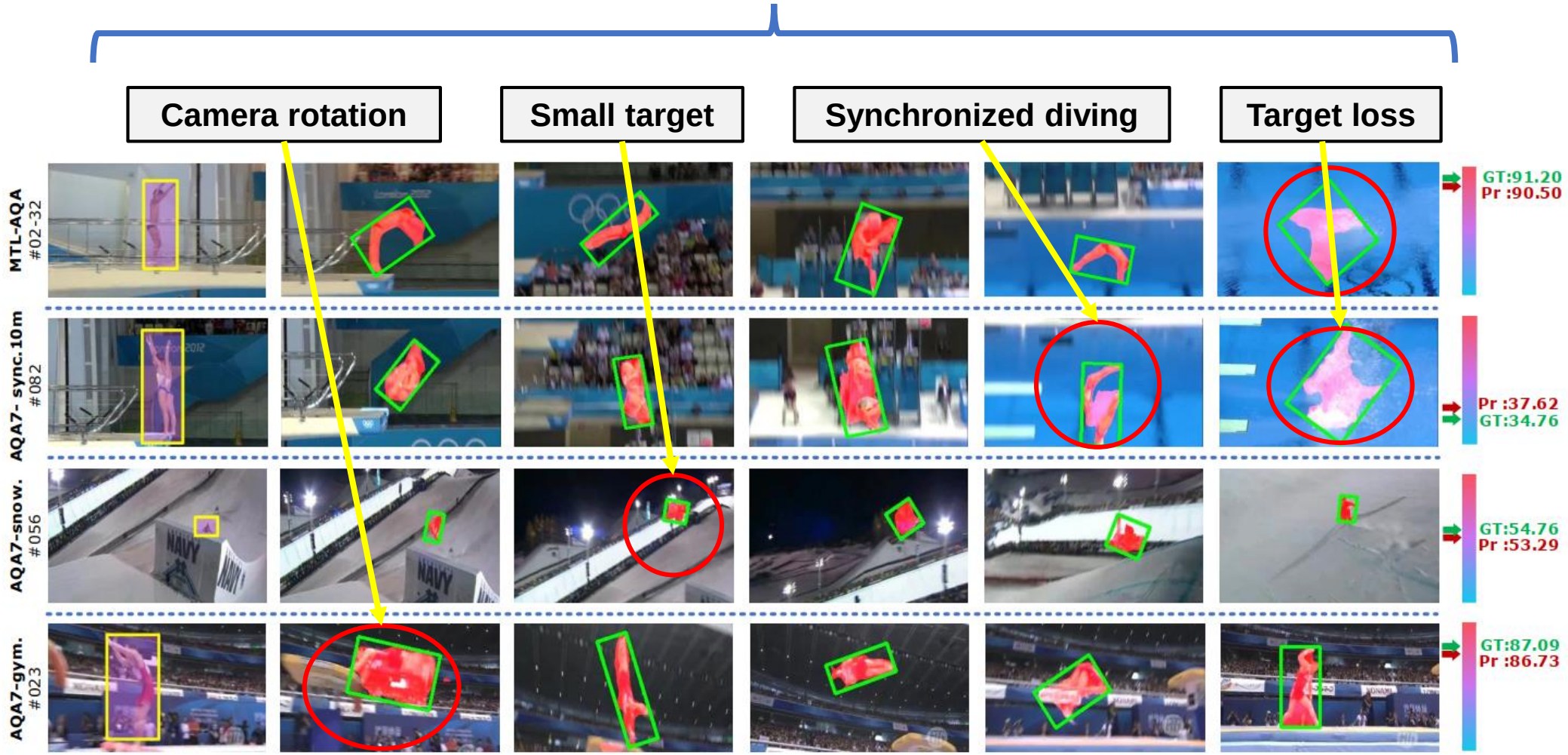
Table 2: Study on different settings of the number of TSA module.

Method	Diving	Gym Vault	Skiing	Snowboard	Sync. 3m	Sync. 10m	Avg. Corr.
TSA-Net	0.8379	0.8004	0.6657	0.6962	**0.9493**	0.9334	0.8476
TSAx2-Net	0.8380	0.7815	**0.6849**	**0.7254**	0.9483	**0.9423**	**0.8526**
TSAx3-Net	**0.8520**	**0.8014**	0.6437	0.6619	0.9331	0.9249	0.8352

TSA-Net achieves SOTA performance on AQA-7 and MTL-AQA.

3. Experimental Results: Validation of Effectiveness

Single object tracking strategy can handle these difficult situations perfectly.



3. Experimental Results: Verification of Efficiency

$$O\left((N \times T \times H \times W) \times (N \times T \times H \times W)\right) \Rightarrow O\left(\left(\sum_c \sum_t |\Omega_{c,t}|\right) \times \left(\sum_c \sum_t |\Omega_{c,t}|\right)\right)$$

Table 3: Comparisons of computational complexity and performance on AQA-7. GFLOPs is adopted to measure the computational cost.

Method	NL-Net	TSA-Net	Comp. Dec.	Corr. Imp.
Diving	2.2G	0.864G	-60.72%	↑0.0083
Gym Vault	2.2G	0.849G	-61.43%	↑0.0066
Skiing	2.2G	0.283G	-87.13%	↓0.0041
Snowboard	2.2G	0.265G	-87.97%	↑0.0106
Sync. 3m	2.2G	0.952G	-56.74%	↑0.0034
Sync. 10m	2.2G	0.919G	-58.24%	↑0.0040
Average	2.2G	0.689G	-68.70%	↑0.0058

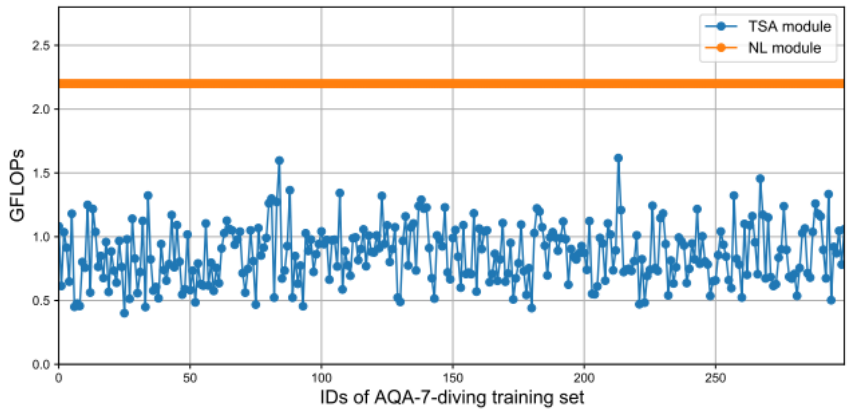
Table 5: Comparisons of computational complexity and performance between NL-Net and the variants of TSA-Net on MTL-AQA.

Method	Sp. Corr.↑	MSE↓	FLOPs↓
NL-Net	**0.9422**	47.83	2.2G
TSA-Net	0.9393	**37.90**	**1.012**G
TSAx2-Net	0.9412	46.51	2.025G
TSAx3-Net	0.9403	47.77	3.037G

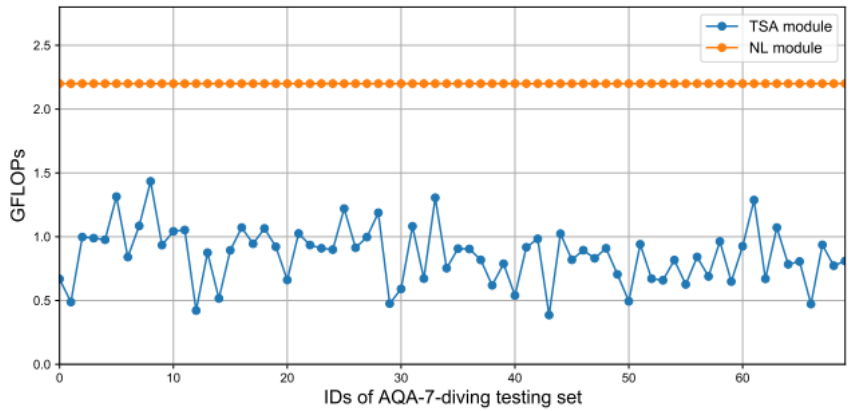
The TSA-Net can obtain better performance while reducing the computational complexity.

3. Experimental Results: Verification of Efficiency

AQA-7 diving

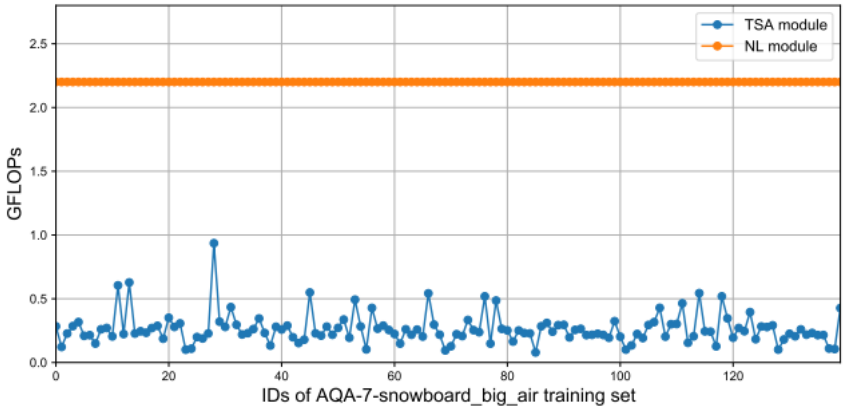


(a) Comparison of computational complexity on AQA-7 *diving* training set.

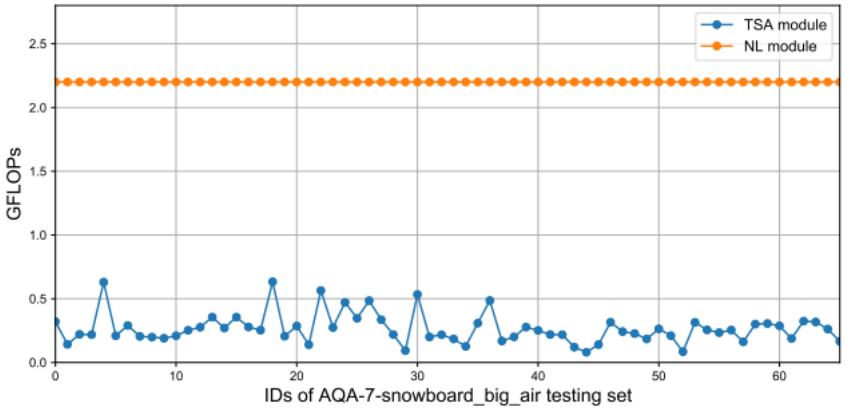


(b) Comparison of computational complexity on AQA-7 *diving* testing set.

AQA-7 snowboard



(g) Comparison of computational complexity on AQA-7 *snowboard* training set.



(h) Comparison of computational complexity on AQA-7 *snowboard* testing set.

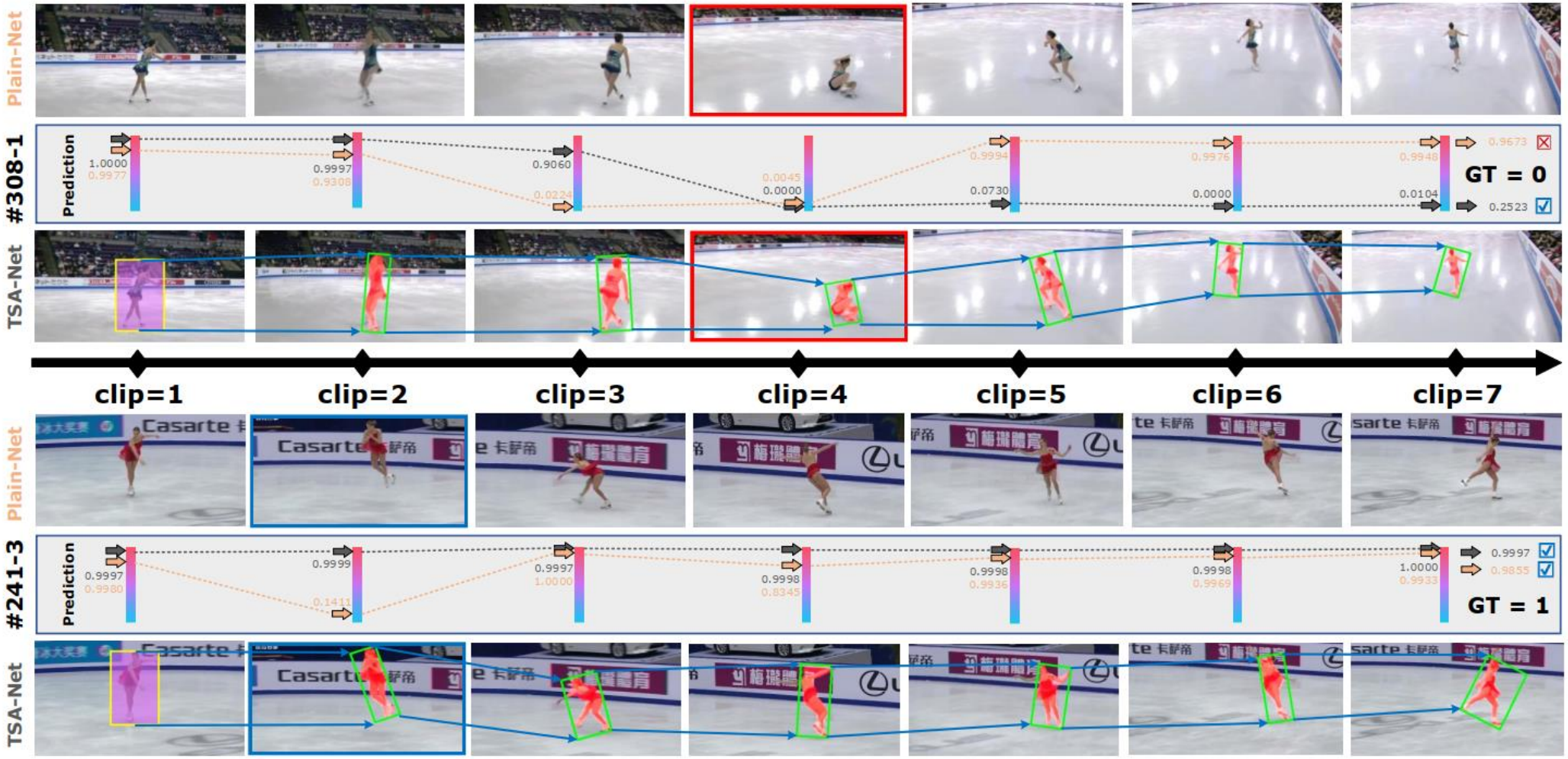
3. Experimental Results: FR-FS

（ **F**all **R**ecognition in **F**igure **S**kating ）

The network with TSA mechanism has higher identification.

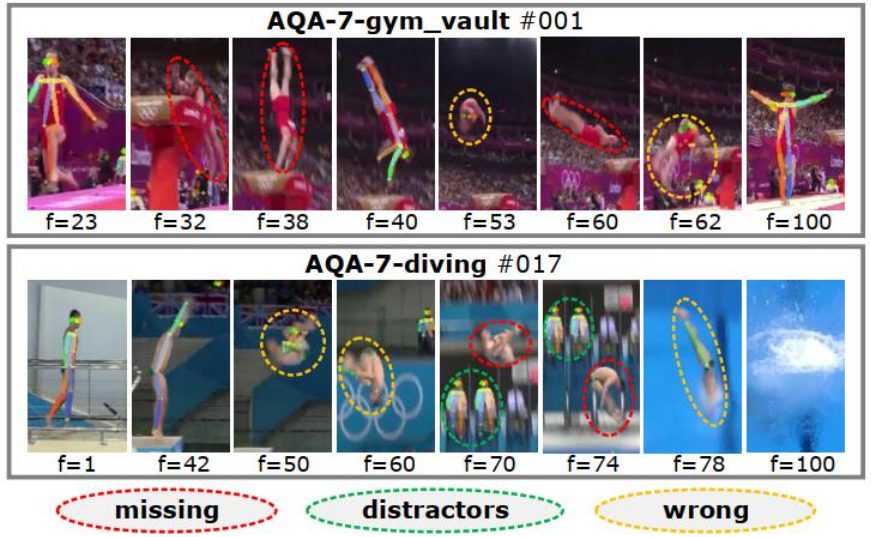
Table 6: Recognition accuracy on FR-FS.

Method	Acc.
Plain-Net	94.23
TSA-Net	**98.56**



4. Conclusions: Discussion and Future Work

1. Why posture information is not adopted in TSA-Net?



Future Work 1：

- **Robust pose estimation algorithm**
- **Higher resolution AQA dataset**

---

2. Invalid feature enhancement caused by small ST-Tube.

Future Work 2：

Adaptive Mechanism of the ST-Tube

4. Conclusions

• We exploit a simple but efficient sparse feature aggregation strategy named Tube Self-Attention (TSA) module.

• We propose an effective and efficient AQA framework named TSA-Net based on TSA module.

• Our approach outperforms state-of-the-arts on the challenging MTL-AQA and AQA-7 datasets and a new proposed dataset named FR-FS.

Thanks!
Q & A