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

1 COMPARISON OF COMPUTATIONAL COMPLEXITY

In the experimental part of the main text, we compare the average computational complexity on AQA-7 and MTL-AQA. The experimental results show that the TSA module can perform feature aggregation more efficiently than the Non-Local (NL) module. In this section, we make a detailed statistical analysis of the computational complexity of the feature enhancement modules on AQA-7 and MTL-AQA. In all figures, the calculation amount of TSA module is marked with blue and NL module with yellow. Since the spatial scales of all videos in this two datasets are resize to 244×244, the size of the feature map output by *Mixed_4e* module is $10 \times 4 \times 14 \times 14$. Therefore, the computational complexity of NL module is fixed at 2.2 GFLOPs.

Analysis on AQA-7. As shown in Figure 1, TSA module can save about 50% of the computation in most items. This is mainly benefited by the application of tube mechanism and sparse self-attention mechanism. Specifically, TSA module has a huge advantage of computing savings in *ski* (Figure 1(e), 1(f)) and *snowboard* (Figure 1(g), 1(h)) in AQA-7 dataset. As analyzed in 4.3 in the main text, this result is caused by the small scale of tracking boxes and spatio-temporal tube (ST-Tube). Small scale limits the effect of feature aggregation performance and leads to performance reduction (0.6657 to 0.6698 of NL-Net on AQA-7 *ski*, and 0.6962 to 0.7109 of USDL on AQA-7 *snowboard*). This result proves the importance of an appropriate ST-Tube and the effectiveness of the TSA module.

Analysis on MTL-AQA. As shown in Figure 2, the amount of calculation of TSA module is evenly distributed around 1.0 GFLOPs in both training set and testing set of MTL-AQA. This is because MTL-AQA only collects diving-related videos. A single data source leads to a stable distribution of computation reduction. We can see that there are several outliers of the TSA module in 2, which are caused by the failure of the VOT tracker. In this scenario, the lost tracking boxes almost cover the whole image, so the computational cost is similar to the NL module.

Through the above quantitative analysis, it can be concluded that the TSA module proposed in this paper can complete the sparse self-attention interaction of features through the construction of ST-Tube, and then generate more representative features with rich spatio-temporal contextual information so as to achieve better action quality assessment performance.

2 VISUALIZATION ANALYSIS

Limited by the length of the main text, we only visually analyze the prediction results of several videos. To explore the generalization performance of our proposed method on all datasets, we select more representative videos from all sub-datasets of AQA-7 and MTL-AQA for visualization analysis. The tracking results and the final prediction scores are shown in Figure 3. Considering the comparability, we select both high score action and low score action. The tracking boxes and predicted scores in Figure 3 show that the VOT tracker can track athletes stably, and TSA-Net can produce reliable results. In the following, the visualization results are explained in groups. Note that all scores are normalized to 0-100 instead of the original score.

MTL-AQA and AQA-7 *diving*: As shown in Figure 3 (a)(b), Because of the high difficulty degree and the small spray, athletes in #02-76 and #021 get high scores. However, in #04-47 and #004, low difficulty degree and big spray result in low scores.

AQA-7 *gym_vault*: In gymnastics, stable landing is very crucial for the final score. Figure 3(c) shows the motion of two athletes. The athlete in #016 lands smoothly at frame 5, resulting in a high score. However, the athlete in #095 makes a mistake during landing and falls to the ground at frame 4 and 5, resulting in a low score.

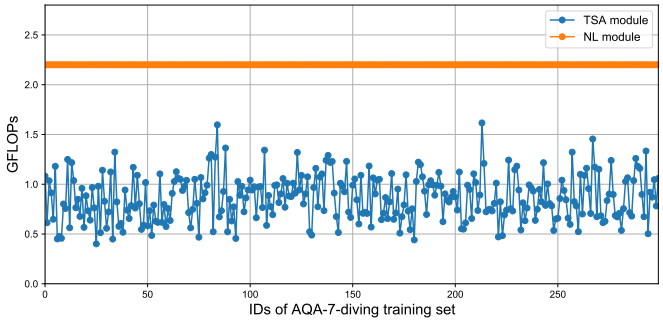
AQA-7 *sky* and AQA-7 *snowboard*: Compared with other sports in AQA-7 and MTL-AQA, the target size in AQA-7 *sky* and AQA-7 *snowboard* changes dramatically, which poses significant challenges to VOT. However, the results Figure 3(d)(e) show that SiamMask can handle these situations. In #120 and #195, the athletes land smoothly and get high scores, while in #103 and #074, the athletes fall in the last two frames, resulting in low scores.

AQA-7 *sync. 3m* and AQA-7 *sync. 10*: As shown in Figure 3(f)(g), due to the appropriate visual angle of videos and the robustness of the VOT tracker, SiamMask can track two athletes in synchronized diving simultaneously. Compared with the athletes in #016 and #011, the athletes in #017 and #042 perform more complex action sequences and produce smaller splashes, so they get higher scores.

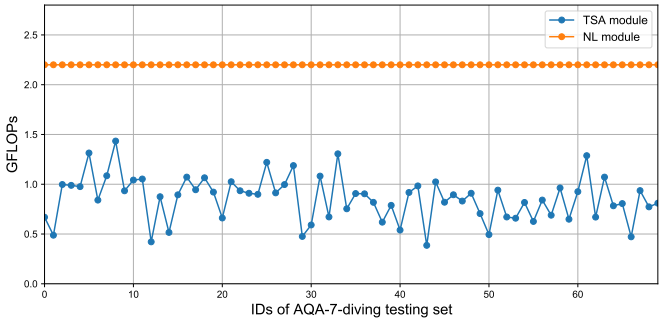
The above results demonstrate that VOT-based strategy plays a vital role in TSA-Net. Based on the accurate ST-Tube generated by stable tracking results, our TSA module achieves efficient feature aggregation through sparse self-attention interaction, and achieves excellent performance.

MM '21, October 20–24, 2021, Virtual Event, China

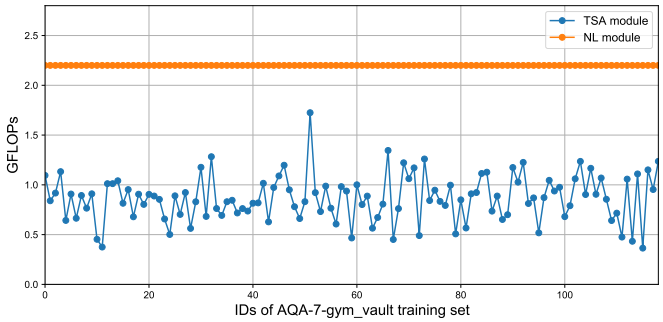
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475438



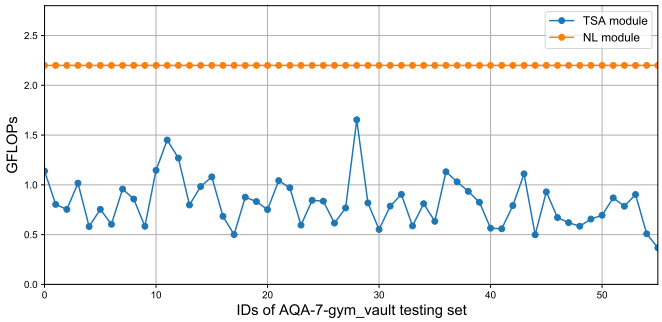
(a) Comparison of computational complexity on AQA-7 *diving* training set.



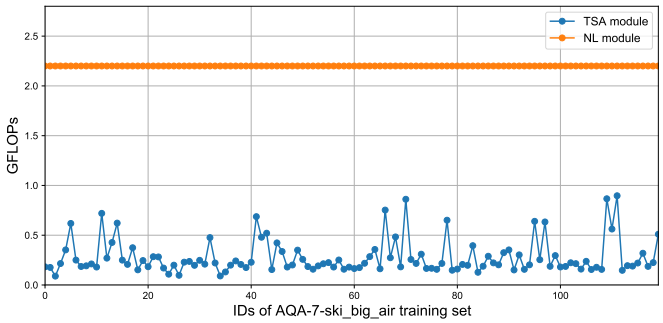
(b) Comparison of computational complexity on AQA-7 *diving* testing set.



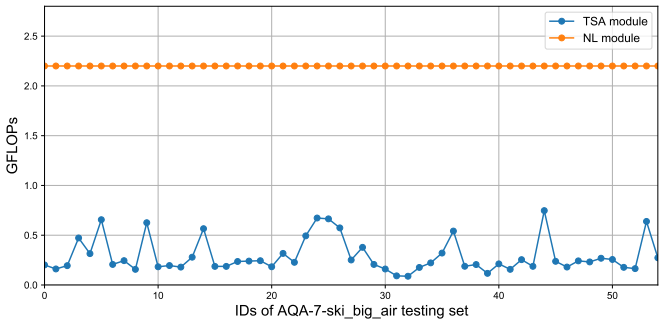
(c) Comparison of computational complexity on AQA-7 *gym_vault* training set.



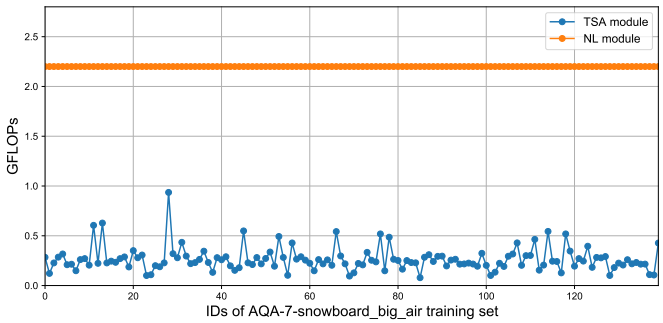
(d) Comparison of computational complexity on AQA-7 *gym_vault* testing set.



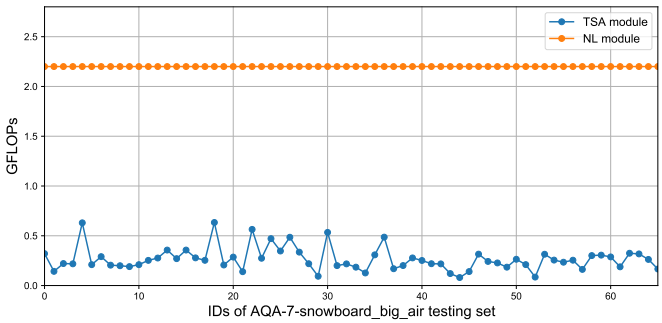
(e) Comparison of computational complexity on AQA-7 *ski* training set.



(f) Comparison of computational complexity on AQA-7 *ski* testing set.

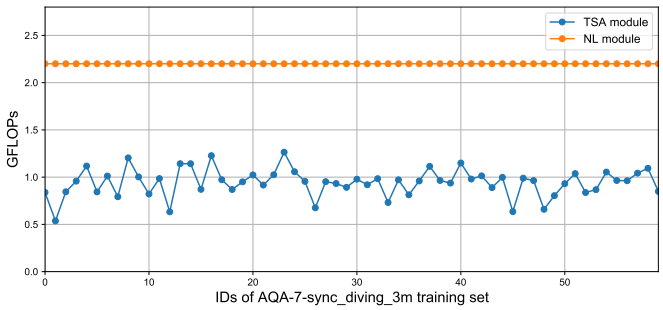


(g) Comparison of computational complexity on AQA-7 *snowboard* training set.

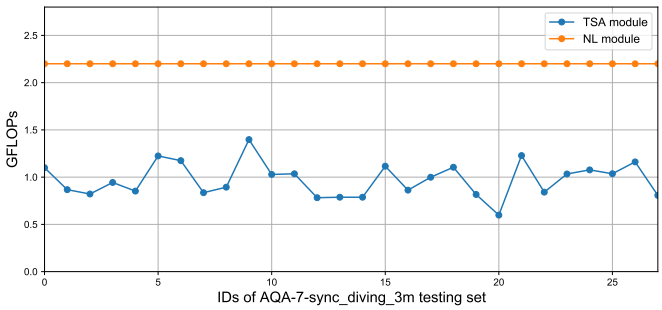


(h) Comparison of computational complexity on AQA-7 *snowboard* testing set.

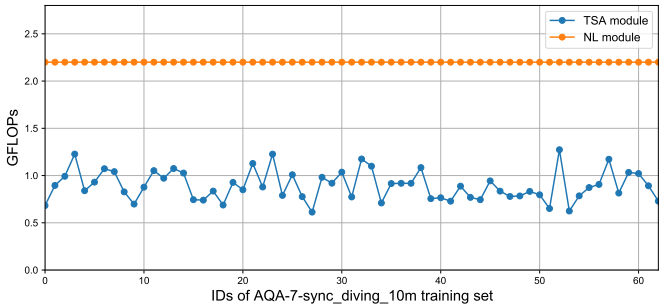
Figure 1: Quantitative analysis of the computational cost of TSA module and NL module on AQA-7.



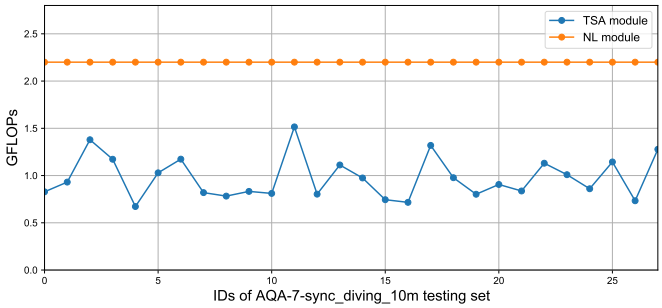
(i) Comparison of computational complexity on AQA-7 *sync. 3m* training set.



(j) Comparison of computational complexity on AQA-7 *sync. 3m* testing set.

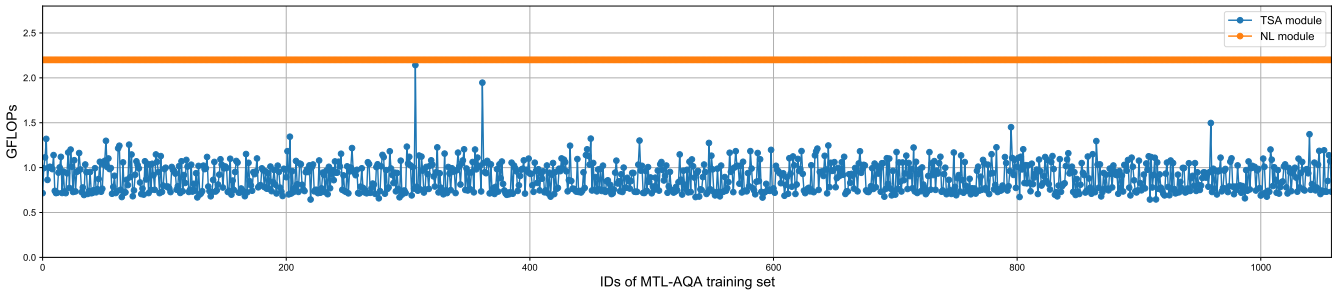


(k) Comparison of computational complexity on AQA-7 *sync. 10m* training set.

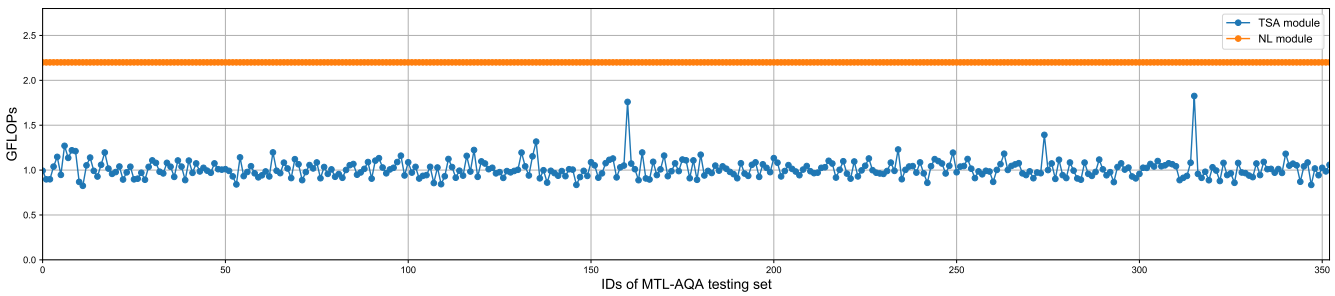


(l) Comparison of computational complexity on AQA-7 *sync. 10m* testing set.

Figure 1: Quantitative analysis of the computational cost of TSA module and NL module on AQA-7.



(a) Comparison of computational complexity on MTL-AQA training set.



(b) Comparison of computational complexity on MTL-AQA testing set.

Figure 2: Quantitative analysis of the computational cost of TSA module and NL module on MTL-AQA.

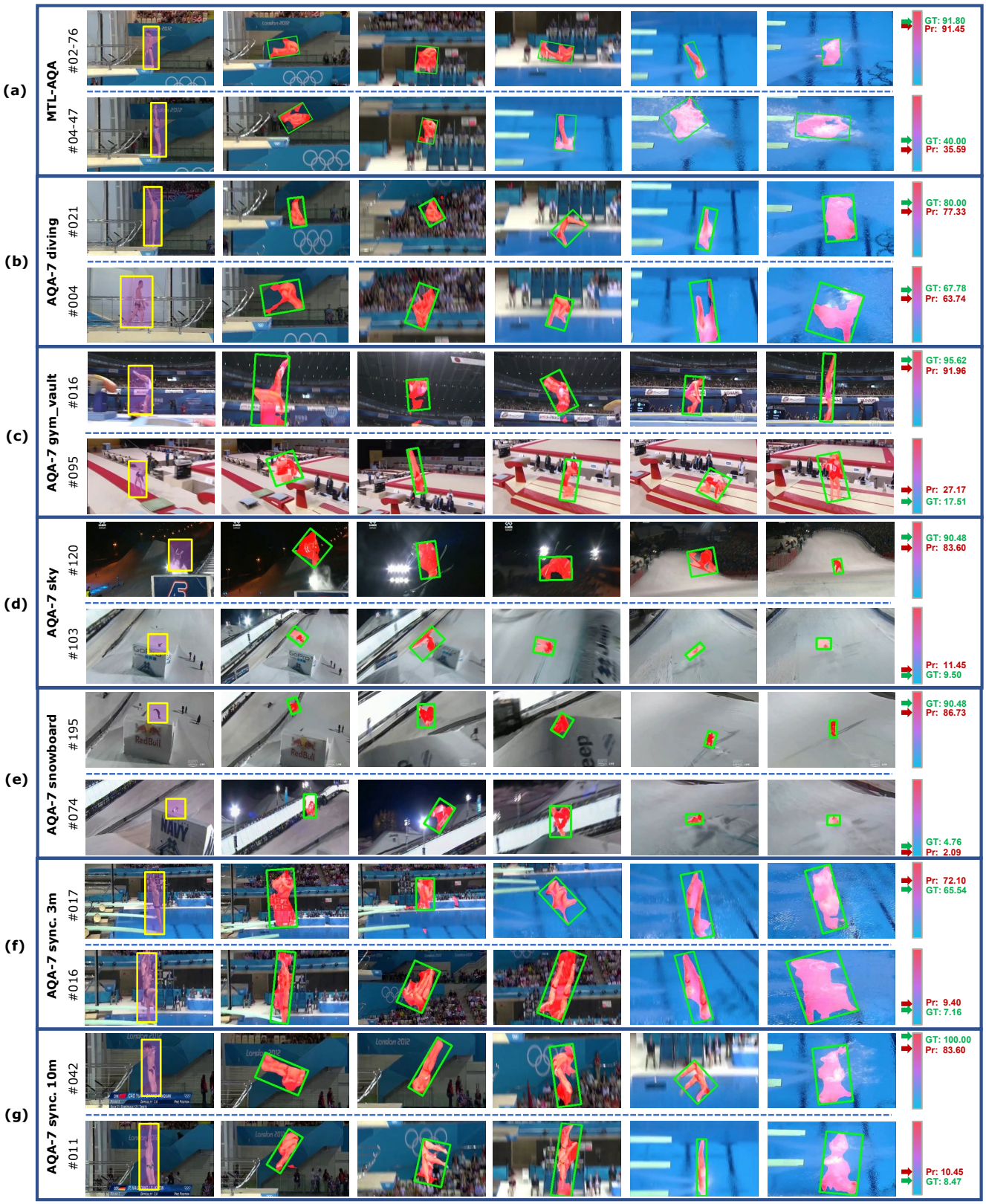


Figure 3: The tracking results and predicted scores of videos selected from AQA-7 and MTL-AQA.