# Adaptive Local-Component-aware Graph Convolutional Network for One-shot Skeleton-based Action Recognition (Supplementary Material)

Anqi Zhu[1], Qiuhong Ke[3], Mingming Gong[2], James Bailey[1]
[1]School of Computing and Information Systems, The University of Melbourne
[2]School of Mathematics and Statistics, The University of Melbourne
[3]Department of Data Science & AI, Monash University
azzh1@student.unimelb.edu.au, qiuhong.ke@monash.edu, {mingming.gong, baileyj}@unimelb.edu.au

## 1. Hyperparameter Selection

During our experiments, we evaluated the selection for two groups of hyperparameters in our ALCA-GCN, which are **the number of spatial/temporal comparing units** for a representation and **the attention configuration** for our Adaptive Dependency Learning module. We use the same episodic training settings and conduct the auxiliary reduction testing according to the one-shot protocol [1] of *NTU-RGB+D 120*, and compare the selection influence with the results in the main paper.

### 1.1. Comparing Unit Division

Table 1 shows the evaluation of using different numbers of comparing units on the spatial or temporal dimension. For spatial-wise division (maintaining temporal comparison on three time sections), except for comparing under the original 4-body-part division or by the average of all body joints (i.e. pure temporal-wise division), we consider the alternative configurations of partitioning body joints into two or six body areas. Fig. 1 provides the detailed division scheme for the mentioned situations. Note that to correspond to the representation scope of the comparing units, the body-part-based local convolution in the encoding network also extracts with the same partitioning scheme to sample each joint's surrounding neighbor features under its new belonging body area. For temporal-wise division (maintaining spatial comparison on four body parts), except for comparing by three time sections or average values (i.e. pure spatial-wise division), we consider the alternative of using five averagely divided time sections.

As shown in the table, for spatial division, increasing the number of body parts performs more refined classification than using average spatial features. Yet, the stability starts to go down when the division is over fragmented to six body parts. Thus we use the four-body-part-based scheme in the main paper. For temporal division, similarly, comparing under multiple time sections performs better than using average temporal features. Even though dividing three time sections provides the second best performance when auxiliary size is 20 or 40, its learning process is more efficient than dividing five sections with the further auxiliary expansion. Note that the performance decay due to the increasing temporal segments is more severe, probably because we used an average division, in which too much segmentation may cause short-term action-critical temporal features forcedly split into different components. Hence we choose to use three time sections in the main paper.

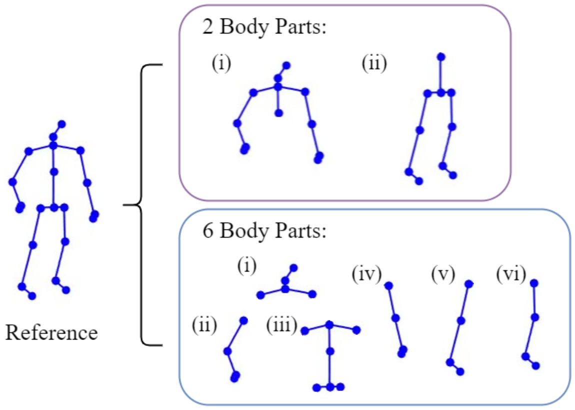


Figure 1. Alternative body partitioning schemes. For two-body-part division, a skeleton is separated into the upper and lower areas. For six-body-part division, the scheme further separates the "hands" and "legs" into individual limb areas.

### 1.2. Adaptive Dependency Learning (ADL)

Table 2 shows the evaluation of using different attention configurations for our ADL module. We examine the learning influence for 20/100-class auxiliary training brought by the changes in the embedding dimension for the query/key matrices and the number of attention heads. The result in-

[1]Pure temporal-wise division.
[2]Pure spatial-wise division.

| # Training Classes | | 20 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|
| Division Dimension | # Segments | | | | | |
| Spatial-wise | 1[1] | 35.5 | 40.2 | 44.2 | 46.5 | 50.0 |
| | 2 | 38.2 | 45.4 | 49.6 | 49.2 | 51.5 |
| | 4 | **38.7** | 46.6 | **51.0** | **53.7** | **57.6** |
| | 6 | 33.9 | **47.0** | 47.5 | 52.5 | 55.3 |
| Temporal-wise | 1[2] | 31.4 | 40.4 | 47.4 | 50.3 | 52.9 |
| | 3 | 38.7 | 46.6 | **51.0** | **53.7** | **57.6** |
| | 5 | **38.9** | **47.8** | 46.7 | 46.3 | 49.7 |

Table 1. Evaluation of different comparing unit division strategies for ALCA-GCN on *NTU-RGB+D 120*.

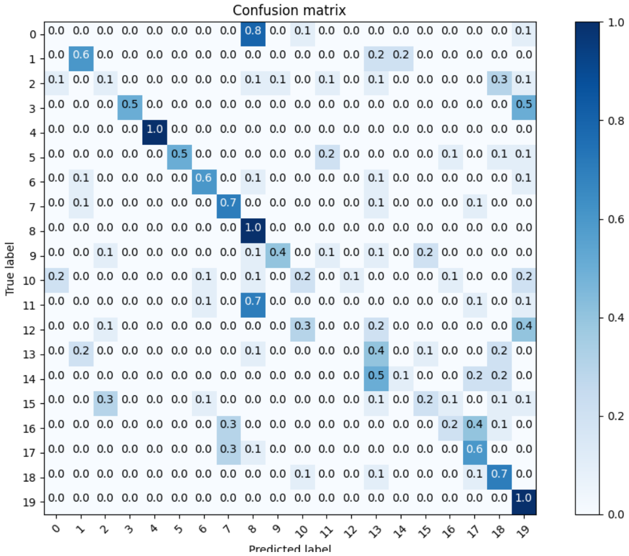
| # Training Classes | | 20 | 100 |
|---|---|---|---|
| Dim of Heads | # Heads | | |
| 64 | 1 | 37.4 | 55.6 |
| 128 | 1 | 37.7 | **58.3** |
| 256 | 1 | **38.7** | 57.6 |
| 128 | 2 | 38.2 | 57.5 |
| 128 | 3 | 37.4 | 57.3 |
| 128 | 4 | 36.5 | 56.9 |

Table 2. Evaluation of different attention configurations for ADL on *NTU-RGB+D 120*.

dicates that using 128 or 256 embedding dimensions provides a comparable best learning efficiency, while applying more attention heads keeps bringing down the performance. This is different from the methods for traditional training, in which having a couple of more attention heads usually helps increase further accuracy. This is probably because the partitioned units decompose the skeleton-level patterns to simple individual component semantics so that single-head learning could already appropriately distribute their action-related importance learnable in the few-shot environment. Eventually, we choose the single-head learning with a 256 embedding dimension for ADL in the main paper experiments.

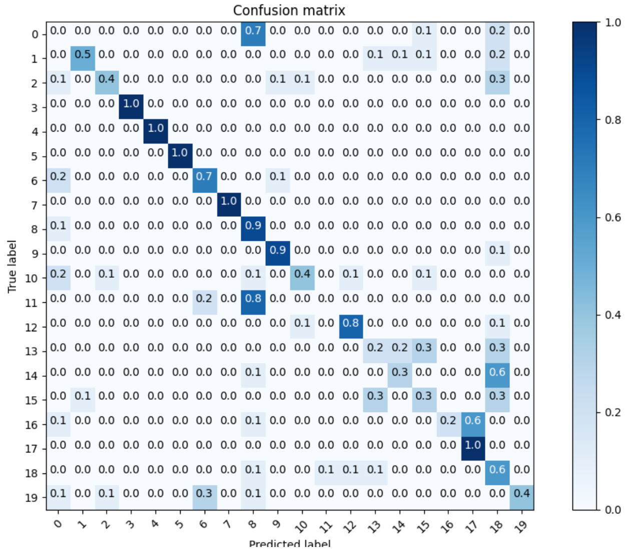
## 2. Visualization of Evaluation Results

We developed quantitative and qualitative visualization for the detailed evaluation learning outcome of our model. Under the full 100-class auxiliary setting with episodic training, we compute the confusion matrices for the performance results of using ST-GCN [2] with global-embedding-based representation (Fig. 2) and ALCA-GCN with action-adaptive local-component-based measurements (Fig. 3). The total evaluation accuracies for the two models are 0.45 and 0.58. We see that our model effectively reaches obvious improvement in the classes of 2, 3, 5, 9, 12, and 17. To separately analyze them, we can conclude that our model achieves a more refined classification on the classes that requires more careful local discrimination around the hands, i.e. class 2 (tear up paper), 3 (take off glasses), 5 (pointing to something with finger), 12 (staple book) and 17 (yawn). On the other hand, the baseline gets confused on these classes by other individual or even mutual classes such as class 8 (use a fan with hand/paper), 18 (grab other person's stuff), 19 (take a photo of other person). Our model also provides more explicit discrimination for the mutual action of class 9 (hugging other person). However, our method mispredicts class 11 (feeling warm) as class 8 (hush) because it requires further refined joint feature discrimination around hand parts, while our method currently focuses on body-part-level matching.



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 |
| 1 | 0.0 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.1 |
| 3 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 0.1 |
| 6 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 |
| 7 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 |
| 8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 9 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.4 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 |
| 10 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.2 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.2 |
| 11 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 |
| 12 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 |
| 13 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 0.0 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 |
| 14 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.1 | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 |
| 15 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.2 | 0.1 | 0.0 | 0.1 | 0.1 |
| 16 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.4 | 0.1 | 0.0 |
| 17 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 |
| 18 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.7 | 0.0 |
| 19 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

Figure 2. Configuration Matrix of the performance results of ST-GCN [2] with global-embedding-based representation on the 20 novel classes in *NTU-RGB+D 120*.



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 |
| 1 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 |
| 2 | 0.1 | 0.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 8 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 |
| 10 | 0.2 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.4 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| 11 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 12 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 |
| 13 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.3 | 0.0 | 0.0 | 0.3 | 0.0 |
| 14 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 |
| 15 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.3 | 0.0 | 0.0 | 0.3 | 0.0 |
| 16 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.6 | 0.0 | 0.0 |
| 17 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 18 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 |
| 19 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 |

Figure 3. Configuration Matrix of the performance results of ALCA-GCN on the 20 novel classes in *NTU-RGB+D 120*.

In Fig. 4 and Fig. 5, we visualize the actual attention prediction generated in a learned ADL (under 100-class auxiliary episodic training) for the support (one-shot class reference) and two query instances of an individual and a mutual action in the novel few-shot classes. For a skeleton sequence, we take a screenshot from each of its three time section, and color each body joint according to the attention score of its belonging body part area under a time section. The body joints with brighter red color have higher attention scores (by percentage). The body joints with grey color reach an attention lower than 2%. We observe that the ADL manages to select intuitively action-critical body parts (from different performers) and time sections as class-agnostic classification focus, and the attention distribution for the instances under the same action class are similar.

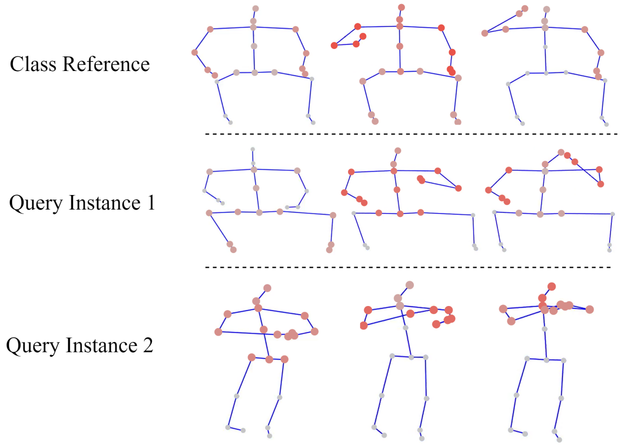


Figure 4. Visualization of the attention prediction for an individual action (sniff) by ADL.

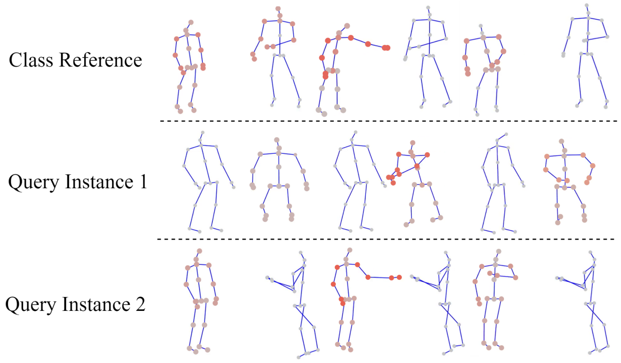


Figure 5. Visualization of the attention prediction for a mutual action (grab other person's stuff) by ADL.

## References

[1] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, 2020.

[2] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, pages 7444–7452. AAAI Press, 2018.