Understanding the Robustness of Skeleton-based Action Recognition under Adversarial Attack-Supplementary Material

He Wang[1,*], Feixiang He[1], Zhexi Peng[2], Tianjia Shao[2,†], Yong-Liang Yang[3], Kun Zhou[2], David Hogg[1]
[1]University of Leeds, UK [2]State Key Lab of CAD&CG, Zhejiang University, China
[3]University of Bath, UK
{h.e.wang, scfh, D.C.Hogg}@leeds.ac.uk, {zhexipeng, tjshao, kunzhou}@zju.edu.cn, y.yang@cs.bath.ac.uk

1. Visual Evaluation of White-box attack

For white-box, we show one motion in NTU attacked using three strategies in Figure 1. 'Wear a shoe' is attacked to 'Sneeze/Cough', 'Writing' and 'punching/Slapping other person' respectively in three attacking strategies. Although the semantics of the action labels are distinctive, the perturbation is hardly noticeable. More examples can be seen in the supplementary video.

2. Perceptual Studies

The 41 subjects are graduate school students and staff, aging between 18 and 37, including 34 males and 7 females. They are trained science and engineering graduate students and researchers, some of whom focus on the research of human motions. During the perceptual study, each subject is given a briefing about the purpose of the study, what experiments should be expected and a few runs to ensure they can comfortably participate the study.

In **Deceitfulness**, the subject was asked 'Please choose the right label that best describes the motion'. In each user study, we randomly choose 100 motions with the ground-truth and after-attack label for 100 trials. In each trial, the video is played for 6 seconds and then the subject is asked to choose which label (between the original and the after-attack one) best describes the motion with no time limit. This is to test whether SMART visually changes the semantics of the motion. In addition, this is also to test whether people can distinguish actions by only observing skeletal motions, because if the subjects could not, their choices would tend to be random and the perceptual study results would not be valid. However, the results conclusively show that the subjects' choices are not random at all and they are able to identify actions. The high success rate of Deceitfulness is, therefore, valid.

Model/Method	SMART	IAA	CIASA
HRNN	100%	98.12%	98.75%
STGCN	99.57%	99.57%	99.56%
2S-AGCN	99.18%	98.77%	98.98%
HRNN	**42.22%**	36.67%	32.22%
STGCN	**90.00%**	87.5%	**90.00%**
2S-AGCN	**80.83%**	35.33%	49.33%

Table 1. Success rate in attack (Upper) and Indistinguishability (Lower). The attack success rate is the best results for SMART, IAA and CIASA.

In **Naturalness**, the subject is asked 'Which motion looks more natural?'. We designed four settings: l2, l2-acc, l2-bone, SMART. l2 is where only the perturbation magnitude of joint locations is used, l2-acc is l2 plus the acceleration profile loss, l2-bone is l2 plus the bone-length loss and SMART is our proposed perceptual loss. In each of the 100 trials, two motions are played together for 6 seconds twice, and then the subject are asked to choose the motion that is more natural, with no time limit.

In **Indistinguishability**, the user is asked 'The left motion is the reference, the right motion might be different from the left. Does the right motion look different?'. Each video is played for 6 seconds then the user is asked to choose if the right motion is a changed version of the left, with no time limit.

3. Comparison

We compare SMART with IAA [1] and CIASA [2]. IAA is designed to attack generic time-series data, not skeletal motions. Since there is no paper published in attacking skeletal motions as far as we know, we include it as a baseline. CIASA is designed for attacking skeletal motions but is only shared on arXiv without code or data. We therefore implemented it ourselves.

In the comparison with other methods, there are two competing factors (attack success and imperceptibility). One can sacrifice one for the other. We need to fix one and compare the other. Given that the success rate of IAA and

[*] https://youtu.be/DeMkN3efp9s
[†] Corresponding author

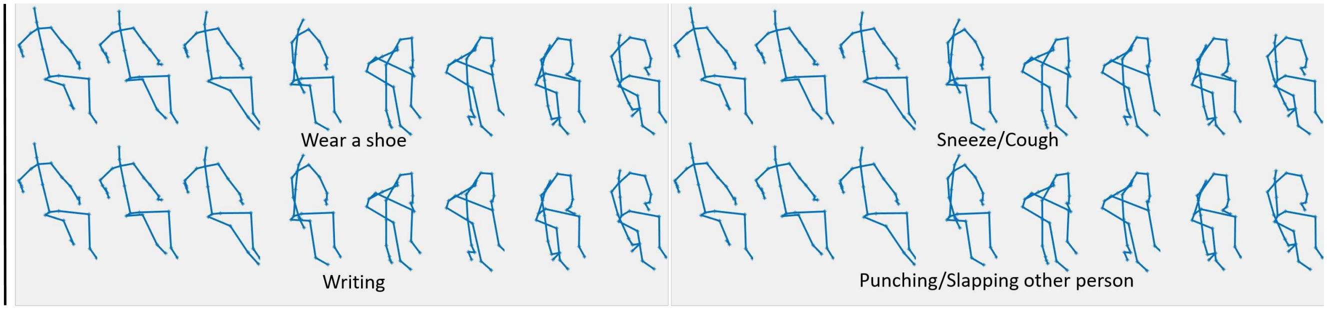

Figure 1. AS-GCN on NTU. Top: Original and AB. Bottom: AB5 and SA. 'Wear a shoe' is the ground-truth label.

CIASA mainly depends on the clipping threshold of the perturbation magnitude, we tune them to achieve similar success rates as SMART. Further, to favor IAA and CIASA, we tune them to achieve slightly lower success rates as this allows us to use smaller clipping threshold values which give better visual quality on IAA and CIASA. Note that this is harsh on SMART.

The results can be found in Table 1. First, we notice that SMART provides the highest imperceptibility rates across different target models. CIASA achieves similar results in STGCN but worse in others. IAA is worst among the three. It is understandable because it does not consider dynamics and generates jittering motions most of the time.

One interesting phenomenon is that the Imperceptibility rates vary across models. STGCN seems to be the most gullible model as all three methods achieve good attack successes while maintaining relatively high imperceptibility. HRNN seems to be the most difficult one in terms of imperceptibility. The reason behind the varying imperceptibility is beyond the scope of this paper. One possible reason could be the structures of the classification boundaries learned by different models and their robustness, which has been studied in image data [3], but not in time series and its correlation with visual indistinguishability. We leave the theoretical analysis for future work.

4. Confusion Matrices and Correlations

Finally, we give detailed confusion matrices and joint correlations for every model on every dataset, to further support our analysis in the vulnerability analysis and the effectiveness of SMART. The joint displacement, displacement-speed and displacement-acceleration correlations in HDM05 are shown in Fig.2-6 for five models. The patterns are very obvious. The higher the speed and acceleration is, the more the joint is attacked. Joint groups with high intra-group correlations are attacked together. We also give detailed confusion matrices of HDM05 in Fig.7-11, where we can see high confusion between motions with significantly different semantics. Similar results can be found in MHAD (Fig.12-21) and NTU (Fig.22-31).

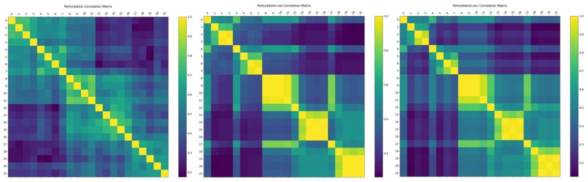

Figure 2. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. HRNN on HDM05.

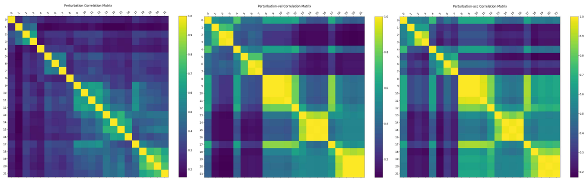

Figure 3. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. STGCN on HDM05.

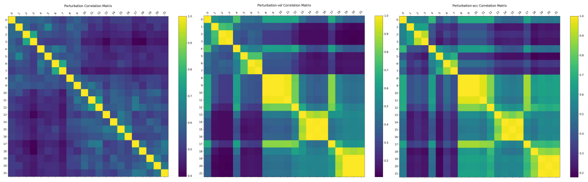

Figure 4. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. ASGCN on HDM05.

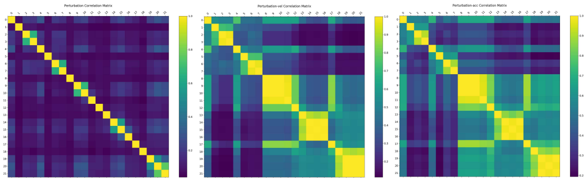

Figure 5. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. DGNN on HDM05.

References

[1] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller. Adversarial attacks on deep neural networks for time series classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 1

[2] Jian Liu, Naveed Akhtar, and Ajmal Mian. Adversarial at-

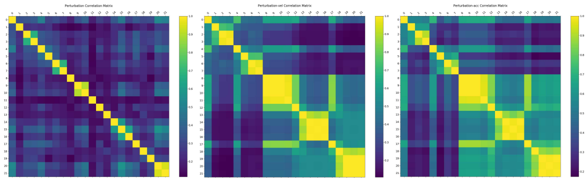

Figure 6. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. 2SAGCN on HDM05.

tack on skeleton-based human action recognition. *arXiv*, abs/1909.06500, 2019. 1

[3] Florian Tramèr, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. The space of transferable adversarial examples. *arXiv*, abs/1704.03453, 2017. 2

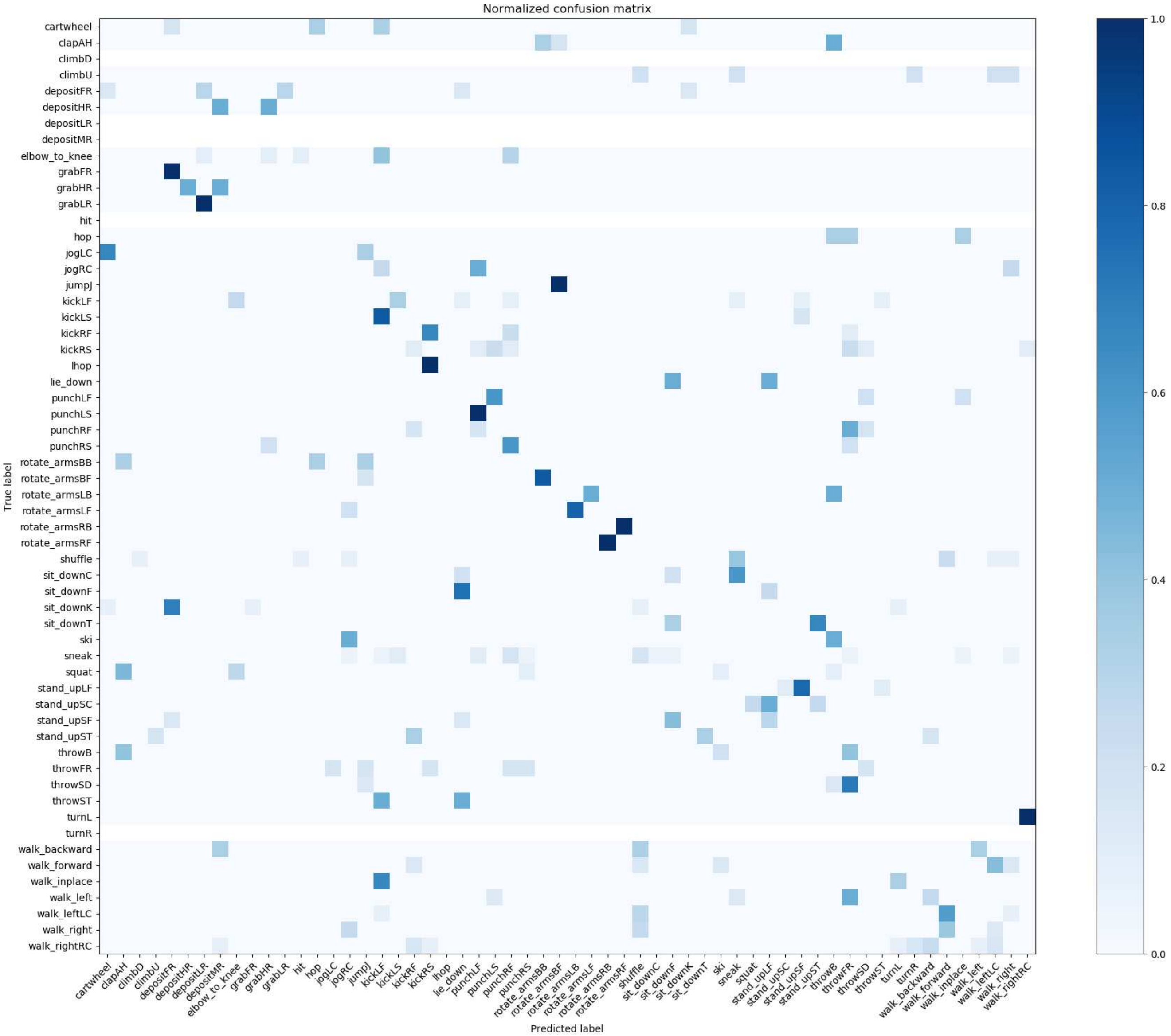

Figure 7. Confusion Matrix. HRNN on HDM05.

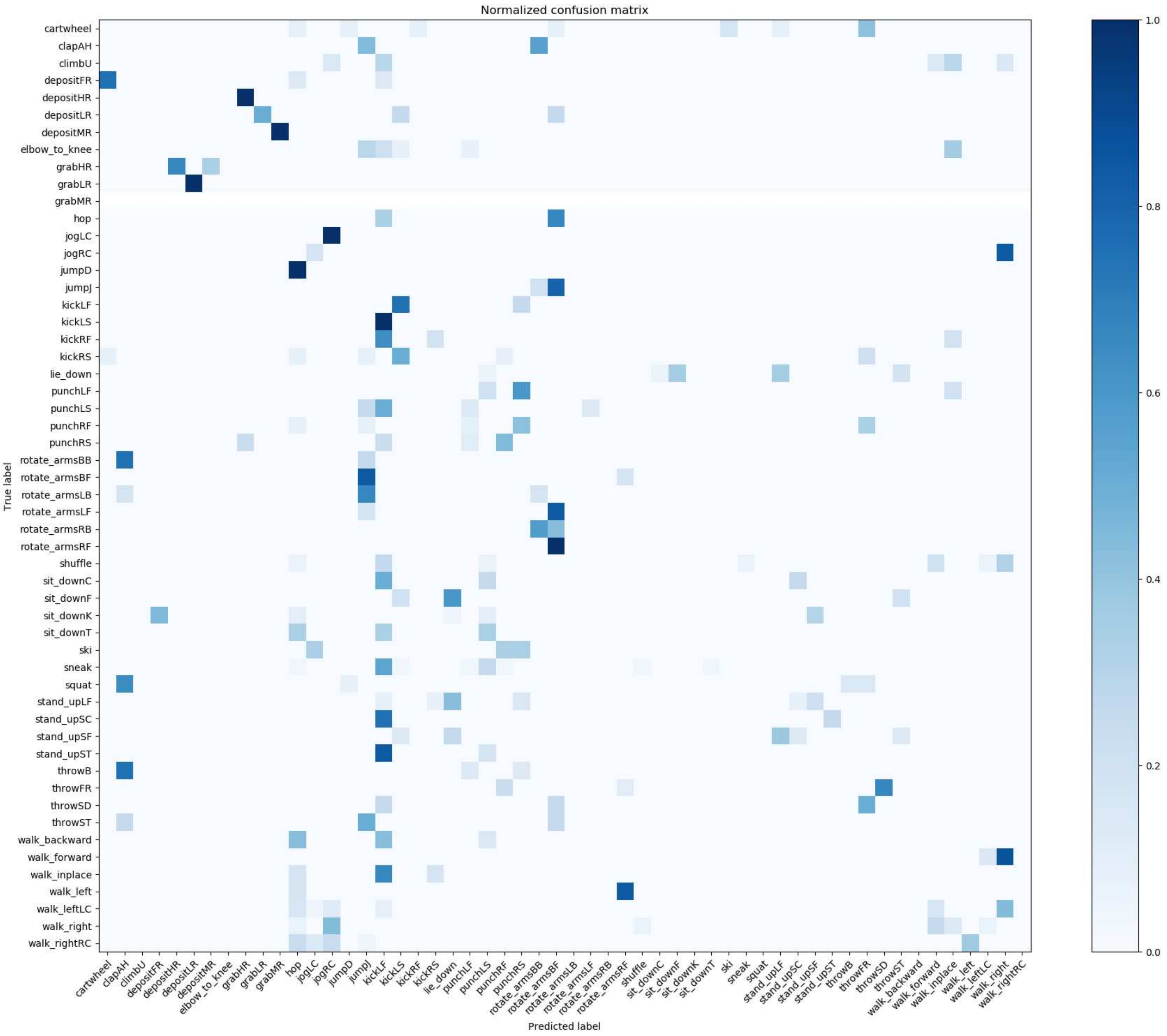

Figure 8. Confusion Matrix. STGCN on HDM05.

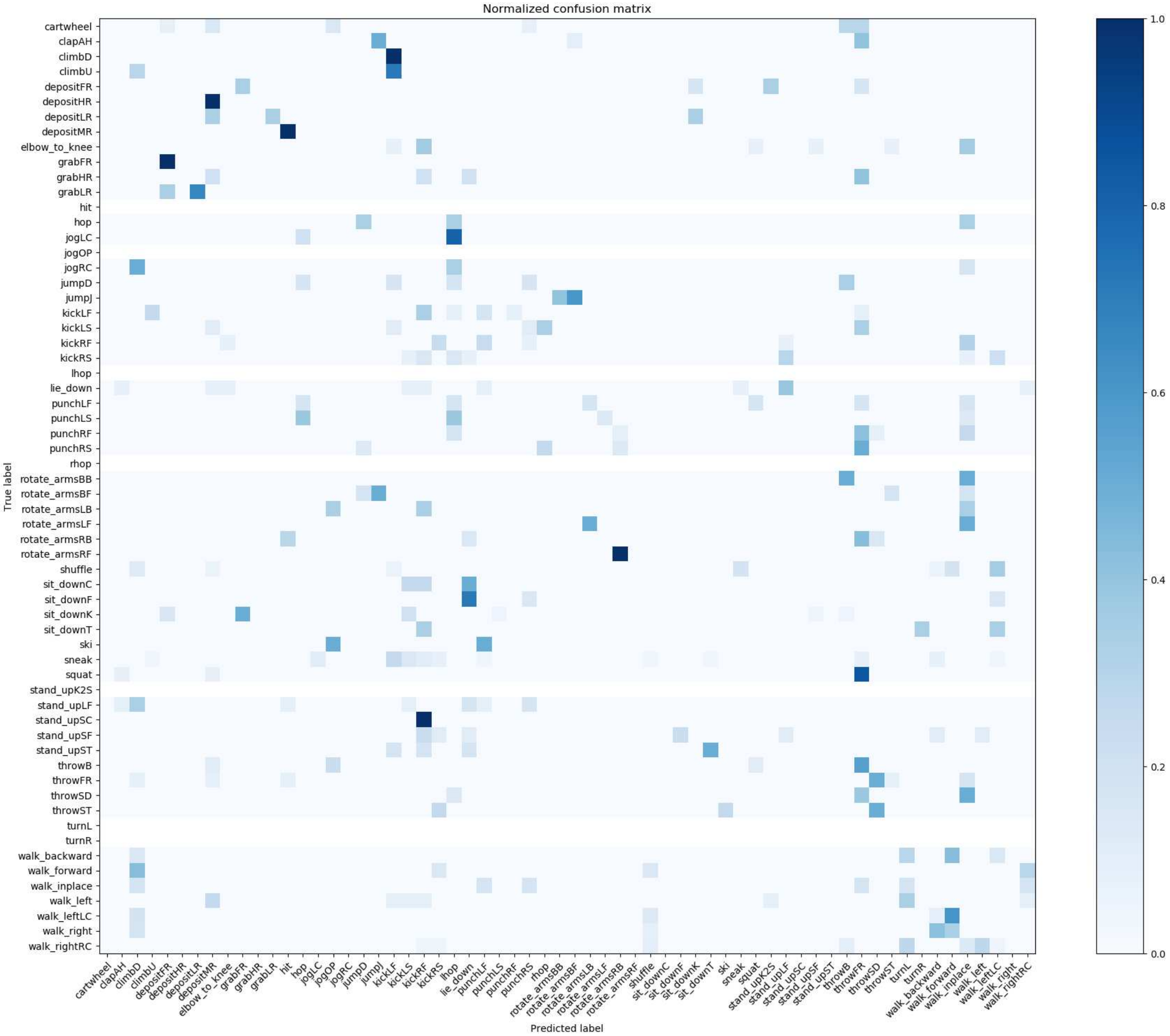

Figure 9. Confusion Matrix. ASGCN on HDM05.

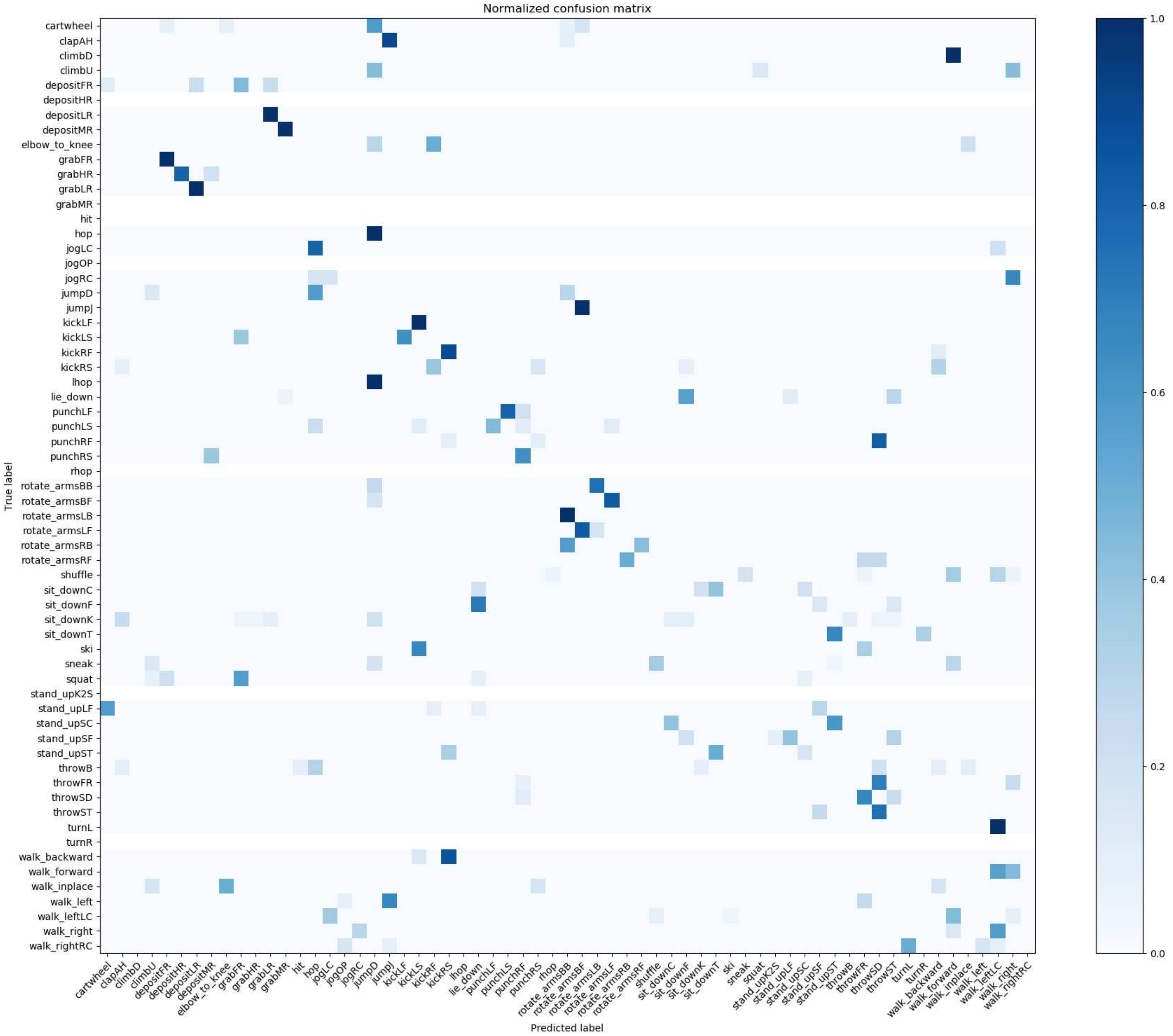

Figure 10. Confusion Matrix. DGNN on HDM05.

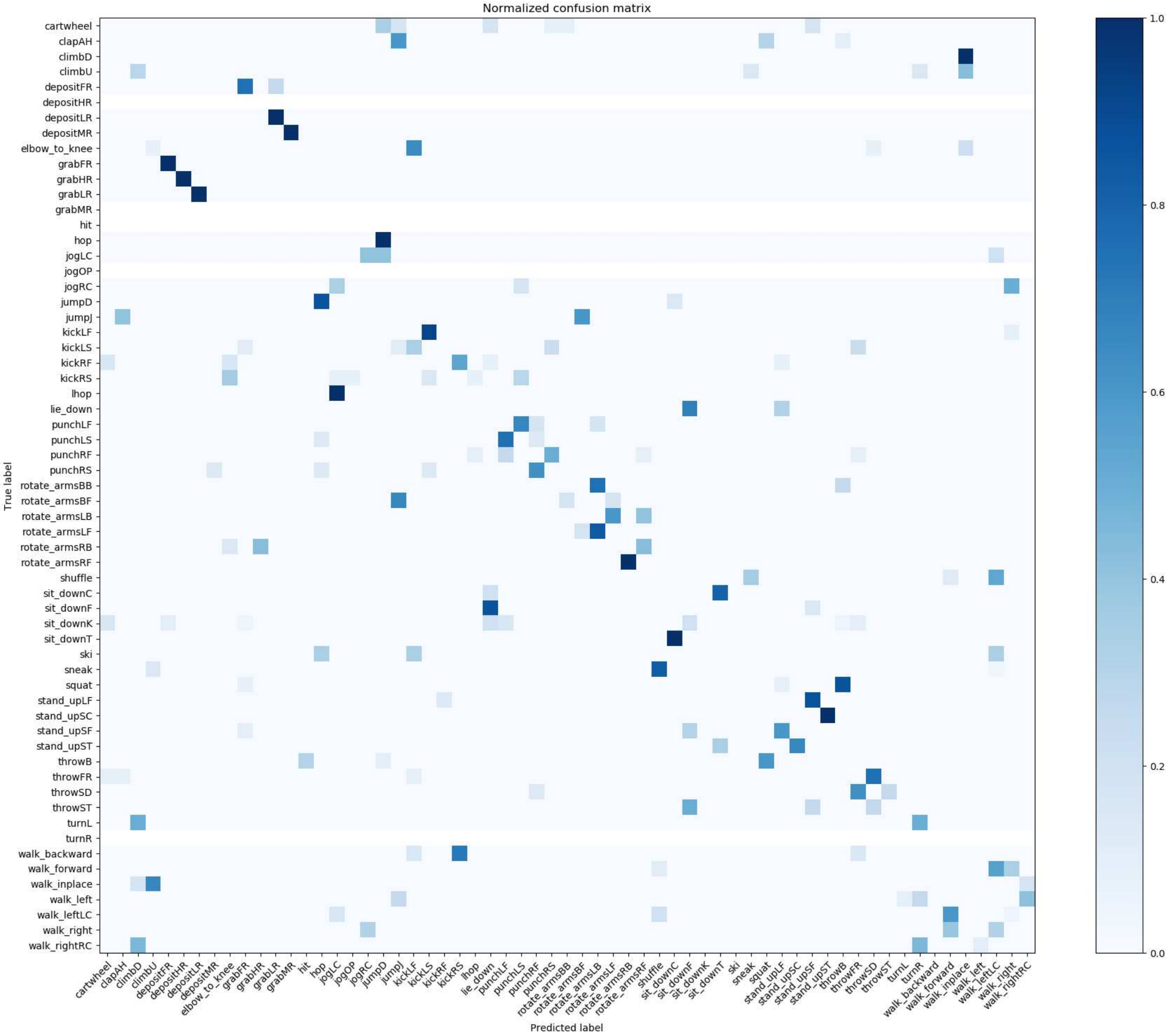

Figure 11. Confusion Matrix. 2SAGCN on HDM05.

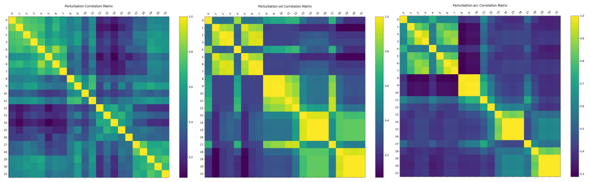

Figure 12. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. HRNN on MHAD.

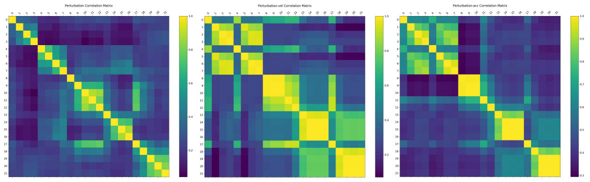

Figure 13. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. STGCN on MHAD.

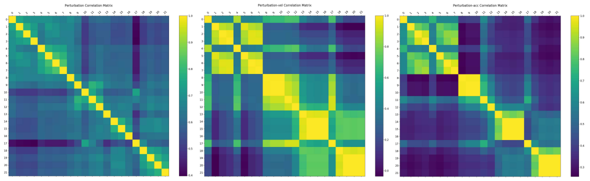

Figure 14. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. ASGCN on MHAD.

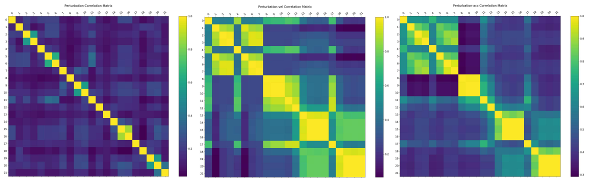

Figure 15. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. DGNN on MHAD.

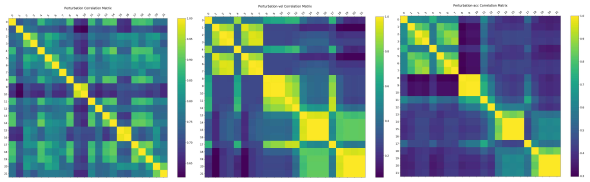

Figure 16. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. 2SAGCN on MHAD.

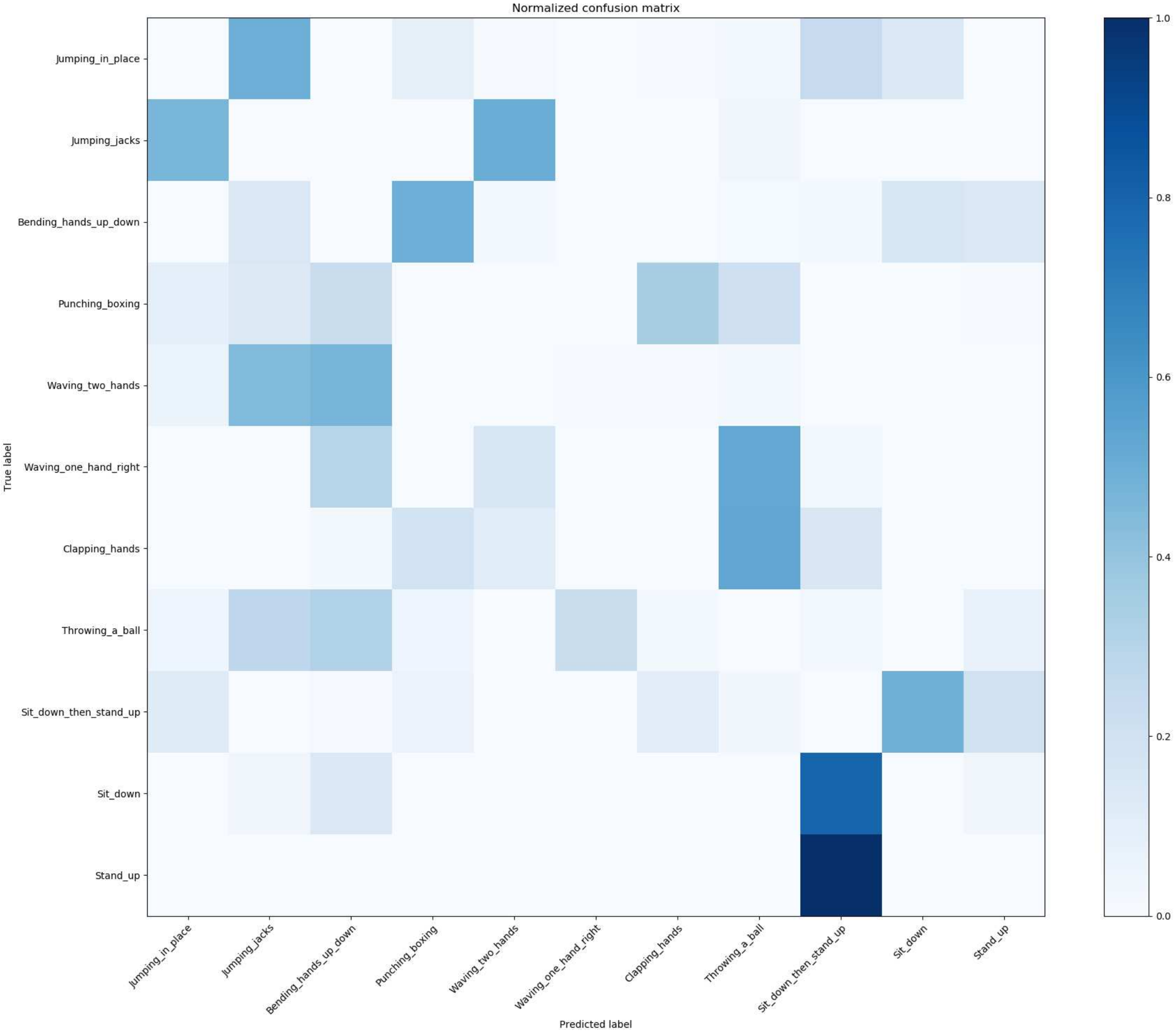

Figure 17. Confusion Matrix. HRNN on MHAD.

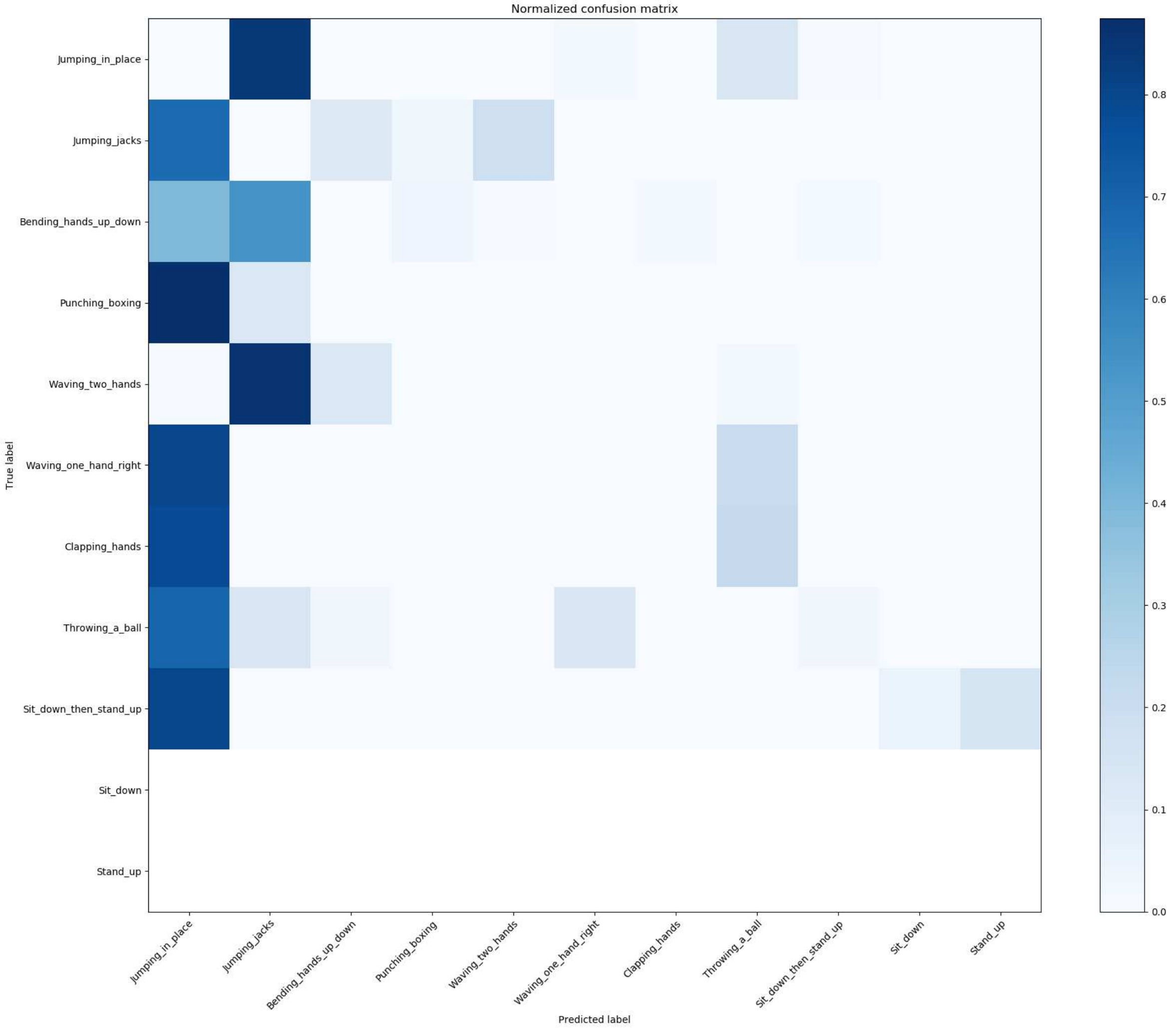

Figure 18. Confusion Matrix. STGCN on MHAD.

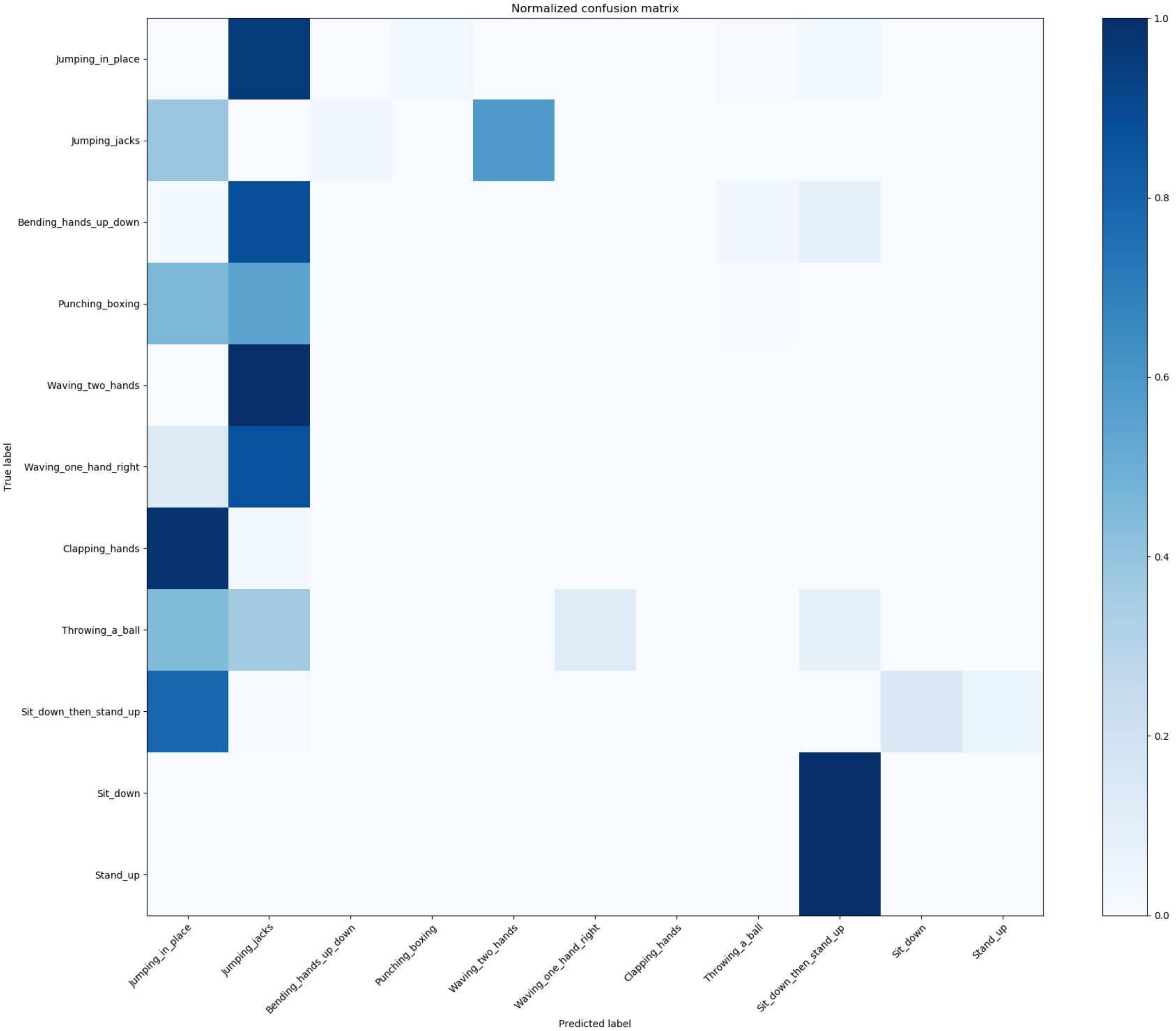

Figure 19. Confusion Matrix. ASGCN on MHAD.

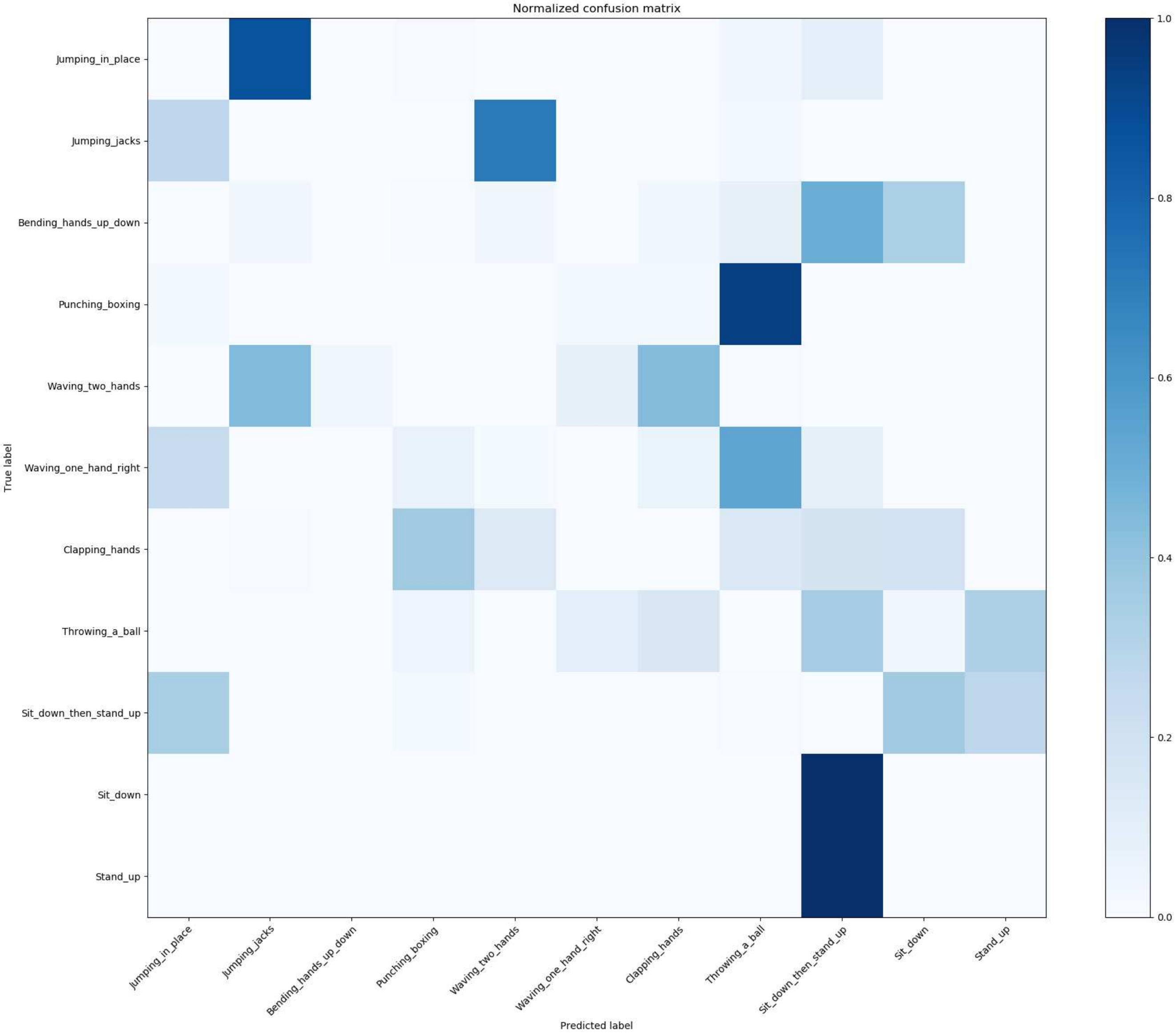

Figure 20. Confusion Matrix. DGNN on MHAD.

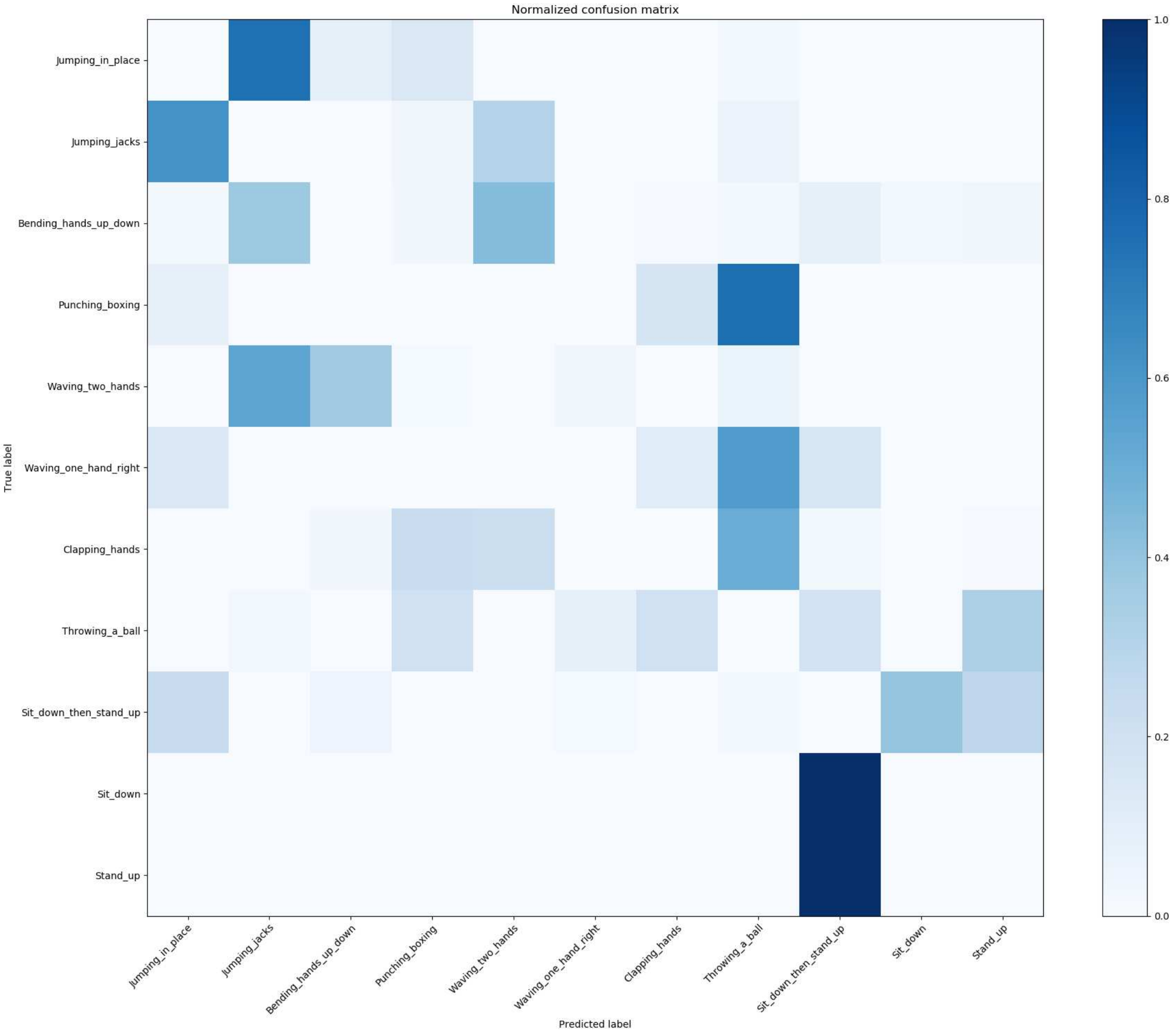

Figure 21. Confusion Matrix. 2SAGCN on MHAD.

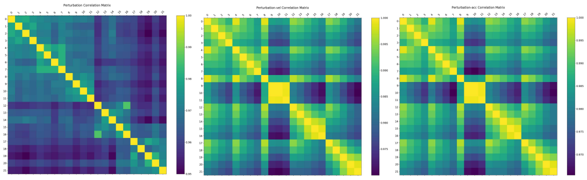

Figure 22. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. HRNN on NTU.

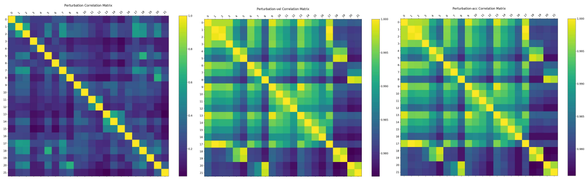

Figure 23. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. STGCN on NTU.

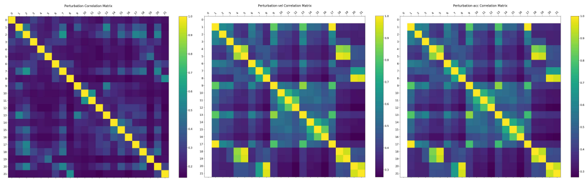

Figure 24. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. ASGCN on NTU.

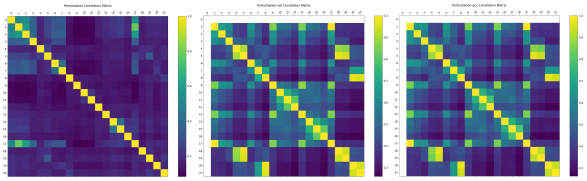

Figure 25. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. DGNN on NTU.

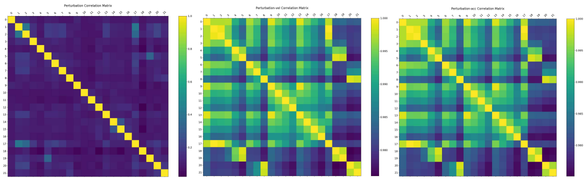

Figure 26. Joint displacement-displacement, displacement-speed and displacement-acceleration correlations. 2SAGCN on NTU.

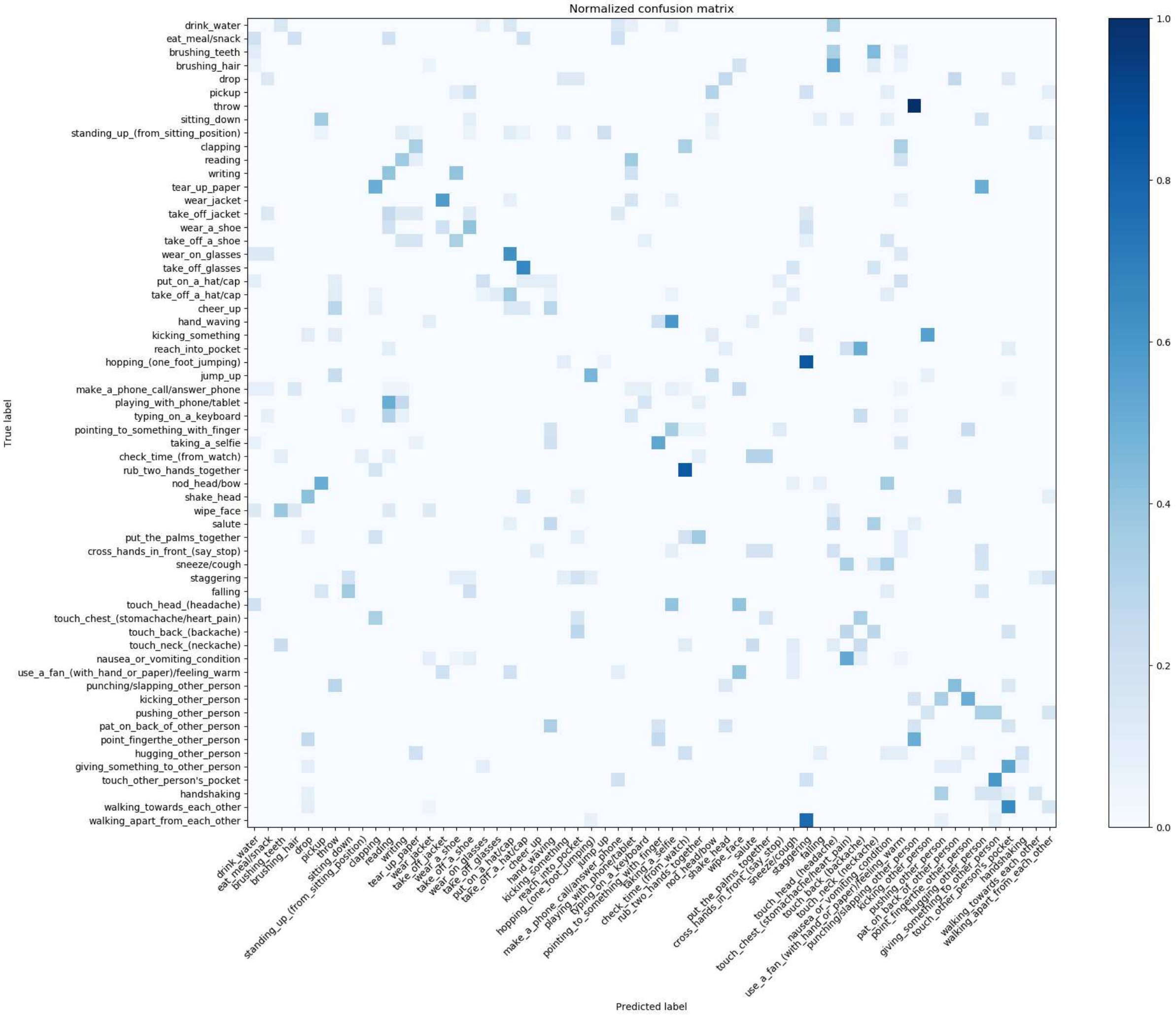

Figure 27. Confusion Matrix. HRNN on NTU.

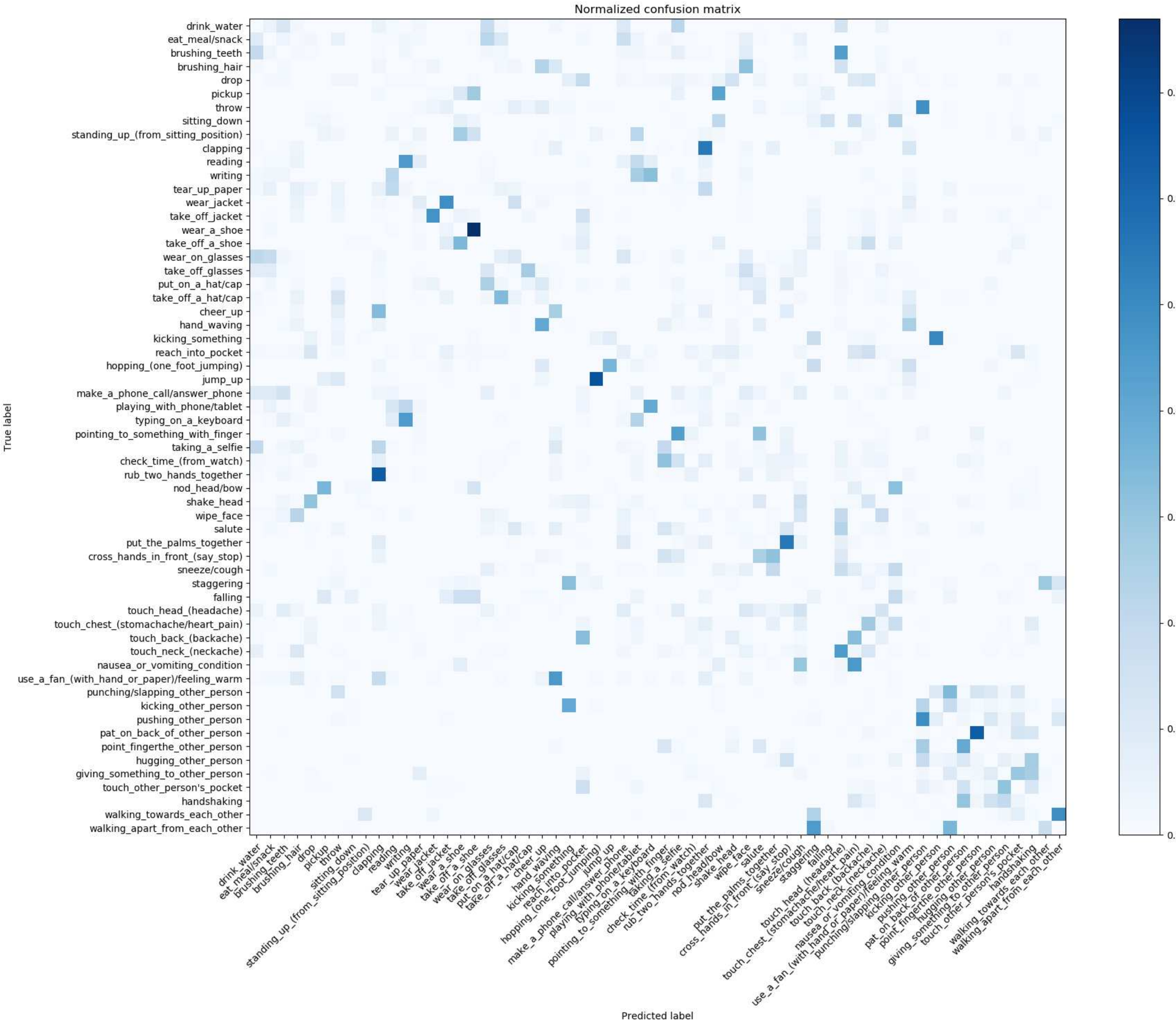

Figure 28. Confusion Matrix. STGCN on NTU.

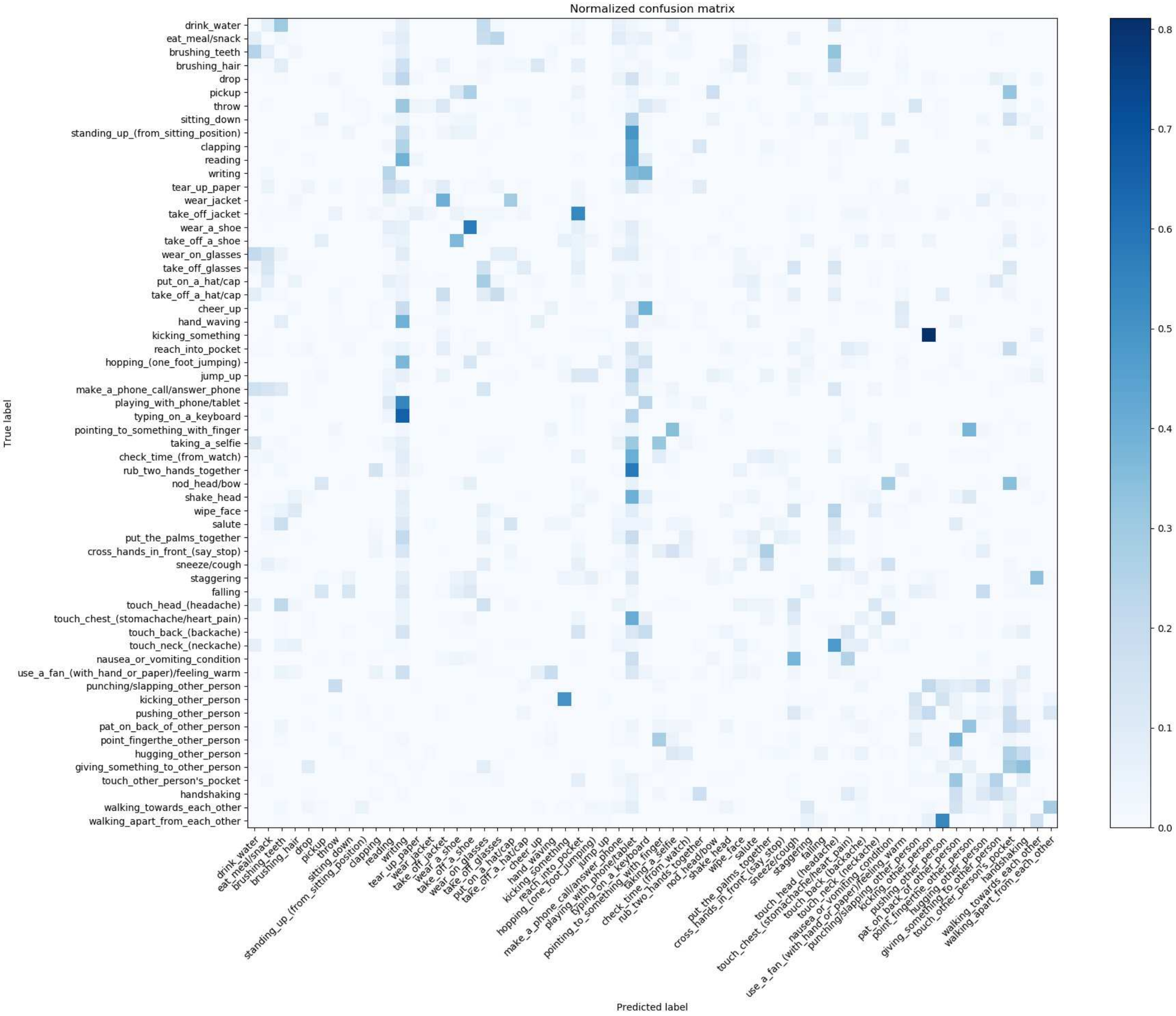

Figure 29. Confusion Matrix. ASGCN on NTU.

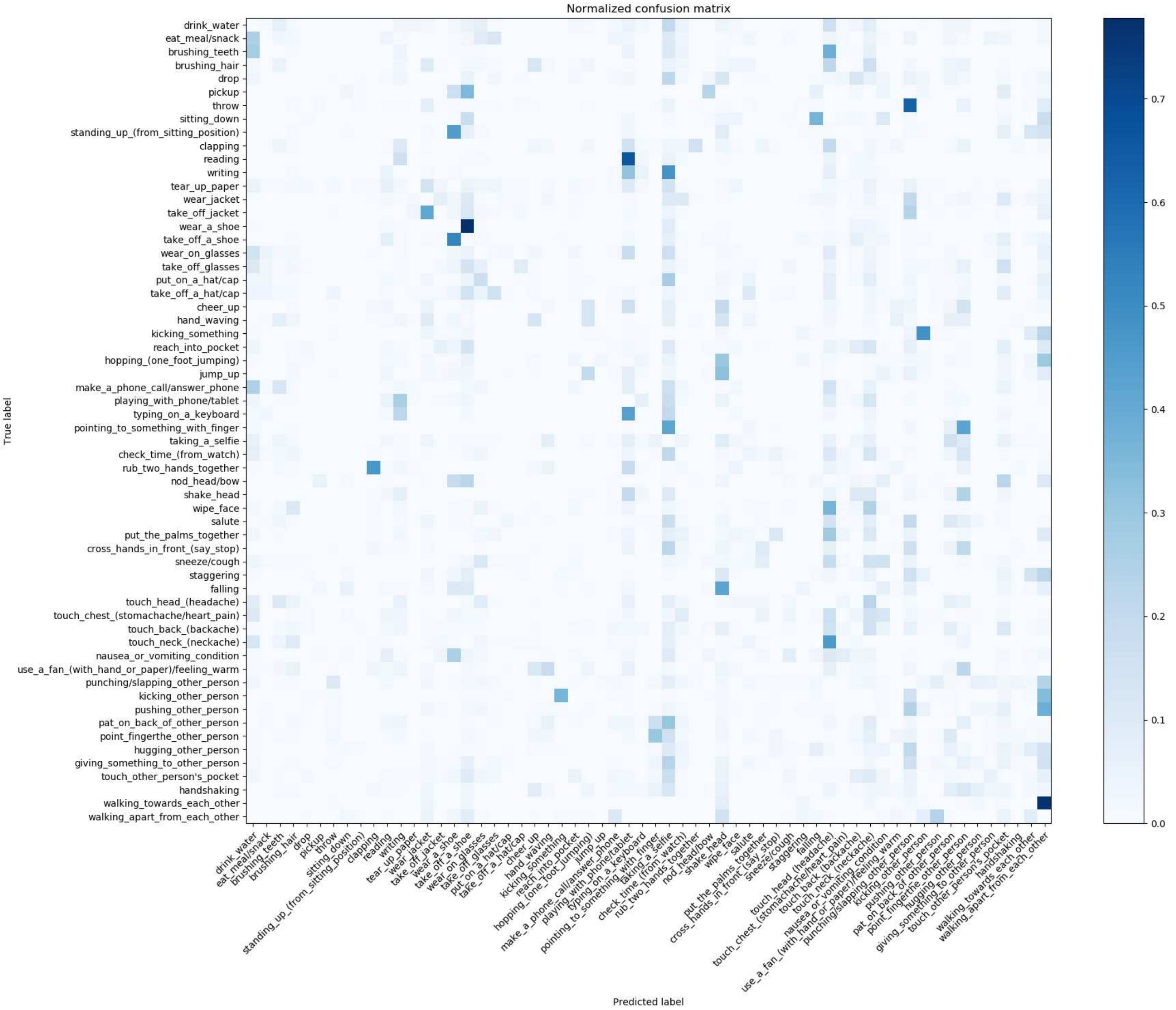

Figure 30. Confusion Matrix. DGNN on NTU.

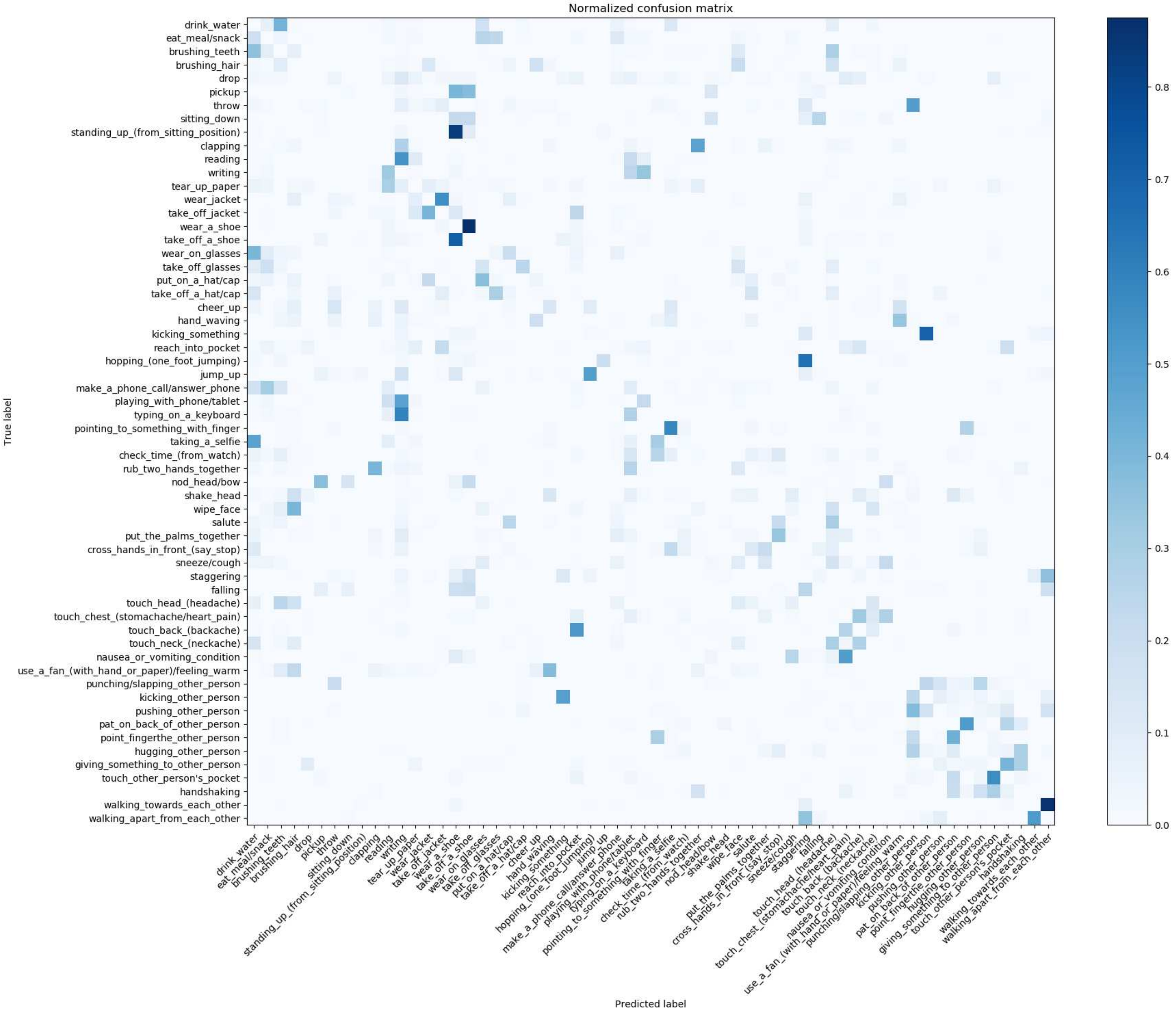

Figure 31. Confusion Matrix. 2SAGCN on NTU.