—— Supplementary Material —— Unsupervised Human Action Recognition with Skeletal Graph Laplacian and Self-Supervised Viewpoints Invariance

Giancarlo Paoletti[1]
giancarlo.paoletti@iit.it
Jacopo Cavazza[1]
jacopo.cavazza@iit.it
Cigdem Beyan[1,2]
cigdem.beyan@unitn.it
Alessio Del Bue[1]
alessio.delbue@iit.it

[1] Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia (IIT)
Genova, Italy

[2] Department of Information Engineering and Computer Science
University of Trento
Trento, Italy

This supplementary material includes the implementation details of our method (Section 1), ablation study performed to investigate how the initialization of Graph Laplacian weight matrix $\mathbf{W}$ affect the performance of proposed AE-L (Section 2), confusion matrices obtained when our AE-L is applied on the testing data of NTU-60 [16] dataset (Section 3) and the comparisons between our AE-L and the state-of-the-art supervised and unsupervised skeleton-based human action recognition methods (Section 4). It is important to note that the results given in Section 4 are equivalent to the results shown in the *main paper Section 4.1*, and the table therein provides the quantitative values as summarised in the figure. We also demonstrate qualitative results comparing the reconstructed skeletons in Section 5.

1 Implementation Details

1.1 Data Pre-processing

We discard the missing time-frames as applied in Predict & Cluster [20]. We normalize each skeleton in terms of bone-length (in the range of [-1, 1]), followed by a regularization of the temporal length of each sample by setting it up to 100 time-frames (cutting frames of longer samples or replicating frames for shorter samples), and finally splitting data *w.r.t.* cross-subject, cross-view and cross-setup settings of benchmarks [11, 16]. This procedure is adapted from Predict & Cluster [20] except from the temporal length of each sample (we choose 100 time-frames instead of 50) and replication of the frames where instead Predict & Cluster [20] uses zero padding for the actions having less than their fixed temporal length.

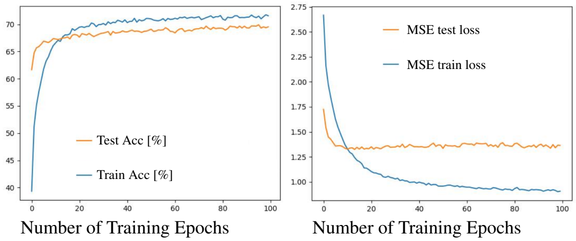


Figure 1: The **learning curves** of our AE model. We provide train/test accuracy values – *left pane* – and MSE loss – *right pane* – of our proposed model trained on NTU-60 in cross-subject protocol. We observe that our model achieves a stable performance at testing time across training epochs: we claim it to be a favorable characteristic, given the plateau in performance across training epochs.

1.2 Model Selection and Hyperparameters

Our model consists of a concatenation of three encoder blocks and three decoder blocks with ReLU activation layers (as depicted in the main paper Section 3.1). Each block is fully-residual, concatenating 1×1, 1×3, and 1×1 padded convolutional layers with stride 1. The decoder blocks use transpose-convolutional layers with the application of batch normalization layers. It is trained for 100 epochs using Adam optimizer with a learning rate of 10^{-3} when the batch size is 128. The hidden representation layer of our model is a fully-connected layer with the size of 2048. We use the features extracted from that layer, which are later given to the classifiers (*i.e.*, 1-NN protocol [20] or Linear Evaluation Protocol [26] (see the main paper Section 3.4 for the definitions). In Figure 1, the learning curves of our model after applying z-normalization are given. As seen in this figure, our model achieves a stable performance at testing time across training epochs. This is an affirmative characteristic, also showing that we are able to learn representations without over-training.

1.3 Implementation of Self-supervised Viewpoints Invariance (SSVI)

For the SSVI experiments (see main paper Section 4.2), we rotated the data along the z-axis. For the fully connected layer of GRL, we apply a sigmoid activation function multiplied by 2π to match the Euler rotation angle. The GRL loss is an $L1$ loss calculated between the original Euler angle of rotations and the predicted Euler angle. Additionally, we use a penalty term included in the GRL loss and a penalty term for the GRL layer (*i.e.*, the alpha value depicted in [4]): both are set to 10^{-3}.

2 The effect of initialization of Graph Laplacian Weight Matrix

In this section, we want to examine how the initialization of Graph Laplacian weight matrix **W** affects our proposed method's performance: AE-L. As mentioned in the main paper, we want to promote the alignment of skeletal joints, connected through a bone (*e.g.*, ***an edge exists if and only if joints are connected***). The reason behind this is to inject the knowledge of skeletal geometry while learning our action representations. This is referred as *fixed W*, a binary and symmetric $n \times n$ skeleton adjacency matrix, including the connectivity between pairs of skeletal joints (as shown in Figure 2 *left*). n is equal to the number of joints of each skeleton (*e.g.*, 25 joints for NTU-60 [16] and NTU-120 [11]). The W_{ij} entries of $\mathbf{W}$ are defined such that $W_{ij} = 1$ if and only if the joints i and j are connected through an edge (in this case a *bone*), 0 otherwise.

A natural alternative to this approach is randomly initializing the weight matrix $\mathbf{W}$ ($n \times n$). Herein we called this *random W*, and its range is $[0, 1]$. We compare the performance of AE-L with *fixed W* (our proposal) against *random W* in Figure 2 rightmost. As seen, *fixed W* achieves better performance than *random W* with a margin of 1.5-2.8% for all settings: cross-subject, cross-view, and cross-setup.

3 Confusion matrices

The confusion matrices for testing our AE-L performance within 1-NN protocol [20] for datasets NTU-60 [16] (cross-subject, cross-view) are given in Figure 3. In the same figure, we also report the accuracy score of each action class in the box-plot form.

3.1 Accuracy-per-action class comparison

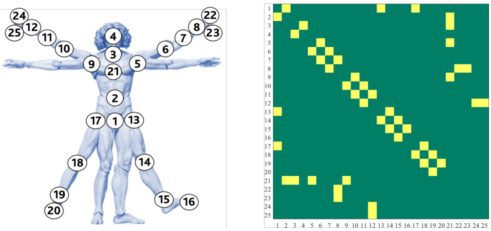
Our AE-L achieves recognition accuracy above 80% for 8 actions (*sitting down*, *standing up from sitting position*, *wearing jacket*, *taking off jacket*, *jumping up*, *falling*, *walking towards each other*, and *walking apart from each other*) in NTU-60 cross-subject setting [16].

In NTU-60 cross-view setting [16], our AE-L performs recognition above 90% accuracy for 13 actions (*throwing*, *sitting down*, *standing up from sitting position*, *wearing jacket*, *taking off jacket*, *cheering up*, *kicking something*, *one foot jumping*, *jumping up*, *salute*, *crossing hands in front*, *staggering*, and *falling*) while class accuracy above 80% is observed for 41 actions.

There are 3 actions: *standing up from sitting position*, *jumping*, and *falling* for which AE-L recognizes with nearly 100% accuracy in cross-view setting of NTU-60 [16].

3.2 Accuracy improvements of Laplacian regularization on cross-subject protocol

As shown in the main paper, AE-L improves the performance of the AE model, showing that Laplacian regularization supplies some advantages.



	NTU-60 [16]		NTU-120 [11]	
	C-Subject	C-View	C-Subject	C-Setup
Fixed W (Ours)	**54.1**	**83.1**	**42.4**	**44.7**
Random W	52.6	80.3	40.9	42.8

Figure 2: **Skeletal Laplacian Regularization.** *Left*: location of the skeletal joints on NTU-60 [16]. *Center*: corresponding adjacency matrix $\mathbf{W}$ (binary). *Right*: ablation table on overall accuracy of AE-L *w.r.t.* using a fixed Laplacian weight matrix, *i.e.*, skeleton adjacency matrix (our proposal) or randomized one.

For NTU-60 cross-subject action classes: *brushing hair, drop, reading, wear on glasses, take off glasses and using a fan* and for NTU-120 cross-subject action classes; *taking off a shoe, wearing on glasses, making a phone call, putting the palms together, patting on back of other person, applying cream on face and kicking backward*; we obtained at least +5% performance gain by involving Laplacian regularization to our AE.

3.3 Accuracy improvements of Laplacian regularization on cross-view and cross-setup protocols

Additionally, our AE-L performs at least +5% better than our AE for NTU-60 cross-view and NTU-120 cross-setup actions.

The NTU60 cross-view actions are: *eating meal, brushing teeth, brushing hair, dropping, clapping, reading, tearing up paper, wearing on glasses, taking off glasses, putting on a hat, taking off a hat, reaching into pocket, hopping, make a phone call, playing with phone, taking a selfie, checking time, rubbing two hands together, wiping face, putting the palms together, sneeze/cough, touching head/chest/back, using a fan, punching other person, patting on back of other person* and *touching other person's pocket*.

In addition, the NTU-120 cross-setup action classes are: *drinking water, eating meal, putting on a hat, taking off a hat, kicking something, making a phone call, putting the palms together, kicking other person, hushing, thumbing up, making victory sign, sniffing, balling up paper, applying cream on face, taking something out of a bag* and *crossing arms*.

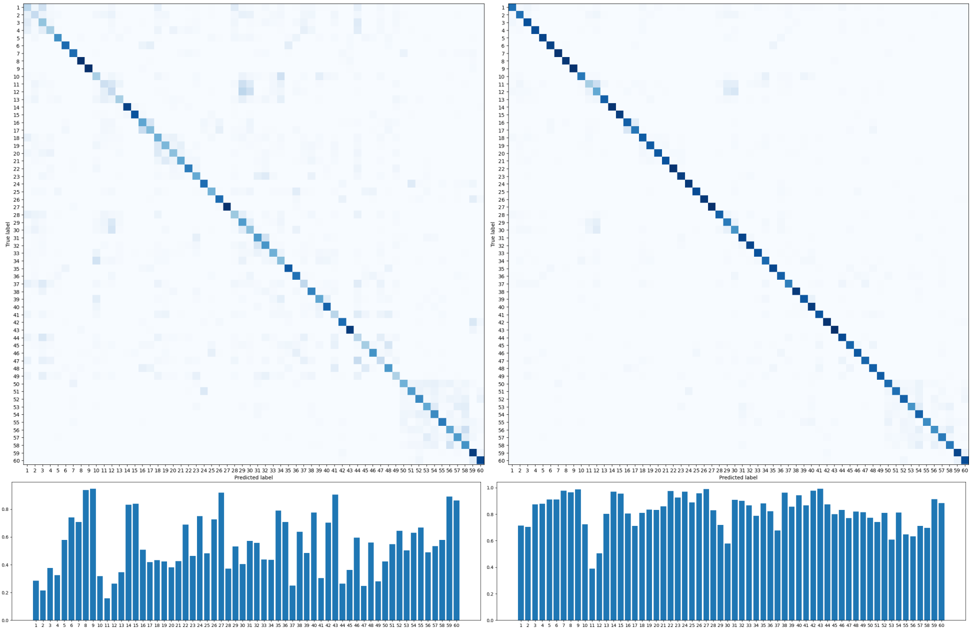


Figure 3: Confusion matrices and the corresponding accuracy scores for each action class obtained when our AE-L is applied with 1-NN protocol on the NTU-60 [16] dataset.

4 Comparisons with state-of-the-art skeleton-based HAR

Performance comparisons between our AE-L and the state-of-the-art unsupervised and supervised skeleton-based HAR methods on NTU-60 dataset [16] are given in Figure 4 (equivalent to Figure 4 in the main paper). The results in Table 1 provide the quantitative values, which are summarised in Figure 4.

Our AE-L outperforms all prior unsupervised skeleton-based approaches on the cross-subject and cross-view settings. Importantly, our learnable representation, although being unsupervised, allows our method to even surpass a few supervised skeleton-based action recognition methods: [1, 6, 10, 13, 15, 24].

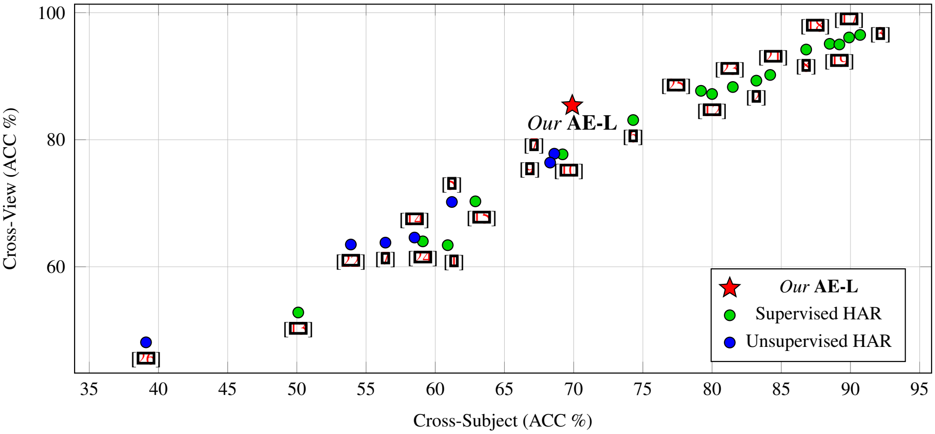


Figure 4: Comparisons between our AE-L and SOTA unsupervised and supervised skeleton-based HAR methods on NTU-60 dataset [16].

	Method	Classifier	Architecture	Cross-Subject	Cross-View
Rahmani et al. (2016) [13]	supervised	*linear SVM*	*3D Spatio-temporal interest points*	<u>50.1</u>	<u>52.8</u>
Cavazza et al. (2019) [1]	supervised	*linear SVM*	*Kernel-approximating random feat maps*	<u>60.9</u>	<u>63.4</u>
Du et al. (2015) [24]	supervised	*softmax*	*Hierarchical RNN*	<u>59.1</u>	<u>64.0</u>
Liu et al. (2016) [10]	supervised	*softmax*	*Spatial Temporal LSTM*	<u>69.2</u>	<u>77.7</u>
Shahroudy et al. (2016) [15]	supervised	*softmax*	*Part-Aware LSTM*	<u>62.9</u>	<u>70.3</u>
Kim et al. (2017) [6]	supervised	*softmax*	*Temporal CNN*	74.3	<u>83.1</u>
Zhang et al. (2017) [25]	supervised	*softmax*	*View-Adaptive LSTM*	79.2	87.7
Liu et al. (2017) [12]	supervised	*softmax*	*Multi-stream CNN*	80.0	87.2
Liu et al. (2017) [2]	supervised	*softmax*	*CNN*	83.2	89.3
Yan et al. (2018) [23]	supervised	*softmax*	*Spatio-Temporal GCN*	81.5	88.3
Wen et al. (2019) [21]	supervised	*softmax*	*Motif GCN + Variable Temporal Dense Block*	84.2	90.2
Li et al. (2019) [8]	supervised	*softmax*	*Actional-structural GCN*	86.8	94.2
Shi et al. (2019) [18]	supervised	*softmax*	*2-stream Adaptive GCN*	88.5	95.1
Si et al. (2019) [19]	supervised	*softmax*	*Attention GCN+LSTM*	89.2	95.0
Shi et al. (2019) [17]	supervised	*softmax*	*Directed Graph Neural Networks*	89.9	96.1
Cheng et al. (2020) [3]	supervised	*softmax*	*Shift GCN*	**90.7**	**96.5**
Holden et al. (2015) [5]	unsupervised	*linear classifier*	*Denoising AE*	<u>61.2</u>	<u>70.2</u>
Zheng et al. (2018) [26]	unsupervised	*linear classifier*	*Adversarial GRU-AE*	<u>39.1</u>	<u>48.1</u>
Kundu et al. (2018) [7]	unsupervised	*linear classifier*	*Variational-AE + poseRNN*	<u>56.4</u>	<u>63.8</u>
Kundu et al. (2018) [7]	unsupervised	*linear classifier*	*Encoder-GAN + poseRNN*	<u>68.6</u>	<u>77.8</u>
Xu et al. (2020) [22]	unsupervised	*linear classifier*	*Contrastive-AE*	<u>53.9</u>	<u>63.5</u>
Rao et al. (2020) [14]	unsupervised	*linear classifier*	*Contrastive-AE*	<u>58.5</u>	<u>64.6</u>
Li et al. (2021) [9]	unsupervised	*linear classifier*	*Contrastive-GCN*	<u>68.3</u>	<u>76.4</u>
Our AE-L	unsupervised	*linear classifier*	*Regularized convolutional, residual AE*	**69.9**	**85.4**

Table 1: Performance comparisons between our AE-L and the state-of-the-art supervised and unsupervised skeleton-based HAR methods on NTU-60 dataset [16] in terms of accuracy (%). The results that our AE-L surpasses are underlined. The best results for the supervised and unsupervised methods individually are shown in black.

5 Visualization of the reconstructed skeletons

Figures 5 and 6 present the visualizations of the reconstructed skeletons obtained by applying our models (AE and AE-L). *Blue* skeletons represent the input data (of action "Drink Water" for Figure 5 and "Standing Up" for Figure 6), *red* and *green* skeletons are reconstructed by AE and AE-L, respectively. In these examples, while the effectiveness of our models is the same, in other words, AE and AE-L both classify the actions correctly, the AE-L makes the reconstructed skeletons smoother compared to AE.

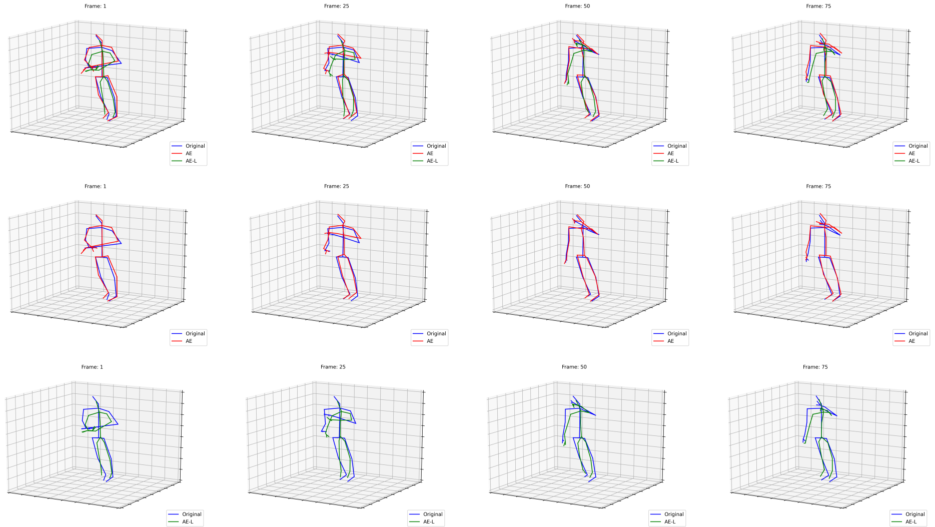


Figure 5: Action class "Drink Water" in NTU-60 [16] cross-view dataset. ***Blue:*** original data, ***Red:*** AE reconstruction, ***Green:*** AE-L reconstruction. Columns correspond to different time-frames. Both AE and AE-L correctly classify this action sample.

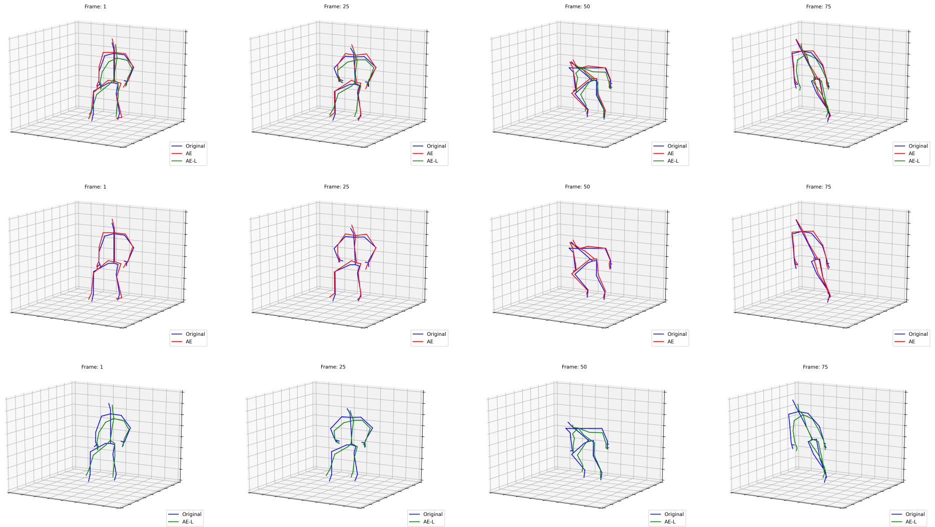


Figure 6: Action class "Standing Up" in NTU-60 [16] cross-view dataset. ***Blue:*** original data, ***Red:*** AE reconstruction, ***Green:*** AE-L reconstruction. Columns correspond to different time-frames. Both AE and AE-L correctly classify this action sample.

References

[1] Jacopo Cavazza, Pietro Morerio, and Vittorio Murino. Scalable and compact 3D action recognition with approximated rbf kernel machines. *Pattern Recognition*, 93:25–35, 2019.

[2] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 597–600, 2017. doi: 10.1109/ICMEW.2017.8026285.

[3] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020.

[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *The International Conference on Machine Learning (ICML)*, 2015.

[5] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, 2015.

[6] T. S. Kim and A. Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1623–1631, 2017. doi: 10.1109/CVPRW.2017.207.

[7] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and R Venkatesh Babu. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *IEEE Winter Conference on Applications of Computer Vision (WACV).*, 2018.

[8] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3590–3598, 2019. doi: 10.1109/CVPR.2019.00371.

[9] Li Linguo, Wang Minsi, Ni Bingbing, Wang Hang, Yang Jiancheng, and Zhang Wenjun. 3d human action representation learning via cross-view consistency pursuit. In *CVPR*, 2021.

[10] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. *arXiv preprint arXiv:1607.07043*, 2016.

[11] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2019. doi: 10.1109/TPAMI.2019.2916873.

[12] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recogn.*, 68:346–362, August 2017. doi: 10.1016/j.patcog.2017.02.030.

[13] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2430–2443, 2016. doi: 10.1109/TPAMI.2016.2533389.

[14] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition, 2020.

[15] A. Shahroudy, J. Liu, T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. doi: 10.1109/CVPR.2016.115.

[16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[17] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[19] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] K. Su, X. Liu, and E. Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9628–9637, 2020. doi: 10.1109/CVPR42600.2020.00965.

[21] Yu-Hui Wen, Lin Gao, Hongbo Fu, Fang-Lue Zhang, and Shihong Xia. Graph cnns with motif and variable temporal block for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8989–8996, Jul. 2019. doi: 10.1609/aaai.v33i01.33018989. URL https://ojs.aaai.org/index.php/AAAI/article/view/4929.

[22] Shihao Xu, Haocong Rao, Xiping Hu, and Bin Hu. Prototypical contrast and reverse prediction: Unsupervised skeleton based action recognition. In *arXiv preprint 2011.07236*, 2020.

[23] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, pages 7444–7452, 2018.

[24] Yong Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. doi: 10.1109/CVPR.2015.7298714.

[25] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[26] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.