Sep 22, 2025

JMM-TGT: Self-supervised 3D action recognition through joint motion masking and topology-guided transformer

PLOS One

DOI
https://dx.doi.org/10.17504/protocols.io.14egnrmjql5d/v1

WenHan[1,2]

[1]School of Computer and Communication Engineering; [2]University of Science and Technology Beijing

wenhan

wen han
University of Science and Technology Beijing

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.14egnrmjql5d/v1

External link: https://github.com/wenhan20201/JMM-TGT.git

Protocol Citation: WenHan 2025. JMM-TGT: Self-supervised 3D action recognition through joint motion masking and topology-guided transformer. **protocols.io** https://dx.doi.org/10.17504/protocols.io.14egnrmjql5d/v1

Manuscript citation:
This manuscript is being prepared for publication in PLOS ONE.

Protocol status: Working
We use this protocol and it's working

Created: September 22, 2025

Last Modified: September 22, 2025

Protocol Integer ID: 227875

Keywords: 3D skeleton action recognition, 3d action recognition through joint motion masking, 3d skeleton action recognition, supervised 3d action recognition, action recognition, joint motion masking, mainstream action recognition model, single motion feature, modeling single motion feature, joint motion masking strategy, subtle joint movement, similarities in joint motion, complex dynamic patterns of joint motion, joint motion, temporal feature modeling, subtle motion variation, model of the action, joint, action, complex spatio, selection of joint, topological relationship between joint

Funders Acknowledgements:
ZengGuangping
Grant ID: 62072031

Disclaimer

none

Abstract

In the field of 3D skeleton action recognition, research on self-supervised learning methods has primarily focused on spatio-temporal feature modeling. However, these methods rely heavily on modeling single motion features, which limits their ability to capture subtle motion variations and complex spatio-temporal relationships. This is a direct result of the fact that understanding the model of the action remains incomplete. To address the above-mentioned issue, this paper proposes the Joint Motion Masking with Topology-Guided Transformer model (JMM-TGT) for action recognition. First, the Joint Motion Masking strategy is applied to enhance the ability of the model to perceive subtle joint movements. This method can generate masking probabilities by combining the differences and similarities in joint motion, thereby guiding the selection of joints to be masked at each time step. Meanwhile, in the transformer-based encoder module, the topological relationship between joints is introduced to adjust the attention mechanism, allowing the model to capture spatio-temporal dependencies and better understand the complex dynamic patterns of joint motion. To verify the performance of the JMM-TGT model, we conducted comparison experiments between it and mainstream action recognition models. Experiments demonstrate that the proposed JMM-TGT achieves performance improvements ranging from 1.5% to 7.9% under different evaluation settings on the NTU RGB+D 60, NTU RGB+D 120, and PUK-MMD datasets.

Image Attribution

none

Guidelines

We have uploaded the original dataset and the processed dataset of our study, along with the model parameters that can reproduce our experimental results. All codes can be accessed through the GitHub link we provided.

Materials

Raw Datasets
PUK-MMD: http://39.96.165.147/Projects/PKUMMD/PKU-MMD.html
NTU RGB+D 60\120: https://rose1.ntu.edu.sg/dataset/actionRecognition

Preprocessed data and The minimal dataset for reproducing the results of this study
https://pan.baidu.com/s/1hCc5iu24rSeW038UXsdqGA?pwd=ipv3 提取码: ipv3

Safety warnings

none

Ethics statement

Our experiments do not involve animal studies.

Before start

none

Requirements

1 python==3.8.13
torch==1.8.1+cu111
torchvision==0.9.1+cu111
tensorboard==2.9.0
timm==0.3.2
scikit-learn==1.1.1
tqdm==4.64.0
numpy==1.22.4

Raw Datasets

2 PUK-MMD: http://39.96.165.147/Projects/PKUMMD/PKU-MMD.html

NTU RGB+D 60\120: https://rose1.ntu.edu.sg/dataset/actionRecognition

Preprocessed data and The minimal dataset for reproducing the results of this study

3 https://pan.baidu.com/s/1hCc5iu24rSeW038UXsdqGA?pwd=ipv3 提取码: ipv3

Protocol references

1. Gayathri T, Mamatha H. How to Improve Video Analytics with Action Recognition: A Survey. ACM Computing Surveys. 2023;57(1). doi:https://doi.org/10.1145/3679011.
2. Armeni I, Sener O, Zamir AR, Jiang H, Brilakis I, Fischer M, et al. 3D Semantic Parsing of Large-Scale Indoor Spaces. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; p. 1534–1543. doi:https://doi.org/10.1109/cvpr.2016.170.
3. Li Y, Yu AW, Meng T, Caine B, Ngiam J, Peng D, et al. Deepfusion: LiDAR-Camera Depth Fusion for Multi-Modal 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; p. 17182–17191. doi:https://doi.org/10.1109/cvpr52688.2022.01667.
4. Zhu C, Jia Q, Chen W, Guo Y, Liu Y. Deep Learning for Video-Text Retrieval: A Review. International Journal of Multimedia Information Retrieval. 2023;12(1):3. doi:https://doi.org/10.1007/s13735-023-00267-8.
5. Yang S, Liu J, Lu S, Hwa EM, Hu Y, Kot AC. Self-supervised 3D action representation learning with skeleton cloud colorization. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;46(1):509–524. doi:https://doi.org/10.1109/tpami.2023.3325463.
6. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI conference on artificial intelligence. 2018;32(1).
7. Yang G, Yang Y, Lu Z, Yang J, Liu D, Zhou C, et al. STA-TSN: spatial-temporal attention temporal segment network for action recognition in video. PloS one. 2022;17(3):e0265115. doi:https://doi.org/10.1371/journal.pone.0265115.
8. Huang KH, Huang YB, Lin YX, Hua KL, Tanveer M, Lu X, et al. GRA: Graph Representation Alignment for Semi-Supervised Action Recognition. IEEE Transactions on Neural Networks and Learning Systems. 2024;35(9):11896–11905. doi:https://doi.org/10.1109/tnnls.2023.3347593.
9. Dai C, Wei Y, Xu Z, Chen M, Liu Y, Fan J. ConMLP: MLP-based self-supervised contrastive learning for skeleton data analysis and action recognition. Sensors. 2023;23(5):2452. doi:https://doi.org/10.3390/s23052452.
10. Li D, Tang Y, Zhang Z, Zhang W. Cross-stream contrastive learning for self-supervised skeleton-based action recognition. Image and Vision Computing. 2023;135:104689. doi:https://doi.org/10.1016/j.imavis.2023.104689.
11. Yang H, Zhang Q, Ren Z, Yuan H, Zhang F. Contrastive Learning with Cross-Part Bidirectional Distillation for Self-supervised Skeleton-Based Action Recognition. HUMAN-CENTRIC COMPUTING AND INFORMATION SCIENCES. 2024;14. doi:https://doi.org/10.22967/HCIS.2024.14.070.
12. Wu W, Hua Y, Zheng C, Wu S, Chen C, Lu A. Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. 2023 IEEE international conference on multimedia and expo workshops (ICMEW). 2023; p.

224–229. doi:https://doi.org/10.1109/icmew59549.2023.00045.
13. Hu J, Hou Y, Guo Z, Gao J. Global and local contrastive learning for self-supervised skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology. 2024;34:10578–10589. doi:https://doi.org/10.1109/tcsvt.2024.3410301.
14. Zhu Y, Han H, Yu Z, Liu G. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023; p. 13913–13922. doi:https://doi.org/10.1109/iccv51070.2023.01279.
15. Cheng K, Zhang Y, He X, et al. Skeleton-based action recognition with shift graph convolutional network. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020; p. 183–192. doi:https://doi.org/10.1109/CVPR42600.2020.00026.
16. Tian H, Ma X, Li X, et al. Skeleton-based action recognition with select-assemble-normalize graph convolutional networks. IEEE Transactions on Multimedia, 2023;25:8527-8538. doi:https://doi.org/10.1109/TMM.2023.3318325.
17. Jang S, Lee H, Kim W J, et al. Multi-scale structural graph convolutional network for skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2024;34(8):7244-7258. doi:https://doi.org/10.1109/TCSVT.2024.3375512.
18. Zhang Y, Yang Y, Gao X. Lightweight Graph Convolutional Network For Efficient Skeleton Based Action Recognition. 2024 International Joint Conference on Neural Networks (IJCNN). 2024; p. 1–8. doi:https://doi.org/10.1109/ijcnn60899.2024.10651467.
19. Ren Z, Luo L, Qin Y. Skeleton-guided and supervised learning of hybrid network for multi-modal action recognition. Journal of Visual Communication and Image Representation. 2025; p. 104389. doi:https://doi.org/10.2139/ssrn.4970121.
20. Xu J, Zhu A, Lin J, Ke Q, Chen C. Skeleton-OOD: An End-to-End Skeleton-Based Model for Robust Out-of-Distribution Human Action Detection. NEUROCOMPUTING. 2025;619. doi:https://doi.org/10.1016/j.neucom.2024.129158.
21. Aouaidjia K, Zhang C, Pitas I. Spatio-temporal invariant descriptors for skeleton-based human action recognition. Information Sciences. 2025;700:121832. doi:https://doi.org/10.1016/j.ins.2024.121832.
22. Wu S, Lu G, Han Z, Chen L. A robust two-stage framework for human skeleton action recognition with GAIN and masked autoencoder. Neurocomputing. 2025;623:129433. doi:https://doi.org/10.1016/j.neucom.2025.129433.
23. Huang H, Xu L, Zheng Y, Yan X. MAFormer: A cross-channel spatio-temporal feature aggregation method for human action recognition. AI Communications. 2024;37(4):735–749. doi:https://doi.org/10.3233/aic-240260.
24. Zhao Z, Liu Y, Ma L. Compositional action recognition with multi-view feature fusion. Plos one. 2022;17(4):e0266259. doi:https://doi.org/10.1371/journal.pone.0266259.

25. Yang H, Wang S, Jiang L, Su Y, Zhang Y. Hierarchical adaptive multi-scale hypergraph attention convolution network for skeleton-based action recognition. Applied Soft Computing. 2025;172:112855. doi:https://doi.org/10.1016/j.asoc.2025.112855.
26. Zhu S, Sun L, Ma Z, Li C, He D. Prompt-supervised dynamic attention graph convolutional network for skeleton-based action recognition. Neurocomputing. 2025;611:128623. doi:https://doi.org/10.1016/j.neucom.2024.128623.
27. Xu Z, Xu J. Spatiotemporal decoupling attention transformer for 3D skeleton-based driver action recognition. Complex & Intelligent Systems. 2025;11(4):1–12. doi:https://doi.org/10.1007/s40747-025-01811-1.
28. Huang B, Wang S, Hu C, Li X. Semi-supervised human action recognition via dual-stream cross-fusion and class-aware memory bank. Engineering Applications of Artificial Intelligence. 2024;136:108937. doi:https://doi.org/10.1016/j.engappai.2024.108937.
29. Lin L, Zhang J, Liu J. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023; p. 2363–2372. doi:https://doi.org/10.1109/cvpr52729.2023.00234.
30. Liu X, Gao B. Reconstruction-driven contrastive learning for unsupervised skeleton-based human action recognition. The Journal of Supercomputing. 2025;81(1):37. doi:https://doi.org/10.1007/s11227-024-06573-0.
31. He Z, Lv J, Fang S. Representation modeling learning with multi-domain decoupling for unsupervised skeleton-based action recognition. Neurocomputing. 2024;582:127495. doi:https://doi.org/10.2139/ssrn.4634150.
32. Liu Z, Lu B, Wu Y, Gao C. Multi-view daily action recognition based on Hooke balanced matrix and broad learning system. Image and Vision Computing. 2024;143:104919. doi:https://doi.org/10.1016/j.imavis.2024.104919.
33. Lin L, Wu L, Zhang J, Liu J. Idempotent Unsupervised Representation Learning for Skeleton-Based Action Recognition. European Conference on Computer Vision. 2024; p. 75–92. doi:https://doi.org/10.2139/ssrn.4634150.
34. Jin Z, Wang Y, Wang Q, Shen Y, Meng H. SSRL: Self-supervised spatial-temporal representation learning for 3D action recognition. IEEE Transactions on Circuits and Systems for Video Technology. 2023;34(1):274–285. doi:https://doi.org/10.1109/tcsvt.2023.3284493.
35. Yao S, Ping Y, Yue X, Chen H. Graph Convolutional Networks for multi-modal robotic martial arts leg pose recognition. Frontiers in Neurorobotics. 2025;18:1520983. doi:https://doi.org/10.3389/fnbot.2024.1520983.
36. Wu C, Wu XJ, Kittler J, Xu T, Ahmed S, Awais M, et al. Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. Proceedings of the AAAI conference on artificial intelligence. 2024;38(6):5949–5957. doi:https://doi.org/10.1609/aaai.v38i6.28409.
37. Moutik O, Sekkat H, Ait Tchakoucht T, El Kari B, Alaoui AEH. A puzzle questions form training for self-supervised skeleton-based action recognition.

Image and Vision Computing. 2024;148:105137. doi:https://doi.org/10.1016/j.imavis.2024.105137.
38. Guan S, Yu X, Huang W, Fang G, Lu H. DMMG: dual min-max games for self-supervised skeleton-based action recognition. IEEE Transactions on Image Processing. 2023;33:395–407. doi:https://doi.org/10.1109/tip.2023.3338410.
39. Guo T, Liu M, Liu H, Wang G, Li W. Improving self-supervised action recognition from extremely augmented skeleton sequences. Pattern Recognition. 2024;150:110333. doi:https://doi.org/10.1016/j.patcog.2024.110333.
40. Wang M, Li X, Chen S, Zhang X, Ma L, Zhang Y. Learning representations by contrastive spatio-temporal clustering for skeleton-based action recognition. IEEE Transactions on Multimedia. 2023;26:3207–3220. doi:https://doi.org/10.1109/tmm.2023.3307933.
41. Wu Y, Xu Z, Yuan M, Tang T, Meng R, Wang Z. Multi-scale motion contrastive learning for self-supervised skeleton-based action recognition. Multimedia Systems. 2024;30(5):267. doi:https://doi.org/10.1007/s00530-024-01463-0.
42. Liu R, Liu Y, Wu M, Xin W, Miao Q, Liu X, et al. SG-CLR: Semantic representation-guided contrastive learning for self-supervised skeleton-based action recognition. Pattern Recognition. 2025; p. 111377. doi:https://doi.org/10.2139/ssrn.4853185.
43. Shahroudy A, Liu J, Ng TT, Wang G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; p. 1010–1019. doi:https://doi.org/10.1109/cvpr.2016.115.
44. Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence. 2019;42(10):2684–2701. doi:https://doi.org/10.1109/tpami.2019.2916873.
45. Liu J, Song S, Liu C, et al. A benchmark dataset and comparison study for multi-modal human action analytics. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2020;16(2):1–24. doi:https://doi.org/10.1145/3365212.
46. Guo T, Liu H, Chen Z, Liu M, Wang T, Ding R. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. Proceedings of the AAAI conference on artificial intelligence. 2022;36(1):762–770. doi:https://doi.org/10.1609/aaai.v36i1.19957.
47. Kim B, Chang HJ, Kim J, Choi JY. Global-local motion transformer for unsupervised skeleton-based action learning. European conference on computer vision. 2022; p. 209–225. doi:https://doi.org/10.1007/978-3-031-19772-7 13.
48. Zhang H, Hou Y, Zhang W, Li W. Contrastive positive mining for unsupervised 3d action representation learning. European Conference on Computer Vision. 2022; p. 36–51. doi:https://doi.org/10.1007/978-3-031-19772-7 3.
49. Mao Y, Zhou W, Lu Z, Deng J, Li H. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. European

Conference on Computer Vision. 2022; p. 734–752. doi:https://doi.org/10.1007/978-3-031-20062-5 42.
50. Dong J, Sun S, Liu Z, Chen S, Liu B, Wang X. Hierarchical contrast for unsupervised skeleton-based action representation learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2023;37(1):525–533. doi:https://doi.org/10.1609/aaai.v37i1.25127.
51. Chen YX, Zhao L, Yuan JB, Tian Y, Xia ZY, Geng SJ, Han LG, Metaxas DN. Hierarchically self-supervised transformer for human skeleton representation learning. European Conference on Computer Vision. 2022; p. 185–202. doi:https://doi.org/10.1007/978-3-031-19809-0 11.
52. Zheng N, Wen J, Liu R, Long L, Dai J, Gong Z. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. Proceedings of the AAAI conference on artificial intelligence. 2018;32(1). doi:https://doi.org/10.1609/aaai.v32i1.11853.
53. Lin L, Song S, Yang W, Liu J. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. Proceedings of the 28th ACM international conference on multimedia. 2020; p. 2490–2498. doi:https://doi.org/10.1145/3394171.3413548.
54. Thoker F M, Doughty H, Snoek C G M. Skeleton-contrastive 3D action representation learning. Proceedings of the 29th ACM international conference on multimedia. 2021; p. 1655–1663. doi:https://doi.org/10.1145/3474085.3475307.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62072031. The authors specially thank the team members for data collection and model optimization.