PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

# Skeleton-based human action recognition with a physics-augmented encoder-decoder network

Hongji Guo, Alexander Aved, Collen Roller, Erika Ardiles-Cruz, Qiang Ji

Hongji Guo, Alexander Aved, Collen Roller, Erika Ardiles-Cruz, Qiang Ji, "Skeleton-based human action recognition with a physics-augmented encoder-decoder network," Proc. SPIE 12525, Geospatial Informatics XIII , 125250K (15 June 2023); doi: 10.1117/12.2664115

SPIE.

Event: SPIE Defense + Commercial Sensing, 2023, Orlando, Florida, United States

# Skeleton-Based Human Action Recognition with A Physics-Augmented Encoder-Decoder Network

Hongji Guo[a], Alexander Aved[b], Collen Roller[b], Erika Ardiles-Cruz[b], and Qiang Ji[a]

[a]Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[b]Air Force Research Laboratory, Rome, NY 13441, USA

## ABSTRACT

Human action recognition is important for many applications such as surveillance monitoring, safety, and healthcare. As 3D body skeletons can accurately characterize body actions and are robust to camera views, we propose a 3D skeleton-based human action method. Different from the existing skeleton-based methods that use only geometric features for action recognition, we propose a physics-augmented encoder and decoder model that produces physically plausible geometric features for human action recognition. Specifically, given the input skeleton sequence, the encoder performs a spatiotemporal graph convolution to produce spatiotemporal features for both predicting human actions and estimating the generalized positions and forces of body joints. The decoder, implemented as an ODE solver, takes the joint forces and solves the Euler-Lagrangian equation to reconstruct the skeletons in the next frame. By training the model to simultaneously minimize the action classification and the 3D skeleton reconstruction errors, the encoder is ensured to produce features that are consistent with both body skeletons and the underlying body dynamics as well as being discriminative. The physics-augmented spatiotemporal features are used for human action classification. We evaluate the proposed method on NTU-RGB+D, a large-scale dataset for skeleton-based action recognition. Compared with existing methods, our method achieves higher accuracy and better generalization ability.

**Keywords:** Skeleton-based action recognition, physics, encoder-decoder

## 1. INTRODUCTION

Skeleton-based action recognition has been an important research topic for a long time. It aims at identifying the action classes from skeletons sequences. It has many applications such as visual surveillance,[1] Internet of Things (IoT),[2] and autonomous driving. Skeleton-based action recognition is challenging since the action may not be well represented without the appearance information. Also, some actions have similar skeletal representations, which are ambiguous to recognize. Also, large amount of data and training are needed to achieve good recognition accuracy.

Most existing methods rely on pure deep learning architectures such as recurrent neural network,[3] graph convolution network,[4] and Transformer.[5] These methods need tremendous amount of training data and are lack of interpretability of human actions. To alleviate these issues, we proposed a physics-augmented encoder-decoder network for skeleton-based action recognition by leveraging the physics principles for modeling the human actions. Different from the existing methods that use only geometric features, we combine the deep learning based features as well as physics-based features for action recognition. In this way, we can leverage both the geometric features and the physics features to improve the performance.

Recently, Physics modeling has been introduced for many computer vision tasks and related applications.[6–8] In this paper, we model the inherent physics that cause the human actions so that the inherent physics representations of human actions can improve the performance and robustness of the action recognition.

In summary, the main contributions of this paper are:

Further author information: (Send correspondence to Qiang Ji)
Qiang Ji: E-mail: jiq@rpi.edu

Geospatial Informatics XIII, edited by Kannappan Palaniappan, Gunasekaran Seetharaman, Joshua D. Harguess, Proc. of SPIE Vol. 12525, 125250K · © 2023 SPIE
0277-786X · doi: 10.1117/12.2664115

- We proposed a physics-augmented encoder-decoder network for skeleton-based action recognition. With a graph-convolution-based encoder and a physics-based decoder, our encoder-decoder network learns physically plausible features.
- By incorporating physics laws into the model, which improves generalization and data-efficiency of the model.
- We evaluated our proposed method on NTU-RGB+D 60 & 120 and NW-UCLA datasets. The competitive recognition accuracy demonstrates the effectiveness of our proposed method.

## 2. RELATED WORK

### 2.1 Skeleton-Based Action Recognition

Skeleton-based action recognition has been a popular research topic for a long period because of the compact and robust representation of human actions. Early works[9–12] adopted hand-crafted features to model the human actions, which had limited performance. Dynamic models such as RNN are also used for skeleton-based action recognition.[13–18]

Then, graph convolution networks became the mainstream for skeleton-based action recognition. Yan *et al.*[19] introduced spatial-temporal graph convolutional networks (ST-GCN) for skeleton-based action recognition. The spatial-temporal convolution models the dynamics of human skeleton sequences. Shi *et al.*[20] introduced two-stream adaptive graph convolutional networks (2s-AGCN) for skeleton-based action recognition. Both the joint information and bone information are considered, and the topology of graph can be either uniformly or individually learned by the BP algorithm in an end-to-end manner. The flexibility of the model for graph construction is increased and it brings more generality to adapt to various data samples. Shi *et al.*[21] introduced directed graph neural networks (DGNN) for skeleton-based action recognition. The skeleton data is represented as directed acyclic graph (DAG) based on the kinematic dependency between the joints and bones in the natural human body. A novel network is designed to extract the information of joints, bones and their relationships and make prediction based on the extract features. Li *et al.*[22] introduced actional-structural graph convolutional networks (AS-GCN) for skeleton-based action recognition, which stacks actional-structural graph convolution and temporal convolution as a basic building block, to learn both spatial and temporal features. A future pose prediction head is added in parallel to the recognition head to help capture more detailed action patterns through self-supervision. Liu *et al.*[23] introduced disentangling and Unifying Graph Convolutions (MS-G3D) for Skeleton-Based Action Recognition. The proposed scheme disentangles the importance of nodes in different neighborhoods for effective long-range modeling. The G3D module leverages dense cross-spacetime edges as skip connections for direct information propagation across the spatial-temporal graph. Cheng *et al.*[24] introduced shift graph convolution network (Shift-GCN) for skeleton-based action recognition. The proposed Shift-GCN is composed of novel shift graph operations and lightweight point-wise convolutions, where the shift graph operations provide flexible receptive fields for both spatial graph and temporal graph. Chen *et al.*[25] introduced channel-wise topology refinement graph convolution (CTR-GCN) for skeleton-based action recognition. The proposed network models channel-wise topologies through learning a shared topology as a generic prior for all channels and refining it with channel-specific correlations for each channel. The proposed refinement method introduces few extra parameters and significantly reduces the difficulty of modeling channel-wise topologies. Recently, Transformer[5] is utilized for skeleton-based action recognition and achieve promising results.[26,27]

### 2.2 Physics Modeling for Computer Vision

Recently, there are increasing approaches incorporating physics knowledge into the model to improve the certain properties of deep learning models.[28–32]

Specifically, some methods utilize the Lagrangian or Hamiltonian mechanics to model the position and momentum.[33–37] To estimate the physical parameters, some work[38,39] uses an autoencoder to predict the physics parameters. In this paper, we also adopt an autoencoder design to learn the physically plausible representations.

## 3. METHOD

In this part, we first show the overall framework of our proposed physics-augmented encoder-decoder network. Then we introduce the encoder and decoder separately. Finally, we discuss the recognizer for skeleton-based action recognition.

### 3.1 Overall Framework

The overall framework is shown in Figure 1. The input of our model is a sequence of 3D human skeletons. Firstly, the input skeletons are fed in the a deformable human mass module to adjust the parameters of the pre-defined human model, which outputs human mass, shape, etc for use by the physics-based decoder. The encoder takes the skeletons as input and output the generalized positions, forces of the corresponding input, as well as deep learning based features. With the fitted human model, the physics-based decoder reconstructs the input skeletons by taking the generalized positions and forces of the joints. Simultaneously, the intermediate physics parameters and features are used to perform the classification. By training the model with the classification loss and the reconstruction loss. The model can well capture the physical properties of the actions and thus improve the performance and robustness.

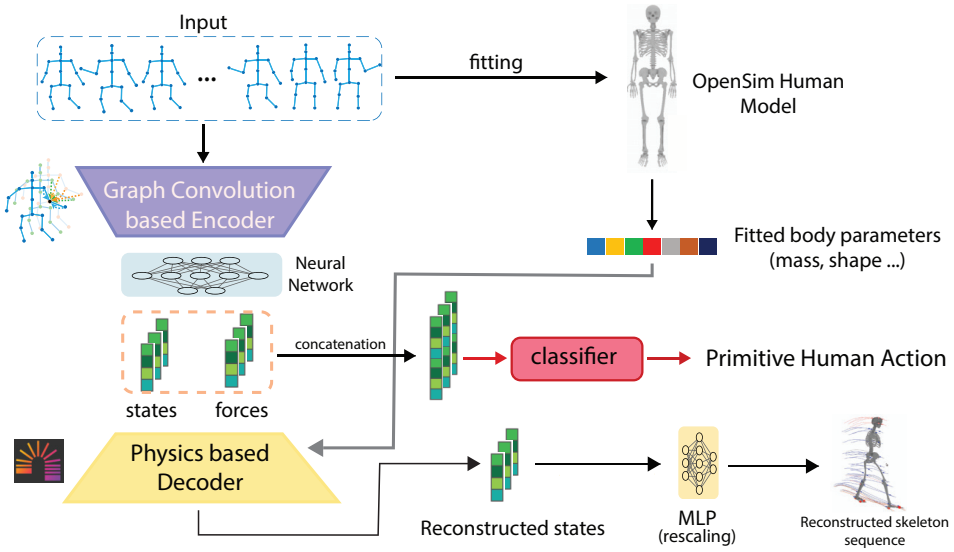



Figure 1. Overall framework of our proposed method. Our model takes a sequence of human skeleton as input. The model firstly fit basic body parameters such as mass to an pre-defined human model, which will be used in the decoder. A graph convolution based encoder predict the generalized positions and forces of human joints. Then these positions and forces are fed to the physics-based decoder to reconstruct the input skeleton sequence. At the same time, the intermediate representations of the encoder-decoder network are used to perform the action recognition.

### 3.2 Encoder

Given the input skeleton sequence, we use a spatial-temporal graph convolution network as the encoder to predict the generalized positions and forces. Specifically, the 3D human skeleton is constructed as an undirected spatial temporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $N$ joints and $T$ frames. The node set $\mathcal{V} = \{v_{ti} | t = 1, ..., T; i = 1, ..., N\}$ includes all the joints in the sequence. Besides the graph, each node is also associated with a feature vector. The feature vector or node $i$ at frame $t$ is denoted as $F(v_{ti})$. And the edge set $\mathcal{E}$ is composed of two types of links: intra-body links and inter-frame links. The intra-body links for each frame are denoted as $\mathcal{E}_S = \{v_{ti}, v_{tj} | (i, j) \in \mathcal{H}\}$, where $\mathcal{H}$ is the set of naturally connected human body joints. The inter-frame link set $\mathcal{E}_F = \{v_{ti}, v_{(t+1)i}\}$ contains the inter-frame links between two consecutive frames. The intra-body connections of joints within a frame are

represented by an adjacency matrix $A$ and an identity matrix $I$. Denote the input feature as $f_{in}$ and the feature after the graph convolution as $f_{out}$. We first perform the spatial convolution as below:

$$f_{out}^{s} = \Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}} f_{in} W \tag{1}$$

where $\Lambda^{ii} = \sum_j (A^{ij} + I^{ij})$ and $W$ is the learnable weight matrix.

After the spatial graph convolution, we perform a temporal graph convolution to obtain the final output feature $f_{out}$. Specifically, the temporal graph convolution is achieved by a $1 \times \Gamma$ convolution along the temporal dimension. The output of the final graph convolution layer is fed into a fully-connected neural network to predict the generalized positions $\boldsymbol{q}$ and the estimated forces $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$, which are then fed to the physics-based decoder.

### 3.3 Decoder

Using the generalized positions $\boldsymbol{q}$ and estimated forces $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ from the encoder as input, the physics-based decoder reconstructs the input skeleton sequence subject to the body movement dynamics satisfying the Euler-Lagrange equation. The physics-based encoder hence plays the same role as the numerical ODE solver for the dynamics regression decoder for 3D skeleton reconstruction. Specifically, given the generalized positions $\boldsymbol{q}_t$ and forces $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$, the physics-based decoder first solves the $\ddot{\boldsymbol{q}}_t$ by:

$$\ddot{\boldsymbol{q}}_t = \boldsymbol{M}^{-1}(\boldsymbol{\beta}, = \boldsymbol{q}_t)(\boldsymbol{J}_C^T \boldsymbol{\lambda_t} + \boldsymbol{\tau}_t - \boldsymbol{C}(\boldsymbol{\beta}, \boldsymbol{q}_t, \dot{\boldsymbol{q}}_t)) \tag{2}$$

where $\boldsymbol{M}$ is the generalized inertia matrix and $\boldsymbol{C}$ is the generalized bias forces.

Then the decoder predict the next position $\hat{\boldsymbol{q}}_{t+1}$ by:

$$\hat{\boldsymbol{q}}_{t+1} = \boldsymbol{q}_t + \dot{\boldsymbol{q}}_t \Delta t \tag{3}$$

where $\Delta t$ is the time interval.

$\boldsymbol{q}_t$ is then resized to the original input skeleton scale to construct a dynamic reconstruction loss as below:

$$\mathcal{L}_{dynamics} = \sum_{t=1}^{T} (\hat{\boldsymbol{q}}_t - \boldsymbol{q}_t)^2 \tag{4}$$

By using the physics-based decoder, the hidden representations (output of the encoder) is constrained to be physically plausible. For implementation, we adopted the nimblephysics[40] as the solver of the physics-based decoder for the prediction of each timestep.

### 3.4 Recognizer

The intermediate physics representations sequence from the encoder-decoder network are highly discriminative of human actions and properties. We combine these physics-based features and the graph convolution features to perform the classification. Specifically, we concatenate the generalized positions and forces to form the physics feature vector. Concatenated with the flattened graph convolution features, the combined features are fed into a fully-connected network for human action classification.

To train the model, our loss function contains two parts: one for the 3D human skeleton reconstruction and the other for the action recognition:

$$\mathcal{L} = \mathcal{L}_{dynamic} - \lambda \sum_{c=1}^{C} P(X = c) log P(X = c) \tag{5}$$

where the first term is the reconstruction loss and the second term is the classification loss (i.e. negative log-likelihood), $C$ is the total number of action classes, and $\lambda$ is the weight of the classification loss. With this loss function, we can ensure the learned intermediate hidden representations are both physically plausible and are discriminative for action recognition, which sets apart our method from that of the existing human action recognition methods. After training, the decoder can be discarded and the outputs of the encoder can be used for action recognition during testing.

# 4. EXPERIMENTS

## 4.1 Datasets

**NTU-RGB+D 60 & 120.**[41,42] NTU-RGB+D 60 dataset is a large-scale dataset for skeleton-based action recognition. The data is represented by 3D human joint positions. It contains 60 action classes. The 3D skeleton data in the dataset was collected by Microsoft Kinect v2, which leads to 25 joints per person at one frame. There are totally 40 different subjects. The common evaluation settings include Cross-Subjects (CS) and Cross-View (CV), which means the training set and testing set are from different people and view angles respectively. NTU-RGB+D 120 dataset[42] is an extension version of NTU-RGB+D 60. It contains 120 action classes.

**Northwestern-UCLA.**[43] NW-UCLA is a dataset for 3D skeleton-based action recognition. It contains 10 action classes with 1494 data samples. The 3D joint positions are captured by Microsoft Kinect sensors. We follow the cross-view evaluation metric.[43] The data from first two cameras are used for training and the data from the third camera are used for testing.

## 4.2 Implementation Details

The implementation of the framework is done in PyTorch. The training and testing was conducted on two Nvidia RTX 3090 Ti GPUs. For NTU-RGB+D 60 & 120 datasets, the number of joints is set to 25. And the number of joints is set to 20 for NW-UCLA dataset. We adopt nimblephysics[40] differentiable physics solver in the decoder.

## 4.3 Experiment Results

We verified our proposed method for skeleton-based action recognition on NTU-RGB+D dataset. The experiment results on NTU-RGB+D 60 and NTU-RGB+D 120 are shown in Table 1 and Table 2 respectively. By comparison, our proposed physics-augmented encoder decoder network achieves competitive performance against state-of-the-art methods. The experiment results on NW-UCLA dataset are shown in Table 3. Our proposed encoder-decoder network also achieves competitive accuracy, which demonstrates its effectiveness.

Table 1. **Experiment results on NTU-RGB+D 60.** Compared with state-of-the-art methods, our proposed physics-augmented encoder-decoder network achieves competitive performance.

| Method | Cross-Subject (%) | Cross-View (%) |
|---|---|---|
| Ind-RNN[17] | 81.8 | 88.8 |
| HCN[44] | 86.5 | 91.1 |
| ST-GCN[19] | 81.5 | 88.3 |
| 2s-AGCN[20] | 88.5 | 95.1 |
| SGN[45] | 89.0 | 94.5 |
| AGC-LSTM[46] | 89.2 | 95.0 |
| DGNN[21] | 89.9 | 96.1 |
| Shift-GCN[24] | 90.7 | 96.5 |
| DC-GCN+ADG[47] | 90.8 | 96.6 |
| PA-ResGCN-B19[48] | 90.9 | 96.0 |
| DDGCN[49] | 91.1 | **97.1** |
| Dynamic GCN[50] | 91.5 | 96.0 |
| MS-G3D[23] | 91.5 | 96.2 |
| CTR-GCN[25] | **92.4** | 96.8 |
| ST-TR[51] | 89.9 | 96.1 |
| STST[27] | 91.9 | 96.8 |
| Encoder-Decoder (ours) | 90.8 | 96.3 |

Table 2. **Experiment results on NTU-RGB+D 120.** Compared with state-of-the-art methods, our proposed physics-augmented encoder-decoder network achieves competitive performance.

| Method | Cross-Subject (%) | Cross-View (%) |
|---|---|---|
| ST-LSTM[14] | 55.7 | 57.9 |
| GCA-LSTM[52] | 61.2 | 63.3 |
| RotClips+MTCNN[53] | 62.2 | 61.8 |
| SGN[45] | 79.2 | 81.5 |
| 2s-AGCN[20] | 82.9 | 84.9 |
| Shift-GCN[24] | 85.9 | 87.6 |
| DC-GCN+ADG[47] | 86.5 | 88.1 |
| MS-G3D[23] | 86.9 | 88.4 |
| PA-ResGCN-B19[48] | 87.3 | 88.3 |
| Dynamic GCN[50] | 87.3 | 88.6 |
| CTR-GCN[25] | **88.9** | **90.6** |
| Encoder-Decoder (ours) | 86.9 | 88.2 |

Table 3. **Experiment results on NW-UCLA dataset.** Compared with state-of-the-art methods, our proposed physics-augmented encoder-decoder network achieves competitive performance.

| Method | Accuracy (%) |
|---|---|
| Lie Group[54] | 74.2 |
| Action Ensemble[55] | 76.0 |
| HBRNN-L[13] | 78.5 |
| Ensemble TS-LSTM[56] | 89.2 |
| AGC-LSTM[46] | 93.3 |
| Shift-GCN[24] | 94.6 |
| DC-GCN+ADG[47] | 95.3 |
| CTR-GCN[25] | **96.5** |
| Encoder-Decoder (ours) | 93.6 |

### 4.4 Qualitative Results

### 4.5 Ablation Studies

**Encoder types.** To further study the physics-augmented encoder-decoder network, we replace the graph-convolution-based encoder with other types of encoders. The experiments results are shown in Table 4.

Table 4. **Encoder types.** We replace the graph-convolution-based encoder with other types of encoders. The experiment results show that the graph-convolution-based encoder gives the best performance.

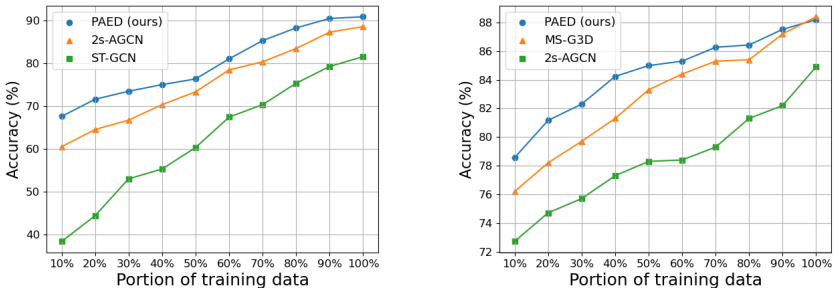
| Encoder Type | NTU-60-CS (%) | NTU-60-CV (%) | NTU-120-CS (%) | NTU-120-CV (%) |
|---|---|---|---|---|
| FFN | 86.3 | 92.0 | 81.4 | 82.1 |
| RNN | 88.9 | 94.5 | 85.1 | 86.6 |
| GCN | **90.8** | **96.3** | **86.9** | **88.2** |

**Decoder types.** To demonstrate the effectiveness of physical modeling of physics-augmented encoder-decoder network, we replace the physics-based decoder with other types of networks. The experiment results are shown in Table 5. The results shown that the physics-based decoder outperforms other types of decoders.

Table 5. **Decoder types.** We replace the physics-based decoder with other types of encoders. The physics-based decoder gives the best performance, which demonstrates the effectiveness of our proposed method.

| Decoder Type | NTU-60-CS (%) | NTU-60-CV (%) | NTU-120-CS (%) | NTU-120-CV (%) |
|---|---|---|---|---|
| FFN | 75.6 | 77.2 | 70.5 | 70.8 |
| RNN | 80.6 | 84.3 | 75.9 | 76.4 |
| GCN | 83.4 | 89.0 | 78.9 | 80.8 |
| Physics-based | **90.8** | **96.3** | **86.9** | **88.2** |

**Training with small-scale data.** To further demonstrate the effectiveness of our proposed, method, we reduce the amount of training data for 100% to 10% and compare with other approaches. The experiment results are shown in Figure 2. The results show that our proposed method is more data-efficient.




(a) NTU-RGB+D 60 (b) NTU-RGB+D 120

Figure 2. Experiment results on NTU-RGB+D 60 & 120 with small-scale training data.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a physics-augmented encoder-decoder network for skeleton-based action recognition. The intermediate hidden states representing the generalized forces as well as features are used to performed the action recognition. Our proposed method achieves competitive performance on benchmark datasets. We conducted ablation studies to demonstrate the effectiveness of physical modeling. The ablation study also shows that our method is more data-efficient with small-scale training data.

Our proposed method is based on the 3D human skeletons, which need to be collected from depth sensors or estimated by 3D pose estimation algorithms. Future work may include how to apply our method on 2D human skeletons and improve the computation efficiency.

## ACKNOWLEDGMENTS

This project is supported, in part, by the U.S. Air Force Research Lab Summer Faculty Fellowship Program.

## REFERENCES

[1] Kim, S., Yun, K., Park, J., and Choi, J. Y., "Skeleton-based action recognition of people handling objects," in [*2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*], 61–70, IEEE (2019).

[2] Xu, W., Wu, M., Zhu, J., and Zhao, M., "Multi-scale skeleton adaptive weighted gcn for skeleton-based human action recognition in iot," *Applied Soft Computing* **104**, 107236 (2021).

[3] Medsker, L. R. and Jain, L., "Recurrent neural networks," *Design and Applications* **5**, 64–67 (2001).

[4] Kipf, T. N. and Welling, M., "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907* (2016).

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need," *Advances in neural information processing systems* **30** (2017).

[6] Ali, S., Basharat, A., and Shah, M., "Chaotic invariants for human action recognition," in [*2007 IEEE 11th International Conference on Computer Vision*], 1–8, IEEE (2007).

[7] Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., and Xu, F., "Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 13167–13178 (2022).

[8] Hu, H., Yi, X., Zhang, H., Yong, J.-H., and Xu, F., "Physical interaction: Reconstructing hand-object interactions with physics," *arXiv preprint arXiv:2209.10833* (2022).

[9] Vemulapalli, R., Arrate, F., and Chellappa, R., "Human action recognition by representing 3d skeletons as points in a lie group," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 588–595 (2014).

[10] Fernando, B., Gavves, E., Oramas, J. M., Ghodrati, A., and Tuytelaars, T., "Modeling video evolution for action recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 5378–5387 (2015).

[11] Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M., "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in [*Twenty-third international joint conference on artificial intelligence*], (2013).

[12] Wang, J., Liu, Z., Wu, Y., and Yuan, J., "Mining actionlet ensemble for action recognition with depth cameras," in [*2012 IEEE Conference on Computer Vision and Pattern Recognition*], 1290–1297, IEEE (2012).

[13] Du, Y., Wang, W., and Wang, L., "Hierarchical recurrent neural network for skeleton based action recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1110–1118 (2015).

[14] Liu, J., Shahroudy, A., Xu, D., and Wang, G., "Spatio-temporal lstm with trust gates for 3d human action recognition," in [*European conference on computer vision*], 816–833, Springer (2016).

[15] Song, S., Lan, C., Xing, J., Zeng, W., and Liu, J., "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in [*Proceedings of the AAAI conference on artificial intelligence*], **31**(1) (2017).

[16] Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., and Zheng, N., "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in [*Proceedings of the IEEE international conference on computer vision*], 2117–2126 (2017).

[17] Li, S., Li, W., Cook, C., Zhu, C., and Gao, Y., "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 5457–5466 (2018).

[18] Cao, C., Lan, C., Zhang, Y., Zeng, W., Lu, H., and Zhang, Y., "Skeleton-based action recognition with gated convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology* **29**(11), 3247–3257 (2018).

[19] Yan, S., Xiong, Y., and Lin, D., "Spatial temporal graph convolutional networks for skeleton-based action recognition," in [*Thirty-second AAAI conference on artificial intelligence*], (2018).

[20] Shi, L., Zhang, Y., Cheng, J., and Lu, H., "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 12026–12035 (2019).

[21] Shi, L., Zhang, Y., Cheng, J., and Lu, H., "Skeleton-based action recognition with directed graph neural networks," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 7912–7921 (2019).

[22] Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., and Tian, Q., "Actional-structural graph convolutional networks for skeleton-based action recognition," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 3595–3603 (2019).

[23] Liu, Z., Zhang, H., Chen, Z., Wang, Z., and Ouyang, W., "Disentangling and unifying graph convolutions for skeleton-based action recognition," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 143–152 (2020).

[24] Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., and Lu, H., "Skeleton-based action recognition with shift graph convolutional network," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 183–192 (2020).

[25] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., and Hu, W., "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 13359–13368 (2021).

[26] Plizzari, C., Cannici, M., and Matteucci, M., "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding* **208**, 103219 (2021).

[27] Zhang, Y., Wu, B., Li, W., Duan, L., and Gan, C., "Stst: Spatial-temporal specialized transformer for skeleton-based action recognition," in [*Proceedings of the 29th ACM International Conference on Multimedia*], 3229–3237 (2021).

[28] Stewart, R. and Ermon, S., "Label-free supervision of neural networks with physics and domain knowledge," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **31**(1) (2017).

[29] Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., et al., "Physics-informed machine learning: case studies for weather and climate modelling," *Philosophical Transactions of the Royal Society A* **379**(2194), 20200093 (2021).

[30] Sanchez-Gonzalez, A., Heess, N., Springenberg, J. T., Merel, J., Riedmiller, M., Hadsell, R., and Battaglia, P., "Graph networks as learnable physics engines for inference and control," in [*International Conference on Machine Learning*], 4470–4479, PMLR (2018).

[31] Jin, P., Zhang, Z., Zhu, A., Tang, Y., and Karniadakis, G. E., "Sympnets: Intrinsic structure-preserving symplectic networks for identifying hamiltonian systems," *Neural Networks* **132**, 166–179 (2020).

[32] Shi, R., Mo, Z., and Di, X., "Physics-informed deep learning for traffic state estimation: A hybrid paradigm informed by second-order traffic models," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **35**(1), 540–547 (2021).

[33] Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S., "Lagrangian neural networks," *arXiv preprint arXiv:2003.04630* (2020).

[34] Greydanus, S., Dzamba, M., and Yosinski, J., "Hamiltonian neural networks," *Advances in neural information processing systems* **32** (2019).

[35] Zhong, Y. D., Dey, B., and Chakraborty, A., "Symplectic ode-net: Learning hamiltonian dynamics with control," *arXiv preprint arXiv:1909.12077* (2019).

[36] Toth, P., Rezende, D. J., Jaegle, A., Racanière, S., Botev, A., and Higgins, I., "Hamiltonian generative networks," *arXiv preprint arXiv:1909.13789* (2019).

[37] Zhong, Y. D., Dey, B., and Chakraborty, A., "Benchmarking energy-conserving neural networks for learning dynamics from data," in [*Learning for Dynamics and Control*], 1218–1229, PMLR (2021).

[38] Yang, T.-Y., Rosca, J. P., Narasimhan, K. R., and Ramadge, P., "Learning physics constrained dynamics using autoencoders," in [*Advances in Neural Information Processing Systems*],

[39] Takeishi, N., Kawahara, Y., and Yairi, T., "Learning koopman invariant subspaces for dynamic mode decomposition," *Advances in neural information processing systems* **30** (2017).

[40] Werling, K., Omens, D., Lee, J., Exarchos, I., and Liu, C. K., "Fast and feature-complete differentiable physics for articulated rigid bodies with contact," *arXiv preprint arXiv:2103.16021* (2021).

[41] Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G., "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1010–1019 (2016).

[42] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C., "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2684–2701 (2019).

[43] Wang, J., Nie, X., Xia, Y., Wu, Y., and Zhu, S.-C., "Cross-view action modeling, learning and recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2649–2656 (2014).

[44] Li, C., Zhong, Q., Xie, D., and Pu, S., "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055* (2018).

[45] Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., and Zheng, N., "Semantics-guided neural networks for efficient skeleton-based human action recognition," in [*proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 1112–1121 (2020).

[46] Si, C., Chen, W., Wang, W., Wang, L., and Tan, T., "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*], 1227–1236 (2019).

[47] Cheng, K., Zhang, Y., Cao, C., Shi, L., Cheng, J., and Lu, H., "Decoupling gcn with dropgraph module for skeleton-based action recognition," in [*European Conference on Computer Vision*], 536–553, Springer (2020).

[48] Song, Y.-F., Zhang, Z., Shan, C., and Wang, L., "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in [*proceedings of the 28th ACM international conference on multimedia*], 1625–1633 (2020).

[49] Korban, M. and Li, X., "Ddgcn: A dynamic directed graph convolutional network for action recognition," in [*European Conference on Computer Vision*], 761–776, Springer (2020).

[50] Ye, F., Pu, S., Zhong, Q., Li, C., Xie, D., and Tang, H., "Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition," in [*Proceedings of the 28th ACM International Conference on Multimedia*], 55–63 (2020).

[51] Plizzari, C., Cannici, M., and Matteucci, M., "Spatial temporal transformer network for skeleton-based action recognition," in [*International Conference on Pattern Recognition*], 694–701, Springer (2021).

[52] Liu, J., Wang, G., Duan, L.-Y., Abdiyeva, K., and Kot, A. C., "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing* **27**(4), 1586–1599 (2017).

[53] Ke, Q., Bennamoun, M., An, S., Sohel, F., and Boussaid, F., "Learning clip representations for skeleton-based 3d action recognition," *IEEE Transactions on Image Processing* **27**(6), 2842–2855 (2018).

[54] Veeriah, V., Zhuang, N., and Qi, G.-J., "Differential recurrent neural networks for action recognition," in [*Proceedings of the IEEE international conference on computer vision*], 4041–4049 (2015).

[55] Wang, J., Liu, Z., Wu, Y., and Yuan, J., "Learning actionlet ensemble for 3d human action recognition," *IEEE transactions on pattern analysis and machine intelligence* **36**(5), 914–927 (2013).

[56] Lee, I., Kim, D., Kang, S., and Lee, S., "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in [*Proceedings of the IEEE international conference on computer vision*], 1012–1020 (2017).