Supplementary Materials for Unsupervised View-Invariant Human Posture Representation

Faegheh Sardari[1]
faegheh.sardari@bristol.ac.uk
Björn Ommer[2]
ommer@uni-heidelberg.de
Majid Mirmehdi[1]
m.mirmehdi@bristol.ac.uk

[1] Department of Computer Science
University of Bristol
UK
[2] Ludwig Maximilian University of Munich
and IWR & HCI, Heidelberg University
Germany

Supplementary Materials

We provide:

- Qualitative results on NTU RGB+D
- More implementation details of downstream tasks
- Training and testing sets on QMAR
- The most recent state-of-the-art approaches on NTU RGB+D from CVPR 2021

1 Qualitative Results on NTU RGB+D

In Fig.1, we illustrate qualitative reconstruction results of our approach when applied on unseen viewpoints for RGB-based Densepose and depth images from the NTU dataset.

2 Action Classification and Human Movement Analysis

Our proposed auto-encoder is learned to represent unsupervised 3D pose features without using any action labels. To encapsulate the temporal element of the action recognition and human movement analysis downstream tasks, we used our view-invariant pose encoder $E_{\odot}$ and added a two-layer bidirectional gated recurrent unit (GRU) followed by one FC layer after it (Please see Fig.2), and the spatio-temporal model was trained on fixed-size 16-frame input sequences with the cross-entropy loss function.

Implementation Details – The size of hidden state of GRU was set on 1024 in our experiments, and we implemented our model in Pytorch and trained it for 50 epochs using Adam [D] with an initial learning rate of 0.0002, and batch size 20. During training, we applied random cropping for data augmentation.

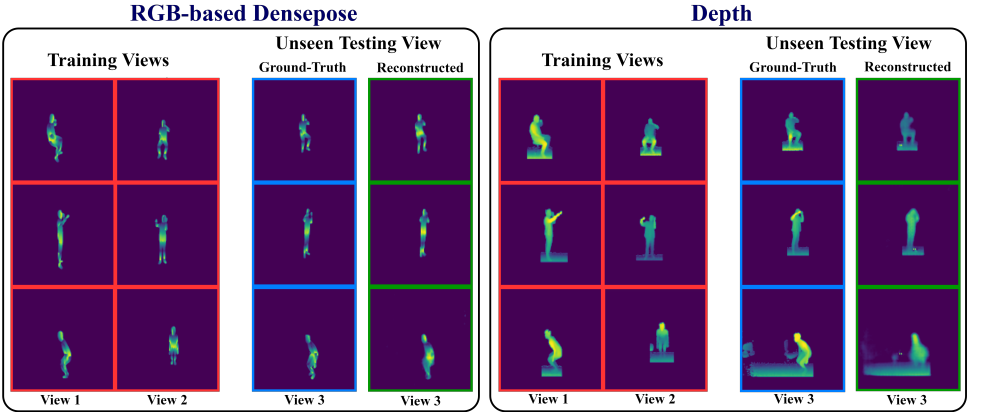

Figure 1: The reconstruction results of our approach on some samples of NTU. Every row shows the simultaneous frames belong to the the same scene from different viewpoints, views 1-3, and for RGB-based Densepose and depth modalities. The red boxes denote training frames and the blue one indicates their corresponding frame captured from unseen viewpoint.

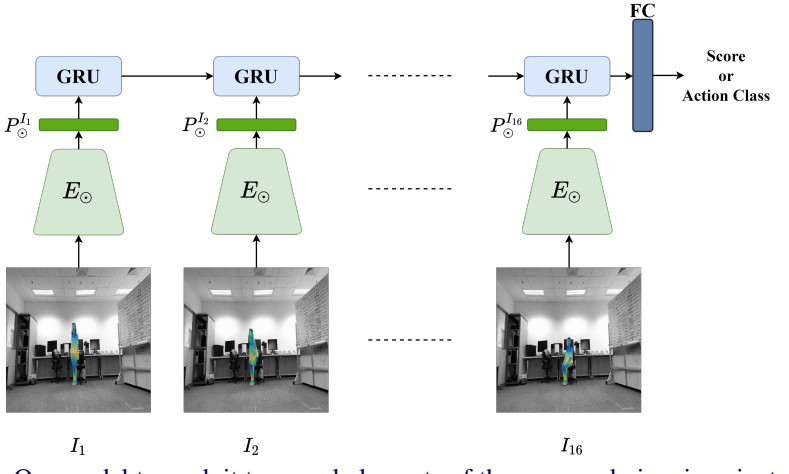

Figure 2: Our model to exploit temporal elements of the proposed view-invariant pose representations for action recognition and human movement analysis downstream tasks.

3 Training and Testing Sets on QMAR

QMAR [4] has been captured from 6 different views, three frontal views at $-45°$, $0°$ and $45°$ and three side views at $45°$, $90°$ and $135°$. The dataset contains RGB images for all views while Depth data are available only for views at $0°$, $90°$. For evaluation, we obtained the RGB results under CV and CS protocols. For CS, we used the same training and testing sets as in [4], and for CV, the data from viewpoints $0°$ and $90°$ were used for training while the rest of viewpoints were applied for testing.

4 The Most Recent State-of-the-art Approaches on NTU RGB+D from CVPR 2021

Table 1 shows results of the most recent state-of-of-the-art representation learning approaches on NTU from CVPR 2021. Among these methods [2, 3] that obtained unsupervised results are based on 3D skeleton data, whereas our proposed method is trained and tested from 2D images. In [5], although the proposed approach is applied on RGB-based 2D pose in inference, it still requires 3D skeleton data in their training process.

Method	Backbone	Pre-train	Input		Supervised (%)		Unsupervised (%)	
			Training	Inference	CV	CS	CV	CS
U-S-VAE [3]	LSTM	NTU	3D Pose	3D Pose	-	-	64.88	50.96
Li et al. [2] ✓	GCN	NTU	3D Pose	3D Pose	**92.5**	**86.2**	**84.3**	**77.8**
CV-MIM [6] ✓	1D CNN	Human3.6M	2D + 3D Pose	2D Pose	<u>89.5</u>	<u>77.8</u>	-	-

Table 1: State-of-the-art action recognition accuracy results on NTU from CVPR 2021. The ✓ symbol highlights view-invariant methods. The best and the second-best results are in **Bold** and <u>underline</u> respectively.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. <u>arXiv preprint arXiv:1412.6980</u>, 2014.

[2] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. <u>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)</u>, 2021.

[3] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. iMiGUE: An Identity-Free Video Dataset for Micro-Gesture Understanding and Emotion Analysis. In <u>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</u>, pages 10631–10642, 2021.

[4] Faegheh Sardari, Adeline Paiement, Sion Hannuna, and Majid Mirmehdi. VI-Net—View-Invariant Quality of Human Movement Assessment. <u>Sensors</u>, 20(18):5258, 2020.

[5] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, and Ting Liu. Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization. <u>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)</u>, 2021.