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

Abstract

Most recent view-invariant action recognition and performance assessment approaches rely on a large amount of annotated 3D skeleton data to extract view-invariant features. However, acquiring 3D skeleton data can be cumbersome, if not impractical, in in-the-wild scenarios. To overcome this problem, we present a novel unsupervised approach that learns to extract view-invariant 3D human pose representation from a 2D image without using 3D joint data. Our model is trained by exploiting the intrinsic view-invariant properties of human pose between simultaneous frames from different viewpoints and their equivariant properties between augmented frames from the same viewpoint. We evaluate the learned view-invariant pose representations for two downstream tasks. We perform comparative experiments that show improvements on the state-of-the-art unsupervised cross-view action classification accuracy on NTU RGB+D by a significant margin, on both RGB and depth images. We also show the efficiency of transferring the learned representations from NTU RGB+D to obtain the first ever unsupervised cross-view and cross-subject rank correlation results on the multi-view human movement quality dataset, QMAR, and marginally improve on the-state-of-the-art supervised results for this dataset. We also carry out ablation studies to examine the contributions of the different components of our proposed network. Our code is available at https://github.com/fsardari/U-VI.

1 Introduction

RGB based deep learning approaches such as [2, 10, 13, 18, 28, 30, 30, 41, 46] have shown an impressive performance in human action recognition and performance assessment. However, as stated in [43], the performance of such approaches drops significantly when they are applied on data that come from unseen viewpoints. To tackle this problem, a simple solution would be to train a network on data from multiple views [43]. However, in practice, capturing a labelled dataset of different views is cumbersome and rare - but two example cases are the commonly used NTU [38] and the recent health-related QMAR [37] datasets, where the granularity of the labelling is still coarse at action class level and at overall performance score level respectively. Ideally, a wholly view-invariant approach would be trained on data

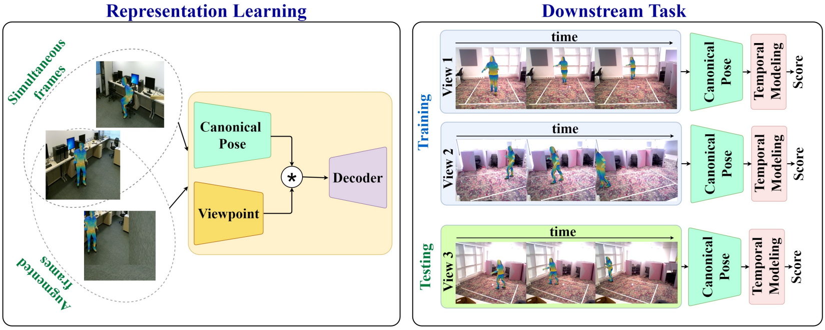

Figure 1: Left: the proposed network learns to disentangle canonical 3D human pose representations and view-dependent features through simultaneous frames from different views and augmented frames from the same view. Right: the unsupervised learned canonical pose representation can be used for downstream tasks.

from as few views as possible and be able to perform well on a single (unseen) view at inference time. We shall use the NTU and QMAR datasets in our downstream tasks for our proposed view-invariant pose representation method towards this ideal.

Most current view-invariant action classification approaches are based on supervised learning, with both training and testing commonly carried out using skeleton data, such as [14, 16, 27, 48, 49]. Others like [6, 7, 31, 36] train on both RGB and 3D joint annotations to facilitate testing using RGB images alone. However, all these approaches rely on a significant amount of annotations (and sometimes camera parameters) during training, the provision of which is expensive and difficult in in-the-wild scenarios. Only very few works, such as [37, 43, 45] deal with training from RGB images only. Moreover, the authors are aware of no other unsupervised view-invariant RGB-only action classification or assessment study, and only of one such work based on RGB and depth [21].

In this paper, we propose a representation learning approach to disentangle canonical (view-invariant) 3D pose representation and view-dependent features from either an RGB-based 2D Densepose human representation map [26] *or* a depth mask image without using 3D skeleton annotations or camera parameters. We design an auto-encoder comprising two encoders and a decoder. The first is a view-invariant 3D pose encoder that learns 3D canonical pose representations from an input image, and the second is a viewpoint encoder that extracts rotation and translation parameters, such that when they are applied on the canonical pose features, it would result in view-dependent 3D pose representation which are fed into the decoder to reconstruct the input image. To train the network, we impose geometrical and positional order consistency constraints on pose representation features through novel view-invariant and equivariance losses respectively. The view-invariant loss is computed based on the intrinsic view-invariant properties of pose features between simultaneous frames from different viewpoints, while the equivariance loss is computed using the equivariant properties between augmented frames from the same viewpoint. After training, the 3D canonical pose representations can be used for downstream tasks such as view-invariant action classification and human movement analysis. Fig. 1 shows the proposed view-invariant pose representation learning framework and its application on a view-invariant downstream task.

Our key contributions can be summarized as follows: (i) we propose a novel unsupervised method that learns view-invariant 3D pose representation from a 2D image without using 3D skeleton data and camera parameters. Our view-invariant features can be applied *directly* by downstream tasks to be resilient to human pose variations in unseen viewpoints, unlike unsupervised 3D pose estimation methods such as [3, 4, 11, 15, 33, 42] which obtain view-specific 3D pose features, and require camera parameters and further steps to align their view-specific features in a canonical space, (ii) we introduce novel view-invariance and equivariance losses that impose on the network to preserve geometrical and positional order consistency of pose features - these losses can benefit the training process in other pretext tasks that exploit landmark representation, (iii) we evaluate the performance of learned pose features on two downstream tasks that demand view-invariancy and achieve state-of-the-art unsupervised cross-view action recognition accuracy on the NTU RGB+D standard benchmark dataset for RGB and depth images at 74.8% and 67.5% respectively, and for the first time we obtain unsupervised cross-view and cross-subject rank correlation results for human movement assessment scores on the QMAR dataset, while exceeding its supervised state-of-the-art results, (iv) we perform ablation studies to explore the impact of our loss functions on our proposed model.

2 Related Works

We now consider *the more recent* related works that deal with unsupervised pose representation and view-invariancy, in particular in relation to our chosen downstream tasks.

Unsupervised Pose Representation – There are several recent examples of RGB-based unsupervised learning approaches to 3D pose estimation, such as [3, 4, 11, 15, 33, 42]. Authors in [3, 4, 42] extract unsupervised pose features from 2D joints generated from RGB data. For example, Chen et al. [3] train a network through a 2D-3D consistency loss, computed after lifting 2D pose to 3D joints and reprojecting 3D onto 2D. Dundar et al. [11] disentangle pose and appearance features from an RGB image by designing a self-supervised auto-encoder that reconstructs an input image into foreground and background with the constraint that the appearance features remain consistent temporally while the pose features change. Honari et al. [15] also relies on temporal information and factorizes the pose and appearance features in a contrastive learning manner. In [33], the authors design a network to encode 3D pose features by predicting a different viewpoint of the input image, but there is no restriction to generate the same pose representation for the simultaneous frames. All these methods are view-specific and do not generate the same (*i.e.* canonical) 3D pose features for different viewpoints, so they cannot be applied to unseen-view downstream tasks, and camera parameters and extra steps are needed to map their view-specific output into a canonical view. Our proposed method learns view-invariant pose representation from the input image such that it can be applied *directly* to unseen-view tasks, such as action recognition. Rhodin et al. [34] use both labelled and unlabelled data to estimate canonical 3D pose. They train a network that maps multiple views into a canonical pose through mean square error, but as using only this constraint may generate random features without any positional order consistency, they also use a small subset of 3D pose annotations to enhance the output. However, our proposed approach achieves positional order consistency without utilizing any labels. Note, positional order consistency is important as it enables us to leverage temporal aspects of corresponding body joints to handle video-based downstream tasks.

Supervised View-Invariant Action Recognition and Performance Assessment – To deal with view-invariancy, most action recognition methods are based on 3D skeleton joints, such as [14, 16, 17, 22, 31, 36, 48, 49]. For example, Zhang et al. [48] present a dual-stream network, one LSTM and one CNN, and fuse the results to predict the action label.

Both streams include a view adaptation network estimating the transformation parameters of skeleton data to a canonical view, followed by a classifier. In general, methods that rely on skeleton annotations must rely on fulsome 3D joint representations which are difficult to come by in in-the-wild scenarios. Recently a few works have developed view-invariant action recognition or analysis approaches from RGB-D images, such as [8, 37, 43, 45]. Varol et al. [43] deploy multi-view synthetic videos for training their network to perform action recognition given novel viewpoints, but still use 3D pose annotations to produce the synthetic data, while the newly generated videos would have to be also labelled by experts if they were to be used for specialist applications such as healthcare.

In view-invariant action performance assessment, we are aware of only one study where Sardari et al. [37] investigated a supervised model to assess and score the quality of movement in subjects simulating Parkinson and Stroke symptoms by evaluating canonical spatio-temporal trajectories derived from body joint heatmaps.

Unsupervised View-Invariant Action Recognition – There are also only a relatively few unsupervised deep learning approaches that challenge view-invariant action recognition, e.g. [5, 21]. For instance, Li et al. [21] introduce an RGB-D based auto-encoder network that extracts unsupervised view-invariant spatio-temporal features from a video sequence. The proposed network is trained to reconstruct two simultaneous source and target view sequences from a given source view video. This method requires both RGB and depth data for training. Cheng et al. [5] propose a 3D skeleton-based unsupervised approach by using motion prediction as a pretraining task to learn temporal dependencies for long video representation with a transformer-based auto-encoder.

3 Proposed Method

Our aim is to learn view-invariant 3D pose representation from 2D RGB or depth images without relying on 3D skeleton annotations and camera parameters. Our method leverages on geometric transformation amongst different viewpoints and the equivariant property of human pose. The proposed auto-encoder includes a view-invariant pose encoder $E_{\odot}$, a viewpoint encoder $E_{\triangleleft}$, and a decoder D arranged as shown in Fig. 2. $E_{\odot}$ learns 3D canonical pose features from a given image which can be either an RGB-based 2D Densepose human representation map [26] or a depth mask image. As the extracted pose features are canonical, they are mapped into a specific viewpoint using the parameters obtained through encoder $E_{\triangleleft}$ before being passed to D to allow the decoder to reconstruct the input image. The network optimises through four losses to generate its view-invariant representation.

Model Architecture and Formulation – The view-invariant pose encoder $E_{\odot}$ learns 3D canonical pose features $P^I_{\odot} = E_{\odot}(I)$ given image $I \in \mathrm{I\!R}^{3\times \mathrm{W} \times \mathrm{H}}$ where $P^I_{\odot} \in \mathrm{I\!R}^{3\times \mathrm{N}}$ and N refers to the number of 3D pose features. $E_{\triangleleft}$ estimates the viewpoint parameters $(R^I, T^I) = E_{\triangleleft}(I)$, i.e. rotation $R^I = (\theta_x, \theta_y, \theta_z)$ and translation $T^I = (t_x, t_y, t_z)$. These viewpoint parameters are applied on the canonical pose features $P^I_{\odot}$ to transfer them into a specific viewpoint $P^I_{\circledast}$, such that $P^I_{\circledast} = R^I \cdot P^I_{\odot} + T^I$ where $P^I_{\circledast} \in \mathrm{I\!R}^{3\times \mathrm{N}}$. Then, the decoder reconstructs the input, $\tilde{I} = D(P^I_{\circledast})$. The network's purpose is therefore that it learns to extract the same canonical 3D pose features for simultaneous frames from different viewpoints while maintaining equivariance for the pose features from their augmented frames (shifted in position - see details under Equivariance Loss below) from the same viewpoint. We train the proposed network by combining four losses, view-invariant $\mathcal{L}_{invar}$, equivariance $\mathcal{L}_{equiv}$, and two reconstruction losses $\mathcal{L}_{rec1}$ and $\mathcal{L}_{rec2}$.

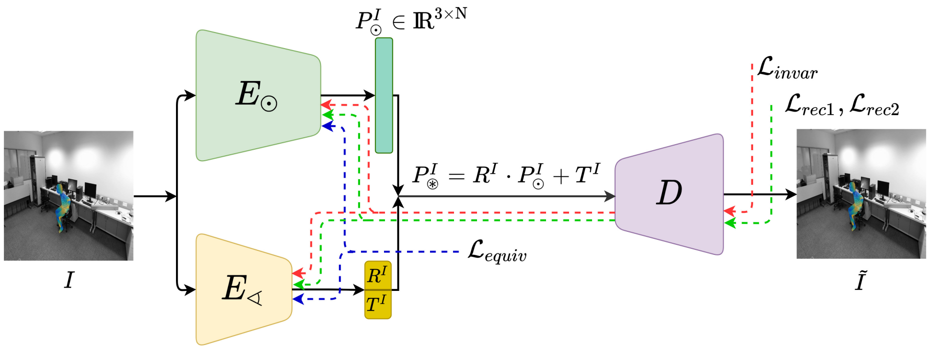
$P^I_{\odot} \in \mathbb{R}^{3\times N}$
$E_{\odot}$
$\mathcal{L}_{invar}$
$\mathcal{L}_{rec1}, \mathcal{L}_{rec2}$
$P^I_{\circledast} = R^I \cdot P^I_{\odot} + T^I$
D
$\mathcal{L}_{equiv}$
$E_{\triangleleft}$
R^I
T^I
I
$\tilde{I}$

Figure 2: The overall schema of the proposed view-invariant posture representation learning architecture.

View-Invariant Loss – We start with two simultaneous frames (I^v_k, I^w_k) from different views v and w of the same scene from their corresponding video sequences at current frame k. These are passed to encoders $E_{\odot}$ and $E_{\triangleleft}$ to extract the canonical 3D pose features $P^{I^{\phi}_k}_{\odot} = E_{\odot}(I^{\phi}_k)$ and viewpoint parameters $(R^{\phi}_k, T^{\phi}_k) = E_{\triangleleft}(I^{\phi}_k)$, for $\phi \in \{v, w\}$.

Each frame k has a distinct translation parameter, while the rotation is the same for all the frames of a sequence captured from the same viewpoint. Thus, if we estimate the rotation parameters from two random frames I^v_m and I^w_n from corresponding sequences and views instead, the network should still retrieve the view-specific pose features. We use this constraint to prevent the model leaking any pose information through $E_{\triangleleft}$ and force it to concentrate on only the viewpoint parameters. Hence, with a probability of 0.5, we randomly select the frame to predict the rotation parameters for the two views,

$$R^v = \begin{cases} R^v_k & \text{if } r \text{ is} < 0.5 \\ R^v_m & \text{else} \end{cases} \quad \text{and} \quad R^w = \begin{cases} R^w_k & \text{if } r \text{ is} < 0.5 \\ R^w_n & \text{else} \end{cases}, \quad r \in U(0,1), \tag{1}$$

where $U(0,1)$ denotes a uniform distribution returning a number between 0 and 1. As we assume $E_{\odot}$ encodes the same canonical 3D pose features for I^v_k and I^w_k, then swapping their pose features while their viewpoint features are retained, the network has to still be able to reconstruct them. Thus, the view-invariant loss is obtained by

$$\tilde{I}'^v_k = D(P'^{I^v_k}_{\circledast}) \;\text{ where }\; P'^{I^v_k}_{\circledast} = R^v \cdot P^{I^w_k}_{\odot} + T^v_k, \tag{2}$$

$$\tilde{I}'^w_k = D(P'^{I^w_k}_{\circledast}) \;\text{ where }\; P'^{I^w_k}_{\circledast} = R^w \cdot P^{I^v_k}_{\odot} + T^w_k, \tag{3}$$

$$\mathcal{L}_{invar} = \sum_{\phi \in \{v,w\}} MSE(I^{\phi}_k, \tilde{I}'^{\phi}_k), \tag{4}$$

where MSE indicates the mean square error.

However, computing only $\mathcal{L}_{invar}$ is not enough to learn the view-invariant pose features, and $E_{\odot}$ still has to reconstruct the simultaneous frames even without swapping their canonical pose features, otherwise the network learns to only assign random latent codes for canonical pose features, so we introduce $\mathcal{L}_{rec1}$ as

$$\mathcal{L}_{rec1} = \sum_{\phi\in\{v,w\}} MSE(I_k^{\phi},\, \tilde{I}_k^{\phi})\,, \tag{5}$$

where $\tilde{I}_k^{\phi} = D(P_{\circledast}^{I_k^{\phi}})$ with $P_{\circledast}^{I_k^{\phi}} = R^{\phi} \cdot P_{\odot}^{I_k^{\phi}} + T_k^{\phi}$ for $\phi \in \{v,w\}$.

Equivariance Loss – The effect of this loss is to help teach the network to preserve the positional order of the pose components. For example, if the i^{th} dimension of the latent variable indicates the right shoulder of a subject, it should be consistent for all the images. We assume that the proposed network generates consistent order of pose features, and x and y axes of view-specific 3D pose space are the same as the x and y directions of the 2D images, so when I_k^v and I_k^w shift by some pixels in the x and y directions, then all components of the view-specific pose $P_{\circledast}^{I_k^{\phi}}$ would shift similarly. Hence, we propose an equivariance loss computed from augmentations of I_k^v and I_k^w, where the augmented images, $\dot{I}_k^v$ and $\dot{I}_k^w$, represent positional changes of the human subject in the scene, for example by c_1 and c_2 pixels respectively, i.e.

$$\mathcal{L}_{equiv} = \sum_{\phi\in\{v,w\},j\in\{1,2\}} MSE(P_{\circledast}^{I_k^{\phi}} + c_j,\, P_{\circledast}^{\dot{I}_k^{\phi}}), \tag{6}$$

where $P_{\odot}^{\dot{I}_k^{\phi}} = E_{\odot}(\dot{I}_k^{\phi})$ and $P_{\circledast}^{\dot{I}_k^{\phi}} = R^{I^{\phi}} \cdot P_{\odot}^{\dot{I}_k^{\phi}} + T^{\dot{I}_k^{\phi}}$ for $\phi \in \{v,w\}$.

$\mathcal{L}_{equiv}$ is computed based on the view-specific pose features while we can also benefit from the reconstruction of the augmented frames to improve on the pose representation, so we introduce $\mathcal{L}_{rec2}$ as.

$$\mathcal{L}_{rec2} = \sum_{\phi\in\{v,w\}} MSE(\dot{I}_k^{\phi},\, \tilde{\dot{I}}_k^{\phi})\,, \tag{7}$$

where $\tilde{\dot{I}}_k^{\phi} = D(P_{\circledast}^{\dot{I}_k^{\phi}})$. The total loss is computed as

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{invar} + \beta \cdot \mathcal{L}_{equiv} + \gamma \cdot (\mathcal{L}_{rec1} + \mathcal{L}_{rec2}) \tag{8}$$

We determine the weights empirically to be $\alpha = 1.0$, $\beta = 0.001$, and $\gamma = 1.0$. After training the proposed network to learn the 3D canonical pose features, $E_{\odot}$ is used for our example view-invariant downstream tasks. Next, in Section 4, we outline our proposed method to model temporal aspects of the canonical pose features for each downstream task.

4 Experiments

Datasets – Capturing multi-view datasets requires elaborate set-ups and is inevitably time-consuming and potentially quite expensive. Hence, very few multi-view action classification or movement assessment datasets exist. We present results on two existing datasets, NTU RGB+D and QMAR.

NTU RGB+D [38] is the main benchmark dataset for view-invariant action recognition, including 60 action classes performed by 40 subjects. NTU (for short) contains 17 different environmental settings captured by three cameras from three viewpoints. Two standard protocols are used to evaluate the performance of view-invariant action recognition methods on NTU, cross-view (CV) and cross-subject (CS). In CV, different views are adopted for training and testing, while in CS different subjects are engaged for training and testing. We followed both protocols by using the same training and testing sets as in [38] for both pretext and downstream tasks.

QMAR [37] is the only RGB+D multi-view dataset known to the authors for quality of movement assessment. It comprises 38 subjects captured from 6 different views, three frontal and three sides. The subjects were trained by an expert to simulate four Parkinsons and Stroke movement tests: walking with Parkinson (W-P), walking with Stroke (W-S), sit-to-stand with Parkinson (SS-P), and sit-to-stand with Stroke (SS-S), and the movements were annotated to determine severity of the abnormality scores. For evaluation, we followed [29, 37] and obtained Spearman's rank correlation (SRC) results under CV and CS protocols. For CS, we used the same training and testing sets as in [37], and for CV, the data from one frontal and one side view were used for training while the rest of the viewpoints were applied for inference (see Supplementary Materials for details).

Implementation Details and Hyper-Parameter Settings – Our auto-encoder is inspired by the U-Net encoder/decoder [9, 12, 33, 35]. The U-Net is a convolutional or spatial latent auto-encoder with skip connections between the encoder and the decoder parts, while we desire a dense one [1] without the skip connections to encode the 3D pose features, so we adapted it for our problem. Table 1 shows details of the proposed network architecture. We implemented our model in Pytorch and trained it for 20 epochs using Adam [20] with a fix learning rate of 0.0002, and batch size 5. During training, we applied random horizontal flipping for data augmentation. The depth mask images of NTU used in our experiments contain bounding box of subjects as released by [38].

Module	**Layers**
$E_\odot$	$\{C2(3\times3,64)$, BN, $ReLU\}\times2$, $MP(2\times2)$, $\{C2(3\times3,128)$, BN, $ReLU\}\times2$, $MP(2\times2)$, $\{C2(3\times3,256)$, BN, $ReLU\}\times2$, $MP(2\times2)$, $\{C2(3\times3,512)$, BN, $ReLU\}\times1$, $\{C2(3\times3,512)$, $ReLU\}\times1$, $\{FC(1024)$, $ReLU\}$, $\{FC(512)$, $ReLU\}$, $\{FC(3\times70)\}$
$E_\triangleleft$	$\{C2(5\times5,128)$, BN, $ReLU\}\times2$, $MP(7\times7)$, $\{C2(5\times5,256)$, BN, $ReLU\}\times2$, $\{FC(512)$, $ReLU$, $Drp\}$, $\{FC(6)\}$
D	$\{FC(16\times16\times512)$, $ReLU$, $Drp\}$, $\{C2(3\times3,256)$, BN, $ReLU\}\times2$, $\{CT2(3\times3,128)$, BN, $ReLU\}\times2$, $\{CT2(3\times3,64)$, BN, $ReLU\}\times2$, $\{CT2(3\times3,3)$, BN, $ReLU\}\times2$, $tanh$

Table 1: The proposed auto-encoder's modules – All modules are 2D. $C2(d\times d,ch)$: $d\times d$ convolution filters with ch channels, $CT2$: transposed convolution filters, MP: max pooling, BN: batch normalization, $FC(O)$: FC layer with O outputs.

To select the 3D canonical pose feature size $P^I_\odot \in \mathbb{R}^{3\times\mathrm{N}}$, we used cross-validation and evaluated the total loss $\mathcal{L}_{total}$ in Eq. 8 for N in the range between 40 and 190 with a step-size of 30. The lower bound was inspired by motion capture systems that use 39 markers, and the upper bound was selected based on Rhodin et al. [33] who set their latent code size at 3×200. As shown in Table 2, the average $\mathcal{L}_{total}$ cross-validation results on the NTU dataset for both CV and CS protocols is best when $N=70$, hence our 3D canonical pose feature size is set at 3×70.

Action Classification – Our proposed auto-encoder can learn unsupervised 3D pose representations without using any action labels. To encapsulate the temporal element of the action recognition downstream task, we added a two-layer bidirectional gated recurrent unit (GRU) followed by one FC layer after our view-invariant pose encoder $E_\odot$, and trained it on

N $\mathcal{L}_{total}$		40	70	100	130	160	190
RGB	CV	0.0088	**0.0080**	0.0084	0.0081	0.0081	0.0082
	CS	0.0064	**0.0061**	**0.0061**	0.0062	0.0062	0.0062
Depth	CV	**0.016**	**0.016**	**0.016**	0.017	0.017	0.017
	CS	**0.015**	**0.015**	0.016	0.016	0.016	0.016

Table 2: Optimising $P^{I}_{\odot}$ - Average $\mathcal{L}_{total}$ cross-validation results on NTU for different canonical pose size ($3 \times N$).

Method		**Backbone**	**Input**	**Supervised** (%)				**Unsupervised** (%)	
				scratch		**fine-tune**			
				CV	**CS**	**CV**	**CS**	**CV**	**CS**
Shuffle & Learn [25]	ECCV 2016	AlexNet	Depth	-	-	-	-	40.9	46.2
Luo et al. [24]	CVPR 2017	VGG + ConvLSTM	Depth	-	-	-	-	53.2	61.4
Vyas et al. [44] ✓	ECCV 2020	3D CNN + LSTM	Depth	-	-	$\underline{78.7}$	$\underline{71.8}$	-	-
Li et al. [21] ✓	NeurIPS 2018	2D ResNet + ConvLSTM	Depth	37.7	42.3	63.9	68.1	53.9	$\underline{60.8}$
Ours ✓		2D ResNet + LSTM	Depth	$\underline{60.4}$	$\underline{63.1}$	75.5	72.7	$\underline{58.3}$	58.0
Ours ✓		2D CNN + GRU	Depth	**76.7**	**75.9**	**82.5**	**78.8**	**67.5**	**64.7**
Luo et al. [24]	CVPR 2017	VGG + ConvLSTM	RGB	-	-	-	-	-	56.0
Vyas et al. [44] ✓	ECCV 2020	3D CNN + LSTM	RGB	-	-	**86.3**	**82.3**	-	-
Li et al. [21] ✓	NeurIPS 2018	2D ResNet + ConvLSTM	RGB	29.2	36.6	49.3	55.5	40.7	48.9
Ours ✓		2D ResNet + LSTM	RGB	$\underline{66.5}$	$\underline{66.7}$	78.2	73.8	$\underline{62.1}$	$\underline{63.0}$
Ours ✓		2D CNN + GRU	RGB	**77.0**	**70.3**	$\underline{83.6}$	$\underline{78.1}$	**74.8**	**68.3**

Table 3: Action classification accuracy on NTU for RGB and depth based representation learning approaches. The ✓ symbol highlights view-invariant methods. Our unsupervised results were obtained after freezing $E_{\odot}$'s parameters during the downstream task while the supervised results were obtained by both fine-tuning $E_{\odot}$ and training it from scratch. The best and the second-best results are in **Bold** and underline respectively.

fixed-size 16-frame input sequences with the cross-entropy loss function. Similar to [50], we subsampled the sequences such that every sequence was divided into 16 segments and one random frame was selected amongst all frames of each segment.

All the representation learning methods on NTU that are compared to ours here can operate on either RGB or depth data for training and inference, except Li et al. [21] which requires both RGB and depth for its training stage. Providing like-to-like evaluations against these relevant methods is difficult since for all such techniques their method defines the nature of their backbone architecture, for example they extract spatio-temporal features while we learn pose representation, *e.g.* [44] uses 3D CNNs whereas ours is integrally a 2D design. In the case of [21] which applies a 2D ResNet with added ConvLSTM [39], we provide results with the closest possible backbone, comprising a 2D ResNet and an LSTM.

Table 3 shows that for the unsupervised scenario, our 2D CNN and GRU backbone significantly improves the state-of-the-art across CV and CS tests, at 74.8%, 68.3% for RGB, and 67.5%, 64.7% for depth data, respectively. The 2D ResNet + LSTM incarnation of our method also exceeds across the board on the state-of-the-art in unsupervised results on NTU, e.g. achieving 62.1% in almost direct comparison to [21]'s 40.7% for cross-view RGB inference. For the supervised learning case, we improve on all other works with depth data whether training from scratch or fine-tuning our network with best results at 82.5% and 78.8% on CV and CS protocols respectively, and attain very competitive results using RGB in comparison to the 3D CNN-based [44].

In Table 4, we report the results of recent state-of-the-art unsupervised pose representation methods that operate on 3D skeleton data. Yao et al. [47] perform better than our method in CV mode and Cheng et al. [5]'s result is marginally better than ours in CS mode. These result vindicate our approach as a viable alternative to skeleton-based methods which are altogether more cantankerous to deal with in real-world applications than RGB or depth derived data.

Method		**Backbone**	**Unsupervised (%)** **CV**	 **CS**
Su et al. [40]	CVPR 2020	GRU	<u>76.1</u>	50.7
Lin et al. [23]	ACM Multimedia 2020	GRU	-	52.5
Yao et al. [47]	ICME 2021	GRU + GCN	**79.2**	54.4
Cheng et al. [5] ✓	ICME 2021	Transformer	72.8	**69.3**
Rao et al. [32] ✓	Information Sciences 2021	LSTM	64.8	<u>58.5</u>

Table 4: State-of-the-art action recognition accuracy results on NTU for skeleton-based representation learning approaches. The ✓ symbol highlights view-invariant methods. The best and the second-best results are in **Bold** and <u>underline</u> respectively.

Ablation Study – We ablate our losses to examine their impact on the learning of our pose features. Table 5 shows the unsupervised action classification accuracy on NTU as we drop each or both of $\mathcal{L}_{invar}$ and $\mathcal{L}_{equiv}$.

Table 5 shows that removing $\mathcal{L}_{equiv}$ from the training process, our results for both CV and CS in both RGB and depth deteriorates. This verifies that positional order consistency is essential in both cases. We also observe that eliminating $\mathcal{L}_{invar}$ causes our method's performance to drop in all cases, except for the cross-subject case with depth as the input modality. The increase in performance in this scenario may be attributed to the removal of the extra geometrical constraints that are imposed on the features by the extra simultaneous frames through the presence of the $\mathcal{L}_{invar}$ computation.

Ours with	**Depth** **CV(%)**	 **CS(%)**	**RGB** **CV(%)**	 **CS(%)**
$\mathcal{L}_{rec1} + \mathcal{L}_{rec2}$	35.4	32.1	35.6	34.1
$\mathcal{L}_{equiv} + \mathcal{L}_{rec1} + \mathcal{L}_{rec2}$	<u>64.1</u>	**65.5**	69.1	<u>64.9</u>
$\mathcal{L}_{invar} + \mathcal{L}_{rec1} + \mathcal{L}_{rec2}$	59.6	52.5	<u>70.3</u>	63.3
$\mathcal{L}_{invar} + \mathcal{L}_{equiv} + \mathcal{L}_{rec1} + \mathcal{L}_{rec2}$	**67.5**	<u>64.7</u>	**74.8**	**68.3**

Table 5: Ablation studies on different combinations of losses used in the unsupervised learning process. The best and the second-best results are in **Bold** and <u>underline</u> respectively.

Human Movement Analysis – Here, we aim to study the efficiency of the learned representation on NTU for quality of movement scoring on QMAR. As in the action recognition task, we added a two-layer bidirectional GRU followed by one FC layer on top of $E_{\odot}$ to deal with temporal analysis. The size of the FC layer is equal to the maximum score for a movement type. However, for movement quality assessment, we require to analyse every single frame of a sequence, so we cannot apply any subsampling strategies for this task. We followed [37] to divide each video sequence into non-overlapping 16-frame video clips. Our network was trained on a random 16-frame clip through the cross entropy loss function, and for inference, all 16-frame clips of a video sequence were processed. The score for a sequence was estimated by averaging the outputs of the last FC layer, as in [37]. Our work offers the first ever unsupervised results on QMAR, and for further direct comparison, we also show the results of our implementation of [21]. Sardari et al. [37] who introduced QMAR present the only other supervised view-invariant results on this dataset. We also provide the performance of two other architectures, taken from [37].

As shown in Table 6, our unsupervised human movement analysis results on QMAR

		Method	Training	Action (SRC)				Average
				W-P	W-S	SS-P	SS-S	(SRC)
CV	Supervised	C3D (after [29])	custom-trained	0.65	0.37	0.21	0.45	0.42
		I3D [2]	fine-tune	0.87	0.71	0.40	0.63	0.65
		VI-Net [37] ✓	scratch	**0.92**	**0.81**	0.46	0.61	0.70
		Li et al. [21]	scratch	0.22	0.15	0.10	0.11	0.14
		Li et al. [21] ✓	fine-tune	0.14	0.38	0.14	0.3.5	0.20
		Ours	scratch	0.81	0.58	0.16	0.53	0.52
		Ours ✓	fine-tune	0.90	0.76	**0.58**	**0.64**	**0.72**
	Unsupervised	Li et al. [21] ✓	-	0.12	0.13	0.06	0.17	0.12
		Ours ✓	-	**0.78**	**0.66**	**0.29**	**0.54**	**0.54**
CS	Supervised	C3D (after [29])	custom-trained	0.50	0.37	0.25	0.54	0.41
		I3D [2]	fine-tune	0.79	0.47	0.54	0.55	0.58
		VI-Net [37] ✓	scratch	0.87	0.52	0.58	0.69	0.66
		Li et al. [21]	scratch	0.55	0.32	0.39	0.64	0.47
		Li et al. [21] ✓	fine-tune	0.57	0.59	0.41	0.75	0.57
		Ours	scratch	0.81	0.51	0.39	0.72	0.60
		Ours ✓	fine-tune	**0.89**	**0.54**	**0.62**	**0.76**	**0.70**
	Unsupervised	Li et al. [21] ✓	-	0.21	0.10	0.24	0.47	0.25
		Ours ✓	-	**0.70**	**0.50**	**0.48**	**0.66**	**0.58**

Table 6: Spearman's rank correlation (SRC) between predicted scores and ground truth labels for cross-view and cross-subject analysis on different actions of QMAR dataset. I3D was pretrained on Kinetic-400 [19], and the ✓ symbol highlights view-invariant methods. The best supervised and unsupervised results are in **Bold** and the second-best supervised results are underlined.

outperforms Li et al. [21], reaching an average SRC of 0.54 and 0.58 for CV and CS respectively. These are broadly already competitive to the supervised results on QMAR, particularly when compared against the Kinetic-400 [19] pretrained, deep I3D network. Finally, the supervised version of our method, where we fine-tune our network weights after transfering the weights learnt through NTU training, exceeds Sardari et al. [37]'s performance on average and achieves 0.72 and 0.70 for CV and CS respectively.

5 Conclusion

Most current *view-invariant* action recognition and performance assessment approaches are based on supervised learning and rely on a large number of 3D skeleton annotations. In this paper, we dealt with these through an unsupervised method to learn view-invariant 3D pose representation from a 2D image. Our experiments show that not only can our learned pose representations be applied on unseen view videos from the same training data, but it can also be used in different domains. Our unsupervised approach is particularly helpful in applications where the use of multi-view data is essential and capturing 3D skeletons is challenging, e.g. in healthcare rehabilitation monitoring at home or in the clinic.

In the pretext stage of our model, we require synchronised multi-view frames to learn view-invariant 3D pose representations. For future work, we will investigate extracting view-invariant pose features from a single view or non-synchronized frames to allow learning to become a simpler process for application to any suitable dataset.

Acknowledgements

The 1st author is funded by a University of Bristol Overseas Postgraduate Student Scholarship.

References

[1] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018.

[2] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[3] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James Rehg. Unsupervised 3D Pose Estimation With Geometric Self-Supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5714–5724, 2019.

[4] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10895–10904, 2019.

[5] Yi-Bin Cheng, Xipeng Chen, Junhong Chen, Pengxu Wei, Dongyu Zhang, and Liang Lin. Hierarchical Transformer: Unsupervised Representation Learning for Skeleton-Based Human Action Recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[6] Srijan Das, Arpit Chaudhary, Francois Bremond, and Monique Thonnat. Where to Focus on for Human Action Recognition? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–80. IEEE, 2019.

[7] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. VPN: Learning Video-Pose Embedding for Activities of Daily Living. In *European Conference on Computer Vision (ECCV)*, pages 72–90. Springer, 2020.

[8] Chhavi Dhiman and Dinesh Kumar Vishwakarma. View-Invariant Deep Architecture for Human Action Recognition Using Two-stream Motion and Shape Temporal Dynamics. *IEEE Transactions on Image Processing*, 29:3835–3844, 2020.

[9] Michael Dorkenwald, Uta Buchler, and Bjorn Ommer. Unsupervised Magnification of Posture Deviations Across Subjects. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8256–8266, 2020.

[10] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The Pros and Cons: Rank-Aware Temporal Attention for Skill Determination in Long Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 7862–7871, 2019.

[11] Aysegul Dundar, Kevin Shih, Animesh Garg, Robert Pottorff, Andrew Tao, and Bryan Catanzaro. Unsupervised Disentanglement of Pose, Appearance and Background from Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[12] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational U-Net for Conditional Appearance and Shape Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8857–8866, 2018.

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

[14] Kumie Gedamu, Yanli Ji, Yang Yang, LingLing Gao, and Heng Tao Shen. Arbitrary-View Human Action Recognition via Novel-View Action Generation. *Pattern Recognition*, page 108043, 2021.

[15] Sina Honari, Victor Constantin, Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised Learning on Monocular Videos for 3D Human Pose Estimation. *arXiv preprint arXiv:2012.01511*, 2021.

[16] Qingqing Huang, Fengyu Zhou, Runze Qin, et al. View Transform Graph Attention Recurrent Networks for Skeleton-based Action Recognition. *Signal, Image and Video Processing*, 15(3):599–606, 2021.

[17] Yanli Ji, Feixiang Xu, Yang Yang, Ning Xie, Heng Tao Shen, and Tatsuya Harada. Attention Transfer (ANT) Network for View-Invariant Action Recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 574–582, 2019.

[18] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late Temporal Modeling in 3D CNN Architectures with Bert for Action Recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020.

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[20] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan Kankanhalli. Unsupervised Learning of View-Invariant Action Representations. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 1254–1264, 2018.

[22] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3595–3603, 2019.

[23] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. MS2L: Multi-Task Self-Supervised Learning for Skeleton Based Action Recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498, 2020.

[24] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised Learning of Long-term Motion Dynamics for Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2203–2212, 2017.

[25] Ishan Misra, Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *European Conference on Computer Vision (ECCV)*, pages 527–544. Springer, 2016.

[26] Natalia Neverova, David Novotny, and Andrea Vedaldi. Correlated Uncertainty for Learning Dense Correspondences from Noisy labels. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[27] Qiang Nie and Yunhui Liu. View Transfer on Human Skeleton Pose: Automatically Disentangle the View-Variant and View-Invariant Information for Pose Representation Learning. *International Journal of Computer Vision (IJCV)*, 129(1):1–22, 2021.

[28] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action Assessment by Joint Relation Graphs. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 6331–6340, 2019.

[29] Paritosh Parmar and Brendan Tran Morris. What and How Well You Performed? A Multitask Learning Approach to Action Quality Assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304–313, 2019.

[30] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano Skills Assessment. *arXiv preprint arXiv:2101.04884*, 2021.

[31] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. Learning a Deep Model for Human Action Recognition from Novel Viewpoints. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(3):667–681, 2018.

[32] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569:90–109, 2021.

[33] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.

[34] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[36] Faegheh Sardari, Adeline Paiement, and Majid Mirmehdi. View-Invariant Pose Analysis for Human Movement Assessment from RGB Data. In *International Conference on Image Analysis and Processing*, pages 237–248. Springer, 2019.

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

[38] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.

[39] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in neural information processing systems*, 28, 2015.

[40] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict and Cluster: Unsupervised Skeleton based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9631–9640, 2020.

[41] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-Aware Score Distribution Learning for Action Quality Assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9839–9848, 2020.

[42] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. PoseNet3D: Learning Temporally Consistent 3D Human Pose via Knowledge Distillation. In *International Conference on 3D Vision (3DV)*, pages 311–321. IEEE, 2020.

[43] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic Humans for Action Recognition from Unseen Viewpoints. *International Journal of Computer Vision (IJCV)*, pages 1–24, 2021.

[44] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multiview Action Recognition Using Cross-View Video Prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[45] Dongang Wang, Wanli Ouyang, Wen Li, and Dong Xu. Dividing and Aggregating Network for Multi-View Action Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 451–467, 2018.

[46] Chao-Yuan Wu and Philipp Krahenbuhl. Towards Long-Form Video Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1884–1894, 2021.

[47] Han Yao, SJ Zhao, Chi Xie, Kenan Ye, and Shuang Liang. Recurrent Graph Convolutional Autoencoder for Unsupervised Skeleton-Based Action Recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[48] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(8):1963–1978, 2019.

[49] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-Guided Neural Networks for Efficient Skeleton-based Human Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1112–1121, 2020.

[50] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained Multi-stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017.