View-Invariant Human Movement Assessment

Faegheh Sardari

A dissertation submitted to the University of Bristol in accordance with the requirements for the degree of Doctor of Philosophy in the Faculty of Engineering

March 16, 2022

43628 words

Abstract

In computer vision, human action or movement assessment is the task of evaluating the quality of a person's movements when they perform specific actions through observing their video. Typically, human movement assessment approaches are view-specific and are not able to assess the quality of human movement when they are applied to camera viewpoints different to their training data, i.e. unseen viewpoints. This thesis explores **view-invariance** in **human movement assessment**, with a particular focus on the healthcare domain. Furthermore, the current approaches in the field of healthcare are based on 3D skeleton data as the features derived from 3D data are rich and can be leveraged to assess a wide range of movements. However, acquiring 3D skeleton data can be cumbersome, if not impractical, in in-the-wild scenarios. This thesis instead focuses on assessing the quality of human movement from **RGB data**. As all existing action quality assessment datasets are single view, this thesis also introduces **two multi-view** human movement assessment **datasets**, SMAD and QMAR, to demonstrate the superior performance of the proposed methods.

To deal with view-invariance, one solution is to develop a method that is trained on data from multiple views. In this scenario, it is important the method's complexity does not increase with the number of training views and the developed approach maintains a high performance on the single views. To achieve this, a pose estimation approach is proposed that estimates high-level pose in a canonical manifold space from RGB images, toward human movement assessment under a multi-view learning scenario.

As capturing a dataset including numerous viewpoints is cumbersome and rare, ideally, a view-invariant approach should be trained on data from as few views as possible while it can operate on arbitrary viewpoints at inference. Thus, this thesis develops an RGB-based approach that learns view-invariant spatio-temporal features by training on only one or two viewpoints and is able to analyse the quality of human movement on novel viewpoints.

This thesis also presents an unsupervised method that learns view-invariant 3D human posture representation from 2D RGB data for unseen view downstream tasks, *e.g.* action recognition and assessment, such that the pose features can be transferred into other domains. The proposed method is particularly helpful in applications where the use of multi-view data is essential and recording 3D skeletons is challenging, e.g. action quality assessment in rehabilitation exercises.

This thesis includes results on SMAD, QMAR, KIMORE, and NTU RGB+D,

and obtains comparative evaluation results against the state-of-the-art approaches where it is possible.

Declaration

I declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. The work is original, except where indicated by special reference in the text, and no part of the dissertation has been submitted for any other academic award.

Any views expressed in the dissertation are those of the author and in no way represent those of the University of Bristol.

The dissertation has not been presented to any other University for examination either in the United Kingdom or overseas.

SIGNED: DATE:

Acknowledgements

I would first like to express my gratitude to my supervisor Professor Majid Mirmehdi for his constant support, guidance and advice throughout my PhD. I have learnt so much about research from working with him and have found his experience and knowledge to be invaluable through-out.

I would also like to thank my collaborators; Professor Björn Ommer, Dr Sion Hannuna, and Dr Adeline Paiement for their valuable contributions; Dr Alan Whone and Dr Harry Rolinski of Southmead Hospital for insightful discussions on the QMAR dataset; my annual PhD reviewer Dr Tilo Burghardt for his helpful feedback and advice; and my thesis committee Professor Antonis Argyros and Dr Raul Santos-Rodriguez, for reading my thesis and their great feedback.

I would particularly like to thank my mum, dad, sisters, and brothers-in-law whose support and encouragement have always driven me forward. They have always been loving and caring despite the long distance and I would not have been able to keep going without their support.

Thank you to everyone in the VIL[1] and MaVi[2] laboratories who has made it a warm and encouraging environment to work: Sam, Jonny, Davide, Laurie, Mike, Will, Vangelis, Janis, Ramon, Abel, Hazel, Zeynel, Perla, Sasha, Young, Toby, Xingrui, Yanan, Erik, Angeliki, Oktay, Igor, Pengcheng, Yuhang, Yao, Obed, Hanyuan, Ruixiong, Xin, Xinyu, Dena, Miguel, Eduardo, Danier, and Bridget. A special thanks goes to Perla, Zeynel, Abel, Sam, Vangelis, and Davide for their empathetic support, and Will for his help in technical matters.

I would also like to thank my uncle Abdolali for his support and encouragement during my PhD; my friends, Lujain, Bushra, Ercan, Seher, Veronica, Marco, Esther, Bhargavi, Luciano, Anthi, and Negar who made my years in Bristol an amazing time of my life; and Mina, Zeynab, Elaheh, Nasim, and Amirsalar for their support even though they were far away.

Finally, I am grateful to the University of Bristol for their generous funding through the Overseas Postgraduate Student Scholarship and full funding of my PhD which brought about this opportunity for me to perform this research. I look forward to future collaboration with the University of Bristol and it's research staff.

[1] Visual Information Laboratory
[2] Machine Learning & Computer Vision

Publications

The work described in this thesis has been presented in the following publications:

1. Faegheh Sardari (F.S.), Adeline Paiement (A.P.), Majid Mirmehdi (M.M.). View-Invariant Pose Analysis for Human Movement Assessment from RGB Data. In *International Conference on Image Analysis and Processing (ICIAP)*, pages 237–248. Springer, 2019. (Chapters 3 and 4)

   Authors Contributions – Methodology, F.S., M.M., and A.P.; investigation, F.S., M.M., and A.P.; coding and validation, F.S. ; formal analysis, F.S, M.M., and A.P.; data curation, A.P., F.S, and M.M.; writing–original draft preparation, F.S.; writing–review and editing, M.M. and A.P.; supervision, M.M. and A.P.; project administration, M.M.

2. Faegheh Sardari (F.S.), Adeline Paiement (A.P.), Sion Hannuna (S.H.), Majid Mirmehdi (M.M.). VI-Net: View-Invariant Quality of Human Movement Assessment. *Sensors*, 20(18):5258, 2020. (Chapters 3 and 5)

   Authors Contributions – Methodology, F.S. and M.M.; investigation, F.S. and M.M.; coding and validation, F.S. ; formal analysis, F.S. and M.M.; data curation, F.S, M.M. and S.H.; writing–original draft preparation, F.S.; writing–review and editing, M.M.; supervision, M.M. and A.P.; project administration, M.M.

3. Faegheh Sardari (F.S.), Björn Ommer (B.O.), Majid Mirmehdi (M.M.). Unsupervised View-Invariant Human Posture Representation. In *British Machine Vision Conference (BMVC)*, 2021. (Chapter 6)

   Authors Contributions – Methodology, F.S., M.M. and B.O.; investigation, F.S. and M.M.; coding and validation, F.S. ; formal analysis, F.S., M.M., and B.O.; data curation, F.S. and M.M.; writing–original draft preparation, F.S.; writing–review and editing, M.M. and B.O.; supervision, M.M. and B.O.; project administration, M.M.

To My Parents,
Fatemeh and Abdolmehdi.

Contents

List of Figures

List of Tables

Acronyms

C3D	Convolutional 3D
CNN	Convolutional Neural Network
CoM	Center of Mass
Conv-LSTM	Convolutional Long Short Term Memory
CSS	Contrastive Self-Supervised
CS	Cross-Subject
CV	Cross-View
DCT	Discrete Cosine Transform
DIs	Dynamic Images
DTW	Dynamic Time Warping
Ex	Exercise
FC	Fully Connected
FN	False Negative
FP	False Positive
FTP	Fourier Temporal Pyramid
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
GL	Graph Laplacian
GMM	Gaussian Mixture Model
GRL	Gradient Reversal Layer
HMM	Hidden Markov Model
HoF	Histogram of Optical Flow
HoG	Histogram of Gradient

HoS	Histogram of Silhouette
I3D	Inflated 3D ConvNets
KDE	Kernel Density Estimation
KIMORE	Kinematic Assessment of Movement and Clinical Scores for Remote Monitoring of Physical Rehabilitation (Dataset)
KL	Kullback-Leibler
LBP	Local Binary Pattern
LSTM	Long Short Term Memory
MDS	Multi-Dimensional Scaling
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Square Error
NCE	Noise Contrastive Estimation
NeRF	Neural Radiance Field
NTU	Nanyang Technological University (Dataset)
PA	Procrustes Analysis
PCA	Principal Component Analysis
pdf	Probability Density Function
QMAR	Multi-View Quality of Movement Assessment for Rehabilitation (Dataset)
RGB	Red Green Blue
RGB-D	Red Green Blue Depth
RNN	Recurrent Neural Network
SMAD	Sphere Multi-View and Multi-Modal Movement Assessment Dataset
SMPL	Skinned Multi-Person Linear
SRC	Spearman's Rank Correlation
SSIM	Structural Similarity Index Matrix
SSM	Self-Similarity Matrix
STGCN	Spatial Temporal Graph Convolutional Network
SVM	Support Vector Machine

SVR Support Vector Regression

TCC Temporal Cycle Consistency

TCN Temporal Convolutional Network

TN True Negative

TP True Positive

VAE Variational Auto-Encoder

Chapter 1

Introduction

Human action or movement assessment includes the automatic analysis of the performance of a participant when performing a specific task, *e.g.* rehabilitation exercises, diving, and rolling pizza dough. To accomplish this task, a variety of input types such as accelerometer data captured from wearable sensors [7, 29, 178], 3D skeleton data extracted by RGB-D cameras or motion capture devices [22, 84, 98, 137], and RGB-D images [31, 102, 108], can be employed.

Action quality assessment differs from action recognition [13, 40, 64, 154]. In action recognition, the goal is to determine the type of a given action amongst different action classes. However, in action assessment, the aim is to evaluate how well people perform a target action. The action recognition task can be accomplished from key frames of a video sequence, whereas action assessment is achieved by considering all video frames [75].

The idea of developing a model to assess actions automatically originates from its potential usage in applications, such as healthcare [3, 9, 83, 128], sports [76, 96, 99, 104, 118, 136], and skill determination for a particular task [31, 86, 105, 151].

Healthcare – In this field, action assessment can benefit in both the diagnosis of diseases and recovery of patients. For instance, to diagnose Parkinson's disease, neurologists require to observe patients when they perform specific actions, such as walking or sitting-to-standing, to establish an objective marker for their level of functional mobility. In case the doctors identify the disease, they need to repeat this process both soon after prescribing medication and longitudinally across weeks and months as the progress of the disease is assessed. In addition, in such diseases, rehabilitation is also essential for the patient's recovery. By automating such mobility disorder assessment and home-care

physical therapy using computer vision, health service authorities can decrease costs, reduce hospital visits, and diminish the variability in clinicians' subjective assessment of patients.

Sports – In the sports domain, action quality assessment can be applied to the judging and scoring process or to improve the performance of athletes. For instance, automatic human movement assessment from video footage can help judges and referees to minimize human error. The athletes can also benefit by practising in front of the camera while they receive real-time scores and feedback [108]. This brings about this opportunity for them to improve their skills faster.

Skill Determination – Video learning has shown significant growth over the last decade since it is a cost-effective training approach and obtains a more convenient and practical learning experience compared to other training materials like text. For example, people can learn many skills such as drawing, cooking, and painting, by observing the experts when performing a specific task from video. Human action evaluation can be employed in video learning platforms by analysing the skill level of both trainers and learners and providing automated feedback for students to improve their skills.

To design an assessment framework, in addition to the accuracy which is essential, we require to consider two important factors, (i) the developed approach should be wholly view-invariant, otherwise the method will fail when it is applied to video data coming from camera views which are not present in training data, i.e. unseen/novel viewpoints, (ii) the proposed approach should be capable to run in in-the-wild scenarios, so it can be used in uncontrolled environments, *e.g.* home, clinic, and stadium.

Previous efforts in action quality assessment, such as [9, 99, 104, 105, 136], are all view-specific, and have not been designed to tolerate explicitly a large degree of view changes. For instance, to assess the performance of sports actions, the authors in [102, 103, 104] use Convolutional 3D (C3D), or Two-Stream Inflated 3D ConvNets (I3D) is employed to design the models in [118, 136, 171]. Raihan et al. [111] propose a network based on 1D Convolutional Neural Network (CNN) to predict a quality score for rehabilitation movements. However, as shown by Piergiovanni and Ryoo [107], current CNNs are not able to extract view-invariant features and their performance drops significantly when they are applied to novel viewpoint data. To the best of the author's knowledge, this thesis presents the first study that investigates view-invariance in human action assessment, with a particular focus on the healthcare domain.

Existing approaches in the field of healthcare, such as [3, 9, 22, 83, 128], are based on 3D skeleton data obtained from RGB-D cameras or motion capture devices since the features

derived from 3D data are rich and can be applied to assess a wide range of movements. For example, authors in [98, 137] train a continuous-state Hidden Markov Model (HMM) from skeleton data generated by an Asus Xmotion camera to analyse the movements of persons walking on stairs, or Elkholy et al. [37] extract a set of handcrafted features from 3D skeleton data to estimate the degree of abnormality in patients performing rehabilitation exercises by a Multiple Linear Regression (MLR) model. Liao et al. [84] also design a deep learning assessment network that utilizes 3D motion capture data. However, RGB-D cameras can estimate 3D pose only in optimal conditions, *i.e.* it is dependent on several parameters, including distance and viewing direction between the subject and the sensor. Although motion capture systems tend to be highly accurate, obtaining 3D pose by such means is expensive and time-consuming, since it requires specialist hardware, software, and setups. Therefore, acquiring 3D skeleton data can be cumbersome, if not impractical, in in-the-wild scenarios. This thesis instead focuses on assessing the quality of human movement from 2D RGB images that may be recorded and used in unconstrained home or clinical settings.

The thesis begins to deal with viewpoint variations through training from multi-view data. Then, it moves on to achieve view-invariance by training on data from as few views as possible. Finally, it explores unsupervised and transfer learning to extract view-invariant human posture representation for human movement assessment.

1.1 Challenges

There are two main challenges when dealing with the view-invariance in human movement assessment from RGB images.

The first challenge is to extract rich view-invariant features from 2D data. In comparison to the methods using 3D data (*e.g.* 3D skeleton and depth images), obtaining canonical features from 2D images is inherently ambiguous due to the lack of depth information, *i.e.* multiple 3D poses may correspond to the same 2D image after projection.

The second challenge is to overcome the above challenge while the method is trained on data from as few views as possible. A highly challenging scenario is to be able to perform well on a single-unseen view at inference while the approach is trained on data from only one viewpoint.

A third challenge is to achieve high performance without requiring a large training dataset. In applications of action quality assessment, particularly in the healthcare domain, capturing and labelling a large amount of data can be impractical and expen-

sive. For instance, recording video from Parkinson or Stroke patients who are at risk of injury when performing the action types (*e.g.* falling), is difficult. In addition, this procedure can become highly costly since we need to hire specialists for both recording and annotating the dataset.

1.2 Contributions

This thesis investigates and tackles view-invariance in human action or movement assessment from RGB data, and it's key contributions can be summarized as follows:

- A method to estimate human pose in a high-level canonical space is proposed such that the extracted pose features facilitate movement analysis from RGB images and are suitable for human movement assessment under multi-view training scenarios.
- A novel method that extracts view-invariant spatio-temporal features for unseen view human movement assessment from RGB video sequences is introduced which tackles the challenge of achieving the view-invariance by training on only one viewpoint.
- An unsupervised view-invariant 3D pose representation method is proposed where the learned view-invariant pose features can be applied for unseen view downstream tasks, *e.g.* action recognition and human movement assessment, and are able to be transferred into other domains.
- The only existing multi-view human movement assessment datasets, SMAD and QMAR, are developed, and QMAR is publicly released.

1.3 Thesis Overview

The rest of this thesis is organized as follows. Chapter 2 elaborates relevant background work and salient concepts to the tasks of view-invariant human movement assessment, and human pose estimation.

Chapter 3 presents two multi-view human movement assessment datasets, SMAD and QMAR, and also reviews the relevant datasets in this area as well as the field of action recognition.

Chapter 4 deals with viewpoint variations through training from multi-view data. It introduces a pose estimation approach that facilitates the assessment process under multi-view learning scenarios. The proposed approach is a CNN-regression network that

estimates a set of high-level pose features in a canonical manifold space from RGB-based body joint heatmaps and limb-maps.

Chapter 5 explores an RGB-based view-invariant approach that is able to assess the quality of human movement from unseen view data while it is trained on data from only one or two viewpoints. To achieve this, an end-to-end two-stage CNN-based network is developed that first learns to extract canonical (view-invariant) body joint trajectories from a single-view video clip, and then exploits the geometric relationship amongst the canonical trajectories to estimate a movement quality score.

Following this, Chapter 6 proposes an unsupervised representation learning approach to extract view-invariant 3D human posture representation from a 2D image toward unseen view human movement analysis. A convolutional auto-encoder is designed that disentangles canonical pose features and viewpoint parameters 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. The learned pose features are applied to two downstream tasks, unseen view action recognition and human movement assessment. The efficiency of transferring the learned representations from action recognition is shown to obtain the first ever unsupervised results for (unseen view) human movement or action assessment.

Finally, Chapter 7 concludes the work presented within this thesis and provides directions for future research.

Chapter 2

Background

This chapter presents a background to the topic of view-invariant human movement assessment and reviews the necessary related works upon which this thesis builds. First, Section 2.1 discusses the relevant works in action quality assessment, then given the lack of existing view-invariant human movement assessment techniques, related view-invariant action recognition methods are considered in Section 2.2. As Chapter 6 tackles view-invariance in human movement assessment by introducing an unsupervised view-invariant 3D pose representation approach, Section 2.3 covers relevant unsupervised 3D pose estimation techniques. Finally, Section 2.4 explains the protocols that are used by this thesis to evaluate the performance of view-invariant approaches in action quality assessment. Figure 2.1 illustrates a taxonomy of the topics presented in this chapter.

2.1 Action Quality Assessment

Action quality assessment aims to automatically evaluate the performance of a participant when performing a particular task. Depending on the scope of the task and the techniques required to analyse it, applications of action quality assessment are 'sports analysis', 'human movement assessment for healthcare', 'skill determination', and many others. In the video understanding domain, action quality assessment is a relatively new field in comparison to other tasks, *e.g.* action recognition and action detection, and it has gained attention recently via public datasets, such as MTL [104], AQA-7 [102], and KIMORE [12], and with the advent of deep neural networks. This section focuses and elaborates on aspect of action quality assessment works in sports analysis (Section 2.1.1) and human movement assessment for healthcare (Section 2.1.2) which are relevant to the scope of this thesis.

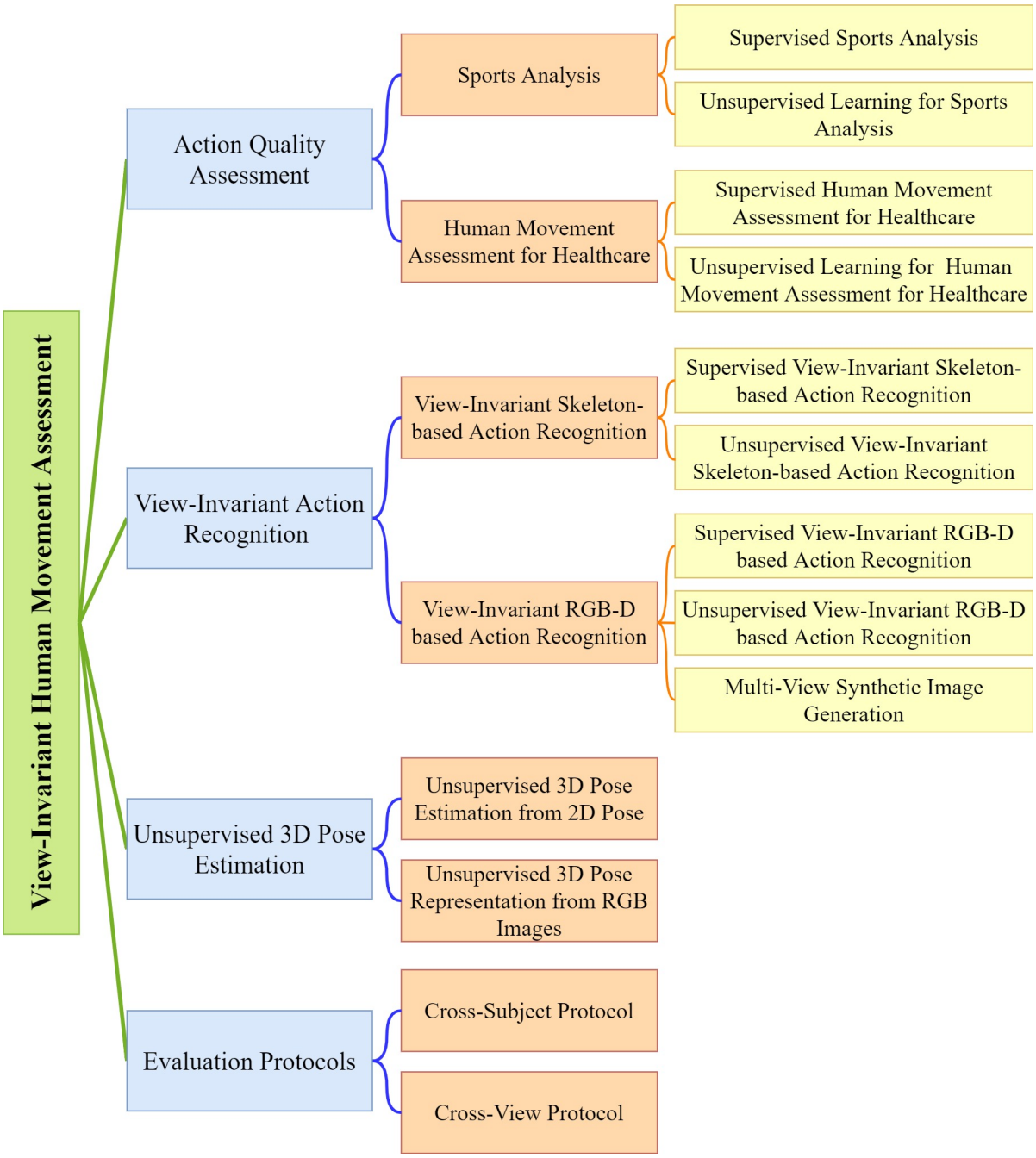


Figure 2.1: *Taxonomy of the topics presented in Chapter 2. This chapter reviews related works on three main topics, action quality assessment, view-invariant action recognition, and unsupervised 3D pose estimation, and also explains the evaluation protocols applied to this thesis. For action quality assessment, it focuses on supervised and unsupervised methods in the sport and healthcare domains. For view-invariant action recognition, it covers relevant supervised and unsupervised works using skeleton or RGB-D data, as well as approaches concentrating on generation of synthetic multi-view images. It also reviews the unsupervised 3D pose estimation methods employing 2D pose as input, and the unsupervised methods learning 3D pose representation from RGB images.*

2.1.1 Sports Analysis

Supervised Sports Analysis – The idea of automatically assessing the quality of sport actions from video is introduced first by Gordon [46]. This paper develops a method to estimate the score of performers in the gymnastic vault task. The scores are given to the athletes from their trajectory by subtracting some points for acting against specific properties defined specifically for the gym vault task.

Pirsiavash et al. [108] propose a regression-based method to score diving and figure skating actions in an Olympic sports dataset (MIT-Olympic), that they also publicly released. They train a Support Vector Regression (SVR) classifier on both low-level edge and velocity features and high-level 2D pose features represented in the frequency domain by the Discrete Cosine Transform (DCT). In addition to score prediction, they obtain feedback for performers to indicate how they can improve their performance. The feedback is generated through the pose estimation framework by predicting the way in which a body joint should be moved to improve the overall performance. Figure 2.2 and 2.3 illustrates qualitatively some sample feedback produced by [108] for diving and figure skating action respectively. The proposed method is also able to narrow down which segments of a video include higher and lower scoring movements. To highlight the most important segments, they remove some segments of the video and observe the changes in estimated scores. Although this method predicts action scores better than human non-experts, it is far from human expert judgment.

The authors in [6, 62] design models to analyse basketball games. For example, Bertasius et al. [6] propose a method to rank a pair of first-person basketball videos. They train

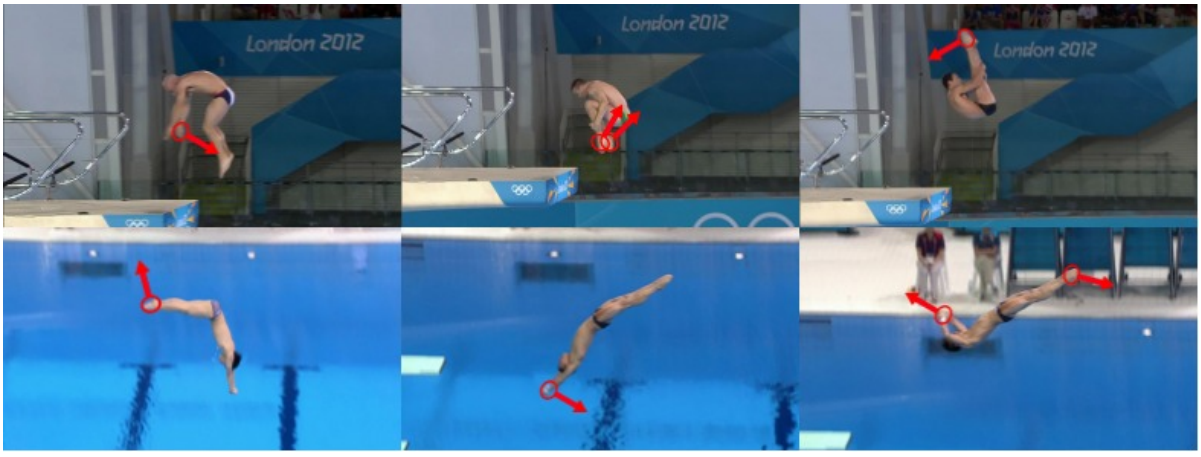


Figure 2.2: *Sample feedback suggested by [108] for some divers. Figure taken from [108].*

Figure 2.3: *Sample feedback suggested by [108] for some figure skaters. Figure taken from [108].*

a Convolutional Long Short Term Memory (Conv-LSTM) [158] to learn the features of different segments of a video by detecting some predefined specific events (*e.g.* possessing the ball and shooting the ball) in a sequence of frames. Then, the extracted features of the various segments are combined and employed by a Gaussian Mixture Model (GMM) to predict the performance of the players. Note, to concentrate on the important parts of the video, before feeding the frames into Conv-LSTM, they use a fully convolutional network [17] to detect and crop the region of the image around the ball.

Parmar and Morris [103] present and compare three different networks to assess sport actions. Each of the networks first applies a C3D network on non-overlapped 16-frame video clips of a video sequence to extract short-term spatio-temporal features. Then, they use different ways to aggregate these features and estimate the final score of the video sequence. In the first network, the video clip features are averaged and then used by an SVR to predict a final score. The second network applies a Long Short Term Memory (LSTM) network to learn the long-term information, and the third one employs both LSTM and SVR to assess the quality of an action. They apply all networks to diving, gym vault and figure skating actions, and their experiments demonstrate that the first network (C3D + SVR) outperforms the other two networks (C3D + LSTM and C3D + LSTM + SVR) and the prior method, Pirsiavash et al. [108], on all action types.

Instead of looking at all frames of a video sequence uniformly, Xiang et al. [156] develop a segment-aware approach. They first use a Temporal Convolutional Network (TCN) [74] to classify all frames of a diving video sequence into five groups, including preparing, jumping, dropping, entering into the water, and ending. Then, five distinct Pseudo-3D Residual (P3D) networks [109] are trained on different segments, and the outputs

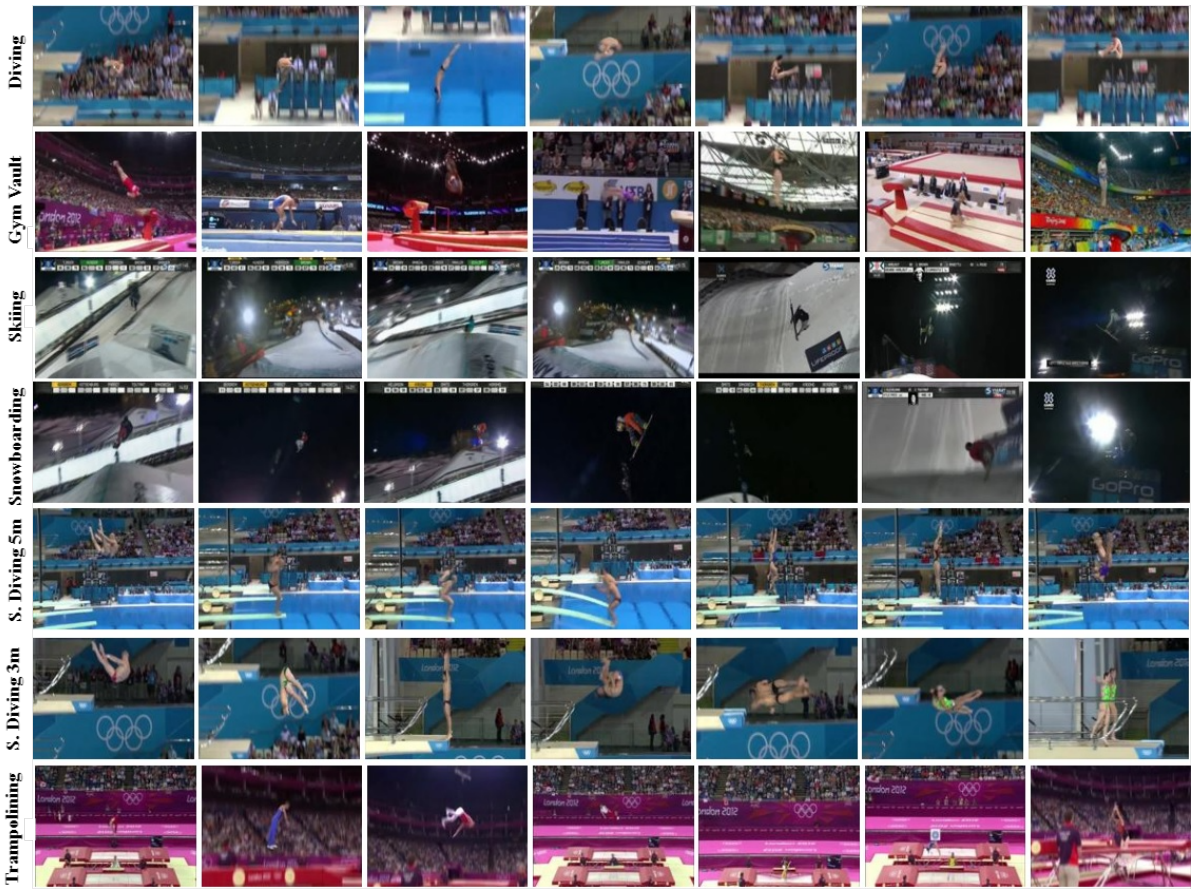


Figure 2.4: *Sample frames of 7 action types of AQA-7 dataset. Figure from [102].*

of all networks are concatenated to predict the overall score. Although this method improves the state-of-the-art results on diving action, all the frames in the training set must be annotated into different components. In addition, as the approach requires running separate P3D networks on different segments, it makes the model unable to be applied to longer actions as the complexity of the method increases with the number of segments.

In [102], the authors introduce a new Olympic action quality assessment dataset (AQA-7) and investigate whether it is possible to transfer knowledge amongst its different action types. The AQA-7 dataset consists of seven action types, gym vault, ski big air, snowboard big air, trampolining, diving, synchronous diving 3m platform, and synchronous diving 10m platform. Figure 2.4 illustrates examples of all action types. Although the actions have various scoring metrics, they share some common elements, such as flipping, twisting, and performing somersaults, so the knowledge or features that are learned through one of the actions can benefit in learning the others. Parmar and Morris [102] study this with several transfer learning approaches and using C3D + LSTM as back-

bone. For instance, instead of training separate networks for each action type, they applied one network on all action classes and observed that the method's performance was improved on separate classes on average, or they trained a network on one action class and then applied it to the rest of the action types (unseen action classes), and observed that in most experiments, the method's performance on unseen action types was poor.

In another work, Parmar and Morris [104] improve the performance of action quality assessment through multi-task learning where two extra tasks, action recognition and commentary generation, are incorporated with action's score prediction. They employ a C3D network to extract the features of separate parts of a video sequence. Then, for action assessment and recognition, the features are averaged to make a video-level presentation, while for commentary generation, the features of different parts are fed into its branch individually since the caption generation is a sequence-to-sequence task (see Figure 2.5). They show that multi-task learning is specifically helpful when a large training set is not available.

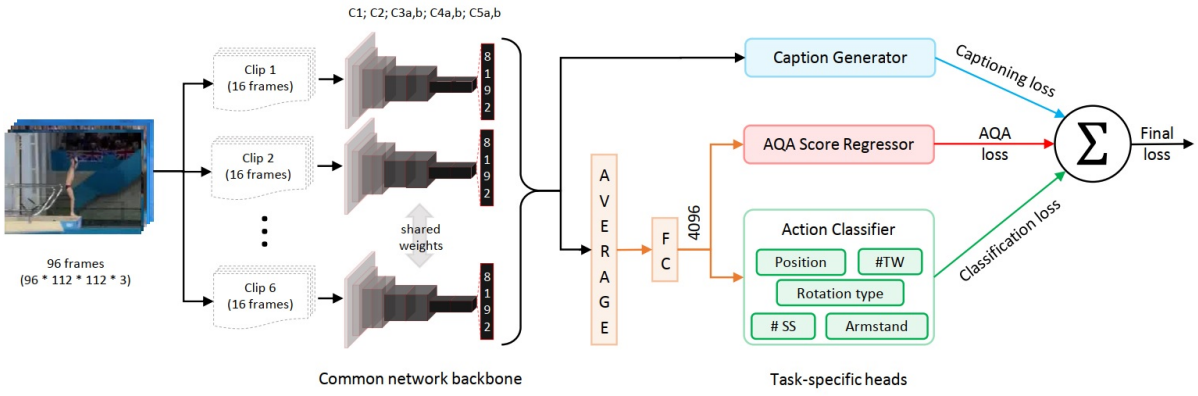


Figure 2.5: *The overall schema of the multi-task learning network proposed in [104]. Figure taken from [104].*

The uncertainty in analysing sport actions is dealt in [136]. All the previous works look at action assessment as a regression problem, while the score labels used for training the network are subjective and ambiguous as they are given by different judges. For example, for diving action, after a diver completes his/her task with the difficulty degree of 3.9, several judges may assign their scores as 8.5, 8.0, 9.0, 9.0, 8.5, 9.0, 7.0. The top two and bottom two scores are discarded, and his/her score is computed as: $3.9 \times (8.5 + 8.5 + 9.0)$. To tackle this issue, instead of estimating a single score, Tang et al. [136] design a network based on I3D to predict a set of Gaussian Distributions ($\psi' = \{\psi'(s_i)\}_{i=1}^{m}$) for a given

input video with score s. They optimize the proposed network through Kullback-Leibler (KL) divergence loss such that the ground truth distributions ($\psi = \{\psi(s_i)\}_{i=1}^{m}$) are generated as:

$$\phi(s_i) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x - s_i)^2}{2\sigma^2}), \tag{2.1}$$

$$\psi(s_i) = \phi(s_i) / \sum_{i=1}^{m} \phi(s_i), \tag{2.2}$$

$$Loss = \sum_{i=1}^{m} \psi(s_i) log \frac{\psi(s_i)}{\psi'(s_i)}, \tag{2.3}$$

where $i \in \{1, 2, ..., m\}$, and the mean of distribution s_i is sampled uniformly from a normal distribution with mean of s. At inference, the score with maximum probability is selected as final score S'.

$$S' = \underset{s_i}{\text{argmax}}\{\psi'(s_1), \psi'(s_2), ..., \psi'(s_m)\}. \tag{2.4}$$

In sports analysis, although there is a large variation in the scores of athletes, exploiting the relations amongst their movements can provide important information in predicting their score. Typically, action quality assessment approaches formulate the assessment process as a simple regression problem that estimates a quantitative score for an athlete from a single video. However, Yu et al. [164] introduce Group-Aware Contrastive Regression (GACR) method that leverages the relations amongst videos to predict the action's score during both training and inference. They design a model that receives two videos belonging to the same action type, main and exemplar, as well as the score of the exemplar video, and is trained to estimate the score difference between the two videos. At inference, the final score of the main video is predicted by averaging the results of the network for several exemplars. For each main video, the exemplars are selected based on some shared properties, such as action type and degree of difficulty. Note, in the sports datasets (*e.g.* AQA-7), for some action types (*e.g.* diving action), in addition to the score annotations, the degree of difficulty is obtained for all video sequences. Their proposed network consists of two modules, an I3D-based network, and a group-aware regression tree. First, the features of two input videos are extracted by I3D. Then, they are concatenated along with the score of the exemplar to fed into a regression tree designed based on Multi-Layer Perceptron (MLP) to produce the final result (see Figure 2.6).

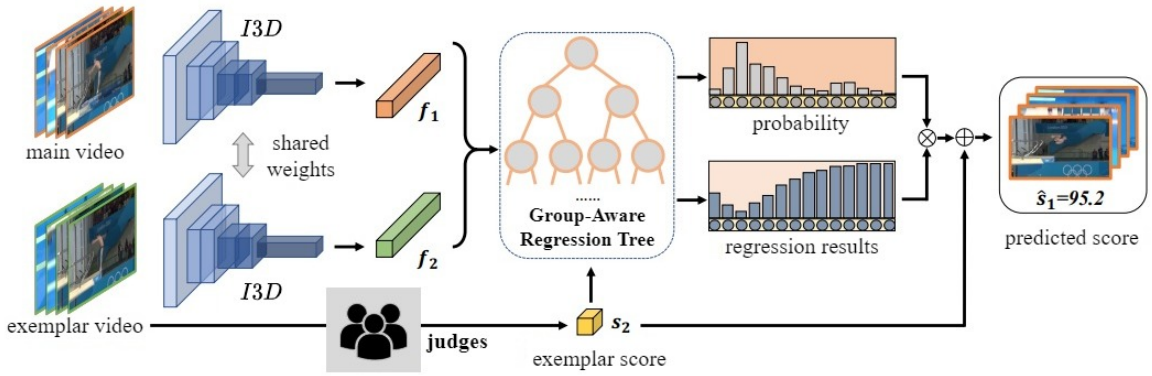


Figure 2.6: *The overall schema of the Group-Aware Contrastive Regression method proposed in [164]. First, an exemplar video is selected for each input video (main video) based on its category and degree of difficulty. Then, the input videos are fed into a shared I3D network to extract their spatio-temporal features. The concatenated features with the score of the exemplar video are given into the group-aware regression tree to estimate the score difference between the videos and compute the final score of the main video. Figure taken from [164].*

Unsupervised Learning for Sports Analysis – Jain and Harit [57] propose a self-supervised method to evaluate the performance of people performing Sun Salutation action by training on data from only experts. They design an LSTM-based auto-encoder to reconstruct a spatio-temporal feature vector that is extracted by applying the K-means algorithm on a sequence of 2D normalized poses. At inference, an input sequence is fed into the model and the quality of the action is computed based on the Levenshtein Distance between the input and output of the network. Jain and Harit [57] showed that the proposed method outperforms the supervised state-of-the-art sport assessment approaches.

Roditakis et al. [118] use Temporal Cycle Consistency (TCC) [35] to improve the performance of estimated scores for diving action. Their method has two learning phases and a temporal alignment step. In the first training phase, TCC is employed to build a self-supervised embedding space which is subsequently utilized to align the video clips temporally. Then, the aligned video clips are fed into the second learning phase to assess the quality of actions. The second stage is supervised and uses the features encoded by the two models, TCC and I3D, to predict the quality scores. During the supervised training, the TCC model is frozen and only the I3D network is optimized through the uncertainty loss function introduced in [136] (Equation 2.3). Figure 2.7 shows the details of the proposed method.

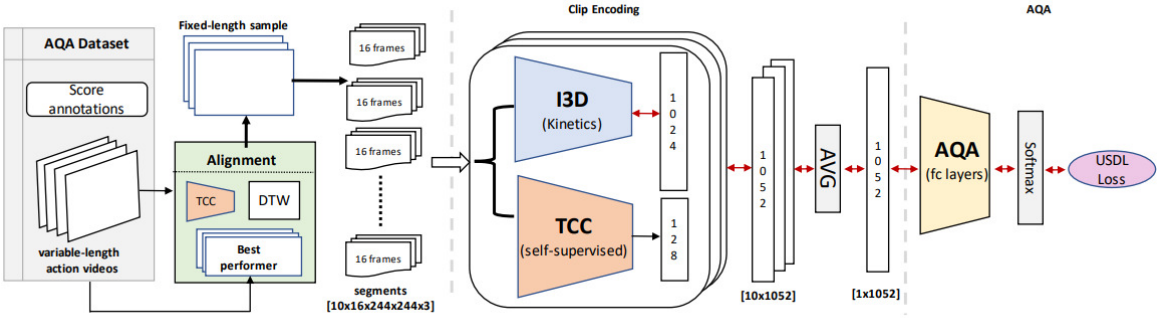


Figure 2.7: *The overall schema of the self-supervised alignment for action assessment. Each video sequence is aligned to a reference video that corresponds to the video of the best performer in the training set, based on the TCC embedding and Dynamic Time Warping (DTW) algorithm. Then the aligned videos are broken into 16-frame segments, which are later fed into TCC and I3D models. The features generated by these two models are concatenated and used by a temporal pooling that averages the features of all segments to obtain video-level representation for quality score prediction. Figure from [118].*

In [171], a semi-supervised approach is introduced where both labelled and unlabelled data are exploited to assess the performance of the actions. Their proposed end-to-end network has three modules, (i) a self-supervised module that learns the video representation without using any labels, (ii) an action assessment module that extracts the video features in a supervised manner, and (iii) a representation distribution alignment module that algins the distribution of video features of labelled and unlabelled data. The self-supervised and action assessment modules contain a shared I3D backbone followed by a shared encoder, while they have separate decoders. The distribution alignment module is designed based on the Gradient Reversal Layer (GRL) to close the gap between the the features learned from labelled and unlabelled videos. To learn the unsupervised representation, Zhang et al. [171] mask one of the video segments and train the network to reconstruct the masked part. This helps the network to leverage the temporal dependency of the unlabeled data. The action assessment module is trained to regress the quality scores. Zhang et al. [171] applied their method on three action assessment datasets MTL [104], Rhy-Gymnastics [166], and JIGSAWS [42] and showed that their results outperform the state-of-the-art semi-supervised methods and the baselines designed from [104, 136].

In summary, the majority of works in assessing the quality of sport actions are not explicitly designed to learn the features of desired actions and extract the spatio-temporal features by applying the 3D convolutional networks on video sequences. Therefore,

Method	Year	USUP	Backbone	Input	Dataset
Gordon [46]	1995	✗		2D Body Center	Gymnastic [46]
Pirsiavash et al. [108]	2014	✗	SVR	2D Pose	MIT-Olympic [108]
Bertasius et al. [6]	2017	✗	Conv-LSTM	RGB	Basketball [6]
Parmar and Morris [103]	2017	✗	C3D + SVR/LSTM	RGB	MIT-Olympic [108], UNLV [103]
Xiang et al. [156]	2018	✗	P3D+TCN	RGB	UNLV [103]
Li et al. [82]	2018	✗	C3D	RGB	UNLV [103], MIT-Olympic [108]
Parmar and Morris [102]	2019	✗	C3D+LSTM	RGB	AQA-7 [102]
Parmar and Morris [104]	2019	✗	C3D	RGB	MTL [104],UNLV [103], MIT-Olympic [108]
Jain and Harit [57]	2019	✔	LSTM	2D Pose	Sun Salutation [57]
Xu et al. [159]	2019	✗	C3D+LSTM	RGB	MIT-Olympic [108], Fis-V [159]
Tang et al. [136]	2020	✗	I3D	RGB	AQA-7 [102], MTL [104], JIGSAWS [42]
Jain et al. [58]	2020	✗	C3D	RGB	MIT-Olympic [108], UNLV [103]
Wang et al. [147]	2020	✗	TCN	RGB	UNLV [103]
Roditakis et al. [118]	2021	✔	I3D	RGB	MTL [104]
Dong et al. [27]	2021	✗	P3D+TCN	RGB	UNLV [103]
Pan et al. [100]	2021	✗	I3D+GCN	RGB+ 2D Pose	UNLV [103]
Zhang et al. [171]	2022	✔	I3D	RGB	AQA-7 [102], JIGSAWS [42], EPIC [30], BEST [31]

Table 2.1: *Overview of the sports assessment approaches. Key works have been explained in Section 2.1.1, while other works are worthy of note. USUP: the method contains unsupervised learning phase. The gray high-lights indicate non deep learning approaches.*

during the training process, they also learn unrelated context (*e.g.* rest of the scene) which has a large impact on their performance. Designing a network that focuses on only the action features is still an open question in this area. In addition, the extracted features by these approaches are not view-invariant, so if they are applied to the data recorded from viewpoints different to the training views, they will fail. Table 2.1 shows an overview of sport assessment approaches.

2.1.2 Human Movement Assessment for Healthcare

Supervised Human Movement Assessment for Healthcare – Paiement et al. [98] develop one of the first methods to automatically assess the quality of human movement in the healthcare domain. They propose an online method that analyses the movement of patients who walk on stairs. Their method uses 3D skeleton data captured by a Kinect camera from frontal view as input and builds two statistical models, pose and dynamic. The former represents the probability of normal poses through a probability density function (pdf) and the latter models temporal information of normal sequences by a continuous-state HMM. At inference, each frame of a sequence is classified to normal and abnormal depending on how far away from these models it is, based on an empirically determined threshold on log-likelihood. Before using the skeleton data for training and testing, this approach applies two preprocessing steps including normalization and dimensionality reduction on the input data. The normalization is used to make the data scale, transnational, and rotational invariant, and due to the curse of dimensionality of skeleton data, Diffusion Maps [21] is applied to the input data. Paiement et al. [98] also extend their work in [137] by applying their method to two other movement types, sitting-to-standing, and gait. To evaluate the performance of the method, they introduce three datasets, SPHERE-Staircase [98], SPHERE-SitStand [137], and SPHERE-Walk [137]. Section 3.1 will present details of these datasets.

To facilitate the process of movement assessment and eliminate the skeleton preprocessing steps in [98], Crabbe et al. [22] develop a convolutional method based on AlexNet [71] that estimates human pose in a low dimensional manifold space from depth images. The ground truth pose annotations are generated by applying the Diffusion Maps [21] on normalized skeleton data. Crabbe et al. [22] examine the performance of the pose features for human movement assessment by employing the SPHERE-Staircase dataset.

Similar to [98], Elkholy et al. [37] propose a statistical-based approach to analyse the quality of patient's movements. They extract handcrafted features from 3D skeleton data to classify a movement into normal and abnormal while for abnormal samples, they also estimate the level of abnormality. They propose three types of feature descrip-

tors including asymmetry, velocity magnitude, and Center of Mass (CoM) trajectory deformation, to model spatio-temporal characteristics of movements. During training, two probabilistic models, GMM and Kernel Density Estimation (KDE) are built upon the descriptors of normal sequences. At inference, a sequence is classified as normal or abnormal by computing its likelihood that is obtained through the trained GMM, and comparing it with a learned threshold. Furthermore, a MLR is developed on the proposed descriptors to estimate the degree of abnormality. To study the performance of the proposed approach for abnormality detection, they employ SPHERE-Staircase, SPHERE-SitStand, and SPHERE-Walk datasets. To demonstrate the efficiency of the method in estimating the abnormality scores, they introduce a new dataset (EJMQA) in which the movement's scores are obtained by a professional physiatrist (see Section 3.1 for more details of EJMQA).

Liao et al. [84] introduce the first deep neural network to predict a quality score for rehabilitation movements. Their proposed network consists of five temporal pyramid sub-networks followed by LSTM layers (see Figure 2.8). Each sub-network has four convolutional blocks that are applied to time-series skeleton data such that the first block uses the information of the whole sequence, and the second to fourth blocks employs the temporal data of ½, ¼, and ⅛ of the sequence respectively. The outputs of the sub-networks are concatenated and fed into the LSTM layers for further processing. The aim of each sub-network is to exploit the spatio-temporal relations amongst body joints of a specific body part (*e.g.* trunk). In a similar fashion to [98], Liao et al. [84] verifies the dimensionality reduction of input data, and implements it through an LSTM-based auto-encoder. They use UI-PRMD [140] which is a motion capture dataset, to evaluate the method's performance (see Section 3.1 for more details of UI-PRMD). Note, Liao et al. [84] annotate and provide the ground truth scores for movements in UI-PRMD by a performance metric based on GMMs.

The authors in [20] compare the strength of two sets of features extracted from 3D skeleton data, in predicting quality scores of rehabilitation movements in KIMORE [12] (see Section 3.1 for more details of KIMORE). They design two networks such that the first one has only an LSTM module, while the second one includes a Graph Convolutional Network (GCN) followed by LSTM layers. The first model is trained through a set of handcrafted features which have been provided manually by physicians from skeleton data, and the second model is trained from raw skeleton data. Their experiments show when LSTM works jointly with GCN, it predicts the movement's scores more accurate than when it is trained alone on the handcrafted features.

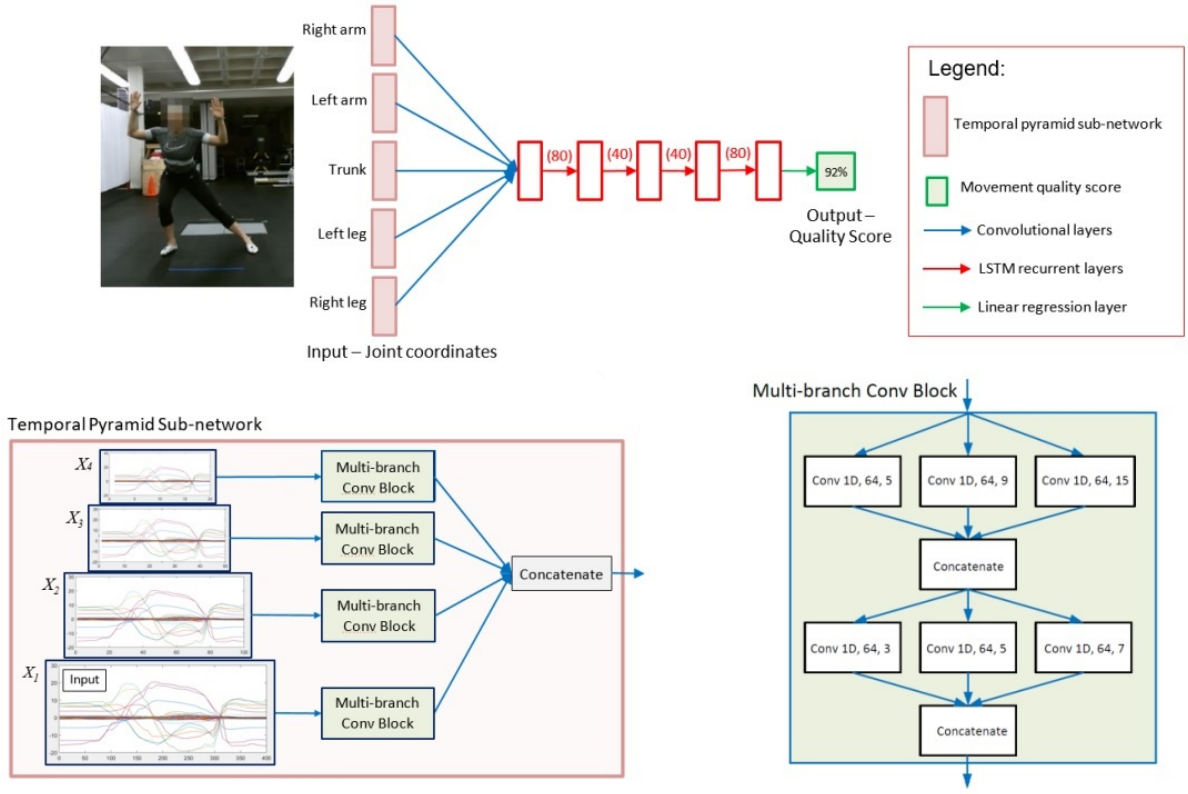


Figure 2.8: *Top: the spatio-temporal network proposed in [84]. Bottom left: temporal pyramid sub-network, and X_1 to X_4 indicates full, ½, ¼, and ⅛ of the input sequence respectively. Bottom right: multi-branch convolutional block unsed in the temporal pyramid sub-network. Conv 1D, ch, d: 1D convolution filter with kernel size d and ch channels. Figure from [84].*

Raihan et al. [111] propose to use evolutionary computation to find the optimal parameters of a rehabilitation assessment network. They design a CNN network where its parameters, such as number of layers, size of filters, and kernels' size are selected by genetic algorithm, and train it on feature vectors that are generated by applying a 1D Local Binary Pattern (LBP) operator [15] on skeleton pose sequences.

To overcome the issue of data imbalance in human movement assessment datasets, Albert et al. [1] propose a Generative Adversarial Network (GAN) based on an LSTM to generate synthetic normal and abnormal skeleton sequences. They also design a 1D-CNN network that learns to classify human movements in normal and abnormal categories by training on both real and synthetic data. To reduce the computational complexity of their proposed method, Albert et al. [1] apply their network on only the most informative joint and axis which are determined from a features ranking function. They use the KIMORE dataset [12] to train the model and perform experiments.

Method	Year	USUP	Task	Backbone	Input	Dataset
Paiement et al. [98]	2014	✗	N & AB	-	Skeleton	SPHERE-Staircase [98]
Crabbe et al. [22]	2015	✗	N & AB	AlexNet	Skeleton + Depth	SPHERE-Staircase [98]
Tao et al. [137]	2016	✗	N & AB	-	Skeleton	SPHERE-SitStand [137], SPHERE-Walk [137]
Elkholy et al. [36]	2017	✗	N & AB	-	Skeleton	SPHERE-Staircase [98], SPHERE-SitStand [137], SPHERE-Walk [137]
Elkholy et al. [37]	2019	✗	N & AB, Scoring	-	Skeleton	SPHERE-Staircase [98], SPHERE-SitStand [98], SPHERE-Walk [137], EJMQA [37]
Bruce et al. [9]	2020	✗	N & AB	GCN	Skeleton	UI-PRMD [140]
Liao et al. [84]	2020	✗	Scoring	1D CNN + LSTM	Skeleton	UI-PRMD [140]
Chowdhury et al. [20]	2021	✗	Scoring	GCN + LSTM	Skeleton	KIMORE [12]
Raihan et al. [111]	2021	✗	Scoring	1D CNN	Skeleton	KIMORE [12]
Albert et al. [1]	2021	✗	N & AB	1D CNN	Skeleton	KIMORE [12]
Bruce et al. [10]	2021	✗	N & AB Scoring	GCN	Skeleton	UI-PRMD [140], EHE [10]
Du et al. [33]	2021	✗	Scoring	GCN + LSTM	Skeleton	UI-PRMD [140]
Nekoui and Cheng [95]	2021	✔	Scoring	C3D + LSTM	2D Pose + RGB	UWA3D [110], UT [155], KIMORE [12], INR [14]
Deb et al. [24]	2022	✗	Scoring	GCN + LSTM	Skeleton	UI-PRMD [140], KIMORE [12]

Table 2.2: *Overview of the human movement assessment approaches in healthcare. Key works have been explained in Section 2.1.2, while other works are worthy of note. USUP: the method contains unsupervised learning phase. N & AB: normal and abnormal. The gray high-lights indicate non deep learning approaches.*

Unsupervised Learning for Human Movement Assessment for Healthcare – To assess the quality of human movement more accurately, Nekoui and Cheng [95] propose to use multi-modal data and design a two-stream network. One of the streams is a C3D-based network that is trained from RGB images, and the other is an LSTM-based model that learns motions from 2D poses. Nekoui and Cheng [95] also found that the performance of the LSTM-based branch can be improved through a self-supervised approach. To achieve this, first, they separately train a GRU-based auto-encoder on pose sequences to learn a sequence of unsupervised features, and then the learned representations are fed into the LSTM-based stream. To learn the unsupervised features, they combine the idea of pace prediction [148] and skeleton sequence in-painting [176]. While they benefit from self-supervised learning, it is applied to only part of the model that uses pose data, and the rest of the network where the RGB features are extracted, is trained in a fully supervised manner.

In a nutshell, most existing human movement assessment methods in the healthcare domain rely on 3D skeleton data, while acquiring skeleton data is challenging in in-the-wild scenarios, and the majority of the approaches also require a prepossessing step including dimensionality reduction and/or normalization to prepare the input data for the assessment process. Furthermore, the current human movement assessment methods are all view-specific and have not been designed explicitly to apply to unseen view scenarios. This thesis instead focuses on view-invariant movement quality assessment from *RGB* data and investigates the extraction of valuable features that can be applied *directly* for human movement analysis without requiring any intermediate steps. Table 2.2 shows an overview of action quality assessment methods for healthcare.

2.2 View-Invariant Action Recognition

Due to the lack of existing view-invariant action quality assessment methods, related view-invariant action recognition works are reviewed in this section since similar to action assessment, action recognition approaches work based on analysing the spatio-temporal features although there are still differences between them, as discussed in Section 1. In addition, in comparison to other video understanding tasks, *e.g.* action localization and prediction, in the action recognition field, view-invariance has been explored more.

As stated in [141], standard action recognition approaches, such as [13, 40, 64, 154], are not be able to deal with view-invariance, and their performance drops significantly if they are applied to data coming from viewpoints not presented in the training data. To tackle this problem, one solution would be to train a network on data from multiple views

[141]. However, in practice, capturing a labelled dataset of different views is cumbersome and rare. Therefore, view-invariant action recognition approaches have been developed such that the proposed models are trained from a few views and are tested on unseen or novel viewpoints.

Before deep learning, view-invariant action recognition approaches, such as [39, 63, 113, 163], usually used handcrafted features. For instance, Farhadi and Tabrizi [39] first model different activities with Histogram of Silhouette (HoS) and Histogram of Optical Flow (HoF), and then, propose a method that is trained to transfer the learned activities from a source view to a target view. Junejo et al. [63] introduce a spatio-temporal self-similarity matrix (SSM) to represent video sequences in spatio-temporal view-invariant maps. The self-similarity maps have different structures and patterns for distinct action types while they similarly represent different viewpoints of the same action. The SSM is generated based on the displacements of a set of descriptors over time such that the descriptors can be extracted from point trajectories, Histogram of Gradient (HoG), and HoF.

Deep learning approaches usually require large datasets, but capturing multi-view datasets requires elaborate set-ups and is inevitably time-consuming and potentially quite expensive. Therefore, multi-view datasets, such as IXMAS [153], N-UCLA [146], and UWA3D [110], were benchmark datasets for view-invariant action recognition till Shahroudy et al. [126] introduced a large multi-view dataset, Nanyang Technological University (NTU) RGB+D dataset, which allows training deep neural networks. Recently, Ji et al. [59] have also captured a large-scale dataset (UESTC) which provides RGB-D videos with entire 360° view angles. Section 2.2.1 and 2.2.2 reviews recent skeleton and RGB-D based deep learning approaches respectively.

2.2.1 View-Invariant Skeleton-based Action Recognition

Supervised View-Invariant Skeleton-based Action Recognition – Liu et al. [88] introduce a skeleton visualization approach where they first transform view-dependent skeleton sequences into a canonical view by applying a rotation matrix. Then, the transformed sequences are represented into a set of colour images that encode the motion generated by different body joints. Finally, all colour images are fed simultaneously into a multi-branch convolutional neural network for action classification. Their results show that their proposed method outperforms other view-specific methods, such as [126], on unseen view data.

Attention learning has also been adapted for view-invariant action recognition. Ji et al.

[60] develop a model consisting of three branches, reference-view, target-view, and attention, by employing Spatial Temporal Graph Convolutional Networks (STGCNs) [161]. The reference-view and target-view networks are trained to learn spatio-temporal features of the skeleton sequences in a reference view and arbitrary views respectively. The attention module connects two feature learning branches to transfer attention from the reference view to the arbitrary ones. The proposed network is optimized by computing three losses, transfer, attention, and classification. The transfer loss works based on KL divergence to measure the difference between the probability distributions of the reference-view and the target-view features, while the two other losses are developed based on cross-entropy.

To regulate the viewpoint variations when observing an action from novel views, Zhang et al. [169] design an end-to-end LSTM-based model including view adaptation and classification networks. The former network is trained to estimate the transformation parameters of 3D skeleton data to a canonical view, and the latter classifies the action from transformed canonical features. They later extend their work in [170] by adding another stream to the network. The new stream also contains the view adaptation and classification modules, but the new modules are modelled by convolutional filters instead of LSTM layers. Each stream is trained separately through cross entropy loss, and at

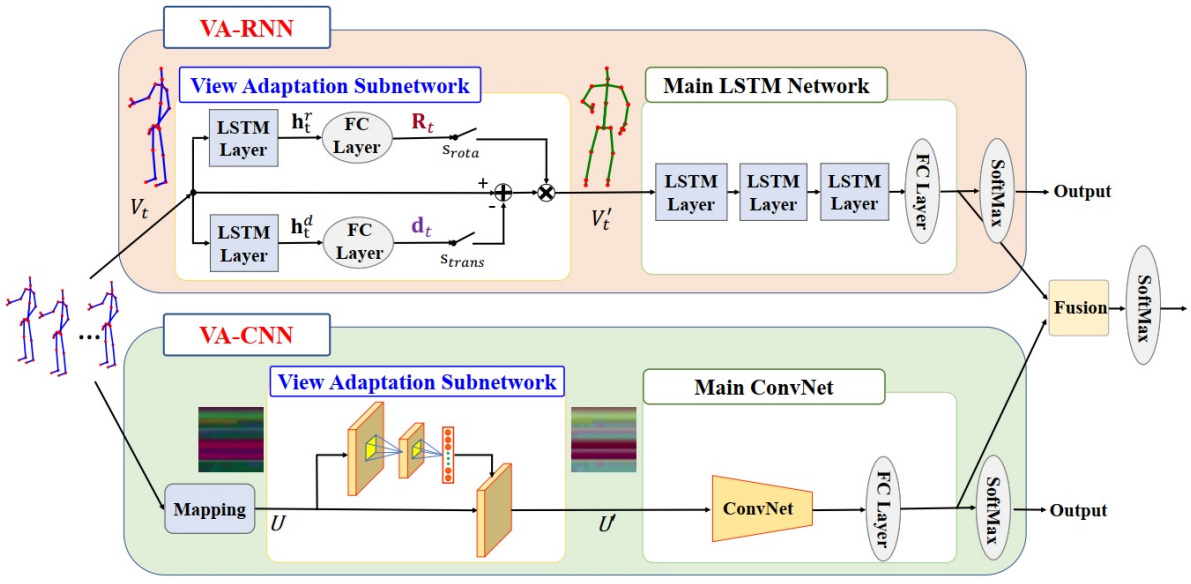


Figure 2.9: *The architecture of the two-stream network proposed in [170]. One stream is LSTM-based (VA-RNN) and the other one is convolutional-based (VA-CNN). In each stream, the view adaptive sub-network estimates the transformation parameters of 3D skeleton data to a canonical viewpoint, and the main network classifies the action from canonical features. Figure taken from [170].*

inference, the results of both streams are fused to make the final decision. Figure 2.9 shows the overall structure of the method.

Unsupervised View-Invariant Skeleton-based Action Recognition – Recently, several methods, such as [79, 114, 174] focus on contrastive learning paradigm to extract unsupervised view-invariant features. For example, Rao et al. [114] develop a framework based on LSTM layers and train it by Noise Contrastive Estimation (NCE) loss. To compute the NCE loss, the positive samples, which are referred to as query and key, are generated from the input sequences by applying augmentation techniques, such as rotation, shear, reverse, gaussian blur, and gaussian noise. The negative samples are obtained through a large dictionary which accumulates the extracted features of all samples, and the dictionary is updated in each iteration by replacing the current positive samples with their old version. The proposed framework has two LSTM networks, query encoder (E_q) and key encoder (E_k) that learns to encode the features of key and query samples respectively. In each iteration, the E_q encoder is updated via the NCE loss, whilst the parameters of the E_k are updated as:

$$\theta_k = \tau\theta_q + (1-\tau)\theta_q, \tag{2.5}$$

where θ_q and θ_k are parameters of E_q and E_k respectively, and $\tau \in [0, 1)$ controls the decay rate. After training the framework, E_q is employed for view-invariant action classification. Their experiments on NTU, UWA3D, and SBU [165] datasets show that their proposed method outperforms handcrafted methods, and also obtains competitive results to the supervised skeleton-based approaches.

Paoletti et al. [101] address viewpoint variants in skeleton-based action recognition by developing an auto-encoder that is designed by fully-residual 2D convolutional blocks. The proposed auto-encoder is trained through three losses, (i) reconstruction loss, (ii) Graph Laplacian (GL) regularization loss $\mathcal{L}_{GLR}$, and (iii) self-supervised view-invariant loss $\mathcal{L}_{SSVI}$. The $\mathcal{L}_{GLR}$ loss is computed by building a skeletal GL to apply Laplacian regularization to the reconstructed space by the decoder. This loss aims to inject the skeletal geometry information into the network. The $\mathcal{L}_{SSVI}$ loss is computed through a regression network that is connected to the encoder via a GRL. The skeleton sequence is augmented by rotation and fed into the network. Then, while the regression network tries to predict the rotation parameters of the augmented input, the encoder tries to mislead the regression network by extracting view-invariant features.

Table 2.3 presents an overview of skeleton-based methods considered in this section, as well as a few other works worth nothing. Although skeleton-based approaches achieve

promising performance on unseen view action recognition, they rely on a significant amount of 3D joint annotations, the provision of which is expensive and difficult in in-the-wild scenarios.

Method	Year	USUP	Backbone	Input	Dataset
Liu et al. [88]	2017	✗	AlexNet	Skeleton	N-UCLA [146], UWA3D [110], NTU [126]
Zhang et al. [169]	2017	✗	LSTM	Skeleton	SBU [165], SYSU [51], NTU [126]
Talha et al. [135]	2018	✗	-	Skeleton	MSRAction3-D [80], UTKinect [155], Florence 3-D [123], Multiview3-D [135]
Zhang et al. [170]	2019	✗	2D ResNet + LSTM	Skeleton	SBU [165], SYSU [51], NTU [126] N-UCLA [146], UWA3D [110]
Ji et al. [60]	2021	✗	STGCN	Skeleton	NTU [126], UESTC [59]
Rao et al. [114]	2021	✔	LSTM	Skeleton	SBU [165], UWA3D [110], NTU [126]
Paoletti et al. [101]	2021	✔	2D CNN	Skeleton	NTU [126], NTU-120 [89]
Li et al. [79]	2021	✔	STGCN	Skeleton	NTU [126], NTU-120 [89]
Gao et al. [41]	2021	✗	CNN + STGCN	Skeleton	NTU [126], NTU-120 [89], UESTC [59]
Gedamu et al. [43]	2021	✗	2D ResNet	Skeleton	NTU [126], NTU-120 [89], UESTC [59]

Table 2.3: *Overview of the view-invariant skeleton-based action recognition approaches. Key works have been explained in Section 2.2.1, while other works are worthy of note. USUP: the method contains unsupervised learning phase. The gray high-light indicates a non deep learning approach.*

2.2.2 View-Invariant RGB-D based Action Recognition

To deal with view-invariance from non-skeleton data, a very few works, such as [44, 143, 144], extract view-invariant features from RGB images alone since obtaining view-invariant features from ambiguous 2D data is highly challenging, while others like [26, 78], use depth images along with RGB data. On the other hand, some works, such as [87, 141], generate multi-view synthetic RGB images for training to increase the performance of view-specific methods on real unseen view data.

Supervised View-Invariant RGB-D based Action Recognition – Wang et al. [144] design a multi-branch deep neural network for view-invariant action recognition. Their proposed network has a shared CNN that learns view-independent features, followed by several view-specific CNN branches, and a novel message-passing module. The message passing module is applied between every two CNN branches to improve the view-specific feature extraction. Then, the refined features are passed through view-specific classifiers whose outputs are fused for action classification (see Figure 2.10). The fusing process is performed by a view-classifier trained on view-independent features. Note, the number of view-specific branches is equal to the number of training views. As a view-specific branch is added to the network for each training view, the complexity of the model increases with the number of training viewpoints.

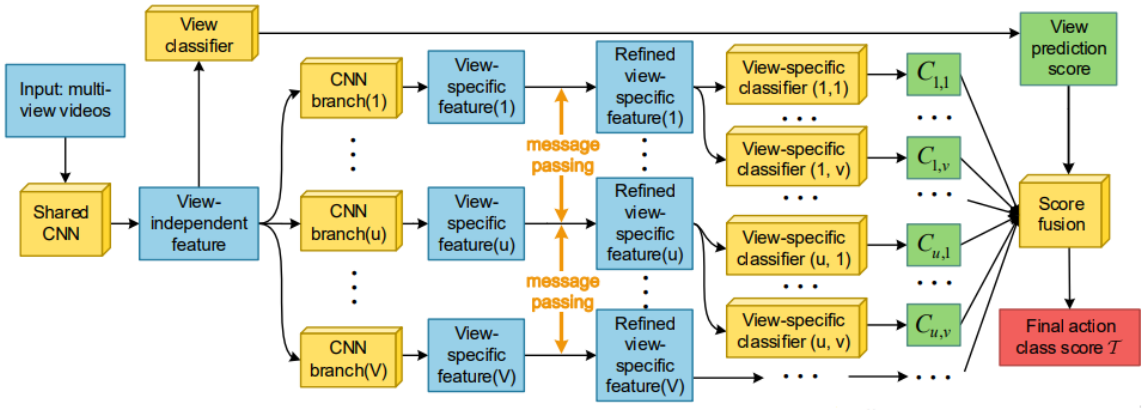


Figure 2.10: *The overall schema of the multi-branch network proposed in [144]. Figure from [144].*

Dhiman and Vishwakarma [26] develop a two-stream view-invariant action recognition framework, which consists of Shape Temporal Dynamic (STD) and motion streams. During training, each stream is optimized separately, and at inference, their outputs are aggregated to classify the actions. STD includes a deep CNN block followed by two

LSTM layers to learn view-invariant shape dynamics of an action over time. It extracts the features through structural similarity index matrix (SSIM) [152] which is applied to a sequence of human depth silhouettes. In the motion stream, first, the appearance and dynamics of a sequence of RGB images are represented as dynamic images (DIs) by a temporal rank pooling function, and then, DIs are sent to a pre-trained Inception-V3 network [133] to predict the action class. Dhiman and Vishwakarma [26] evaluate their approach on several multi-view datasets including NTU, and show the importance of different components of their method.

Ghorbel et al. [44] introduce a fast view-invariant approach that is trained on human skeleton sequences generated from RGB images by applying the VNect method [93] where the body joints and their location heatmaps are estimated by a CNN network. First, Ghorbel et al. [44] extract a set of view-invariant features from the pose data by exploiting the geometric relationship between different body parts, and the skeleton alignment based on joint positions, velocity, and acceleration. Then, they train a Support Vector Machine (SVM) on the learned view-invariant features to classify the actions. This method outperforms the state-of-the-art methods on the N-UCLA and IXMAS datasets, but they do not provide their proposed method's results on the NTU dataset to allow comparison against the recent state-of-the-art view-invariant approaches.

Unsupervised View-Invariant RGB-D based Action Recognition – Li et al. [78] design an end-to-end recurrent convolutional auto-encoder to learn video representation from 3D scene flow and depth data which are later used for unseen view action recognition. Their proposed network has one encoder and two decoders (reconstruction and cross-view) and learns to obtain the view-invariant representation through the encoder by reconstructing the input sequence into two different views. The reconstruction decoder reconstructs the same source view input sequence from the view-invariant representation while the cross-view decoder predicts the input sequence in a target view. The input of the encoder and the output of the reconstruction decoder are 3D scene flows while the output of the cross-view decoder is depth images. Li et al. [78] also boost the view-invariant representations by adding a view-adversarial classifier that encourages the network to generate features that are invariant to view changes. They show the efficiency of the proposed approach by training and testing on NTU RGB+D, and also fine-tuning their pre-trained network on NTU for action classification on the N-UCLA [146] and MSR-DailyActivity3D [145] datasets. However, this method relies on 3D scene flow that obtaining this type of data is computationally expensive.

Vyas et al. [143] use video prediction and develop an unsupervised representation learning framework to address multi-view action recognition. Their proposed framework receives

several video clips of an action from different times and viewpoints as input to learn a representation through representation learning network (RL-NET). A video rendering network (VR-NET) takes the learn representation to predict an unseen view and time video clip. The end-to-end proposed framework is optimized by computing the Mean Squared Error (MSE) between the predicted video clip and the ground truth. After learning the unsupervised representation, a classifier network (CL-NET) consisting of 2D convolutional layers followed by fully connected layers, is added on top of RL-NET and trained through cross entropy loss for supervised action recognition. Figure 2.11 shows more details of the proposed network.

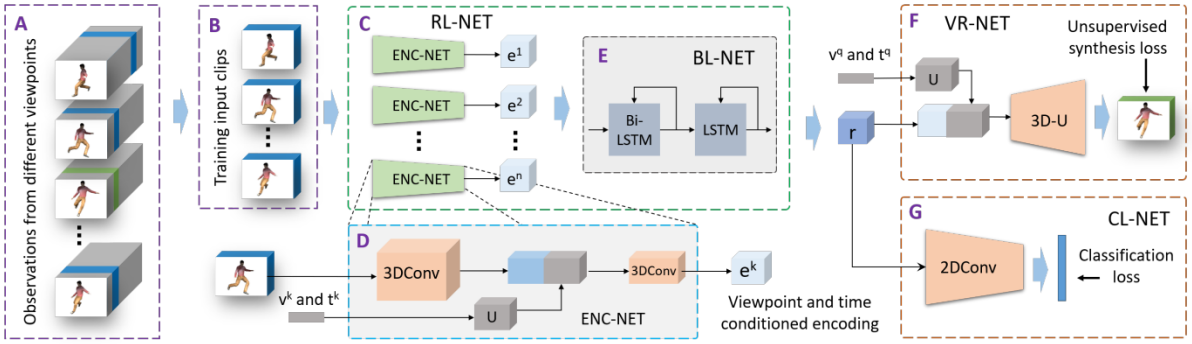


Figure 2.11: *Video prediction for view-invariant action recognition. A: simultaneous video sequences from different viewpoints $S = \{s^i\}_{i=1}^m$, for a given action. B: training video clips from different times and viewpoints $VC = \{vc^k\}_{k=1}^n$, that are collected from S. C: Representation learning network (RL-NET) that takes VC to learn the representation r. D: ENC-NET is embedded in RL-NET to learn the individual video representation e^k that is conditioned on its time t^k and viewpoint v^k. E: The blending network (BL-NET) aggregates the video clip representations $\{e^k\}_{k=1}^n$, to make the final r. F: The unified representation r is applied to predict an unseen view v^q and time t^q video clip from S by VR-NET. G: The representation r is also used by CL-NET to classify the action types. 3D-U indicates to 3D convolutions filters that are combined with up-sampling. Figure taken from [143].*

Multi-view Synthetic Image Generation – In [87], synthetic multi-view RGB data is employed to train a CNN-based network that extracts view-invariant pose features for action recognition. Liu et al. [87] generate the synthetic data by fitting 3D human body shapes with motion capture data and applying the graphic techniques. They also use a GAN to improve the quality of the synthetic frames. For action classification, Liu et al. [87] model the temporal information of view-invariant pose features with the Fourier Temporal Pyramid (FTP) and SVM. Although in this work, the authors facilitate the action recognition process using RGB images only, their method relies on a large amount

of 3D data.

Similar to [87], Varol et al. [141] produce synthetic multi-view data to deal the view-invariance. They train a 3D ResNet-50 [48] on both synthetic and real video sequences to classify the actions. To generate the synthetic data, they apply VIBE [70] and HMMR [65] on real single view action videos to estimate 3D pose sequence, camera, and body shape parameters, and then the artificial data are generated by augmenting the camera and body shapes parameters using graphics techniques. In their work, the synthetic data can improve their method's performance on unseen view data. However, when their network is trained on only synthetic data, it performs poorly which indicates that the synthetic videos do not represent realistic motion.

In conclusion, RGB-D based view-invariant approaches remove the need for 3D joint annotations. However, due to the inherent ambiguity of this kind of data, they usually rely on synchronized multi-view images, Table 2.4 summarises the RGB-D based approaches presented in this section, and a few other works worth noting.

2.3 Unsupervised 3D Human Pose Estimation

Although the majority of current RGB-based 3D human pose estimation approaches, *e.g.* [45, 77, 106, 149, 160], are fully supervised and rely on a large amount of 3D joint annotations, recently several works have been introduced that remove or decrease the need for labelled data, such as [16, 18, 34, 50, 115, 139]. Among these approaches, authors in [16, 18, 52, 55, 139] extract unsupervised pose features from 2D joints generated from RGB data, while others, such as [34, 50, 72, 115], learn 3D pose representations directly from RGB images.

Unsupervised 3D Pose Estimation from 2D Pose – Chen et al. [16] train a model by lifting 2D pose to 3D joints and reprojecting 3D onto 2D through a geometrical self-consistency loss that allows the network to learn in a self-supervised manner. The proposed loss is designed based on the hypothesis that any 2D projection of a learned 3D pose should generate the same 3D pose. They show that self-consistency alone is not sufficient to learn the 3D pose, and improve the performance of the network by adding a spatial and a temporal 2D pose discriminator. The spatial discriminator is trained to classify the projected 2D pose from real ones, and the temporal discriminator learns to classify the differences in 2D poses in subsequent frames of a sequence as real or fake. Their experiments demonstrate that incorporating the additional temporal discriminator improves their method's performance by 7%.

Method	Year	USUP	Backbone	Input	Dataset
Zhang et al. [168]	2018	✔	-	RGB	IXMAS [153], WVU [112]
Liu et al. [90]	2018	✔	-	RGB	IXMAS [153], WVU [112], MuHAVi [130], N-UCLA [146]
Wang et al. [144]	2018	✗	TSN [150]	RGB	N-UCLA [146], NTU [126]
Li et al. [78]	2018	✔	2D ResNet + LSTM	RGB + Depth	N-UCLA [146], MSR [145] NTU [126]
Ghorbel et al. [44]	2019	✗	2D ResNet + SVM	RGB	IXMAS [153], N-UCLA [146]
Liu et al. [87]	2019	✗	2D CNN + SVM	RGB* + MoCap	N-UCLA [146], UWA3D [110], NTU [126]
VI-DA [26]	2020	✗	Inception + LSTM	RGB + Depth	N-UCLA [146], NTU [126] NTU [126]
Vyas et al. [143]	2020	✔	3D CNN + LSTM	RGB	N-UCLA [146], NTU [126]
Varol et al. [141]	2021	✗	3D ResNet	RGB* + 3D Pose	NTU [126], UESTC [59]

Table 2.4: *Overview of the view-invariant RGB-D based action recognition approaches. Key works have been explained in Section 2.2.2, while other works are worthy of note. USUP: the method contains unsupervised learning phase. *: the training data includes both real and synthetic images. The gray high-lights indicate non deep learning approaches.*

In [139], knowledge distillation is used to estimate both 3D pose representation and Skinned Multi-Person Linear (SMPL) model parameters from 2D joints. The proposed framework has a temporal backbone that is followed by a teacher and a student branch. The input of the network is a sequence of 2D joints generated from RGB images, and the output of the teacher and student sub-network is model-free 3D poses and SMPL body shape parameters respectively. The proposed framework learns the pose representations in two stages. First, the backbone and teacher branch are trained by a 3D to 2D projection pose loss, temporal and bone length consistency losses, and an adversarial loss coming from a temporal discriminator. Then, the backbone and the teacher are frozen, and only the weights of the student branch are updated. In this stage, a knowledge distillation loss is defined by computing the distance between the model-free 3D poses estimated by the teacher network, and the 3D joints obtained via the SMPL model. They

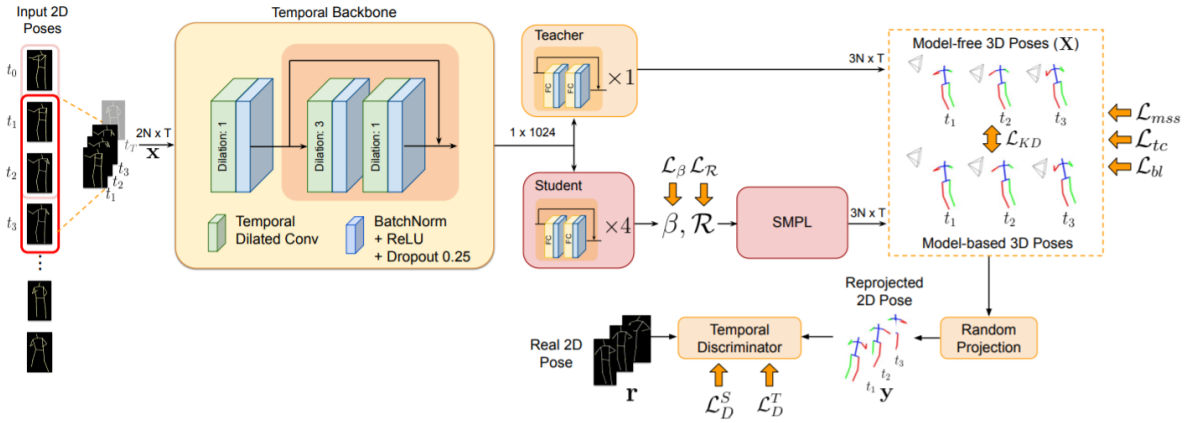


Figure 2.12: *Unsupervised 3D pose estimation via knowledge distillation. $\mathcal{L}_{mss}$: projection 3D-2D loss. $\mathcal{L}_{tc}$: temporal consistency loss. $\mathcal{L}_{bl}$: bone length consistency loss. $\mathcal{L}_{KD}$: knowledge distillation loss. $\mathcal{L}_R$, $\mathcal{L}_B$: regularization losses. $\mathcal{L}_D^T$, $\mathcal{L}_D^S$: adversarial losses for teacher and student branches respectively. Figure from [139].*

also use two simple regularization losses for the pose parameters to prevent over-twisting, and the same as the teacher branch, the estimated 3D joints via the SMPL model are given to the temporal discriminator. Figure 2.12 shows details of the proposed framework in [139].

To learn unsupervised 3D skeletons, Iqbal et al. [55] propose to leverage 2.5D poses [54] instead of 2D body joints. The 2.5D pose representation is scale and translation invariant and is defined as $P = \{(x_j, y_j, d_j)\}_{j=1}^{J}$, where x_j and y_j are the 2D coordinates of the joint j, and d_j is its metric depth with respect to the root joint. Iqbal et al. [55] design a convolutional framework that takes a set of multi-view RGB images and estimates 2D heatmaps generated from 2.5D pose representation. To train the network, in addition to the reconstruction loss which computes the difference between the ground truth and the estimated heatmaps, they use multi-view and limb length consistency losses, and a regularization loss. The multi-view consistency loss is computed based on the transformation of the estimated 3D poses between different viewpoints, while the limb length loss imposes the kinematic constraints of the human body to estimate the pose representation, and the regularization loss aims to help the network concentrate on the foreground.

In [116], the authors aim to develop a network that is able to extract canonical 3D pose features driectly from RGB images. To achieve this, [116] train a CNN network that maps multiple views into a canonical pose through MSE, 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.

Unsupervised 3D Pose Representation from RGB Images – Rhodin et al. [115] train an auto-encoder that learns to disentangle 3D pose representation and the appearance features from RGB images such that the pose features encode 3D geometry, and the pretrained encoder is later used for supervised 3D pose estimation. As the 3D geometry is already encoded in the learned representations, estimating 3D pose from them is much easier and can be obtained through a smaller training set than fully supervised approaches. To enforce the network to separate the appearance and pose representations, Rhodin et al. [115] divide the latent vector into appearance and pose feature vectors and use two different frames of the same subject for training. They assume that the encoder learns the same appearance representation for these frames, so the network should still reconstruct them if their appearance features are swapped. Note, different images of the same subject are selected from distinct frames of the same video sequence. The 3D geometry is learned by using the simultaneous images of the same person captured from several views while the auto-encoder is trained to reconstruct the image captured from one viewpoint from the image coming from another viewpoint. Their experiments show that the proposed method improves the supervised state-of-the-art results and outperforms the other semi-supervised methods by training on only 1% of the 3D pose annotations.

The approach proposed in [115] requires multi-view data and camera parameters for training. Zhang et al. [172] instead develop a method that can be trained from monoc-

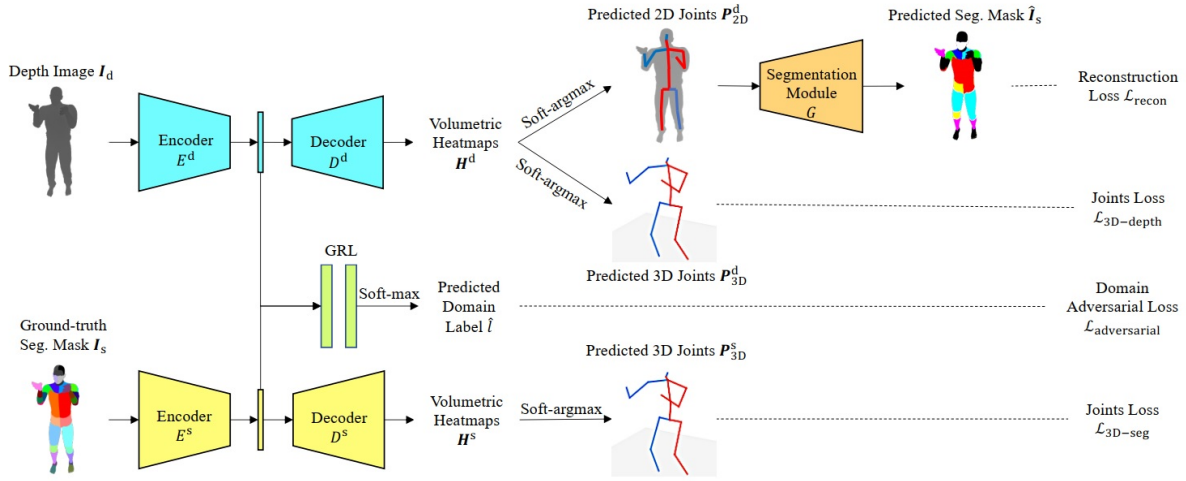


Figure 2.13: *An overview of the first learning stage of [172] where the proposed network is trained on synthetic data in a supervised manner. Figure from [172].*
m

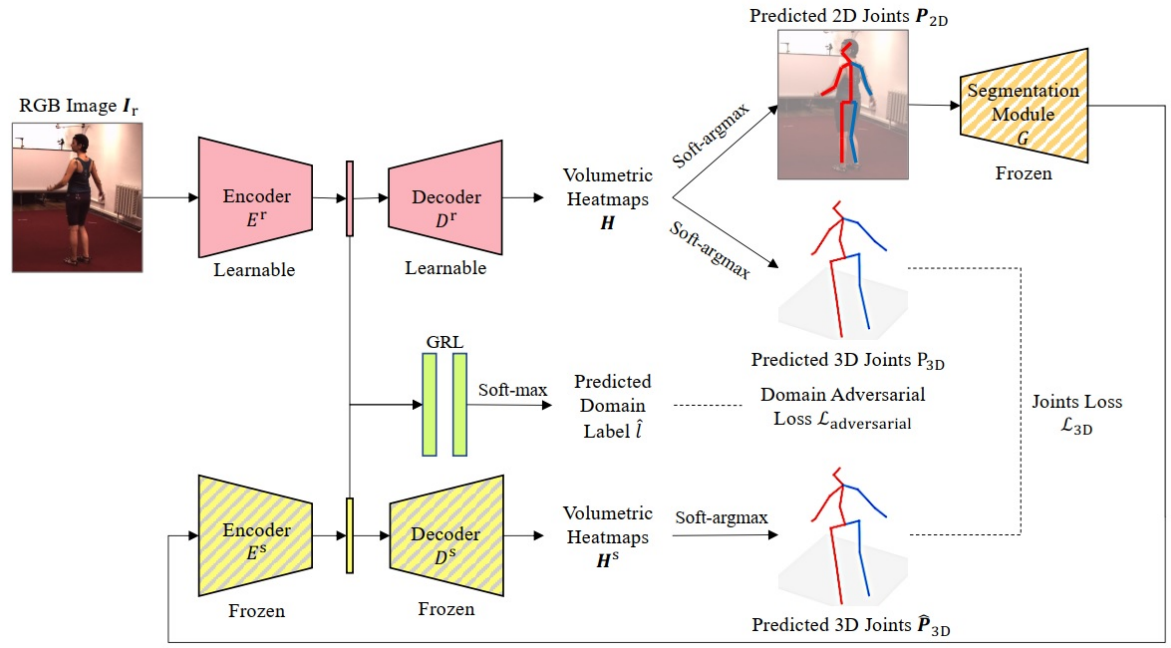


Figure 2.14: *An overview of the second learning stage of [172] where the proposed network learns to estimate 3D pose from real data in an unsupervised manner by transferring the learned knowledge from synthetic data. Figure from [172].*

ular RGB data and without relying on viewpoint information. To do this, they apply domain adaptation and knowledge transfer to leverage the pose information of synthetic depth images to estimate 3D pose in real RGB data. Their proposed approach has two training stages, (i) learning 3D pose from synthetic data in a supervised manner, and (ii) transferring the learned knowledge from the first stage for unsupervised 3D pose estimation from real data. In the first stage, two auto-encoders, $E^r - D^r$ and $E^s - D^s$, are trained simultaneously to estimate 3D pose from depth images and body segmentation masks respectively. $E^r - D^r$ also regresses 2D poses which are used to train the body segmentation module (G). To encourage the auto-encoders to encode the same hidden features for depth and body segmentation masks, a domain classifier is also added to the model. Figure 2.13 shows details of the first learning stage. Then, in the second learning stage, while $E^s - D^s$ and G are frozen, the model is optimized on real data without using 3D pose annotations (see Figure 2.14). Finally, at inference, $E^r - D^r$ is utilized for pose estimation.

Honari et al. [50] leverage contrastive self-supervised (CSS) learning to encode the 3D pose representation from single-view RGB videos. CSS approaches aim to learn the image features by pulling the features of positive samples close to each other in embedding space while pushing the features of the negative ones apart. CSS-based methods, such as

Method	Year	Training				Sup.	Dataset
		MV	PRM	Pose	VI		
Rhodin et al. [116]	2018	✔	✗	3D	✔	✗	Human3.6M [53], MPII [92], Ski[116]
Rhodin et al. [115]	2018	✔	✔	✗	✗	✔	Human3.6M [53]
Zhang et al. [172]	2019	✗	✗	3D	✗	✗	Human3.6M [53], ITOP [47], UBC3V [125]
Chen et al. [16]	2019	✗	✗	2D	✗	✗	Human3.6M [53], MPII [92], LSP [61]
Tripathi et al. [139]	2020	✗	✗	2D	✗	✗	Human3.6M [53], MPII [92], 3DPW [142]
Iqbal et al. [55]	2020	✗	✗	2D	✗	✗	Human3.6M [53], MPII [92]
Honari et al. [50]	2021	✗	✗	✗	✗	✔	Human3.6M [53], MPII [92], Diving[50], Ski[116]

Table 2.5: *Overview of the unsupervised 3D pose estimation approaches presented in Section 2.3. MV: multi-view. PRM: Camera parameters. VI: View-invariant. Sup.: Supervision to map latent into 3D pose.*

[124], apply this strategy to the whole feature vector. However, this idea does not work properly if it is used on the frames of a video sequence since when a person moves, part of the features like pose, changes over time, while the others, such as appearance features are constant. To overcome this problem, Honari et al. [50] break the latent vector into time-variant and time-invariant components and apply CSS on only the time-variant one. Then, both vectors are concatenated and fed into a decoder to reconstruct the input image. To train the model via CSS, they also proposed a novel distance-based similarity loss that computes the similarity between any pairs of samples depending on their temporal distance. The time-variant features that represent pose information are later used for 3D pose estimation and tested on several datasets, such as Human3.6M [53] and MPII [92]. The results show that the proposed CSS-based approach outperforms other unsupervised single-view methods. Table 2.5 provides an overview of the unsupervised 3D pose estimation approaches presented in this section.

In summary, typically the unsupervised 3D pose estimation and representation approaches 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 (*e.g.* human movement assessment) and camera parameters and extra steps are

needed to map their view-specific output into a canonical view. The method proposed in [116] estimates 3D pose in a canonical view, but it requires both labelled and unlabelled data and camera parameters for training. In Chapter 6, this thesis proposes a method that learns view-invariant 3D pose representation from input images and without using any 3D joint annotations and viewpoint parameters such that the pose features can be applied *directly* to unseen-view downstream tasks.

2.4 Evaluation Protocols for View-Invariant Action Assessment

To demonstrate the performance of the proposed methods, this thesis follows the state-of-the-art view-invariant action recognition approaches, such as [78, 126, 143, 144], and applies two standard protocols, cross-subject and cross-view, where possible.

Cross-Subject Protocol – The cross-subject protocol (CS) aims to evaluate if a method maintains a high performance when it is trained on several viewpoints. To do this, distinct subjects are engaged for training and testing, while data from all viewpoints are used in both phases.

Cross-View Protocol – The cross-view evaluation criteria (CV) aims to assess the performance of the methods on novel viewpoints, so data from different viewpoints are applied during training and testing, while all subjects are applied in both training and testing sets.

2.5 Conclusion

This chapter presented an overview of the related works upon which this thesis builds. Since this thesis presents work that relies upon action quality assessment, the related works in this area, including sports analysis approaches and human movement assessment methods in the healthcare domain, were first reviewed. Then, due to the lack of existing view-invariant action quality assessment method, related view-invariant action recognition works were discussed by covering both skeleton and RGB-D based view-invariant techniques. Finally, as this thesis proposes an unsupervised approach to learn view-invariant 3D human pose representation towards movement assessment, recent unsupervised 3D pose estimation and representation works were discussed.

Chapter 3

Datasets

In this thesis, several datasets are used; two multi-view human movement assessment datasets, SMAD and QMAR, that have been developed in-house, and QMAR is publicly released; a single-view human movement assessment dataset, KIMORE; and the most popular multi-view action recognition dataset, NTU RGB+D. KIMORE and NTU RGB+D are both available in the computer vision community.

SMAD is utilized to train and evaluate the proposed method in Chapter 4. QMAR, along with KIMORE are involved in experiments of Chapters 5 and 6. NTU RGB+D is also employed in Chapter 6.

This chapter begins by providing a summary of movement assessment datasets in healthcare and discusses the need for new datasets in Section 3.1. Then, in Section 3.2, it first introduces SMAD and QMAR, and then reviews the details of the KIMORE and NTU RGB+D datasets.

3.1 Human Movement Assessment Datasets

Paiement et al. [98] have collected one of the first human action analysis datasets, SPHERE-Staircase. In another work [137], they introduce two other datasets, SPHERE-Walking and SPHERE-SitStand. All three datasets include both depth and skeleton data and have been recorded from frontal view. SPHERE-Staircase has been captured with 12 participants and includes 48 abnormal gaits in going up the stairs with lower musculoskeletal conditions (*e.g.* freezing of gait). In this dataset, all the frames of the sequences have been labelled as normal or abnormal by an experienced physiotherapist. SPHERE-Walking contains 40 video sequences of 10 subjects walking on a flat surface. It

includes videos of normal walking and walking while the participants simulated Stroke and Parkinsons ailments. In SPHERE-SitStand, Tao et al. [137] record 101 video sequences of 10 subjects performing sitting-to-standing. This dataset has normal and two types of abnormal movements containing (i) freezing and (ii) restricted knee and hip flections.

Although SPHERE-Staircase, SPHERE-Walking, and SPHERE-SitStand datasets allow exploring automatic movement assessment, they are not large enough for training the most recent deep learning backbones. To overcome this problem, authors in [10, 140] have collected bigger skeletal datasets. Vakanski et al. [140] develop UI-PRMD using

Dataset	Year	MV	Data Type	Annotation Type	#Movement Types	#Subjects
SPHERE-Staircase [98]	2014	✗	Depth, Skeleton	N & AB	1	12
SPHERE-Walking [137]	2016	✗	Depth, Skeleton	N & AB	3	10
SPHERE-SitStand [137]	2016	✗	Depth, Skeleton	N & AB	3	10
UI-PRMD [140]	2018	✗	Motion Capture, Skeleton	Action Classes	10	10
SMAD [120]	2019	✔	Motion Capture, RGB, Depth, Skeleton	N & AB	4	19
EJMQA [37]	2019	✗	Depth, Skeleton	Score	3	43
KIMORE [12]	2019	✗	RGB, Depth, Skeleton	Score	5	78
EHE [10]	2021	✗	Skeleton	Action Classes	6	25
QMAR [121]	2021	✔	RGB, Depth, Skeleton	Score	6	38

Table 3.1: *Details of known human movement assessment datasets reviewed in Section 3.1, as well as action quality assessment datasets employed in experiments of this thesis (gray highlights). MV: multi-view. N & AB: normal and abnormal.*

a VICON system and a Kinect camera for physical rehabilitation exercises. It involves 10 healthy subjects who performed 10 types of exercises (*e.g.* deep squat, side lunge, and standing shoulder extension) in both correct and incorrect fashion. The participants have been asked to repeat 10 times each movement type. Bruce et al. [10] have recorded EHE dataset from 25 older participants with Alzheimer's disease completing six routine exercises that patients usually perform in the elderly home. In total, the dataset includes 869 sequences.

As opposed to the datatests described above that annotated the movements with only the types of abnormality or action classes, Elkholy et al. [37] introduce a single-view dataset, EJMQA, for which the depth and skeleton sequences have been annotated by a physiatrist to reflect the severity of the abnormality. The EJMQA dataset has been recorded in a clinic from 41 participants, including both patients and healthy individuals, performing walking, standing up, and sitting down action types.

Table 3.1 provides a summary of known human movement assessment datasets. The existing datasets for human movement analysis are all single-view, and most datasets do not include RGB images, while as outlined in Chapter 1, this thesis aims to deal with the view-invariance when assessing the quality of human movement from RGB images. Therefore, the current human movement assessment datasets are not suitable for this purpose, and recording a multi-view dataset, including RGB images, is essential to evaluate the performance of the proposed methods.

3.2 Datasets Used in this Thesis

3.2.1 SMAD: Sphere Multi-View and Multi-Modal Movement Assessment Dataset

This thesis introduces the first known multi-view human action or movement assessment dataset, SMAD [120] which combines, motion capture, skeletons, depth and RGB images. It should be noted that SMAD has been collected by the researchers at the Digital Health Group, University of Bristol, and has been post-processed by this author for use in this thesis. Both stages are described in detail next.

3.2.1.1 Data Recording

SMAD has been recorded from 19 healthy subjects, 6 female and 13 male, who have been trained by a specialist physiotherapist to perform a turn-walk action, *i.e.* a return walk to approximately the original position, in both normally and with three types of

abnormalities: Stroke, Parkinson, and short-limp. Figure 3.1 illustrates examples of all movement types. The normal movement has been repeated five times, while the abnormal ones have been performed only once.

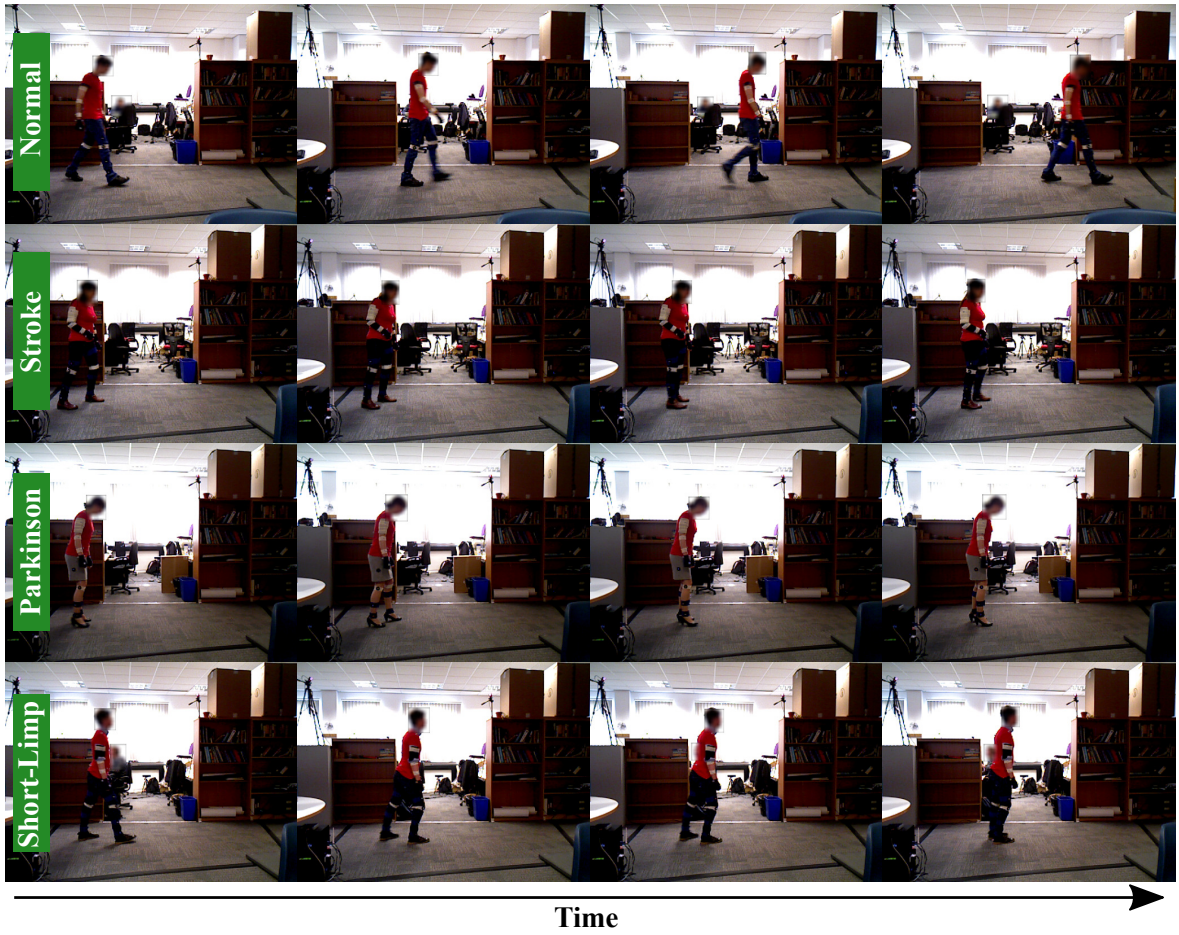


Figure 3.1: *Examples of all movement types of SMAD. Top row: normal turn-walk action. Second row: turn-walk with Stroke. Third row: turn-walk with Parkinson. Bottom row: turn-walk with short-limp.*

SMAD has been captured by motion capture[1] and four RGB-D cameras, three Primesense and one Kinect, from four viewing directions for the entirety of each walk: towards one camera and back to the opposite camera, one side view, and one downward view of the scene, *i.e.* views 1, 2, and 3 are at $\approx 0°$, $\approx 90°$, and $\approx 180°$ respectively, and view 4 is around $\approx 45°$ above view 3. Note that all the cameras have been synchronized. Figure 3.2 illustrates the position of RGB-D cameras in SMAD, and Table 3.2 details the number of frames and sequences for each movement type.

[1]The Optitrack Flex 3 acquisition system has been used for human motion capturing

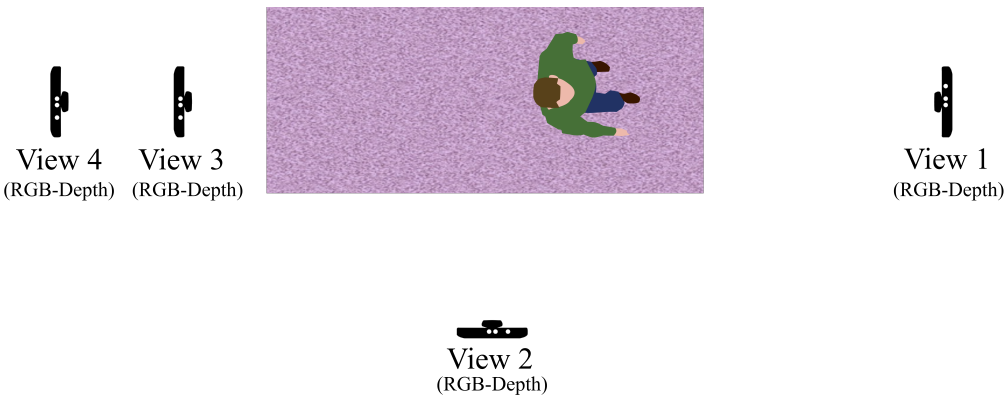


Figure 3.2: *Typical camera views in the SMAD dataset.*

Turn-Walk Action	# Sequences	#Frames/Video Min-Max	Total Frames
Normal	85	172-394	103026
Stroke	8	675-1023	27512
Limp	8	720-1351	31644
Parkinson	12	600-1740	47948

Table 3.2: *Details of movements in the SMAD dataset. The dataset contains a total of 210130 frames representing 113 sequences of videos.*

RGB and depth images have been collected for all four viewpoints, while the skeleton data have been captured from only the Kinect camera, and the motion capture data are available for only the normal movement types. The videos have been recorded in an uncontrolled environment, a part of the old site of the Visual Information Laboratory at the University of Bristol, so there are no control over the light and background of the views. Figure 3.3 shows some RGB sample frames from four viewpoints of SMAD. To track the body joint movements with the motion capture system, 39 markers have been mounted on the subjects' bodies. The locations for attaching the markers are shown in Figure 3.4. Note that in addition to 3D skeleton data, 3D motion capture data have been also collected in SMAD since motion capture systems are more accurate than the RGB-D cameras that can estimate 3D pose only in optimal conditions, *i.e.* the pose efficiency of the RGB-D cameras is depended on their distance and viewing direction from the subject.

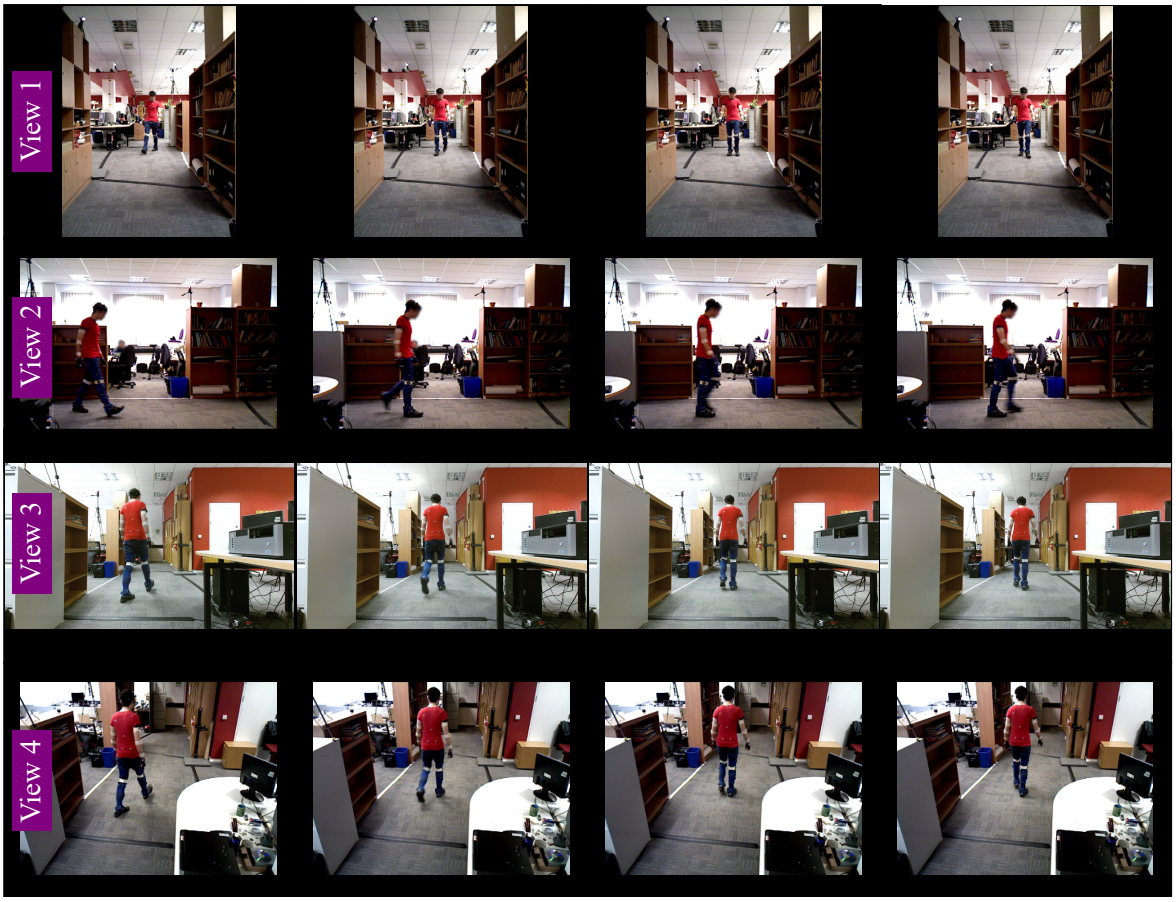


Figure 3.3: *Sample frames from SMAD for all views. Each column shows a scene captured from 4 different camera views.*

3.2.1.2 Data Post-Processing

Depth images and motion capture data require some post-processing steps before they are applied in experiments.

Improving Quality of Depth Images – The depth images acquired by RGB-D cameras frequently contain hole regions, *i.e.* unfilled areas where depth values are missing. To overcome this problem, the hole filling algorithm proposed in [22] has been used to improve the quality of depth data.

Labeling Motion Capture Data – After recording the motion capture data, a post-processing phase including marker labelling and gap-filling is essential. First, the human body skeletons have to be constructed from raw markers, so a marker labelling step is required. Second, as during the motion capturing, some markers may disappear for some frames due to noise or occlusion, a gap-filling step is also necessary. The data post-processing in SMAD has been performed with VICON's NEXUS skeleton building software. Figures 3.5 shows a sample of motion capture data before and after labelling

and filling the gaps.

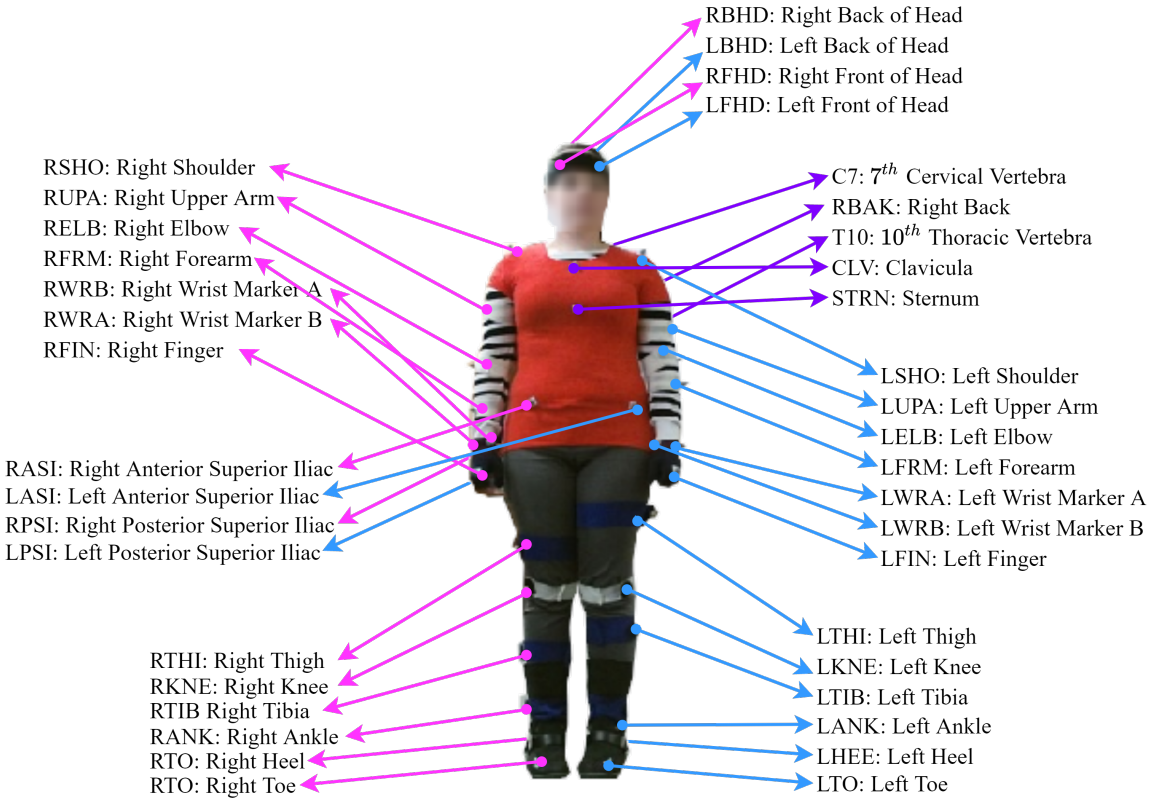
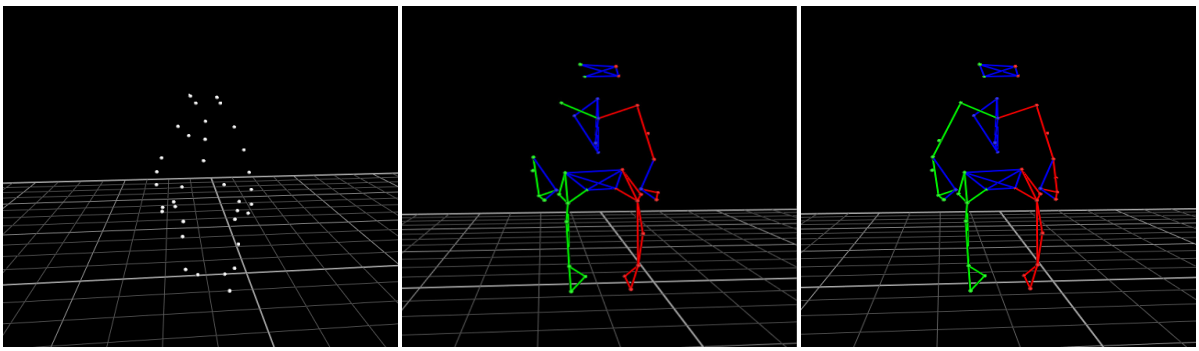


Figure 3.4: *Locations for attaching the motion capture markers on the human body in the SMAD dataset.*



a: raw motion capture data

b: motion capture data after labelling

c: motion capture data after labelling and gap-filling

Figure 3.5: *a, b, and c show a sample of motion capture data before joint labelling and gap filling, after labelling and before filling the gaps, and after labelling and gap-filling respectively.*

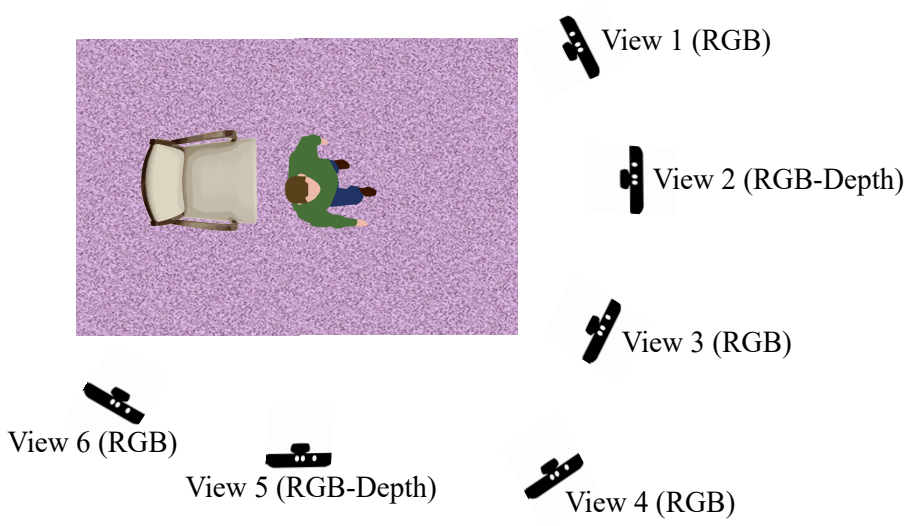


Figure 3.6: *Typical camera views in the QMAR dataset with each one placed at a different height.*

3.2.2 QMAR: Multi-View Quality of Movement Assessment for Rehabilitation Dataset

To better explore view-invariance in assessing the quality of human movement, this thesis also introduces the second multi-view human action or movement assessment dataset, QMAR[2] [121], larger than the SMAD dataset. Note, this dataset has been recorded and labelled specifically for this thesis.

Although SMAD allows studying view-invariance, it is limited by the number of action types, sequences, and variety of participants. However, the QMAR dataset provides further exploration opportunity by capturing more viewpoints and complex actions with double the number of participants. In further contrast to the SMAD dataset where movements have been annotated into normal and abnormal, the movements in QMAR have been scored by the severity of the abnormality. All these features make QMAR a suitable dataset for view-invariant human movement assessment to compare the strengths and shortcomings of different approaches, and the size of dataset also allows training deep neural networks.

3.2.2.1 Data Recording

To capture QMAR, the author of this thesis developed the necessary software and set up the associated hardware. The QMAR dataset has been recorded using six Primesense

[2]The QMAR dataset is published at https://data.bris.ac.uk/data/dataset/1y37kc9a8y47y2cen7j907bpm7

Figure 3.7: *Sample frames from QMAR for all views. Each column shows a scene captured from 6 different camera views.*

cameras in an uncontrolled environment, in the Visual Information Laboratory at the University of Bristol. Figures 3.6 and 3.7 show the position of the six cameras and some RGB sample frames from each of the viewpoints respectively. The dataset includes RGB, depth, and skeleton data. As capturing depth data from six Primesense cameras is not possible due to infrared interference, the depth and skeleton data have been retained

from only view 2 at $\approx 0°$ and view 5 at $\approx 90°$. Note, all the six cameras have been synchronized.

QMAR has been captured with 38 healthy subjects, 8 female and 30 male. The subjects have been trained by a physiotherapist to perform two different types of movements while simulating two ailments, resulting in four overall possibilities: a return walk to

Figure 3.8: *Examples of all movement types of QMAR. First column: walking with Parkinsons (W-P). Second column: walking with Stroke (W-S). Third column: sit-stand with Parkinsons (SS-P). Fourth column: sit-stand with Stroke.*

approximately the original position while simulating Parkinsons (W-P), and Stroke (W-S), and standing up and sitting down with Parkinsons (SS-P) and Stroke (SS-S). Figure 3.8 illustrates examples of all movement types.

In QMAR, the movements have been scored by the severity of the abnormality, and the score ranges are 0 to 4 for W-P, 0 to 5 for W-S and SS-S, and 0 to 12 for SS-P. A score of 0 in all cases indicates a normally executed action. It should be noted that the author of this thesis has been trained by the physiotherapist to annotate QMAR, and after recording and annotating the dataset, the movements and scores have been approved by the clinical experts. Table 3.3 details the quality score or range and the number of frames and sequences for each action type. Table 3.4 details the number of sequences for each score.

Action		Quality Score	# Sequences	#Frames/Video Min-Max	Total Frames
W	Normal	0	41	62-179	12672
W-P	Abnormal	1-4	40	93-441	33618
W-S	Abnormal	1-5	68	104-500	57498
SS	Normal	0	42	28-132	9250
SS-P	Abnormal	1-12	41	96-558	41808
SS-S	Abnormal	1-5	74	51-580	47954

Table 3.3: *Details of movements in the QMAR dataset.The dataset contains a total of 202800 frames representing 306 sequences of videos.*

Action	Score											
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**#11**	**#12**
W-P	4	8	16	12	-	-	-	-	-	-	-	-
W-S	10	14	19	15	10	-	-	-	-	-	-	-
SS-P	1	1	6	8	4	4	4	3	3	1	2	4
SS-S	3	19	19	13	20	-	-	-	-	-	-	-

Table 3.4: *Details of abnormality score ranges in the QMAR dataset*

3.2.2.2 Data Post-Processing

As with SMAD, depth data in QMAR also contains hole regions and requires a post-processing phase to improve their quality. To perform this task, the hole filling algorithm introduced in [22] has been applied to depth data.

3.2.3 KIMORE Dataset

KIMORE [12] is a single-view rehabilitation movement dataset for which the quality of movements have been annotated for quantitative scores. KIMORE includes RGB, depth, and skeleton joints positions and has 78 subjects (44 healthy, and 34 real patients suffering from Parkinson, Stroke, and back pain) performing five types of rehabilitation exercises for lower-back pain. Exercise 1 (Ex #1) contains the movement of the upper limbs, exercises 2 to 4 (#Ex 2 to Ex #4) involve movement of the trunk, and Exercise 5 (Ex #5) contains the movements of the lower limbs. All videos are frontal view (see sample frames in Figure 3.9).

In KIMORE [12], clinicians have defined a score with values in the range of 0 to 50 for each exercise. The scores are computed by the sum of two sub-scores, PO_S and CF_S which represent the exercise goal achievement and physical constraints during the exercise respectively.

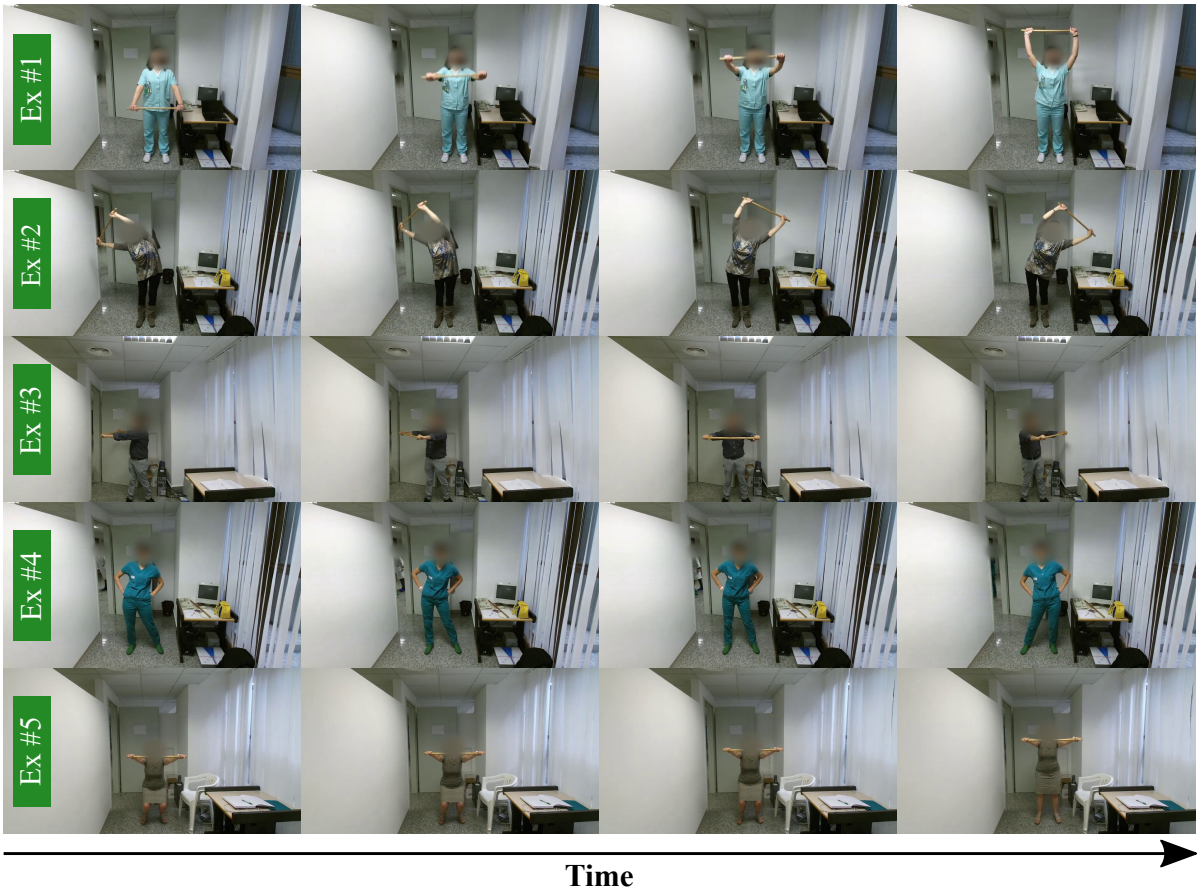


Figure 3.9: *Sample frames of KIMORE for five different exercises. Each row shows frames belonging to a video sequence.*

3.2.4 NTU RGB+D Dataset

NTU RGB+D[126] is the most popular multi-view action recognition dataset since it is a large-scale dataset that allows training deep neural networks, whereas the other available multi-view action recognition datasets, such as [110, 146, 153] are limited by the size of training data, the number of classes, viewpoints, and participants. NTU has 56000 video sequences that have been recorded from 40 subjects performing 60 distinct action types. It contains 17 different environmental settings captured by three cameras from three different viewpoints, view 1 $\approx 0°$, view 2 $\approx 45°$ and view 3 $\approx 90°$. Figure 3.10 shows three samples of the NTU dataset for three views that have been recorded under three different settings.

Figure 3.10: *Sample frames of NTU RGB+D [126] for three different actions: put on a hat, brush hair, and put on a shoe. Each column shows a different viewpoint while each row represents a distinct environmental setting.*

3.3 Conclusion

This chapter provided a summary of human movement assessment datasets in the healthcare domain. As reviewed, all existing datasets are single-view and do not allow exploring view-invariance in human movement assessment. In addition, this thesis aims to tackle this challenging task from RGB images, whereas most datasets include skeleton and/or depth data. To allow studying view-invariance from RGB sequences, this chapter introduced two multi-view human movement assessment datasets SMAD and QMAR. It also presented the details of KIMORE and NTU RGB+D which are employed in this thesis.

Chapter 4

Multi-View Training for Human Movement Assessment

This chapter explores an approach that estimates 3D pose in a high-level low-dimensional manifold space from *RGB images* such that the pose features are suitable for human movement assessment under *multi-view learning scenarios*. The aim is to propose a method where the complexity does not increase with the number of training views, and it maintains a high performance on single-views. In addition, the proposed method allows assessing the quality of human movement directly from the extracted pose features without requiring any intermediate skeleton-based step. The work presented in this chapter has been published in [120].

Section 4.1 gives an overview of the proposed pose estimation approach to human movement assessment in multi-view learning scenarios. Subsequently, the proposed approach is described in Section 4.2. Using the SMAD dataset, Section 4.3 obtains experiments for the proposed pose estimation method and evaluate the performance of the learned pose features for human movement assessment. Conclusions are in Section 4.4.

4.1 Overview of Proposed Method

Current action assessment methods in the healthcare domain are commonly based on 3D skeleton data, such as [3, 9, 67, 98, 128], since the features extracted from 3D poses are rich and can be leveraged to analyse a wide range of human movements. However, capturing skeleton data by RGB-D cameras (*e.g.* Kinect) or motion capture devices is challenging in in-the-wild scenarios, e.g. in healthcare rehabilitation monitoring at home or in the clinic.

Skeleton-based approaches usually require two pre-processing steps, (i) normalization and (ii) dimensionality reduction. The skeleton data should be normalized since different subjects come in various shapes and sizes and they also do not perform actions at the same world coordinates. In addition, due to the curse of dimensionality, a dimensionality reduction step becomes necessary to reduce the redundancy presented in this data. For instance, Paiement et al. [98] use manifold learning techniques, or the authors in [36, 37] manually select a subset of body joints based on the movement type (see Section 2.1.2 for more details).

To reduce the dimensionality of data, manifold learning techniques show promising performance since they discover a low dimensional space while preserving the intrinsic geometrical structure of the original data. However, the approaches that use these algorithms cannot be extended easily for multi-view learning scenarios, since almost all the existing manifold learning techniques, such as [5, 21, 138, 173], work under single view settings. To overcome this problem, there are two solutions (i) generating one manifold per view and training the model on each independently, and (ii) operating iterative algorithms, such as [167, 175], on single-view manifolds to exhaustively seek a multi-view manifold space. In the former, the viewpoint information of the input data is required to be mapped into the proper manifold space at inference. In the latter approach, the complexity of the model increases with the rising number of views.

To tackle the challenges mentioned above, this chapter proposes a CNN-regression model that estimates high-level 3D pose features in a canonical (view-invariant) manifold space from RGB images towards human movement assessmet in multi-view learning senarios. Figure 4.1 illustrates an overview of the proposed pose estimation approach and its application to multi-view human movement assessment.

The inputs of the proposed CNN-regression model are body joint heatmaps and body limb-maps, derived from RGB images to help the network exploit geometric relationships amongst different body parts to estimate the pose features more accurately. The ground truth poses come from a canonical manifold that is generated from normalized 3D motion capture data.

Although the proposed method requires 3D skeleton data for training, it facilitates human movement assessment at inference by *eliminating* the need for *3D data*. It also *removes* the need for the pose-based *pre-processing* steps, *i.e.* normalization and dimensionality reduction, as the extracted features are in a high-level manifold. Furthermore, as the embedding space is canonical, *i.e.* the proposed method estimates the same pose features for all simultaneous frames captured from different viewpoints, it can be ap-

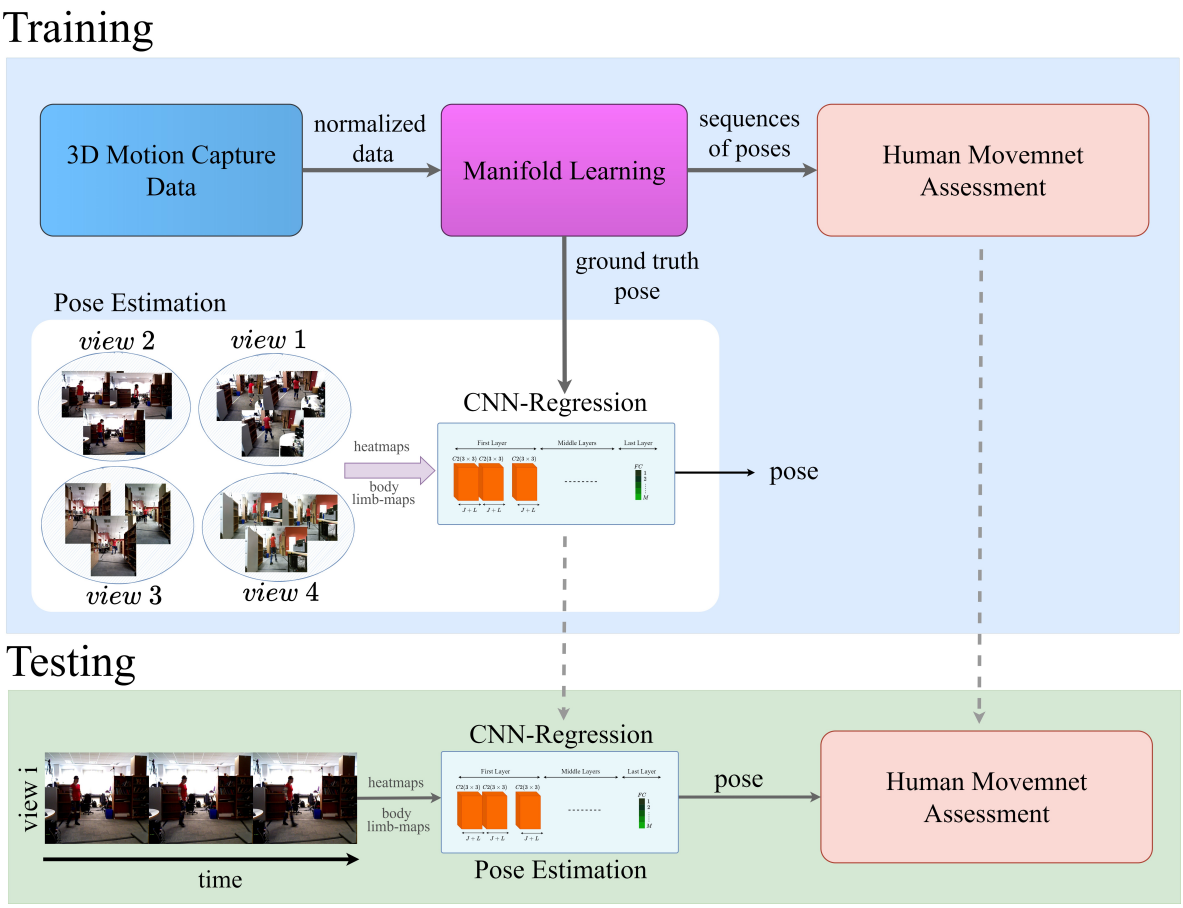


Figure 4.1: *Top: first a canonical manifold is generated from normalized 3D motion capture data. Then, (i) the proposed pose estimation method is trained through the manifold space and images from different views, and (ii) the movement assessment approach is trained on sequences of high-level canonical poses in the manifold space. Bottom: for inference, the proposed pose estimation approach is applied to each frame of a sequence and then the estimated poses are fed into the assessment method.*

plied easily for the *multi-view* learning scenarios while the method's complexity does not increase with the number of training views. The proposed pose estimation approach is described in detail next.

4.2 High-Level 3D Pose Estimation in Multi-View Learning Scenarios

This section presents the proposed high-level 3D pose estimation approach for human movement assessment in multi-view learning scenarios. It first describes the process of

generating the ground-truth pose annotations (Section 4.2.1), and then it elaborates details of the proposed network and the loss function used for training (Section 4.2.2).

4.2.1 Ground Truth Pose Generation

The ground-truth pose features are obtained from 3D motion capture data such that they are first normalized, and a manifold learning technique is then applied to them.

4.2.1.1 Data Normalization

The normalization aims to remove variations in scale, translation, and rotation from the pose data since the subjects have various shapes and sizes and they also do not perform actions at the same world coordinates. To do this, several normalization methods, such as [98, 137], were applied to the data of this chapter, and the best result was conducted through the following proposed approach:

Translation Invariance – To normalise for translation, given a pose $P \in \mathbb{R}^{3\times J}$ where J refers to the number of joints, its hip centre $P^{\odot}$ is considered as the origin of the coordinate system and the other joint positions are normalized relative to it as

$$P^j = P^j - P^{\odot} \ , \tag{4.1}$$

where $j = \{1, 2, \ldots, J\}$ and $J = 39$ for the motion capture system used in this thesis.

Scale Invariance – To normalise for scaling, first a model skeleton $\boxdot$ is defined as a template and then its torso, hand and leg sizes are applied to normalise pose P as

$$r_o = \mathfrak{O}(\boxdot)/\mathfrak{O}(P), \quad P^i = P^i \times r_o \ , \tag{4.2}$$
$$r_{h_\cap} = \mathfrak{H}_\cap(\boxdot)/\mathfrak{H}_\cap(P), \quad P^{c_1} = P^{c_1} \times r_{h_\cap} \ , \tag{4.3}$$
$$r_{h_\cup} = \mathfrak{H}_\cup(\boxdot)/\mathfrak{H}_\cup(P), \quad P^{c_2} = P^{c_2} \times r_{h_\cup} \ , \tag{4.4}$$
$$r_{l_\cap} = \mathfrak{L}_\cap(\boxdot)/\mathfrak{L}_\cap(P), \quad P^{k_1} = P^{k_1} \times r_{l_\cap} \ , \tag{4.5}$$
$$r_{l_\cup} = \mathfrak{L}_\cup(\boxdot)/\mathfrak{L}_\cup(P), \quad P^{k_2} = P^{k_2} \times r_{l_\cup} \ , \tag{4.6}$$

where $i \in \mathit{torso}$, $c_1 \in \mathit{upper\ hand}$, $c_2 \in \mathit{lower\ hand}$, $k_1 \in \mathit{upper\ legs}$, and $k_2 \in \mathit{lower\ legs}$. $\mathfrak{O}$, $\mathfrak{H}_\cap$, $\mathfrak{H}_\cup$, $\mathfrak{L}_\cap$, and $\mathfrak{L}_\cup$ functions return the size of torso, upper hand, lower hand, upper leg, and lower leg respectively by computing the Euclidean distance between the upper and lower joints of that body part.

Rotation Invariance – To normalise for rotation, Procrustes Analysis (PA) is applied to data. Note, PA is not employed for translation and scaling since for translation invariance, the centre computed by PA is different from the centre of the human body shape, and for scale invariance, PA scales the whole shape at once, while different scale ratios must be considered for different body parts.

4.2.1.2 Manifold Learning

Let $D = \{P'_n\}_{n=1}^{N}$ be the set of normalized 3D motion capture body joints obtained through the previous section, where $P'_n \in \mathbb{R}^{3\times J}$ and J refers to the number of body joints. To reduce the dimensinality of this data, following [22, 98, 137], Diffusion Maps Φ [21] as a non-linear manifold learning technique is applied to D,

$$\tilde{D} = \Phi(D), \tag{4.7}$$

where $\tilde{D} = \{\tilde{P}_n\}_{n=1}^{N}$, $\tilde{P}_n \in \mathbb{R}^{M}$ and $M \ll 3 \times J$. Note, M refers to the new dimension of pose data.

Diffusion Maps represent a dataset in a weighted graph and use the spectral properties of the graph Laplacian to embed the high dimensional data into a lower-dimensional space [2]. This method shows several advantages over classical dimensionality reduction approaches, such as Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS) [2, 21], since it is able to deal with data points that rely on nonlinear manifolds, and preserves the local geometric structure of the original data. It is also robust to noise.

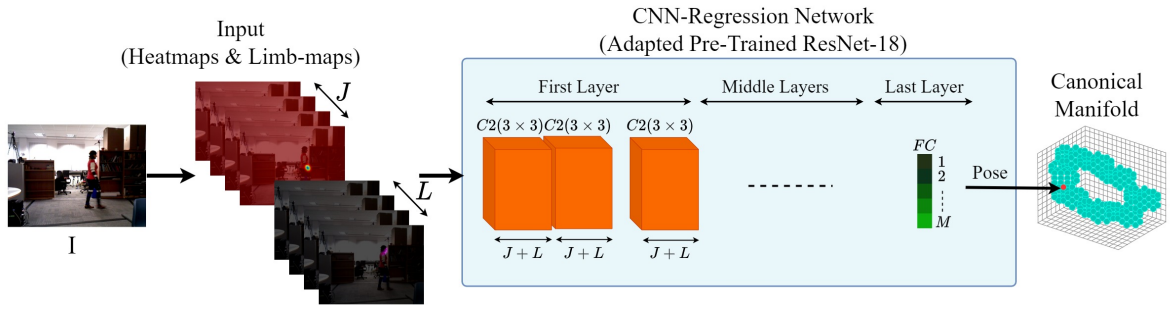


Figure 4.2: *The overall schema of the proposed network to estimate high-level canonical 3D human pose.*

4.2.2 Proposed Network

The overall schema of the proposed CNN-regression is shown in Figure 4.2. The proposed network exploits geometric relationships amongst 2D body parts to learn 3D pose features in a high-level low-dimensional manifold space. To this end, a set of body joint heatmaps and limb-maps are derived from RGB images and used as input to inject priors on the position and structure of the human body into the network. Note that utilizing these maps instead of raw RGB images also prevents over-fitting on the subject appearances during the network training.

4.2.2.1 Network Inputs

Body Joint Heatmap – Given an RGB image, joint heatmap H_j represents the probability of joint j occurring at each pixel position $x \in \mathbb{R}^2$ of the image as:

$$\mathbf{H}_j(x) = exp(-\frac{\|x - x^{\oplus}\|_2^2}{\sigma^2}) \ , \tag{4.8}$$

where $x^{\oplus} \in \mathbb{R}^2$ is the ground-truth position of joint j, σ determines the spread of the peak.

Body Limb-Map – A body limb-map is a set of 2D vectors encoding the orientation and location of a body limb. Given an RGB image, the value of limb-map $\mathbf{B}_l$ at each pixel position $x \in \mathbb{R}^2$ of the image is computed as:

$$\mathbf{B}_l(x) = \begin{cases} v & \text{if } x \text{ is on limb } l \\ 0 & \text{otherwise} \end{cases} \quad , \text{ where } \quad v = \frac{x^{\circledast} - x^{\circledcirc}}{\|x^{\circledast} - x^{\circledcirc}\|} \ , \tag{4.9}$$

$\mathrm{x}^{\circledast} \in \mathbb{R}^2$ and $x^{\circledcirc} \in \mathbb{R}^2$ are the ground-truth pixel positions of body joints defining limb l.

To generate heatmaps and body limb-maps, state-of-the-art OpenPose [11][1] is applied to RGB data to produce 26 body joint heatmaps $H = \{H_j\}_{j=1}^{26}$ and 52 body limb-maps $B = \{B_l\}_{l=1}^{52}$. Figures 4.3 and 4.4 show some body joint heatmaps and limb-maps generated from a sample RGB image of SMAD by OpenPose.

[1] It should be noted that this work was carried out between 2018 and 2019 and at that time OpenPose [11] was the stat-of-the-art 2D pose estimation method.

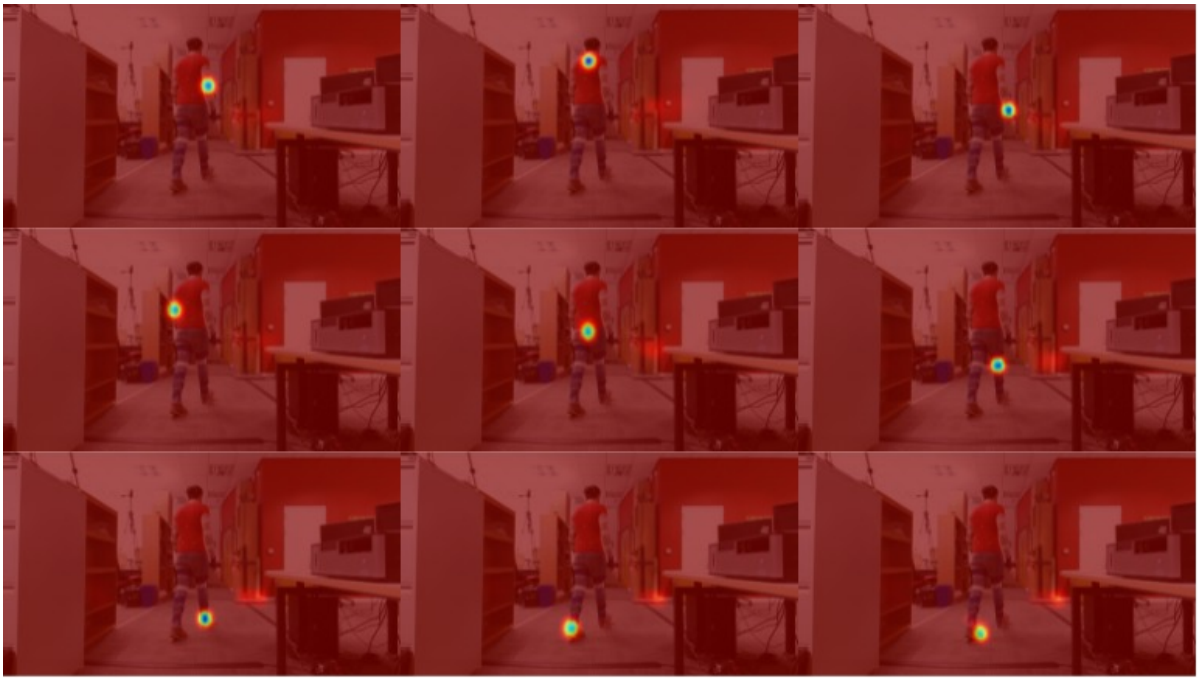


Figure 4.3: *Sample body joint heatmaps generated by OpenPose [11] for an RGB images of SMAD.*

Figure 4.4: *Sample body limb-maps generated by OpenPose [11] for an RGB images of SMAD.*

4.2.2.2 Network Architecture and Training

To design the proposed network, AlexNext [71], VGG [129], and ResNet [49] were investigated and it was found that ResNet obtains the best performance. This may be attributed to addressing the vanishing gradient problem by ResNet architecture. Although ResNet is a deep CNN network, using shortcut connections [8, 49, 117] in its design, which operate as gradient superhighways, allows the gradients flow into earlier layers during backward pass. Therefore, the proposed network is implemented by adapting a pre-trained ResNet-18 [49][2] such that the first layer of ResNet-18 is replaced with a convolutional layer, with a depth of $J + L$ where J and L denote the number of body joint heatmaps and limb-maps respectively. The last layer of ResNet-18 is also replaced with a regression layer with the size of the manifold dimension M.

The network is trained through MSE loss function that computes the difference between the ground-truth pose $P \in \mathbb{R}^{\mathrm{M}}$ in manifold space and the pose $\hat{P} \in \mathbb{R}^{\mathrm{M}}$ estimated by the proposed method:

$$Loss(P, \hat{P}) = \frac{1}{M} \sum_{i=1}^{M} \|p_i - \hat{p}_i\|_2^2 \,, \tag{4.10}$$

where p_i and $\hat{p}_i$ are the i^{th} coordinates of P and $\hat{P}$ respectively.

4.3 Experiments and Results

This section reports the results of experiments that evaluate the performance of the proposed pose estimation approach on the SMAD dataset. It also assesses the efficiency of the learned pose features for the human movement analysis task under a multi-view learning scenario, *i.e.* cross-subject protocol (see Section 2.4).

Comparison to Baseline – For evaluation against the closest possible approach, the proposed method is compared against a baseline using the method of Crabbe et al. [22] that was originally developed to estimate the pose features in the manifold space from depth silhouette inputs for a single-view scenario. However, as their dataset is not multi-view and does not contain RGB data, the only possible comparative analysis is to apply their method on depth data of SMAD and adapt it for the multi-view learning scenario. Crabbe et al. [22] uses AlexNet [71] to design their model, but for a fairer comparison, ResNet-18 [49] was employed for all the experiments.

[2] ResNet-18 [49] is pre-trained on ImageNet [25].

Sections 4.3.1 and 4.3.2 contain implementation details and evaluation metrics respectively. Section 4.3.3 compares the proposed method to the baseline and examines the importance of different input types in estimating the high-level pose features. Section 4.3.4 probes the robustness of the proposed method on multi-view training, and Section 4.3.5 examines the method's performance on unseen viewpoints. Finally, movement quality assessment based on the learned pose features is presented in Section 4.3.6.

CNN-Regression Network (Adapted ResNet-18)	
First layer:	$\{C2(3 \times 3, J + L)\} \times 64$, BN, ReLU, $\{$MP(3×3)$\}$
Middle layers:	As in ResNet-18
Last layer:	$\{FC(M)\}$

Table 4.1: *Details of the proposed CNN-regression network:* $\{C2(3 \times 3, J + L)\} \times 64$*: 64 2D convolution filters with size* 3 *and* $J + L$ *channel size,* $MP(3 \times 3)$*: 2D max pooling with size* 3*,* $FC(M)$*: fully connected (FC) layer with* M *outputs.* J *and* L *are the number of body joint heatmaps and limb-maps respectively, and* M *is the size of the manifold dimension.*

4.3.1 Implementation Details

Details of Network Architecture – Details of the proposed CNN-regression network are shown in Table 4.1.

Training and Testing Details – Size of manifold dimension for turn-walk action of SMAD was set to $M = 5$ since the first 5 dimensions of the manifold generated by the Diffusion Maps algorithm were able to represent 95% of the total variance of the original data. This exceeds the three dimensions used in [22], since the more complex actions of SMAD require more dimensions to be described. For pose estimation experiments, all models were trained for 20 epochs using mini-batch stochastic gradient descent with an initial learning rate of 0.001 that decreased by a factor of 10 every 5 epochs, momentum of 0.9, and a batch size of 10. In addition, for all the experiments, the input images were resized into 244×244 pixels.

Dataset Splits – As size of SMAD was large enough to conduct both types of experiments, pose estimation and human movement assessment, only one training and one testing set were selected randomly, *i.e.* 60 normal sequences were selected for training randomly and the remaining normal movements including 25 sequences were applied for

testing. Therefore, no mean, standard deviation and/or p-value is required to be computed for these experiments. In addition, as the cross-subject protocol was employed to evaluate the results, the normal training and testing sequences were selected from distinct subjects. The movement quality assessment models were additionally tested on 26 abnormal movements including 8 Stroke, 8 limp, and 12 Parkinson sequences. Note, the pose estimation networks were trained and tested on the frames comeing from only normal movements since as explained in Section 3.2.1, in SMAD, the motion capture data were available for only normal sequences.

4.3.2 Evaluation Metrics

To evaluate the performance of the proposed pose estimation approach, MSE between the ground-truth poses and estimated human poses were used. For human movement assessment, as each frame of a sequence was classified into normal or abnormal, true positive (TP), true negative (TN), false positive (FP), false negative (FN), true negative rate (specificity), true positive rate (sensitivity), precision, and recall were used for evaluation.

4.3.3 Pose Estimation in a Multi-View Learning Scenario

This section presents the results of the proposed method in a multi-view learning scenario where all four viewpoints of SMAD (views 1 to 4) are utilized for training and testing. It also outlines ablation studies to examine the impact of the network inputs on learning of the pose features.

Ablation Study – To ablate the network inputs, the proposed method was trained on (i) RGB bounding box of subject (RGB BB), (ii) body joint heatmaps generated from RGB data, (iii) body limb-maps extracted from RGB data, and (iv) combined heatmaps and limb-maps. Furthermore, to train the baseline [22], in addition to depth silhouette, the depth bounding box of subject (Depth BB) was also used as input and compared against the proposed method since extracting the depth silhouette is challenging for the cluttered environments in which SMAD has been recorded.

Comparative results on SMAD are shown in Table 4.2. The proposed method has the least error in estimating the high-level poses when it is trained on combined heatmaps and body limb-maps. It also outperforms the baseline [22] at 0.67, 0.67, and 0.66 where it uses heatmaps, limb-maps and combined heatmap and limb-maps as input respectively. The result from Crabbe et al. [22] using depth silhouettes is poorer than when Depth BB is used potentially due to the general difficulty of inaccurate silhouette extraction.

Note that for the rest of the experiments, the proposed method is trained with heatmaps and body limb-maps, and to compare against Crabbe et al. [22]'s work, instead of depth silhouettes, the simpler Depth BB is put through the network.

	Crabbe et al. [22] (Depth)		**Proposed Method** (RGB)			
	BB	**Silhouette**	**BB**	**Heatmap**	**Limb-map**	**Heatmap & Limb-map**
MSE	0.70	0.72	0.72	0.67	0.67	**0.66**

Table 4.2: *MSE between the ground-truth and estimated pose on SMAD under multi-view learning scenario for the baseline and the proposed method, and for different input types.*

Testing Set	**Training Set**	**Crabbe et al. [22]**	**Proposed Method**
View 1	View 1	0.73	0.67
	Views 1-4	0.70	**0.66**
View 2	View 2	0.74	0.72
	Views 1-4	0.72	**0.71**
View 3	View 3	0.73	0.66
	Views 1-4	0.70	**0.64**
View 4	View 4	0.70	0.65
	Views 1-4	0.69	**0.63**

Table 4.3: *MSE between estimated pose and ground-truth on SMAD for single-view testing sets where single and all views are used for training.*

4.3.4 Robustness of Proposed Method for Multi-View Training

This section investigates if the proposed method can maintain a high performance on the single views where multiple views are applied for training. Table 4.3 shows the performance of the proposed method when only a single view is used for training in comparison to when all views are employed. As shown by the results in Table 4.3, not only can the proposed method learn well to distinguish between views when multiple views are provided in the training set, but also the accuracy of the method is improved as the network's generalization increases by training on more diverse data. The MSE of

the proposed method is lower than Crabbe et al. [22] at 0.66, 0.71, 0.64, and 0.63 for views 1 to 4 respectively.

4.3.5 Performance of Proposed Method on Unseen View Data

Different from the previous sections that evaluate the efficiency of the proposed method on the scenarios where the testing data come from the viewpoints that are present in training data, this section examines the generalization ability of the proposed method on unseen views by deploying distinct views in training and testing sets.

Table 4.4 shows that the performance of both methods drop significantly when they are applied to data coming from an unseen viewpoint. This deterioration is especially noticeable when the methods are trained on data from viewpoint 2. From these results, it can be concluded that although the networks are trained to map the input images into a view-invariant (canonical) pose, they cannot tolerate the appearance variations as they have not been designed explicitly to address the viewpoint variations.

Training Set	Testing Set	Crabbe et al. [22]	Proposed Method
View 1	View 1	**0.73**	**0.67**
	View 2	1.40	1.38
	View 3	1.23	1.18
	View 4	1.20	1.07
Average		1.14	**1.07**
View 2	View 1	1.35	1.27
	View 2	**0.74**	**0.72**
	View 3	1.40	1.29
	View 4	1.45	1.34
Average		1.23	**1.15**
View 3	View 1	1.14	1.09
	View 2	1.42	1.41
	View 3	**0.73**	**0.66**
	View 4	1.05	0.97
Average		1.08	**1.03**
View 4	View 1	1.13	0.95
	View 2	1.42	1.47
	View 3	1.10	0.94
	View 4	**0.70**	**0.65**
Average		1.08	**1.00**

Table 4.4: *MSE between estimated pose and ground-truth on SMAD where networks are trained on single-views and tested on single-(un)seen views.*

4.3.6 Cross-Subject Human Movement Assessment

To evaluate the performance of learned pose features for human movement assessment, the statistical method proposed in [98] that is a frame-by-frame movement assessment approach was utilized. This method has two models, pose and dynamic, that are trained on poses of normal movements.

Pose Model – Let p be a vector representing human pose in the high-level manifold space. The pose model embodies normal poses by learning their pdf $f_P(p)$ using a Parzen window estimator [98].

Dynamic Model –The dynamic model leverages temporal information of normal sequences through a continuous-state HMM [98]. The hidden states of the HMM are modeled by a random variable S_t with value $s_t \in [0, 1]$ that represents the progression of movement at frame t, where S_t is set to 0 at the first frame and it increases linearly to 1 at the last frame.

The observation model of the HMM is trained as

$$f_{P_t}(p_t|s_t) = \frac{f_{P_t,S_t}(p_t, s_t)}{f_{S_t}(s_t)}, \tag{4.11}$$

where p_t denotes pose p in manifold space at frame t, and $f_{P_t,S_t}(p_t, s_t)$ and $f_{S_t}(s_t)$ are computed by employing a Parzen window estimator.

The transition model of the HMM is defined as

$$f_{S_t}(s_t|s_{t-1}) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{\Delta s_t - \nu\Delta\eta_t}{\sigma})), \tag{4.12}$$

where $\nu = \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta s_i}{\Delta\eta_i}$, $\Delta s_t = s_t - s_{t-1}$, $\Delta\eta_t = \eta_t - \eta_{t-1}$, η_t is time at frame t, and N is number of frames of a sequence.

The dynamic model classifies p_t by a conditional pdf as

$$f_{P_t}(p_t|p_1, ..., p_{t-1}) \approx f_{P_t}(p_t|\hat{s}_t) f_{S_t}(\hat{s}_t|\hat{s}_{t-1}), \tag{4.13}$$

where $\hat{s}_t$ denotes the optimum value for S_t that satisfies the constraints on the hidden state, *i.e.* the hidden state linearly increases during a normal sequence. The derivation of Equation 4.13 is presented in Appendix A.

To perform the experiments of this section, the pose and dynamic models were trained on the high-level canonical poses derived from normal sequences of motion capture data. At inference, each frame of a sequence, which can belong to normal or abnormal movements

is given to the proposed pose estimation method, and its output is classified as normal or abnormal depending on how far away from the pose and dynamic models it is, based on two thresholds on the pose and dynamic log-likelihoods. The thresholds were empirically set to 4.0 and 4.2 for the pose and dynamic models respectively. Figure 4.1 shows the overall schema of the assessment framework.

		Normal		**Stroke**		**Limp**		**Parkinson**	
	Method	TN	FP	TP	FN	TP	FN	TP	FN
View 1	Crabbe et al. [22]	4245	3620	**4782**	**1971**	**5009**	**2902**	**7586**	**4401**
	Proposed Method	**4856**	**3009**	4337	2416	4351	3560	6200	5787
View 2	Crabbe et al. [22]	4079	3768	**4529**	**2224**	**5087**	**2824**	**7961**	**4026**
	Proposed Method	**4936**	**2929**	4244	2509	4419	3492	6948	5039
View 3	Crabbe et al. [22]	4625	3240	**4487**	**2266**	**4933**	**2978**	**7415**	**4572**
	Proposed Method	**4880**	**2985**	4259	2494	4213	3698	6648	5339
View 4	Crabbe et al. [22]	3981	3884	**4400**	**2353**	**4923**	**2988**	**7520**	**4467**
	Proposed Method	**4964**	**2901**	4259	2495	4630	3281	6706	5281

Table 4.5: *Results of frame classification on SMAD for normal and abnormal sequences. TN: true negative, FT: false positive, TP: true positive, FN: false negative.*

Tables 4.5 reports the results of the classification of individual frames for different views separately under the cross-subject protocol. It presents TN and FP for normal sequences, and TP and FN for abnormal sequences including Stroke, limp and Parkinson. Table 4.6 reports true negative rate (specificity) and true positive rate (sensitivity) for normal and abnormal movements respectively, and precision and recall values for all sequences. Similar to the previous sections, the performance of the proposed method is also compared against the baseline using the method of Crabbe et al. [22] for pose estimation.

Table 4.6 shows that the specificity for the proposed method to estimate pose of normal sequences for views 1 to 4 is higher at 0.61, 0.62, 0.62, and 0.63 than Crabbe et al. [22] at 0.53, 0.52, 0.58, and 0.50 which implies that the estimated poses are close to motion-

	Method	Specificity	Sensitivity			All Seqs	
		Normal	Stroke	Limp	Parkinson	Precision	Recall
View 1	Crabbe et al. [22]	0.53	**0.70**	**0.63**	**0.66**	0.82	**0.65**
	Proposed Method	**0.61**	0.64	0.55	0.51	**0.83**	0.55
View 2	Crabbe et al. [22]	0.52	**0.67**	**0.64**	**0.66**	0.83	0.65
	Proposed Method	**0.62**	0.62	0.55	0.57	**0.84**	0.58
View 3	Crabbe et al. [22]	0.58	**0.66**	**0.62**	**0.61**	**0.83**	**0.63**
	Proposed Method	**0.62**	0.63	0.53	0.55	**0.83**	0.56
View 4	Crabbe et al. [22]	0.50	**0.65**	**0.62**	**0.62**	0.81	**0.63**
	Proposed Method	**0.63**	0.63	0.58	0.55	**0.84**	0.58

Table 4.6: *Performance of human movement assessment on SMAD for normal and abnormal sequences.*

captured data. Crabbe et al. [22] tends to yield more abnormal pose outcomes than the proposed method, in line with the results of previous experiments (Tables 4.2, 4.3, and 4.4). This may contribute to explaining its poorer classification results on normal sequences in Table 4.6 and its better results on the abnormal sequences.

The percentage of the frames that are classified as normal by the pose and dynamics models of the movement assessment approach [98] is shown in Table 4.7. It can be seen that the movement analysis modelling mostly finds pose to be normal, while the dynamics is particularly abnormal in all abnormal sequences. This is in line with the scenarios where all three abnormality types mostly imply abnormal dynamics with relatively normal poses.

Note that from Table 4.6, the overall performance of the movement quality assessment method compared to the previous uses of it in [22, 98, 137] is lower. This may not necessarily indicate a poor performance of the pose estimation, but rather be due in large part to the method being designed for modelling and assessing the quality of single movements, *i.e.* walking action. However, in this chapter a more complex action made up of two distinct basic movements, walking and turning, is considered.

		Normal	Stroke	Limp	Parkinson
View 1	Crabbe et al. [22]	80% / 62%	72% / 31%	80% / 38%	85% / 38%
	Proposed Method	86% / 65%	76% / 37%	85% / 49%	87% / 51%
View 2	Crabbe et al. [22]	81% / 58%	74% / 34%	82% / 37%	83% / 34%
	Proposed Method	82% / 64%	77% / 41%	83% / 47%	84% / 44%
View 3	Crabbe et al. [22]	83% / 66%	71% / 36%	79% / 39%	81% / 39%
	Proposed Method	86% / 67%	75% / 38%	83% / 51%	85% / 49%
View 4	Crabbe et al. [22]	83% / 56%	76% / 36%	80% / 40%	81% / 40%
	Proposed Method	88% / 67%	80% / 40%	84% / 44%	85% / 46%

Table 4.7: *Percentage of the frames classified as normal by the pose/dynamics models on SMAD.*

4.4 Conclusion

This chapter introduced a pose estimation method that obtains high-level 3D pose features from *RGB images* for human movement assessment suitable for *multi-view learning scenarios*. The proposed approach also facilitates the movement analysis process by removing the skeleton-based pre-processing steps, *e.g.* normalization and dimensionality reduction. However, the current method has some limitations which are outlined below.

Requiring Motion Capture Data – Although this chapter proposes an approach that facilitates human movement assessment using RGB images alone during the testing stage, it relies on 3D motion capture data for training. Obtaining 3D data through motion capture systems is expensive and time-consuming since these systems require specialist hardware, software, and environmental setups. Chapters 5 and 6 will explore the human movement assessment approaches that are trained and tested only on RGB images.

Human Movement Assessment on Unseen View data – This chapter tackles the problem of viewpoint variations by introducing a method that can be trained on multiple views while maintaining a high performance on the single-views. Its complexity also does not increase with the number of viewpoints. However, the proposed method

fails on unseen view data since it has not been designed explicitly to address the changes in the viewpoints which are not presented in the training data. Chapters 5 and 6 will develop the methods that also perform well on arbitrary unseen view data while they are trained on only one or two viewpoints.

Chapter 5

Unseen View Human Movement Assessment

Chapter 4 introduced an approach to extract high-level 3D human pose in a low dimensional manifold space from RGB images towards human movement assessment in multi-view learning scenarios. Results showed that the proposed method outperforms the adapted baseline, and it can maintain a high accuracy on single view data when multiple viewpoints are employed during the training. However, the proposed method has a poor performance on unseen view data. In addition, although the proposed method facilitates the testing process through RGB images alone, it still requires 3D motion capture data for training.

This chapter instead proposes a movement quality assessment approach that is able to analyse human movement from *unseen* or *novel viewpoints*. Furthermore, the proposed method *does not require* any knowledge about *camera parameters* and works based on *RGB* sequences *alone* during both training and testing. The work in this chapter has been published in [121].

Section 5.1 discusses unseen view human movement assessment. The proposed method to assess the quality of human movement from unseen views is described in Section 5.2. Then, in Section 5.3, experiments are conducted using QMAR and KIMORE [12]. Finally, conclusions are presented in Section 5.4.

5.1 Assessing Quality of Human Movement from Unseen View Data

To tackle view-invariance in human movement assessment, one solution is to train a network on data from multiple views, *e.g.* the approach proposed in Chapter 4. However, in practice, capturing a labelled dataset of different views is cumbersome and rare. Ideally, a wholly view-invariant approach should be trained on data from as few views as possible and be able to perform well on a single (unseen) view at inference time.

To the best of the author's knowledge, there is no unknown view-invariant human movement analysis method, but in the action recognition domain, most works, such as [60, 88, 169, 170], employ 3D data (*e.g.* 3D skeleton data and depth images) to extract view-invariant features to achieve some degree of invariance (see Chapter 2 for further detail). However, as discussed in Section 4.1, acquiring 3D pose data is challenging in in-the-wild scenarios. Other works, such as [87, 141], deal with this issue by deploying multi-view synthetic videos to train their networks to perform action recognition given novel viewpoints (see Chapter 2 for more detail). Although in these approaches, the action classification task is performed from RGB data, they still use 3D pose annotations to produce the synthetic data, and the newly generated videos have to be also labelled by experts if they were to be used for specialist applications, such as healthcare (see Chapter 2).

This chapter proposes an end-to-end View-Invariant Network (VI-Net) that assesses the quality of human movement from a sequence of body joint heatmaps generated from RGB data, and argues that using temporal pose information learned from 2D RGB images, can be repurposed, instead of 3D data, for view-invariant movement quality assessment. To achieve this, the proposed VI-Net first attempts to extract view-invariant (canonical) spatio-temporal trajectory descriptors for all body joints, and then exploits the relationship amongst the joint trajectories to estimate a score for the quality of movement.

Recent works, such as [73, 78, 81, 144], provide unseen view results only when their network is trained on multiple views, while as recently noted by Varol et al. [141], a highly challenging scenario in view-invariant action understanding would be to obtain unseen view results by training from only one viewpoint. Not only does this chapter present unseen view results of VI-Net using a prudent set of two viewpoints only within a multi-view training scenario, but also it rises to the challenge to provide unseen view results by training solely from a single viewpoint.

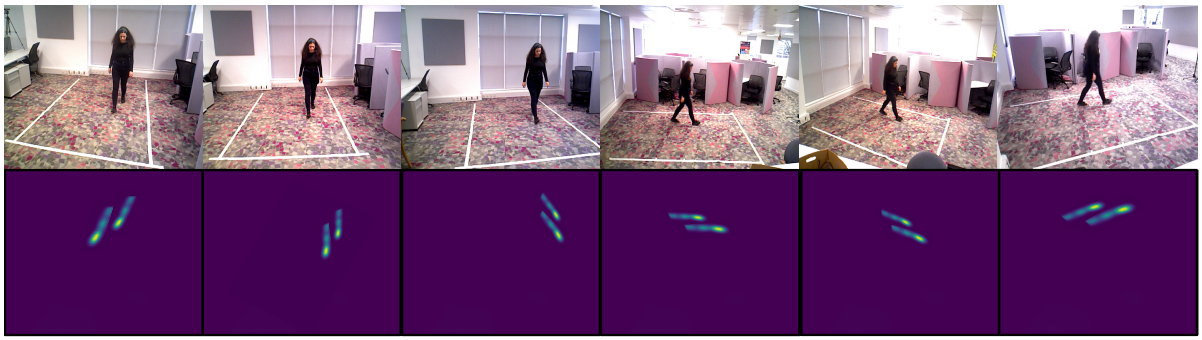


Figure 5.1: *Sample images of a person walking in all six views, and the corresponding trajectory maps of her feet for each view.*

5.2 VI-Net: View-Invariant Network to Assess the Quality of Human Movement

Although its appearance changes significantly when we observe an instance of human movement from different viewpoints, the 2D spatio-temporal trajectories generated by body joints in a sequence can be assumed affine transformations of each other. For example, see Figure 5.1, where the trajectory maps of just the feet joints appear different in orientation, spatial location and scale. This chapter proposes a view-invariant model by relying on this hypothesis that by extracting body joint trajectory maps that are translation, rotation, and scale invariant, the quality of human movement can be assessed from arbitrary viewpoints one may encounter in-the-wild.

The proposed end-to-end View-Invariant Network (VI-Net) has a view-invariant trajectory descriptor module (VTDM) that feeds into a subsequent movement score module (MSM) as shown in Figure 5.2. The input of the network is a sequence of human body joint heatmaps generated from RGB images. The aim of VTDM is to generate a view-invariant (canonical) spatio-temporal trajectory descriptor map for each body joint where later the canonical descriptors from all body joints are stacked and fed as input into MSM. The MSM module learns to obtain a score for the overall quality of movement by exploiting the relationship amongst the joint trajectories.

Details of how VTDM produces the view-invariant body joint trajectories are given in Section 5.2.1. Section 5.2.2 describes how the trajectories are applied by MSM to assess the quality of human movement. Finally, Section 5.2.3 explains the training and testing process of VI-Net.

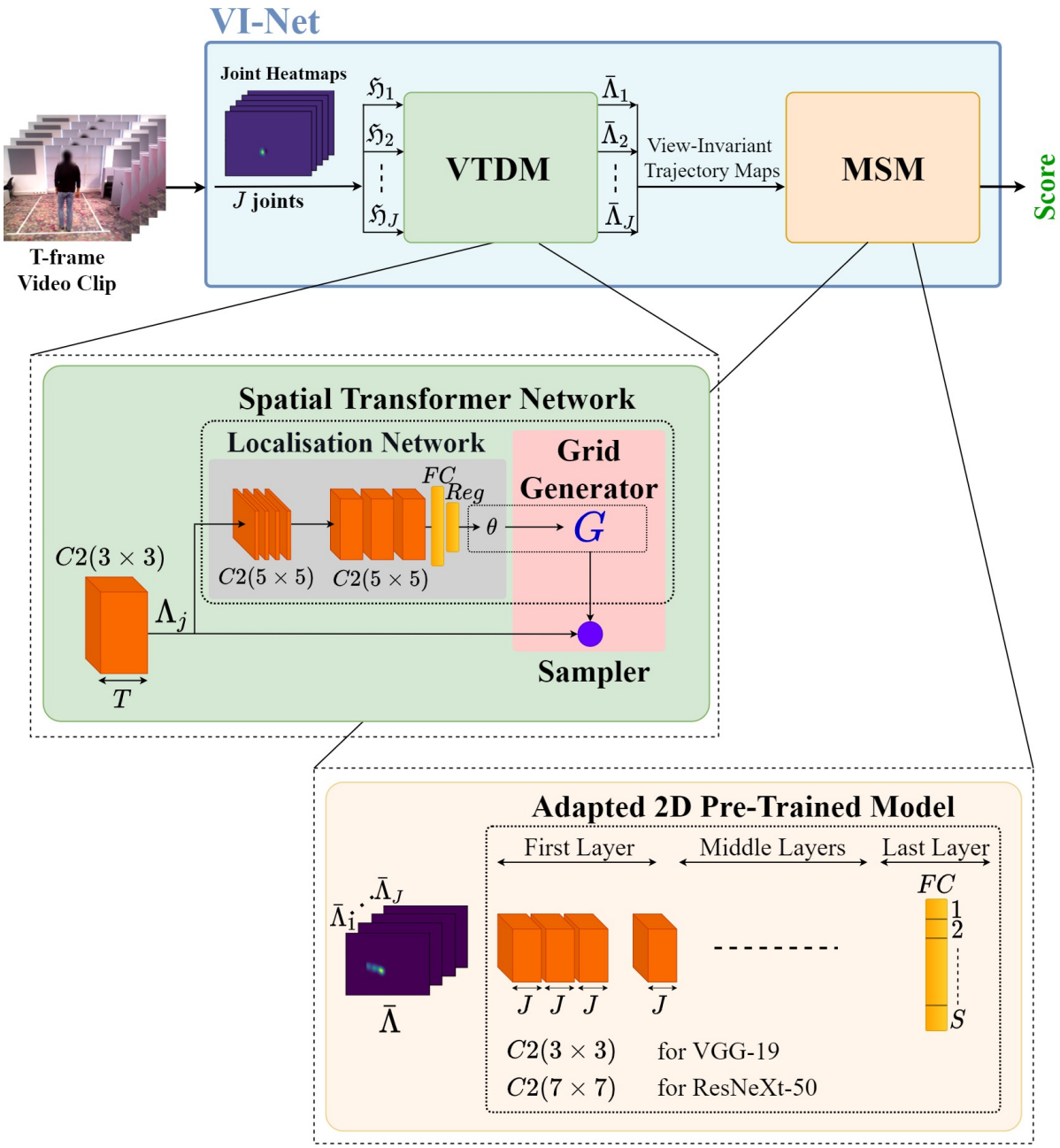


Figure 5.2: *The overall schema of VI-Net. It has a view-invariant trajectory descriptor module (VTDM) and a movement score module (MSM) where the classifier output corresponds to a quality score.*

5.2.1 VTDM: View-Invariant Trajectory Descriptor Module

In VTDM, first a 2D convolution filter is applied to stacked heatmaps of each body joint over the video clip frames to aggregate the spatial poses over time and generate a trajectory descriptor map per body joint. Then, the Spatial Transformer Network

(STN) [56] is applied to the trajectory descriptor to make it view-invariant. This process is detailed next.

Body Joint Heatmap – Given an RGB image I^t at timestamp t, body joint heatmap $H_j^t \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}}$ represents the probability of joint j occurring at each pixel position $x \in \mathbb{R}^2$ of I^t as

$$H_j^t(x) = exp(-\frac{\|x - x^{\oplus}\|_2^2}{\sigma^2}) \ , \tag{5.1}$$

where $x^{\oplus} \in \mathbb{R}^2$ is the ground-truth position of joint j, and σ determines the spread of the peak.

Generating a Joint Trajectory Descriptor – For each body joint j where $j \in \{1, 2, \ldots, J\}$, its heatmaps over the T-frame video clip $\{H_j^t\}_{t=1}^T$, are stacked to get the 3D heatmap $\mathfrak{H}_j \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times\mathrm{T}}$ which then becomes the input to the VTDM module. To obtain a body joint's trajectory descriptor Λ_j, the processing in VTDM starts with the application of a convolution filter Φ on $\mathfrak{H}_j$ to aggregate its spatial poses over time, *i.e.*

$$\Lambda_j = \mathfrak{H}_j * \Phi \ , \tag{5.2}$$

where $\Lambda_j \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 1}$.

Note, to implement this part of the network, both 2D and 3D convolutions were experimented with, and it was observed that a 3×3 2D convolution filter yields the best results.

Forging a View-Invariant Trajectory Descriptor – In the next step of the VTDM module, the Spatial Transformer Network (STN) [56] is deployed to forge a view-invariant trajectory descriptor out of Λ_j. STN can be applied to feature maps of CNN's layers as normalization to make them translation, rotation, scale, and shear invariant. Note, to make the trajectory descriptor view-invariant, STN [56], DCN [23, 177], and ETN [134] networks were investigated, and it was observed that STN obtains the best performance.

The STN network [56] is composed of three stages. At first, a CNN-regression network, referred to as the localisation network, is applied to the joint trajectory descriptor Λ_j to estimate the parameters for a 2D affine transformation matrix θ,

$$\theta = f_{loc}(\Lambda_j). \tag{5.3}$$

Then, in the second stage, to estimate each pixel value of the view-invariant trajectory descriptor $\bar{\Lambda}_j$, a sampling kernel is applied to specific regions of Λ_j, where the centres of these regions are defined on a sampling grid. This sampling grid $\Gamma_\theta(G)$ is generated from a general grid $G = \{(x_i^g, y_i^g)\}, i \in \{1, \ldots, W^{'} \times H^{'}\}$ and the predicted transformation parameters, such that

$$\begin{pmatrix} x_i^{\Lambda_j} \\ y_i^{\Lambda_j} \end{pmatrix} = \Gamma_\theta(G_i) = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix} \times \begin{pmatrix} x_i^g \\ y_i^g \end{pmatrix} , \tag{5.4}$$

where $\Gamma_\theta(G) = \{(x_i^{\Lambda_j}, y_i^{\Lambda_j}), i \in \{1, \ldots, W^{'} \times H^{'}\}$ are the centers of the regions of Λ_j the sampling kernel is applied to, in order to generate the new pixel values of the output feature map $\bar{\Lambda}_j$.

Jaderberg et al. [56] recommend the use of different types of transformations to generate the sampling grid $\Gamma_\theta(G)$ based on the problem domain. In VTDM, the 2D affine transformations shown in Equation 5.4 is applied. Finally, the sampler takes both Λ_j and $\Gamma_\theta(G)$ to generate a view-invariant trajectory descriptor $\bar{\Lambda}_j$ from Λ_j at the grid points by bilinear interpolation.

5.2.2 MSM: Movement Score Module

In the final part of VI-Net (see Figure 5.2-MSM), the collection of view-invariant trajectory descriptors $\bar{\Lambda}_j$ for joints $j \in \{1, 2, \ldots, J\}$, are stacked into a global descriptor $\bar{\Lambda}$ and passed through a pre-trained 2D CNN network (*e.g.* VGG and ResNeXt) in the MSM module to assess the quality of movement of the joints.

The pre-trained network needs to be adapted such that its first layer is replaced with a new 2D convolutional layer for which its kernel size remains unchanged, while its channel size is changed to J (instead of 3 used for RGB input images). The last fully connected (FC) layer is also modified to allow movement quality scoring through classification where each score is considered as a class, *i.e.* for a movement type with S possible scores, the last FC layer of VI-Net has S output units.

5.2.3 VI-Net Training and Testing

The proposed network is trained from scratch for each movement type (*e.g.* W-P), and in both the training and testing phases, video sequences ($V = \{V_n\}_{n=1}^{N}$) are divided into T-frame clips (without overlaps), $V_n = \{C_n^m\}_{m=1}^{M}$ where $C_n^m \in \mathbb{R}^{\mathrm{W} \times \mathrm{H} \times \mathrm{T}}$, and $length(V_n) = M.T$.

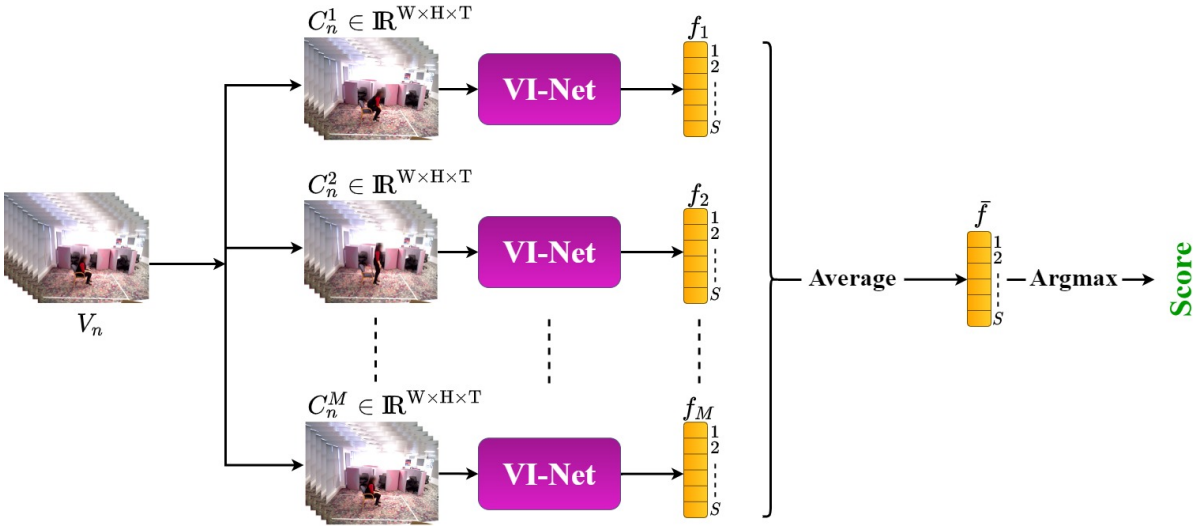


Figure 5.3: *Scoring process of VI-Net for a full video sequence in testing phase.*

Training – In the training stage, the clips are selected randomly from amongst all video sequences of the training set, and passed to VI-Net. Then, the weights were updated following a cross-entropy loss,

$$L_{C_n^m}(f, s) = -log(\frac{exp(f(s))}{\sum_{k=0}^{S} exp(f(k))}) \ , \tag{5.5}$$

where for given C_n^m as input, $f(.)$ is the S dimensional output of the last fully connected layer of VI-Net and s is the video clip's ground truth label/score.

Testing – In the testing phase, every T-frame clip of a video sequence is passed to VI-Net. After averaging the outputs of the last fully connected layer across each class for all the clips, then the score for the whole video sequence is set as the maximum of the clip scores (see Figure 5.3), that is,

$$s = \underset{k}{\text{argmax}}(\bar{f}(k) = \frac{1}{M}\sum_{m=1}^{M} f_m(k)) \ , \tag{5.6}$$

where $k \in \{1, 2, \dots S\}$ and M is the number of clips of the video.

5.3 Experiments and Results

This section first reports on two sets of experiments on QMAR to evaluate the performance of VI-Net to assess quality of movement, based around cross-subject and cross-view protocols. Then, to show the efficiency of VI-Net on other datasets and movement types, it also presents the results of VI-Net on the single-view KIMORE dataset.

Comparison to Baselines – As there is no prior work on view-invariant human movement assessment, the performance of the VI-Net network is evaluated against[1] (i) a C3D baseline (fashioned after Parmar and Morris [104]) by combining the outputs of the C3D network to score a sequence in the test phase in the same fashion as VI-Net, and (ii) the pre-trained, fine-tuned I3D [13].

Ablation Study – This section also provides an ablation study for all scenarios by removing STN from VI-Net to analyse the impact of this part of the proposed method, and it also studies the effect of different pre-trained models applied in the MSM module of VI-Net.

Implementation details and the applied evaluation metric are explained in Sections 5.3.1 and 5.3.2. Then, Sections 5.3.3, 5.3.4, and 5.3.5 compare cross-subject, cross-view, and single-view results of the proposed method to baselines respectively, while they also outline the ablation studies and qualitative results of VI-Net.

5.3.1 Implementation Details

Details of Network Architecture – Table 5.1 shows details of the proposed VI-Net network architecture. To design the localization network in VTDM module, instead of the original CNN in [56], which applied two 32-filter 5×5 convolutional layers followed by two FC layers, the localisation network is made up of two 10-filter 5×5 convolutional layers followed by two FC layers. The rationale for this is that the proposed trajectory descriptor maps are not as complex as RGB images, and hence fewer filters are sufficient to extract their features. The flexibility of MSM is illustrated by implementing two different pre-trained networks[2], VGG-19 [129] and ResNeXt-50 [157], and their results will be compared in the next sections. VGG-19 and ResNeXt-50 were chosen for their state-of-the-art performances, popularity, and availability.

Training and Testing Details – Similar to [13, 102, 104], the inputs to the model are 16-frame video clips with size 128 × 128 pixels. Human body joint heatmaps are extracted by applying OpenPose[11][3]. To reduce computational complexity, the first 15 joint heatmaps of the BODY-25 version of OpenPose are retained. This is further motivated by the fact, highlighted in [98], that the remaining joints only provide repeti-

[1] It should be noted that between 2019 and 2020 when VI-Net was proposed, the approach proposed by Parmar and Morris [104] was state-of-the-art for human movement assessment, and I3D [13] was state-of-the-art for action recognition.

[2] VGG-19 [129] and ResNeXt [157] are pre-trained on ImageNet [25].

[3] Other methods (*e.g.* [69]) which estimate body joint heatmaps from RGB images can equally be used.

	VTDM	MSM (Adapted VGG-19 or ResNeXt-50)
VI-Net	**First layer:** $\{C2(3\times3,T)\}\times1$, BN, ReLU **Localisation Network:** $\{C2(5\times5,1)\}\times10$,$\{MP(2\times2)\}$, ReLU, $\{C2(5\times5,10)\}\times10$, $\{MP(2\times2)\}$, ReLU, $\{FC(32)\}$, ReLU, $\{FC(4)\}$	**First layer VGG-19**: $\{C2(3\times3,J)\}\times64$, BN, ReLU **First layer ResNeXt-50:** $\{C2(7\times7,J)\}\times64$, $\{MP(3\times3)\}$, ReLU **Middle layers:** As in VGG-19/ResNeXt-50 **Last layer:** $\{FC(S)\}$

Table 5.1: *Details of VI-Net's modules:* $\{C2(d\times d,ch)\}\times n$*: n 2D convolution filters with size d and ch channel size,* $MP(d\times d)$*: 2D max pooling with size d,* $FC(N)$*: FC layer with N outputs. T is the number of clip frames, J is the number of joints and S is the number of possible scores for a movement type.*

tive information. All models in this section were trained for 20 epochs using mini-batch stochastic gradient descent. The initial learning rate was set to 0.001 and was decayed by a factor of 10 every 5 epochs. The momentum and batch size were set to 0.9, and 5 respectively.

Dataset Imbalance – It can be seen from Tables 3.3 and 3.4 that the number of sequences for score 0 (normal) is many more than the number of sequences for other individual scores, so 15 normal sequences for W-P, W-S, SS-S movements and 4 normal sequences for SS-P were randomly selected to mix with abnormal movements to perform all the experiments. To further address the imbalance, offline temporal cropping was applied to add new sequences.

5.3.2 Evaluation Metric

To evaluate the performance of the proposed method, similar to other action assessment approaches, such as [82, 99, 104], Spearman's rank correlation (SRC) was used to measure the relationship between estimated and ground-truth scores. The SRC gives a value between 1 and -1, with 1 indicating a perfect positive correlation, -1 showing a perfect negative correlation, and 0 representing no correlation.

5.3.3 Cross-Subject Human Movement Assessment

This section provides cross-subject results on QMAR where all available views are used in both training and testing, while the subjects performing the movements are distinct.

For this scenario, as size of QMAR is small, the results are obtained by applying k-fold cross validation where k is the number of scores for each movement type. Note, following state-of-the-art action quality assessment approaches, such as [31, 32], that report only average results for cross-valuation without reporting standard deviation and/or p-values, only average results are provided for this section.

Comparison to Baselines – Table 5.2 shows SRC results for VI-Net and baselines for each movement type. The VI-Net network outperforms networks based on C3D (after [104]) and I3D [13] for all types of movements, regardless of whether VGG-19 or ResNeXt-50 are used in the MSM module. While I3D results are mostly competitive, C3D performs less well due to its shallower nature, and larger number of parameters, exacerbated by QMAR's relatively small size. Section 5.3.5 will show that C3D performs significantly better on a larger dataset.

Ablation Study – To test the effectiveness of STN, VI-Net's results are presented with and without engaging STN in Table 5.2. It can be observed that the improvements with STN are not necessarily consistent across the actions since when all viewpoints are used in training, the MSM module gets trained on all trajectory orientations such that the effect of STN is often overridden. Table 5.2 also shows that on average VI-Net performs better with adapted ResNeXt-50.

Method			Training	Action (SRC)				Average
				W-P	W-S	SS-P	SS-S	(SRC)
C3D (after [104])			scratch	0.50	0.37	0.25	0.54	0.41
I3D[13]			fine-tune	0.79	0.47	0.54	0.55	0.58
VI-Net	VTDM+MSM (VGG-19)	w/o STN	scratch	0.81	0.49	<u>0.57</u>	**0.74**	<u>0.65</u>
		w STN	scratch	<u>0.82</u>	<u>0.52</u>	0.55	<u>0.73</u>	<u>0.65</u>
	VTDM+MSM (ResNeXt-50)	w/o STN	scratch	**0.87**	**0.56**	0.48	0.72	<u>0.65</u>
		w STN	scratch	**0.87**	<u>0.52</u>	**0.58**	0.69	**0.66**

Table 5.2: *SRC between predicted scores and ground truth labels for cross-subject analysis on different actions of QMAR. I3D was pretrained on Kinetic-400 [66]. The best and the second-best results are in* ***Bold*** *and* <u>underline</u> *respectively.*

Qualitative Results – Figure 5.4 shows output scores of VI-Net on some sample of QMAR for cross-subject scenario where VI-Net contained STN, and ResNeXt-50 was used to design the MSM module.

Figure 5.4: *Example scores estimated by VI-Net on QMAR under cross-subject protocol for all four movement types. First column: walking with Parkinsons (W-P). Second column: walking with Stroke (W-S). Third column: sit-stand with Parkinsons (SS-P). Fourth column: sit-stand with Stroke. Each row shows a distinct viewpoint.*

5.3.4 Cross-View Human Movement Assessment

The generalization ability of VI-Net on unseen views is evaluated by using the cross-view protocol, that is, distinct training and testing views of the scene, while data from all subjects is utilised. The experiments of this section are performed under the assumption that each test set contains a balanced variety of scores from low to high (see Section 5.3.1 on the data imbalance issue). Note in this scenario, each set of experiments only has one possible training and one testing set, as such no mean, standard deviation, or p-value is computed.

Recent view-invariant action recognition approaches, such as [60, 88, 143, 169, 170], provide cross-view results only when their network is trained on multiple views. However, this chapter rises to the challenge introduced by Varol et al. [141] and also provides unseen view results *by training solely from a single viewpoint*. Therefore, the training and testing for each movement type are performed such that (i) only one view was used for training and all other views were applied for testing, and in the next experiment, (ii) a combination of one frontal view (views 1 to 3) and one side view (views 4 to 6) were used for training and all other available views were applied for testing. Since for the latter case there are many combinations of views in QMAR, results for only selected views are obtained: view 2 $\approx 0^\circ$ with all side views, and view 5 $\approx 90^\circ$ with all frontal views.

5.3.4.1 Cross-View Results by Training from One Viewpoint

Table 5.3 shows the results of VI-Net, C3D baseline (after [104]), and pre-trained, fine-tuned I3D [13] for each movement type of QMAR when only one view is used for training.

Comparison to Baselines and Ablation Study – The results show that when VI-Net has STN with adapted ResNeXt, it performs best on average, and outperforms the baselines on average rank correlation, at 0.70, 0.62, 0.39, and 0.43 for W-P, W-S, SS-P and SS-S action types respectively.

From Table 5.3, it can also be observed that for walking movements W-P and W-S, VI-Net is able to assess the movements from unseen views well, with the best results reaching 0.73 and 0.66 rank correlation respectively (green highlights), and only relatively affected by short term occlusions. However, for sit-to-stand movements SS-P and SS-S, the long-term occlusions during these movements affect the integrity of the trajectory descriptors and the performance of VI-Net is not as strong, with the best results reaching 0.52 and 0.56 respectively (green highlights).

Action	Training View	VTDM+MSM (VGG-19)		VTDM+MSM (ResNeXt-50)		C3D (after [104])	I3D [13]
		w/o STN	w STN	w/o STN	w STN		
W-P	1	0.51	**0.67**	0.64	**0.67**	0.18	0.60
	2	0.69	0.66	0.58	**0.72**	0.18	0.61
	3	0.62	0.66	0.63	**0.70**	0.21	0.62
	4	0.67	0.64	**0.72**	**0.72**	0.23	0.50
	5	0.67	0.67	0.68	**0.71**	0.24	0.57
	6	0.69	0.72	0.69	**0.73**	0.21	0.55
	Average	0.64	0.67	0.65	**0.70**	0.20	0.60
W-S	1	0.51	0.43	0.60	**0.64**	0.14	0.49
	2	0.47	0.54	0.55	**0.62**	0.10	0.44
	3	**0.64**	0.56	0.61	0.59	0.23	0.52
	4	0.60	0.59	0.60	**0.66**	0.20	0.54
	5	0.62	0.60	0.62	**0.63**	0.17	0.45
	6	0.46	0.40	0.53	**0.60**	0.17	0.54
	Average	0.55	0.52	0.58	**0.62**	0.16	0.53
SS-P	1	0.30	**0.32**	0.25	0.25	0.10	0.18
	2	0.27	0.31	0.31	**0.32**	0.10	0.21
	3	0.16	0.23	0.36	**0.43**	0.12	0.25
	4	0.10	0.34	0.44	**0.49**	0.17	0.20
	5	0.50	**0.52**	0.43	0.45	0.12	0.37
	6	0.41	0.24	**0.48**	0.44	0.09	0.18
	Average	0.29	0.32	0.37	**0.39**	0.11	0.23
SS-S	1	0.36	**0.49**	0.44	0.45	0.26	0.43
	2	0.47	0.40	**0.56**	**0.56**	0.30	0.49
	3	0.37	**0.52**	0.38	0.43	0.25	0.40
	4	0.38	0.34	0.41	**0.54**	0.32	0.50
	5	0.26	**0.50**	**0.50**	0.48	0.20	0.46
	6	0.21	**0.28**	0.13	0.16	0.18	0.20
	Average	0.34	0.42	0.40	**0.43**	0.25	0.42

Table 5.3: *SRC between predicted scores and ground truth labels for cross-view analysis where only one view is used for training on QMAR. I3D was pretrained on Kinetic-400 [66]. The **Bold** and underline numbers show the best and the second-best results for each view of each action type respectively. The green highlights indicate best results for each action type amongst all views.*

Qualitative Results – Figure 5.5 shows output scores of VI-Net on some unseen viewpoints and for all movement types of QMAR where only one viewpoint was used for training. Note, to obtain these results VI-Net contained STN, and ResNeXt-50 was used to design the MSM module.

Figure 5.5: *Example scores estimated by VI-Net on unseen views of QMAR for all four movement types, where one viewpoint was employed for training. First column: walking with Parkinsons (W-P). Second column: walking with Stroke (W-S). Third column: sit-stand with Parkinsons (SS-P). Fourth column: sit-stand with Stroke. The top row shows the viewpoint used during training, and the rest of the rows show VI-Net's results on simultaneous frames on novel viewpoints at inference.*

5.3.4.2 Cross-View Results by Training from Two Viewpoints

Table 5.4 reports cross-view results when one side view and one frontal view are combined for training.

Comparison to Baselines – It can be seen from Table 5.4 the VI-Net network performs better than the baselines on average rank correlation at 0.83, 0.80, 0.54, and 0.59 for W-P, W-S, SS-P and SS-S action types respectively, and VI-Net's performance improves compared to the single-view experiment in Table 5.3 with the best results reaching 0.92 and 0.83 for W-P and W-S movements (green highlights) and 0.61 and 0.67 for SS-P and SS-S movements (green highlights), because the network is effectively trained with both short-term and long-term occluded trajectory descriptors.

Ablation Study – The results also show that on average VI-Net performs better with adapted ResNeXt-50 for walking movements (W-P and W-S) and with adapted VGG-19 for sit-to-stand movements (SS-P and SS-S). This is potentially because ResNext-50's variety of filter sizes are better suited to the variation in 3D spatial changes of joint trajectories inherent in walking movements compared to VGG-19's 3×3 filters which can tune better to the more spatially restricted sit-to-stand movements.

It should also be noted that the fundamental purpose of STN in VI-Net is to ensure efficient cross-view performance is possible when the network is trained from a single-view only. It would therefore be expected and plausible that STN's effect would diminish as more views are used since the MSM module gets trained on more trajectory orientations (which it was verified experimentally by training with multiple views in Table 5.4).

Qualitative Results – Figure 5.6 shows example output scores of VI-Net on unseen views of QMAR for four movement types, W-P, W-S, SS-P and SS-S, where two views (one frontal view and one side view) were used for training, and VI-Net has STN, and ResNeXt-50 was applied to implement the MSM module.

5.3.5 Single-View Human Movement Assessment

To examine the performance of VI-Net on an independent publicly available dataset, the KIMORE dataset is considered as the best possible candidate. To perform the experiments, the network was trained to predict a final score for each action type, and as the size of KIMORE was large enough, only one training and one testing set were selected randomly, *i.e.* 70% of the subjects were used for training and the remaining 30% were applied for testing ensuring each set contains a balanced variety of scores from low to high.

Action	Training Views	VTDM+MSM (VGG-19)		VTDM+MSM (ResNeXt-50)		C3D (after [104])	I3D [13]
		w/o STN	w STN	w/o STN	w STN		
W-P	2, 4	0.77	0.81	0.87	**0.89**	0.65	0.85
	2, 5	0.72	0.75	0.90	**0.92**	0.65	0.87
	2, 6	0.75	0.76	0.73	0.77	0.69	**0.80**
	1, 5	0.70	0.76	**0.80**	0.75	0.63	0.79
	3, 5	0.73	0.79	**0.87**	0.84	0.57	0.80
	Average	0.73	0.77	**0.83**	**0.83**	0.63	0.82
W-S	2, 4	0.58	0.72	**0.81**	0.73	0.42	0.76
	2, 5	0.74	0.74	0.80	**0.81**	0.37	0.71
	2, 6	0.64	0.67	**0.74**	0.68	0.33	0.73
	1, 5	0.70	0.68	**0.83**	0.81	0.48	0.71
	3, 5	0.66	0.66	**0.82**	0.79	0.45	0.70
	Average	0.66	0.69	**0.80**	0.76	0.41	0.72
SS-P	2, 4	**0.55**	0.52	0.41	0.46	0.25	0.48
	2, 5	**0.60**	0.53	0.49	0.46	0.21	0.40
	2, 6	**0.48**	0.35	0.36	0.42	0.30	0.47
	1, 5	0.46	**0.55**	039	0.52	0.38	0.45
	3, 5	**0.61**	0.40	0.43	0.47	0.37	0.39
	Average	**0.54**	0.47	0.41	0.46	0.30	0.43
SS-S	2, 4	0.57	**0.64**	0.52	**0.64**	0.53	0.55
	2, 5	0.62	0.56	**0.63**	0.61	0.45	0.60
	2, 6	0.50	**0.62**	0.48	0.46	0.44	0.54
	1, 5	**0.64**	0.53	0.48	0.58	0.30	0.62
	3, 5	0.62	0.60	0.63	**0.67**	0.35	0.65
	Average	**0.59**	**0.59**	0.55	0.58	0.41	0.58

Table 5.4: *SRC between predicted scores and ground truth labels for cross-view analysis where two views are used for training on QMAR. I3D was pretrained on Kinetic-400 [66]. The **Bold** and underline numbers show the best and the second-best results for each combination of views of each action type respectively. The green highlights indicate best results for each action type amongst all view combinations.*

Figure 5.6: *Example scores estimated by VI-Net on unseen views of QMAR for all four movement types, where two viewpoints were employed for training. First column: walking with Parkinsons (W-P). Second column: walking with Stroke (W-S). Third column: sit-stand with Parkinsons (SS-P). Fourth column: sit-stand with Stroke. The top rows show the viewpoints used during training, and the rest of the rows show VI-Net's results on simultaneous frames on novel viewpoints at inference.*

Comparison to Baselines and Ablation Study – Table 5.5 shows the results of C3D baseline (after [104]), pre-trained, fine-tuned I3D [13] and VI-Net on KIMORE. It can be observed that VI-Net outperforms the other methods for all movement types except for Exercise #3. VI-Net with adapted VGG-19 performs better than with ResNeXt-50 for all movement types. This may be because, similar to sit-to-stand movements in QMAR where VI-Net performs better with VGG-19, all movements types in KIMORE are also performed at the same location and distance from camera, and thus carry less variation in 3D trajectory space. This shows that VI-Net's results are consistent in this sense across both datasets.

In addition, although all sequences in both training and testing sets have been captured from the same view, VI-Net's performance on average improves with STN. This can be attributed to STN improving the network generalization on different subjects. Also, unlike in QMAR's cross-subject results where C3D performed poorly, the results on KIMORE for C3D are promising because KIMORE has more data to help the network train more efficiently.

Qualitative Results – Figure 5.7 illustrates example scores predicted by VI-Net for all movement types of KIMORE, where VI-Net has STN, and ResNeXt-50 was applied to implement the MSM module.

Method			Training	Action Ex (SRC)					Average
				#1	#2	#3	#4	#5	(SRC)
C3D (after [104])			scratch	<u>0.66</u>	<u>0.64</u>	**0.63**	0.59	0.60	0.62
I3D [13]			fine-tune	0.45	0.56	<u>0.57</u>	<u>0.64</u>	0.58	0.56
VI-Net	VTDM+MSM (VGG-19)	w/o STN	scratch	0.63	0.50	0.55	**0.80**	**0.76**	<u>0.64</u>
		w STN	scratch	**0.79**	**0.69**	<u>0.57</u>	0.59	<u>0.70</u>	**0.66**
	VTDM+MSM (ResNeXt-50)	w/o STN	scratch	0.55	0.42	0.33	0.62	0.57	<u>0.49</u>
		w STN	scratch	0.55	0.62	0.36	0.58	0.67	0.55

Table 5.5: *SRC between predicted scores and ground truth labels for different action types of the single-view KIMORE dataset. I3D was pretrained on Kinetic-400 [66]. The* ***Bold*** *and* <u>underline</u> *numbers show the best and the second-best results for each scenario of each action type respectively.*

Figure 5.7: *Example scores estimated by VI-Net on the single-view KIMORE dataset for all five movement types, Ex #1, Ex #2, Ex #3, Ex #4, and Ex #5.*

5.4 Conclusion

This chapter introduced a view-invariant human movement assessment approach and tackled a highly challenging scenario in this field, that is, assessing the quality of human movement from *novel viewpoints*, where the method is trained on *only one viewpoint*. As opposed to the proposed solution in Chapter 4 that requires 3D skeleton data for training, the proposed approach in this chapter applies only *RGB* data while it does not rely on any knowledge about camera viewpoints.

This chapter has evaluated the performance of the proposed approach on the multi-view QMAR and the single-view KIMORE datasets and demonstrated that the proposed method is applicable to multi-view (cross-subject) and unseen view (cross-view) scenarios, and it can work across different datasets and movement types (see Sections 5.3.3 to 5.3.5). It was also shown that the proposed method outperforms the baselines on average on all these scenarios. However, the proposed approach still has a few limitations that highlight potential avenues for future work.

Occlusion – The proposed method's performance drops in situations where long-term occlusions occur, since OpenPose [11] fails in such cases. Thus, future work could explore how to produce sufficiently consistent heatmaps when the occlusion happens.

Movement Quality Assessment Per Action Type – Another limitation of the proposed approach is that it needs to be trained separately for each action type (*e.g.* W-P). Future work could investigate to develop a multitask learning model such that the network can recognize the action type and estimate its score simultaneously.

Transfer Learning – In healthcare applications, *e.g.* rehabilitation monitoring at home or in the clinic, capturing and annotating data is challenging and expensive. On the other hand, there are several multi-view datasets in other domains, *e.g.* NTU [126] that is a large-scale multi-view dataset for action recognition that can benefit learning view-invariant features for human movement assessment. Chapter 6 takes this direction and proposes an unsupervised method to extract 3D view-invariant (canonical) human posture representation for unseen view downstream tasks, *e.g.* action recognition and human movement assessment, and shows the learned view-invariant features can also be transferred into a different domain.

Chapter 6

Unsupervised View-Invariant Human Movement Assessment

Chapters 4 and 5 presented *supervised* approaches to assess the quality of human movement. These methods have certain limitations that are addressed in this chapter:

- The method proposed in Chapter 4 estimates 3D human pose for human movement assessment. However, during training, the proposed approach requires not only 3D joint annotations, but also action labels since it works based on generating separate manifolds for distinct action types. Furthermore, it cannot be applied to novel-view data.
- The movement quality assessment method proposed in Chapter 5 is capable of analysing movements recorded from camera viewpoints that are not present in training data, and the proposed method does not require 3D skeleton annotation for training. However, as it assesses the movements by extracting spatio-temporal features that are derived from action types and their abnormality scores, the learned features through this method cannot easily be transferred amongst different action types, tasks or domains.

This chapter introduces an *unsupervised* 3D human posture representation approach for *unseen view downstream tasks*, *e.g.* movement assessment and action recognition. The proposed method learns to extract view-invariant 3D pose features from a 2D image without using 3D pose annotations and action type labels such that the learned representations can be *transferred* into *other domains*. The work in this chapter has been published in [122].

Section 6.1 discusses the need for an unsupervised view-invariant human posture repre-

sentation method. Sections 6.2 and 6.3 detail the components of the proposed approach and the temporal models that are applied to the unsupervised learned pose features for two downstream tasks, action recognition and human movement assessment. Section 6.4 conducts comparative experiments for cross-view and cross-subject action recognition on NTU RGB+D. It also shows the efficiency of transferring the learned representations from NTU RGB+D and action recognition to obtain unsupervised cross-view, cross-subject and single-view human movement assessment results on QMAR and KIMORE. Finally, conclusions are in Section 6.5.

6.1 Unsupervised View-Invariant Human Posture Representation

Most unsupervised RGB-based 3D pose estimation approaches, such as [16, 18, 34, 50, 115, 139], 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. In such cases, camera parameters are needed to map their view-specific output into a canonical view (Chapter 2 provides a full overview of these approaches). On the other hand, there are a few works, such as [116, 132, 174], that obtain view-invariant pose features from RGB data, but they are fully or weakly supervised and require 3D skeleton data during training.

This chapter tackles the above challenges, and proposes a representation learning approach that disentangles canonical (view-invariant) 3D pose representation and view-dependent features from either an RGB-based 2D Densepose human representation map or a depth mask image without using 3D skeleton data and camera parameters such that the learned view-invariant features can be applied *directly* by downstream tasks to be resilient to human pose variations in unseen viewpoints.

Fig. 6.1 shows the proposed view-invariant pose representation learning framework and its application on a view-invariant downstream task. The proposed network is an auto-encoder comprising two encoders and a decoder. The first encoder is a view-invariant 3D pose encoder that learns 3D canonical pose representations from an input image, and the second one is a viewpoint encoder that extracts rotation and translation parameters, such that when they are applied to the canonical pose features, it would result in view-dependent 3D pose representations which are fed into the decoder to reconstruct the input image.

To train the proposed network, geometrical and positional order consistency constraints

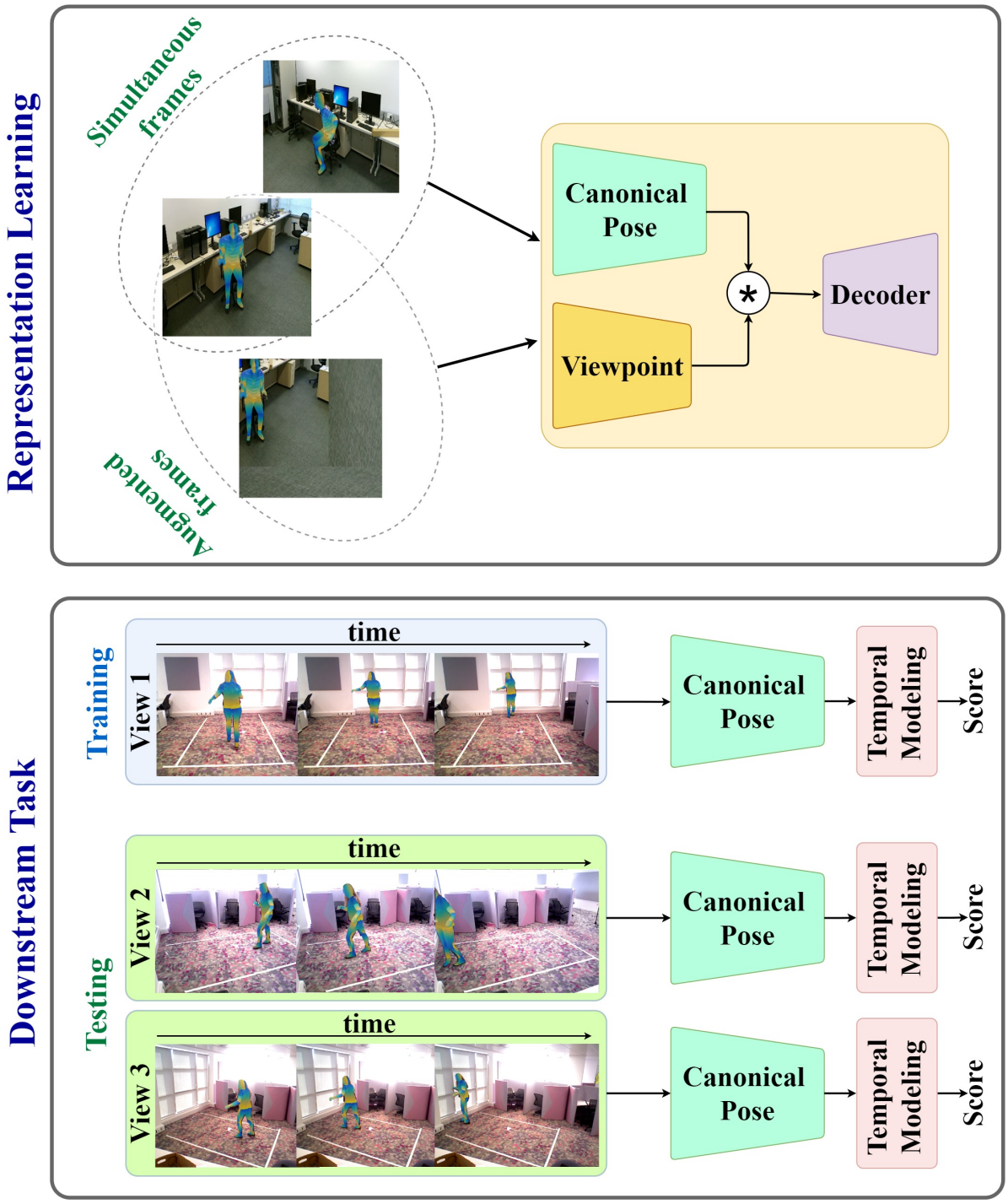


Figure 6.1: *Top: the proposed network learns to disentangle canonical 3D human pose representations and view-dependent features through simultaneous frames from different views and augmented frames from the same view. Bottom: the unsupervised learned canonical pose representation can be used for downstream tasks.*

are imposed 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 recognition and human movement assessment. The method is described in detail next.

6.2 Proposed Method

This section introduces the method that learns view-invariant 3D pose representation from 2D images (RGB or depth) *without* relying on *3D skeleton annotations* and *camera parameters*. The proposed method leverages on geometric transformation amongst different viewpoints and the equivariant property of human pose. The proposed method is an auto-encoder that includes a view-invariant pose encoder $E_{\odot}$, a viewpoint encoder $E_{\triangleleft}$, and a decoder D arranged as shown in Figure 6.2. $E_{\odot}$ learns 3D canonical pose features from a given image which can be either an RGB-based 2D Densepose human representation map [97] 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.

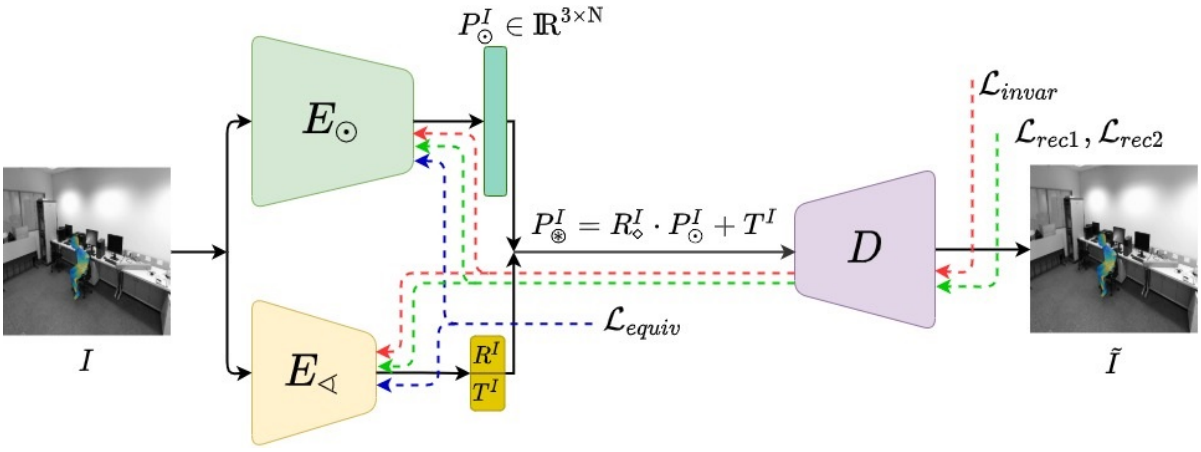


Figure 6.2: *The overall schema of the proposed view-invariant posture representation learning network.*

6.2.1 Model Architecture and Formulation

The view-invariant pose encoder $E_{\odot}$ learns 3D canonical pose features $P^I_{\odot} = E_{\odot}(I)$ given image $I \in \mathbb{R}^{3\times W\times H}$ where $P^I_{\odot} \in \mathbb{R}^{3\times N}$, 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 to the canonical pose features $P^I_{\odot}$ to transfer them into a specific viewpoint $P^I_{\circledast}$, such that $P^I_{\circledast} = R^I_{\diamond} \cdot P^I_{\odot} + T^I$ where $P^I_{\circledast} \in \mathbb{R}^{3\times N}$, and $R^I_{\diamond}$ is computed as

$$R^I_{\diamond} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta_x) & -sin(\theta_x) \\ 0 & sin(\theta_x) & cos(\theta_x) \end{bmatrix} \times \begin{bmatrix} cos(\theta_y) & 0 & sin(\theta_y) \\ 0 & 1 & 0 \\ -sin(\theta_y) & 0 & cos(\theta_y) \end{bmatrix} \times \begin{bmatrix} cos(\theta_z) & -sin(\theta_z) & 0 \\ sin(\theta_z & cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (6.1)$$

Then, decoder D 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 in Section 6.2.3) from the same viewpoint. The proposed network is trained by combining four losses, view-invariant $\mathcal{L}_{invar}$, equivariance $\mathcal{L}_{equiv}$, and two reconstruction losses $\mathcal{L}_{rec1}$ and $\mathcal{L}_{rec2}$.

6.2.2 View-Invariant Loss

Let's 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 the rotation parameters are estimated 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. This constraint is used 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, the frame is randomly selected to predict the rotation parameters for the two views,

$$R^v = \begin{cases} R^v_k & \text{if } r \text{ is } < 0.5 \\ R^v_m & \text{else} \end{cases} \quad \text{and} \quad R^w = \begin{cases} R^w_k & \text{if } r \text{ is } < 0.5 \\ R^w_n & \text{else} \end{cases}, \quad (6.2)$$

where $r \in U(0,1)$, and $U(0,1)$ denotes a uniform distribution returning a number between 0 and 1.

As it is assumed $E_{\odot}$ encodes the same canonical 3D pose features for I_k^v and I_k^w, then swapping their pose features while their viewpoint features are retained (as depicted in Figure 6.3), the network has to still be able to reconstruct them. Thus, the view-invariant loss is obtained by

$$\tilde{I'}_k^v = D(P'^{I_k^v}_{\circledast}) \quad \text{where} \quad P'^{I_k^v}_{\circledast} = R_{\diamond}^v \cdot P_{\odot}^{I_k^w} + T_k^v, \tag{6.3}$$

$$\tilde{I'}_k^w = D(P'^{I_k^w}_{\circledast}) \quad \text{where} \quad P'^{I_k^w}_{\circledast} = R_{\diamond}^w \cdot P_{\odot}^{I_k^v} + T_k^w, \tag{6.4}$$

$$\mathcal{L}_{invar} = \sum_{\phi \in \{v,w\}} MSE(I_k^{\phi},\ \tilde{I'}_k^{\phi}) \ . \tag{6.5}$$

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 $\mathcal{L}_{rec1}$ is introduced as a reconstruction loss, such that

$$\mathcal{L}_{rec1} = \sum_{\phi \in \{v,w\}} MSE(I_k^{\phi},\ \tilde{I}_k^{\phi}) \ , \tag{6.6}$$

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

6.2.3 Equivariance Loss

The effect of equivariance 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. It is assumed 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 (see Figure 6.4). Hence, an equivariance loss is proposed by computing 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.

$$\mathcal{L}_{equiv} = \sum_{\phi \in \{v,w\}, j \in \{1,2\}} MSE(P_{\circledast}^{I_k^{\phi}} + c_j,\ P_{\circledast}^{\dot{I}_k^{\phi}}), \tag{6.7}$$

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

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

$$\mathcal{L}_{rec2} = \sum_{\phi \in \{v,w\}} MSE(\dot{I}_k^{\phi},\ \tilde{\dot{I}}_k^{\phi}) \ , \tag{6.8}$$

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

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{invar} + \beta \cdot \mathcal{L}_{equiv} + \gamma \cdot (\mathcal{L}_{rec1} + \mathcal{L}_{rec2}) \tag{6.9}$$

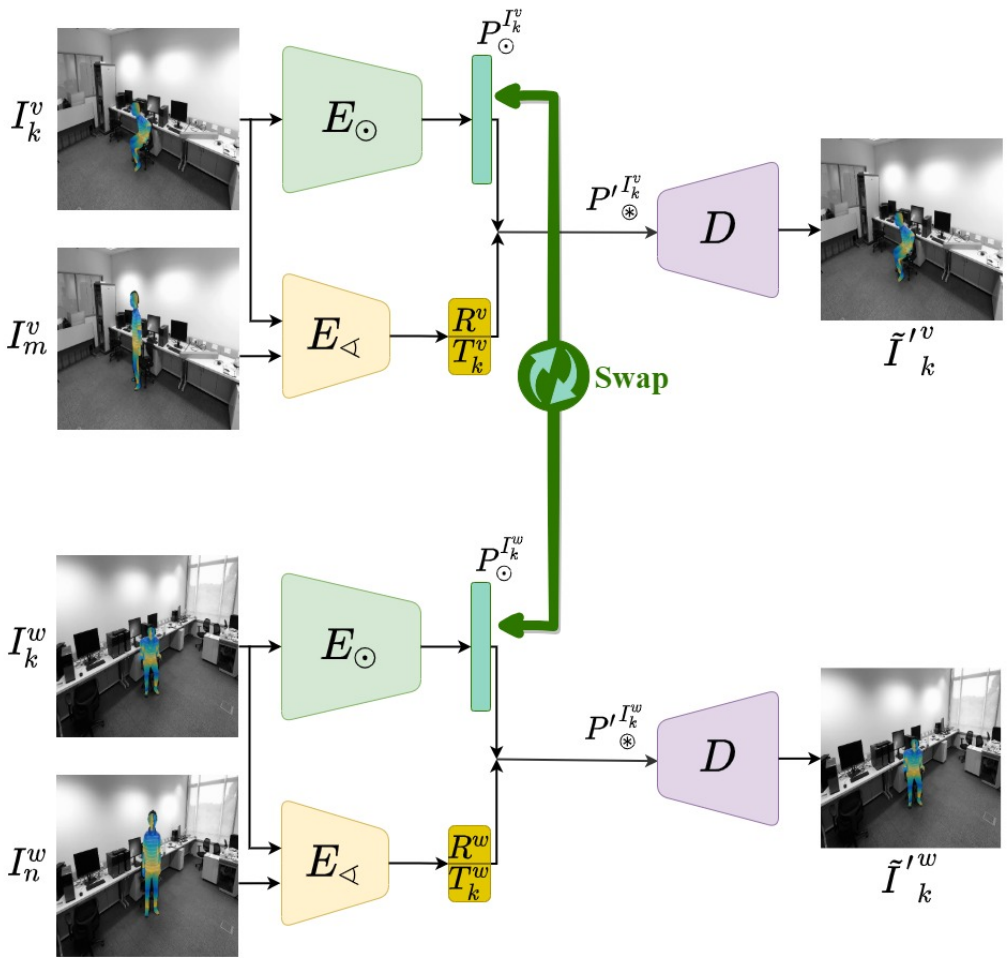


Figure 6.3: *Learning view-invariant pose features through simultaneous frames. Two simultaneous frames I_k^v and I_k^w from different views v and w of the same scene are reconstructed such that their canonical pose features are swapped while their viewpoint parameters are retained.*

The weights are determined 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 the example view-invariant downstream tasks.

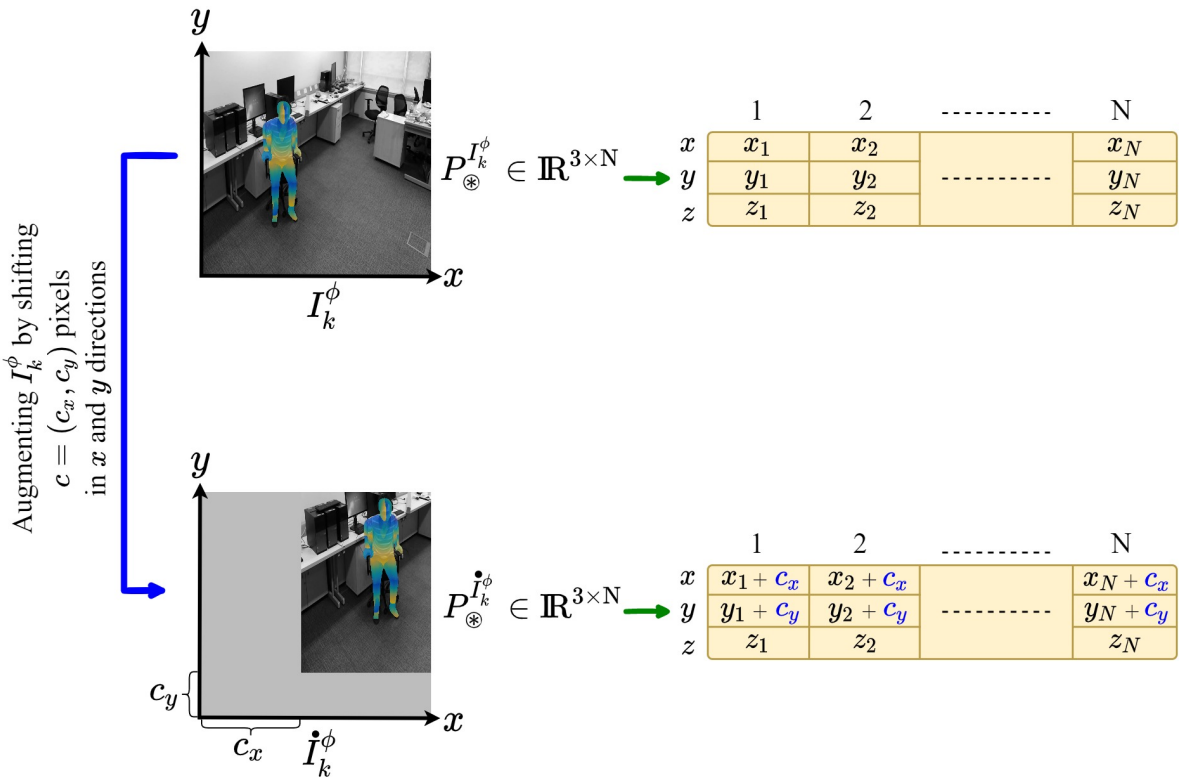


Figure 6.4: *The equivariance loss is computed by assuming that the proposed network extracts consistent order of pose features. Thus, if the subject is shifted by some pixels in the x and y directions, all pose components of the view-specific pose would shift similarly.*

6.3 Downstream Tasks

This section outlines the proposed method to model temporal aspects of the canonical pose features for two downstream task, action recognition and human movement assessment.

6.3.1 Action Recognition

The 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, a bidirectional gated recurrent unit (GRU), followed by one FC for which its size is equal to the number of action classes, is added after the view-invariant

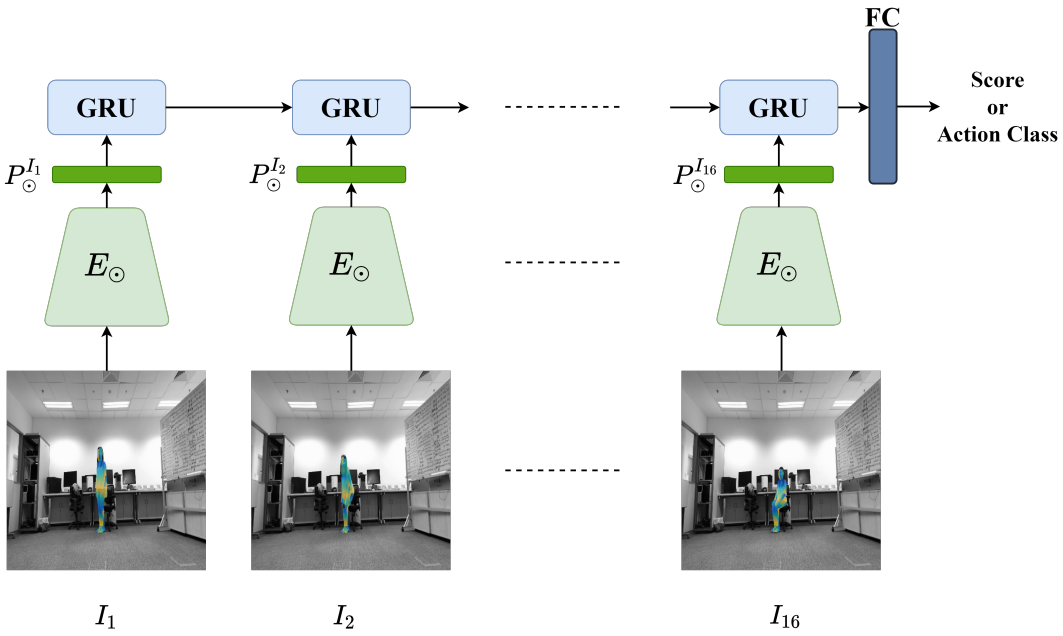


Figure 6.5: *The proposed model to exploit temporal elements of the learned view-invariant pose representations for action recognition and human movement assessment downstream tasks.*

pose encoder $E_{\odot}$ (see Figure 6.5). Then, the network is trained on fixed-size 16-frame input sequences with the cross-entropy loss function. Similar to [179], the sequences are subsampled such that every sequence is divided into 16 segments and one random frame is selected amongst all frames of each segment.

6.3.2 Human Movement Assessment

To study the efficiency of the learned representation for quality of human movement, as in the action recognition task, a bidirectional GRU followed by one FC layer is added on top of $E_{\odot}$ to deal with temporal analysis (see Figure 6.5). The size of the FC layer is equal to the number of possible scores for a movement type. However, as discussed in Chapter 1, for movement quality assessment, every single frame of a sequence should be analysed, so no subsampling strategies can be applied for this task, and the network is trained and tested on consecutive frames. To do this, following Section 5.2.3 that divides each video sequence into non-overlapping 16-frame video clips, the network is trained on a random 16-frame clip through the cross entropy loss function. For inference, all 16-frame clips of a video sequence are processed, then the score for a sequence is estimated by averaging the outputs of the last FC layer, as in Section 5.2.3.

6.4 Experiments and Results

As the learned pose features are unsupervised and in an unknown high-level canonical space, there is no ground-truth to evaluate them directly, so the comparison is performed only indirectly on downstream tasks. This section first presents the experiments and results of applying the unsupervised learned view-invariant pose features for action recognition. Then, the efficiency of transferring the learned pose representation from action recognition for human movement assessment is investigated. The contribution of the different components of the proposed method is also evaluated.

The datasets used in the experiments are outlined in Section 6.4.1. The implementation details and evaluation metrics respectively are described in Sections 6.4.2 and 6.4.3. Action recognition experiments including quantitative and qualitative results, and ablation studies are then presented in Section 6.4.4. Finally, Section 6.4.5 outlines the experiments and results for the human movement assessment task.

6.4.1 Datasets

To learn unsupervised view-invariant pose representation, NTU RGB+D [126] was applied (see Section 3.2.4 for details of NTU RGB+D). Then, the performance of the view-invariant pose features were evaluated on NTU (for short), based around cross-view and cross-subject protocols (see Section 2.4 for further details). For both pretext and downstream tasks, the same training and testing sets as in [126] were used.

The QMAR and KIMORE datasets were also applied to evaluate the efficiency of transferring the learned pose representations from NTU and action recognition to obtain cross-view, cross-subject, and single-view human movement assessment results. To perform the experiments on QMAR and KIMORE, the same training and testing sets as in Section 5.3 were used.

6.4.2 Implementation Details

Details of Network Architecture – The proposed auto-encoder is inspired by the U-Net encoder/decoder [28, 38, 115, 119]. The U-Net is a convolutional or spatial latent auto-encoder with skip connections between the encoder and the decoder parts, while a dense one [4] without the skip connections is desired to encode the 3D pose features, so it was adapted for the pose representation problem. Table 6.1 shows details of the proposed auto-encoder architecture and the extra layers added after it for modeling the temporal information in downstream tasks.

Module		Layers
Pose Representation (Auto-encoder)	$E_{\odot}$	$\{C2(3\times3,64),\ BN,\ ReLU\}\times2,\ MP(2\times2),$ $\{C2(3\times3,128),\ BN,\ ReLU\}\times2,\ MP(2\times2),$ $\{C2(3\times3,256),\ BN,\ ReLU\}\times2,\ MP(2\times2),$ $\{C2(3\times3,512),\ BN,\ ReLU\}\times1,\ \{C2(3\times3,512),\ ReLU\}\times1,$ $\{FC(1024),\ ReLU\},\ \{FC(512),\ ReLU\},\ \{FC(3\times70)\}$
	$E_{\triangleleft}$	$\{C2(5\times5,128),\ BN,\ ReLU\}\times2,\ MP(7\times7),$ $\{C2(5\times5,256),\ BN,\ ReLU\}\times2,\ \{FC(512),\ ReLU,\ Drp\},\ \{FC(6)\}$
	D	$\{FC(16\times16\times512),\ ReLU,\ Drp\},\ \{C2(3\times3,256),\ BN,\ ReLU\}\times2,$ $\{CT2(3\times3,128),\ BN,\ ReLU\}\times2,\ \{CT2(3\times3,64),\ BN,\ ReLU\}\times2,$ $\{CT2(3\times3,3),\ BN,\ ReLU\}\times2,\ tanh$
Temporal Modeling		b-GRU(2, 1024), FC(S)

Table 6.1: *Details of the proposed network's modules – All modules are 2D. $C2(d\times d,ch)$: $d\times d$ convolution filters with ch channels, $CT2$: transposed convolution filters, $b-GRU(l,h)$: l-layer bidirectional gated recurrent unit with hidden state size h, MP: max pooling, BN: batch normalization, $FC(O)$: fully connected layer with O outputs. S is the number of classes and possible scores for a movement type for action recognition and human movement assessment downstream tasks respectively.*

Training and Testing Details – The proposed model was implemented in Pytorch. For the pretext task, it was trained for 20 epochs using Adam [68] with a fixed learning rate of 0.0002, and batch size 5. During training, random horizontal flipping was applied for data augmentation. For downstream tasks, the proposed network was trained for 50 epochs using Adam [68] with batch size 20 and an initial learning rate of 0.0002 that was decayed by a factor of 10 every 10 epochs. During training, random cropping was applied for data augmentation. Note, the depth mask images of NTU used in the experiments contain bounding box of subjects as released by [126].

Hyper-Parameter Settings – To select the 3D canonical pose feature size $P^{I}_{\odot}\in\mathbb{R}^{3\times N}$, cross-validation was used and the total loss $\mathcal{L}_{total}$ in Eq. 6.9 was evaluated 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. [115] who set their latent code size at 3×200. As shown in Table 6.2, the average $\mathcal{L}_{total}$ cross-validation results on the NTU dataset for both CV and CS protocols is best when $N=70$, hence the 3D canonical pose feature size is set at 3×70.

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

Table 6.2: *Optimising $P^I_\odot$ - Average $\mathcal{L}_{total}$ cross-validation results on NTU for different canonical pose size ($3 \times N$). The* **Bold** *numbers show the best results.*

6.4.3 Evaluation Metrics

To evaluate the performance of the pose features for action recognition task, classification accuracy was used, and for human movement assessment, SRC was employed [82, 99, 104] (see Section 5.3.2 for more details of SRC).

6.4.4 Action Recognition

This section evaluates the performance of the learned pose representations by the proposed method for view-invariant action recognition. Tables 6.3 and 6.4 report the proposed method's results for cross-subject (CS) action recognition accuracy on the NTU dataset where RGB and depth data are applied as input, and Tables 6.5 and 6.6 present cross-view (CV) results on NTU for RGB and depth data. These tables contain the results of both supervised and unsupervised experiments such that the supervised results were obtained by both fine-tuning $E_\odot$ and training it from scratch during the downstream task, while the unsupervised results were obtained after freezing $E_\odot$'s parameters.

Tables 6.3 to 6.6 also present the CS and CV results of the state-of-the-art RGB and depth based representation learning approaches on NTU. All the representation learning methods on NTU that are compared to the proposed method here can operate on either RGB or depth data for training and inference, except Li et al. [78] 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 the proposed network learns pose representation, *e.g.* [143] uses 3D CNNs whereas the proposed method is integrally a 2D design. In the case of [78] which applies a 2D ResNet with added ConvLSTM [127], the results with the closest possible backbone, comprising a 2D ResNet and an LSTM are provided.

Method	Year	Backbone	Supervised (%)		Unsupervised (%)
			scratch	fine-tune	
Luo et al. [91]	2017	VGG + ConvLSTM	-	-	56.0
Li et al. [78] ✔	2018	2D ResNet + ConvLSTM	36.6	55.5	48.9
Vyas et al. [143] ✔	2020	3D CNN + LSTM	-	**82.3**	-
Proposed Method ✔	2021	2D ResNet + LSTM	<u>66.7</u>	73.8	<u>63.0</u>
Proposed Method ✔	2021	2D CNN + GRU	**70.3**	<u>78.1</u>	**68.3**

Table 6.3: *Cross-subject action recognition accuracy on NTU for RGB 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.*

Method	Year	Backbone	Supervised (%)		Unsupervised (%)
			scratch	fine-tune	
Misra et al. [94]	2016	AlexNet	-	-	46.2
Luo et al. [91]	2017	VGG + ConvLSTM	-	-	61.4
Li et al. [78] ✔	2018	2D ResNet + ConvLSTM	42.3	68.1	<u>60.8</u>
Vyas et al. [143] ✔	2020	3D CNN + LSTM	-	<u>71.8</u>	-
Proposed Method ✔	2021	2D ResNet + LSTM	<u>63.1</u>	72.7	58.0
Proposed Method ✔	2021	2D CNN + GRU	**75.9**	**78.8**	**64.7**

Table 6.4: *Cross-subject action recognition accuracy on NTU for depth 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.*

Method	Year	Backbone	Supervised (%)		Unsupervised (%)
			scratch	fine-tune	
Li et al. [78] ✔	2018	2D ResNet + ConvLSTM	29.2	49.3	40.7
Vyas et al. [143] ✔	2020	3D CNN + LSTM	-	**86.3**	-
Proposed Method ✔	2021	2D ResNet + LSTM	<u>66.5</u>	78.2	<u>62.1</u>
Proposed Method ✔	2021	2D CNN + GRU	**77.0**	<u>83.6</u>	**74.8**

Table 6.5: *Cross-view action recognition accuracy on NTU for RGB 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.*

Method	Year	Backbone	Supervised (%)		Unsupervised (%)
			scratch	fine-tune	
Misra et al. [94]	2016	AlexNet	-	-	40.9
Luo et al. [91]	2017	VGG + ConvLSTM	-	-	53.2
Li et al. [78] ✔	2018	2D ResNet + ConvLSTM	37.7	63.9	53.9
Vyas et al. [143] ✔	2020	3D CNN + LSTM	-	<u>78.7</u>	-
Proposed Method ✔	2021	2D ResNet + LSTM	<u>60.4</u>	75.5	<u>58.3</u>
Proposed Method ✔	2021	2D CNN + GRU	**76.7**	**82.5**	**67.5**

Table 6.6: *Cross-view action recognition accuracy on NTU for depth 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.*

As shown in Tables 6.3 to 6.6, for the unsupervised scenario, the proposed method with 2D CNN and GRU backbone significantly improves the state-of-the-art across CS and CV tests, at 68.3%, 74.8% for RGB, and 64.7%, 67.5% for depth data, respectively. The 2D ResNet + LSTM incarnation of the proposed 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 [78]'s 40.7% for cross-view RGB inference. For the supervised learning case, the proposed method improves on all other works with depth data, whether training from scratch or fine-tuning the proposed network with best results, at 78.8% and 82.5% on CS and CV protocols respectively, and attains very competitive results using RGB in comparison to the 3D CNN-based [143].

Table 6.7 reports the results of recent state-of-the-art unsupervised pose representation methods that operate on 3D skeleton data and compares them with the proposed method's results. Cheng et al. [19]'s result is marginally better than the proposed method in CS mode, and Yao et al. [162] perform better than the proposed method in CV mode. These result vindicate the proposed 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	**Year**	**Backbone**	**Input**	**Unsupervised (%)**	
				CS	**CV**
Su et al. [131]	2020	GRU	Skeleton	50.7	<u>76.1</u>
Lin et al. [85]	2020	GRU	Skeleton	52.5	-
Yao et al. [162]	2021	GRU + GCN	Skeleton	54.4	**79.2**
Cheng et al. [19] ✔	2021	Transformer	Skeleton	**69.3**	72.8
Rao et al. [114] ✔	2021	LSTM	Skeleton	58.5	64.8
Proposed Method ✔	2021	2D CNN + GRU	Depth	64.7	67.5
Proposed Method ✔	2021	2D CNN + GRU	RGB	<u>68.3</u>	74.8

Table 6.7: *State-of-the-art action recognition accuracy results on NTU for skeleton-based representation learning approaches vs. the proposed method's results. The ✔ symbol highlights view-invariant methods. The best and the second-best results are in **Bold** and <u>underline</u> respectively.*

Qualitative Results – Figures 6.6 and 6.7 show some qualitative results of the proposed approach for the cross-subject scenario on RGB-based Densepose and depth modalities respectively. The qualitative results of the network for cross-view scenario on RGB-based Densepose and depth data are also illustrated in Figures 6.8 and 6.9.

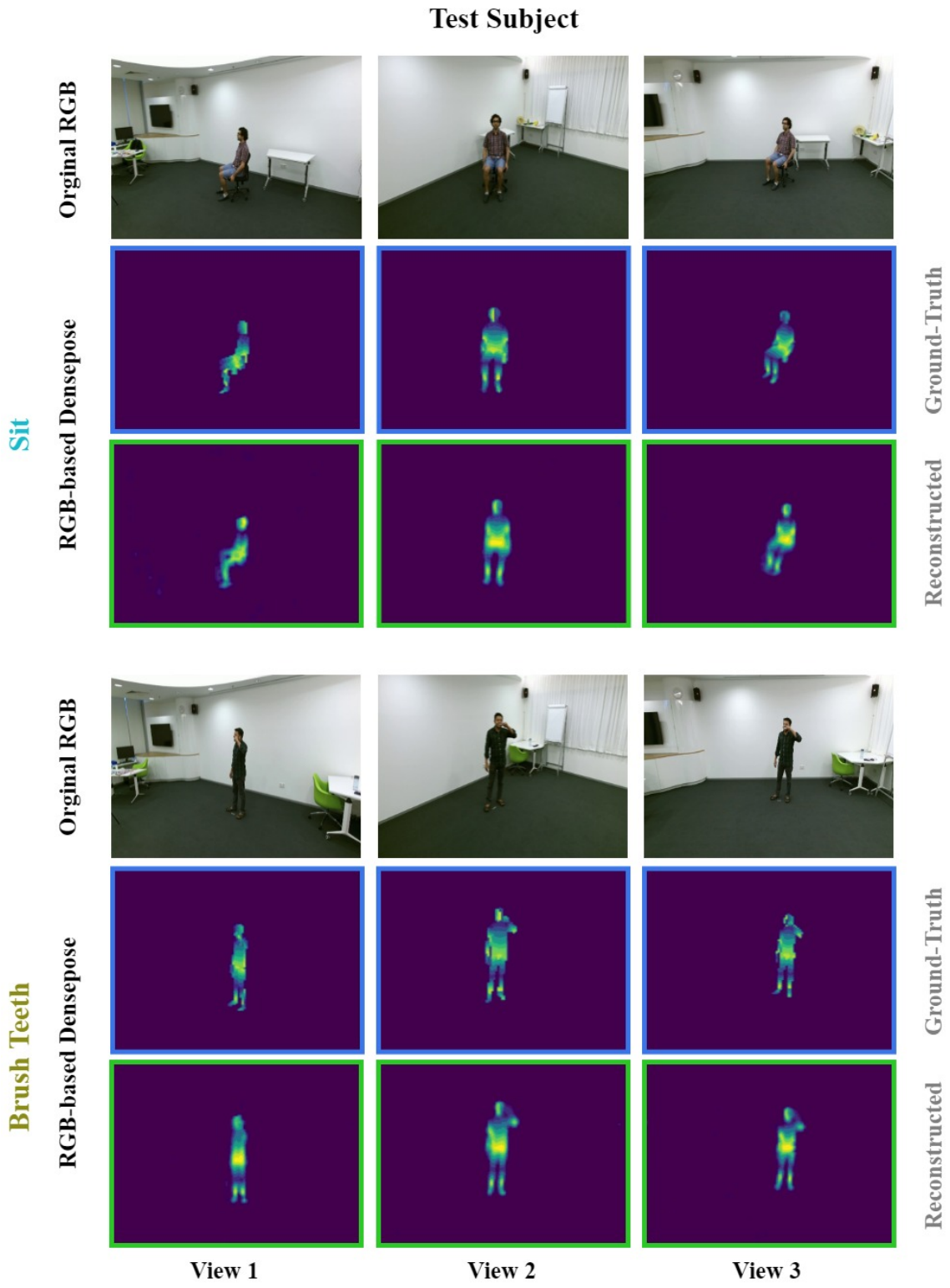


Figure 6.6: *The reconstruction results of the proposed approach on unseen subject data for two samples of NTU belonging to sit and brush teeth actions for RGB-based Densepose modality. Views 1 to 3 show the simultaneous frames belonging to the same scene captured from different viewpoints. The blue boxes denote testing frames, and the green ones indicate their corresponding reconstructed frames.*

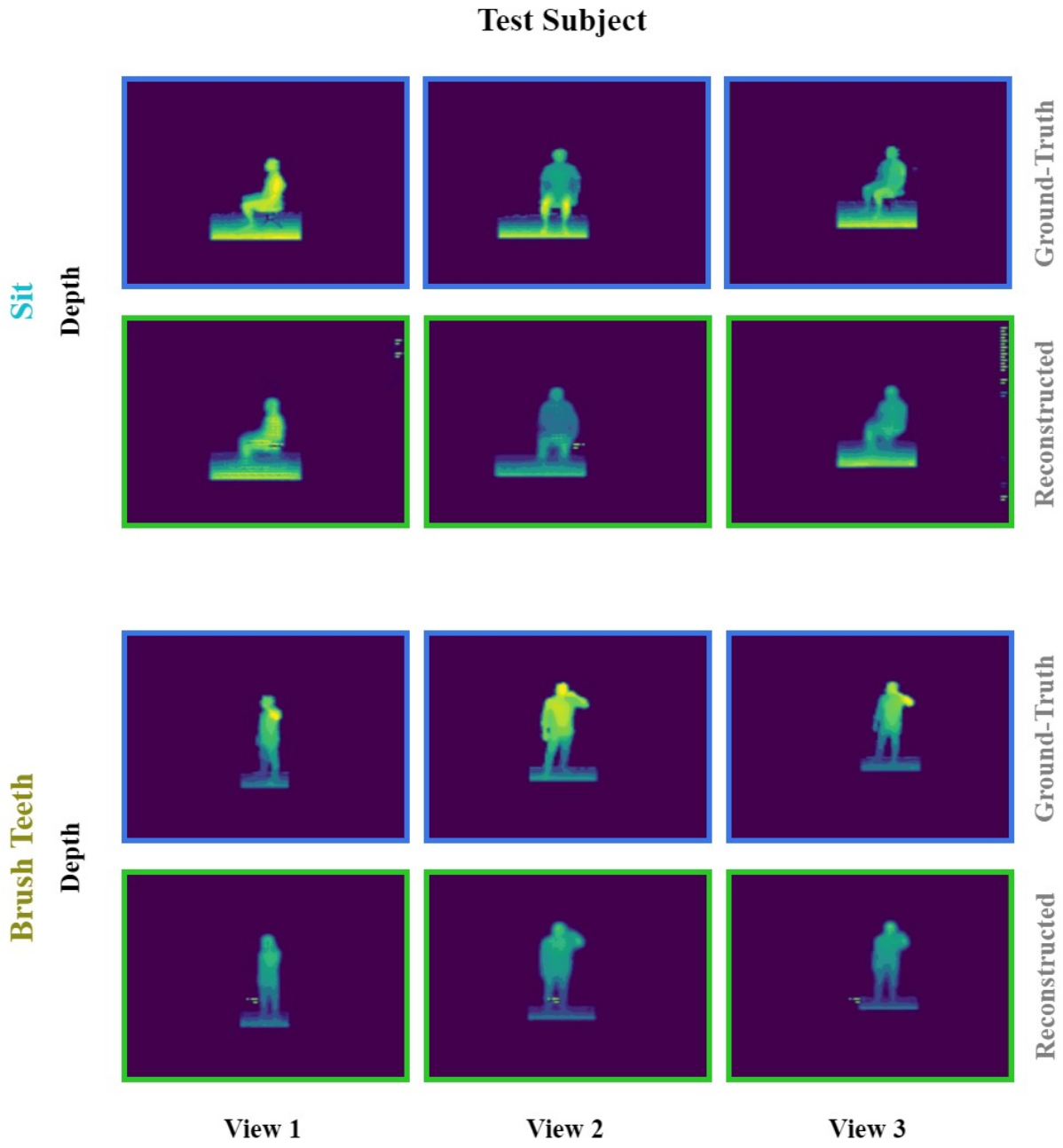


Figure 6.7: *The reconstruction results of the proposed approach on unseen subject data for two samples of NTU belonging to sit and brush teeth actions for depth modality. Views 1 to 3 show the simultaneous frames belonging to the same scene captured from different viewpoints. The blue boxes denote testing frames, and the green ones indicate their corresponding reconstructed frames.*

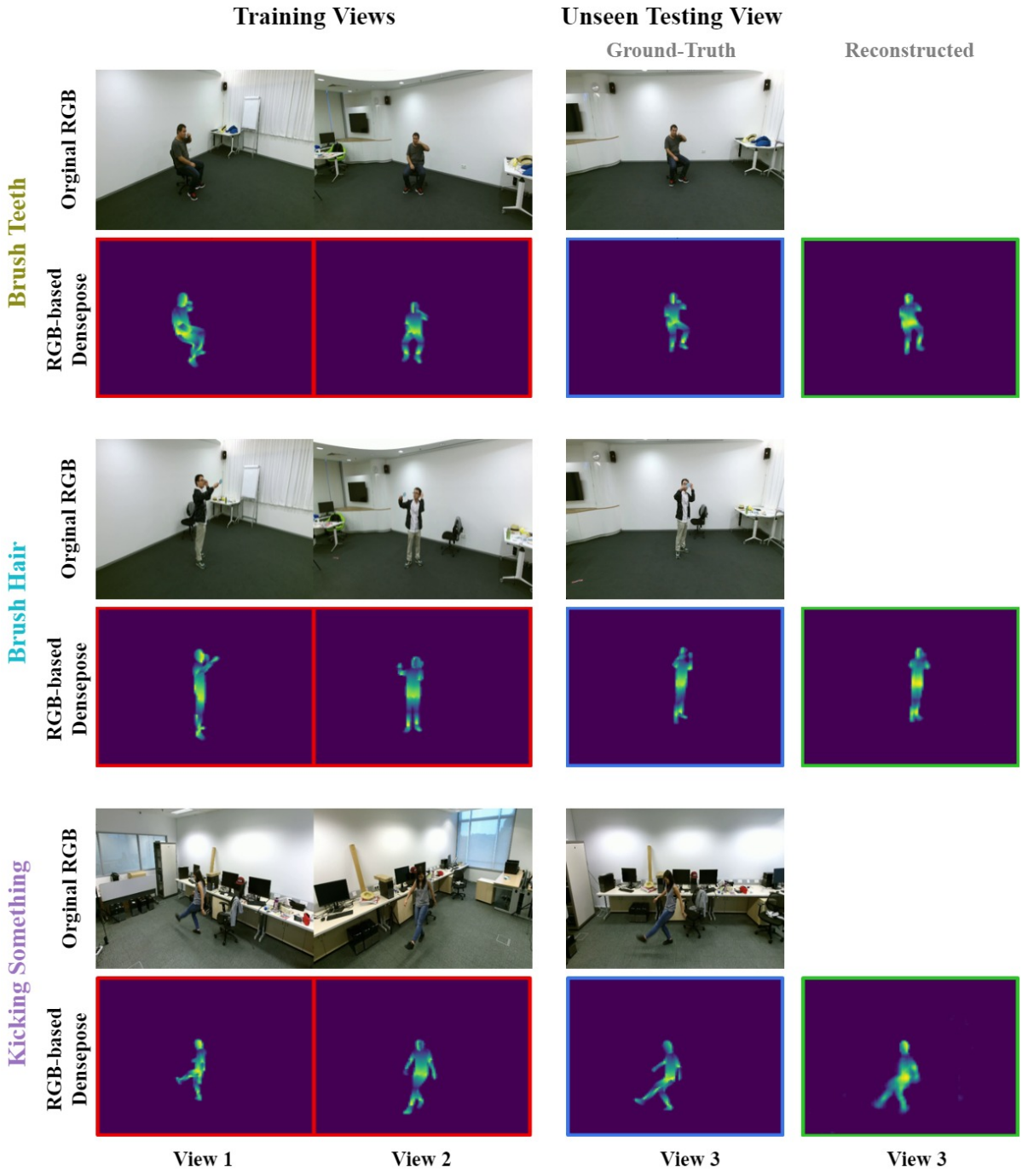


Figure 6.8: *The reconstruction results of the proposed approach on unseen view data for three samples of NTU belonging to brush teeth, brush hair, and kicking something actions and for RGB-based Densepose modality. For each sample, Views 1 to 3 show the simultaneous frames belonging to the same scene captured from different viewpoints. The red boxes denote training frames, the blue ones indicate their corresponding testing frame captured from the unseen viewpoint, and the green ones present the reconstruction result of the testing frame.*

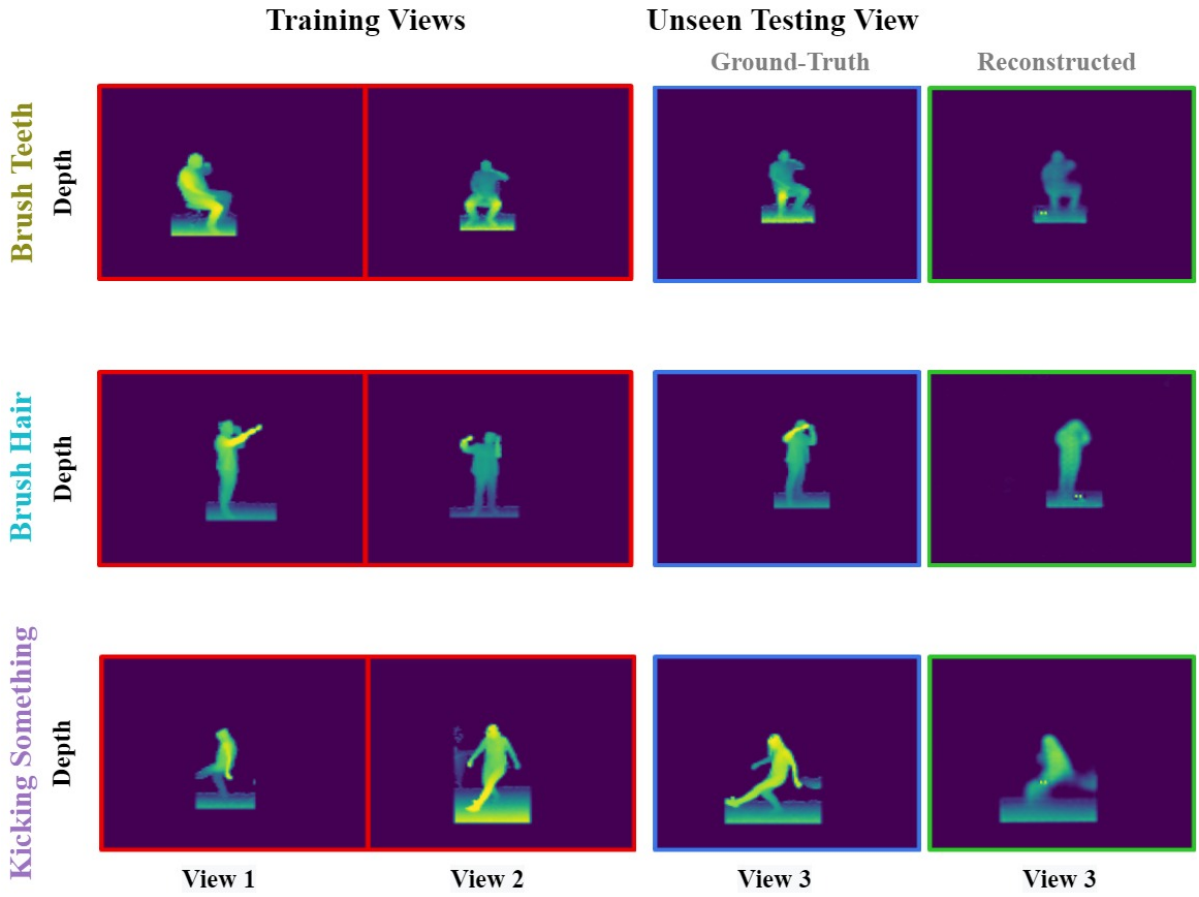


Figure 6.9: *The reconstruction results of the proposed approach on unseen view data for three samples of NTU belonging to brush teeth, brush hair, and kicking something actions and for depth modality. For each sample, Views 1 to 3 show the simultaneous frames belonging to the same scene captured from different viewpoints. The red boxes denote training frames, the blue ones indicate their corresponding testing frame captured from the unseen viewpoint, and the green ones present the reconstruction result of the testing frame.*

Ablation Study – This section also ablates the losses to examine their impact on the learning of the unsupervised pose features. Table 6.8 shows the unsupervised action recognition accuracy on NTU as each or both of $\mathcal{L}_{invar}$ and $\mathcal{L}_{equiv}$ are dropped. It can be seen that removing $\mathcal{L}_{equiv}$ from the training process, the results for both CV and CS in both RGB and depth deteriorates. This verifies that positional order consistency is essential in both cases. It is also observed that eliminating $\mathcal{L}_{invar}$ causes the 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.

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

Table 6.8: *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.*

6.4.5 Human Movement Assessment

This section aims to study the efficiency of transferring the learned representation on NTU and action recognition task for quality of movement scoring on QMAR and KIMORE. It reports supervised and unsupervised results of cross-view, cross-subject, and single-view human movement assessment. The 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.

As this chapter offers the first ever unsupervised cross-subject and cross-view results on the QMAR dataset, for further direct comparison with an unsupervised approach, the results of [78] were obtained for this dataset. Note, similar to the proposed method, to provide the unsupervised and fine-tuning results, [78] was pretrained on NTU. For the supervised scenario, the performance of the proposed network is also evaluated against the proposed method in Chapter 5 (VI-Net) and two other baselines introduced in Section 5.3. For single-view experiments on the KIMORE dataset, in addition to [78] and the baselines, it includes the supervised and unsupervised results of the work presented in [95].

Cross-Subject Human Movement Assessment – Table 6.9 shows the cross-subject results on QMAR. It is observed that the unsupervised human movement analysis results of the proposed method on QMAR outperforms Li et al. [78], reaching an average SRC of 0.58. These are broadly already competitive to the supervised results on QMAR, particularly when compared against the Kinetic-400 [66] pretrained, deep I3D network. The supervised version of the proposed method, where the network weights are fine-tuned after transfering the weights learnt through NTU training, exceeds VI-Net's performance on average and achieves 0.70.

Method	Year	Training	Action (SRC)				Average
			W-P	W-S	SS-P	SS-S	(SRC)
Supervised							
I3D [13]	2017	fine-tune	0.79	0.47	0.54	0.55	0.58
Li et al. [78]	2018	scratch	0.55	0.32	0.39	0.64	0.47
Li et al. [78] ✔	2018	fine-tune	0.57	0.59	0.41	0.75	0.57
C3D (after [104])	2019	scratch	0.50	0.37	0.25	0.54	0.41
VI-Net (VGG) ✔	2020	scratch	0.82	<u>0.52</u>	0.55	<u>0.73</u>	0.65
VI-Net (ResNeXt) ✔	2020	scratch	<u>0.87</u>	<u>0.52</u>	<u>0.58</u>	0.69	<u>0.66</u>
Proposed Method	2021	scratch	0.81	0.51	0.39	0.72	0.60
Proposed Method ✔	2021	fine-tune	**0.89**	**0.54**	**0.62**	**0.76**	**0.70**
Unsupervised							
Li et al. [78] ✔	2018	-	0.21	0.10	0.24	0.47	0.25
Proposed Method ✔	2021	-	**0.70**	**0.50**	**0.48**	**0.66**	**0.58**

Table 6.9: *SRC between predicted scores and ground truth labels for cross-subject analysis on different actions of QMAR dataset. I3D was pretrained on Kinetic-400 [66], and the ✔ symbol highlights view-invariant methods. The best supervised and unsupervised results are in* ***Bold*** *and the second-best supervised results are <u>underlined</u>.*

Cross-View Human Movement Assessment – Similarly to Section 5.3.4, two sets of experiments are performed here, (i) only one view is used for training and the rest of the viewpoints are applied for testing, shown in Tables 6.10 and 6.11, (ii) a combination of one frontal view and one side view is used for training and all other available views are applied for testing, shown in Tables 6.12 and 6.13 (see Section 5.3.4 for further details of training and testing sets).

As shown in Tables 6.10 to 6.13, the unsupervised results of the proposed method on QMAR outperforms Li et al. [78]. When two viewpoints are employed for training, it reaches an average SRC of 0.77, 0.64, 0.39, and 0.56 for W-P, W-S, SS-P, and SS-S respectively which are broadly already competitive to the supervised results on QMAR, particularly when compared against the Kinetic-400 [66] pretrained, deep I3D network.

	Method	Year	Training	Training View (SRC)						Avg
				1	2	3	4	5	6	(SRC)
W-P	**Supervised**									
	I3D [13]	2017	fine-tune	0.60	0.61	0.62	0.50	0.57	0.55	0.60
	Li et al. [78]	2018	scratch	0.19	0.19	0.17	0.21	0.14	0.26	0.19
	Li et al. [78] ✔	2018	fine-tune	0.13	0.08	0.05	0.23	0.18	0.15	0.13
	C3D (after [104])	2019	scratch	0.18	0.18	0.21	0.23	0.24	0.21	0.20
	VI-Net (VGG-19) ✔	2020	scratch	0.67	0.66	0.66	0.64	0.67	0.72	0.67
	VI-Net (ResNeXt-50)✔	2020	scratch	0.67	**0.72**	**0.70**	**0.72**	**0.71**	**0.73**	**0.70**
	Proposed Method	2021	scratch	0.21	0.42	0.45	0.43	0.45	0.51	0.41
	Proposed Method ✔	2021	fine-tune	**0.70**	0.66	0.64	0.53	0.39	0.69	0.55
	Unsupervised									
	Li et al. [78] ✔	2018	-	0.08	0.12	0.05	0.14	0.17	0.08	0.10
	Proposed Method ✔	2021	-	**0.38**	**0.26**	**0.36**	**0.50**	**0.46**	**0.40**	**0.39**
W-S	**Supervised**									
	I3D [13]	2017	fine-tune	0.49	0.44	0.62	0.64	0.45	0.54	0.53
	Li et al. [78]	2018	scratch	0.16	0.07	0.28	0.07	0.07	0.10	0.12
	Li et al. [78] ✔	2018	fine-tune	0.13	0.12	0.23	0.08	0.16	0.05	0.12
	C3D (after [104])	2019	scratch	0.14	0.10	0.23	0.20	0.17	0.17	0.16
	VI-Net (VGG-19) ✔	2020	scratch	0.43	0.54	0.56	0.59	0.60	0.40	0.52
	VI-Net (ResNeXt-50)✔	2020	scratch	**0.64**	**0.62**	**0.59**	**0.66**	**0.63**	**0.60**	**0.62**
	Proposed Method	2021	scratch	0.35	0.25	0.24	0.33	0.10	0.33	0.26
	Proposed Method ✔	2021	fine-tune	0.41	0.37	0.43	0.60	0.61	0.51	0.48
	Unsupervised									
	Li et al. [78] ✔	2018	-	0.05	0.11	0.12	0.08	0.04	0.07	0.07
	Proposed Method ✔	2021	-	**0.33**	**0.30**	**0.27**	**0.51**	**0.42**	**0.20**	**0.33**

Table 6.10: *SRC between predicted scores and ground truth labels for cross-view analysis on W-P and W-S actions of the QMAR dataset, where one viewpoint is used for training. I3D was pretrained on Kinetic-400 [66], and the ✔ symbol highlights view-invariant methods. The best supervised and unsupervised results for each view are in* ***Bold*** *and the second-best supervised results for each view are underlined. The green highlights indicate the best results for W-P and W-S actions amongst all views.*

	Method	Year	Training	Training View (SRC)						Avg
				1	2	3	4	5	6	(SRC)
SS-P	**Supervised**									
	I3D [13]	2017	fine-tune	0.18	0.21	0.25	0.20	0.37	0.18	0.23
	Li et al. [78]	2018	scratch	0.10	0.20	0.10	0.06	0.11	0.09	0.11
	Li et al. [78] ✔	2018	fine-tune	0.23	0.07	0.05	0.05	0.11	0.09	0.10
	C3D (after [104])	2019	scratch	0.10	0.10	0.12	0.17	0.12	0.09	0.11
	VI-Net (VGG-19) ✔	2020	scratch	0.32	0.31	0.23	0.34	**0.52**	0.24	0.32
	VI-Net (ResNeXt-50)✔	2020	scratch	0.25	0.32	**0.43**	**0.49**	0.45	**0.44**	**0.39**
	Proposed Method	2021	scratch	0.10	0.10	0.27	0.25	0.19	0.13	0.21
	Proposed Method ✔	2021	fine-tune	**0.38**	**0.39**	0.37	0.32	0.31	0.32	0.34
	Unsupervised									
	Li et al. [78] ✔	2018	-	0.22	0.13	0.13	0.09	0.06	0.15	0.13
	Proposed Method ✔	2021	-	**0.30**	**0.20**	**0.31**	**0.22**	**0.23**	**0.27**	**0.25**
SS-S	**Supervised**									
	I3D [13]	2017	fine-tune	0.43	0.49	0.47	0.50	0.46	0.20	0.42
	Li et al. [78]	2018	scratch	0.25	0.13	0.21	0.38	0.08	0.15	0.20
	Li et al. [78] ✔	2018	fine-tune	0.09	0.23	0.40	0.26	0.16	0.08	0.20
	C3D (after [104])	2019	trained	0.26	0.30	0.25	0.32	0.20	0.18	0.25
	VI-Net (VGG-19) ✔	2020	scratch	**0.49**	0.40	**0.52**	0.34	**0.50**	**0.28**	0.42
	VI-Net (ResNeXt-50)✔	2020	scratch	0.45	**0.56**	0.43	**0.54**	0.48	0.16	**0.43**
	Proposed Method	2021	scratch	0.21	0.37	0.37	0.31	0.30	0.11	0.28
	Proposed Method ✔	2021	fine-tune	0.39	0.42	0.43	0.41	0.32	**0.28**	0.37
	Unsupervised									
	Li et al. [78] ✔	2018	-	0.09	0.18	0.05	0.23	0.14	0.10	0.13
	Proposed Method ✔	2021	-	**0.38**	**0.33**	**0.40**	**0.26**	**0.25**	**0.24**	**0.31**

Table 6.11: *SRC between predicted scores and ground truth labels for cross-view analysis on SS-P and SS-S actions of the QMAR dataset, where one viewpoint is used for training. I3D was pretrained on Kinetic-400 [66], and the ✔ symbol highlights view-invariant methods. The best supervised and unsupervised results for each view are in* ***Bold*** *and the second-best supervised results for each view are underlined. The green highlights indicate the best results for SS-P and SS-S actions amongst all views.*

	Method	Year	Training	Training Views (SRC)					Avg
				2, 4	2, 5	2, 6	1, 5	3, 5	(SRC)
W-P	**Supervised**								
	I3D [13]	2017	fine-tune	0.85	0.87	0.80	0.80	0.80	0.82
	Li et al. [78]	2018	scratch	0.34	0.22	0.18	0.19	0.15	0.21
	Li et al. [78] ✔	2018	fine-tune	0.37	0.14	0.17	0.15	0.23	0.21
	C3D (after [104])	2019	scratch	0.65	0.65	0.69	0.63	0.57	0.63
	VI-Net (VGG-19) ✔	2020	scratch	0.81	0.75	0.76	0.76	0.79	0.77
	VI-Net (ResNeXt-50)✔	2020	scratch	0.89	**0.92**	0.77	0.75	0.84	0.83
	Proposed Method	2021	scratch	0.84	0.82	0.82	0.81	0.80	0.81
	Proposed Method ✔	2021	fine-tune	**0.93**	**0.92**	**0.83**	**0.87**	**0.91**	**0.88**
	Unsupervised								
	Li et al. [78] ✔	2018	-	0.27	0.12	0.11	0.12	0.11	0.14
	Proposed Method ✔	2021	-	**0.84**	**0.81**	**0.70**	**0.75**	**0.78**	**0.77**
W-S	**Supervised**								
	I3D [13]	2017	fine-tune	**0.76**	0.71	**0.73**	0.71	0.70	0.72
	Li et al. [78]	2018	scratch	0.08	0.15	0.11	0.10	0.19	0.12
	Li et al. [78] ✔	2018	fine-tune	0.06	0.38	0.13	0.17	0.17	0.18
	C3D (after [104])	2019	trained	0.42	0.37	0.33	0.48	0.45	0.41
	VI-Net (VGG-19) ✔	2020	scratch	0.72	0.74	0.67	0.68	0.66	0.69
	VI-Net (ResNeXt-50)✔	2020	scratch	0.73	**0.81**	0.68	0.81	**0.79**	0.76
	Proposed Method	2021	scratch	0.59	0.68	0.63	0.71	0.68	0.65
	Proposed Method ✔	2021	fine-tune	**0.76**	0.76	**0.73**	**0.85**	0.77	**0.77**
	Unsupervised								
	Li et al. [78] ✔	2018	-	0.14	0.13	0.10	0.07	0.14	0.11
	Proposed Method ✔	2021	-	**0.61**	**0.65**	**0.62**	**0.70**	**0.66**	**0.64**

Table 6.12: *SRC between predicted scores and ground truth labels for cross-view analysis on W-P and W-S actions of the QMAR dataset, where two views are used for training on. I3D was pretrained on Kinetic-400 [66], and the ✔ symbol highlights view-invariant methods. The best supervised and unsupervised results for each combination of views are in* ***Bold*** *and the second-best supervised results for each combination of views are underlined. The green highlights indicate the best results for W-P and W-S actions amongst all view combinations.*

	Method	Year	Training	Training Views (SRC)					Avg
				2, 4	2, 5	2, 6	1, 5	3, 5	(SRC)
SS-P	**Supervised**								
	I3D [13]	2017	fine-tune	0.48	0.40	**0.47**	0.45	0.39	0.43
	Li et al. [78]	2018	scratch	0.11	0.10	0.09	0.11	0.11	0.10
	Li et al. [78] ✔	2018	fine-tune	0.16	0.14	0.20	0.15	0.10	0.15
	C3D (after [104])	2019	scratch	0.25	0.21	0.30	0.38	0.37	0.30
	VI-Net (VGG-19) ✔	2020	scratch	**0.52**	0.53	0.35	**0.55**	0.40	0.47
	VI-Net (ResNeXt-50)✔	2020	scratch	0.46	0.46	0.42	0.52	0.47	0.46
	Proposed Method	2021	scratch	0.16	0.17	0.27	0.20	0.25	0.21
	Proposed Method ✔	2021	fine-tune	0.50	**0.58**	**0.47**	0.52	**0.61**	**0.53**
	Unsupervised								
	Li et al. [78] ✔	2018	-	0.14	0.06	0.11	0.09	0.12	0.10
	Proposed Method ✔	2021	-	**0.48**	**0.34**	**0.37**	**0.33**	**0.47**	**0.39**
SS-S	**Supervised**								
	I3D [13]	2017	fine-tune	0.55	0.60	0.54	0.62	0.65	0.58
	Li et al. [78]	2018	scratch	0.19	0.11	0.18	0.30	0.08	0.17
	Li et al. [78] ✔	2018	fine-tune	0.27	0.35	0.19	0.21	0.22	0.24
	C3D (after [104])	2019	scratch	0.53	0.45	0.44	0.30	0.35	0.41
	VI-Net (VGG-19) ✔	2020	scratch	**0.64**	0.56	**0.62**	0.53	0.60	0.59
	VI-Net (ResNeXt-50)✔	2020	scratch	**0.64**	0.61	0.46	0.58	0.67	0.58
	Proposed Method	2021	scratch	0.59	0.59	0.56	0.52	0.39	0.53
	Proposed Method ✔	2021	fine-tune	**0.64**	**0.66**	**0.62**	**0.63**	**0.69**	**0.64**
	Unsupervised								
	Li et al. [78] ✔	2018	-	0.13	0.17	0.13	0.13	0.18	0.14
	Proposed Method ✔	2021	-	**0.52**	**0.59**	**0.56**	**0.58**	**0.57**	**0.56**

Table 6.13: *SRC between predicted scores and ground truth labels for cross-view analysis on SS-P and SS-S actions of the QMAR dataset, where two views are used for training on. I3D was pretrained on Kinetic-400 [66], and the ✔ symbol highlights view-invariant methods. The best supervised and unsupervised results for each combination of views are in* ***Bold*** *and the second-best supervised results for each combination of views are underlined. The green highlights indicate the best results for SS-P and SS-S actions amongst all view combinations.*

In the supervised scenario, when one view is used for training (Tables 6.10 and 6.11), VI-Net performs better than the proposed network. However, when the proposed network is trained on two viewpoints (Tables 6.12 and 6.13) and its weights are fine-tuned after transferring the weights learnt through NTU training, the proposed method exceeds VI-Net's performance on average and achieves rank correlation of 0.88, 0.77, 0.53, and 0.64 for W-P, W-S, SS-P, and SS-S action types respectively.

Single-View Human Movement Assessment – The results of the proposed method on the KIMORE dataset are available in Table 6.14. This table shows that in the unsupervised mode where all the convolutional filters are frozen during training, the results of the proposed method are broadly competitive with the deep supervised networks (*e.g.* I3D), especially for Exercises #3 to #5, its performance is even higher than the deep supervised VI-Net network. Furthermore, the proposed method outperforms the

Method	**Year**	**Training**	**Action Ex** (SRC)					**Avg**
			#1	**#2**	**#3**	**#4**	**#5**	(SRC)
Supervised								
I3D [13]	2017	fine-tune	0.45	0.56	0.57	0.64	0.58	0.56
Li et al. [78]	2018	scratch	**0.80**	**0.77**	**0.65**	<u>0.74</u>	0.82	<u>0.75</u>
Li et al. [78]	2018	finetune	0.76	0.67	0.52	0.71	<u>0.83</u>	0.69
C3D (after [104])	2019	scratch	0.66	0.64	<u>0.63</u>	0.59	0.60	0.62
VI-Net (VGG-19)	2020	scratch	<u>0.79</u>	<u>0.69</u>	0.57	0.59	0.70	0.66
VI-Net (ResNeXt-50)	2020	scratch	0.55	0.62	0.36	0.58	0.67	0.55
Nekoui and Cheng [95]	2021	scratch	0.63	0.60	0.54	0.56	0.57	0.58
Nekoui and Cheng [95]	2021	finetune	0.66	0.60	0.61	0.59	0.61	0.61
Proposed Method	2021	scratch	0.74	0.63	0.58	0.67	0.66	0.65
Proposed Method	2021	finetune	0.75	**0.77**	0.61	**0.78**	**0.91**	**0.76**
Unsupervised								
Li et al. [78]	2018	-	<u>0.67</u>	**0.64**	<u>0.63</u>	<u>0.76</u>	**0.76**	<u>0.69</u>
Nekoui and Cheng [95]	2021	-	0.60	0.56	0.37	0.59	0.54	0.53
Proposed Method	2021	-	**0.70**	<u>0.62</u>	**0.65**	**0.78**	<u>0.75</u>	**0.70**

Table 6.14: *SRC between predicted scores and ground truth labels for different action types of the single-view KIMORE dataset. I3D was pretrained on Kinetic-400 [66]. The* ***Bold*** *and* <u>underline</u> *numbers show the best and the second-best results for each scenario of each action type respectively.*

other methods on average, reaching the best rank correlation of 0.76 and 0.70 for supervised and unsupervised scenarios respectively. It can also be observed that the proposed method performs better than [95] that employs a sequences of 2D poses as input for all movement types and all scenarios.

6.5 Conclusion

This chapter presented a view-invariant human posture representation approach that *does not rely* on any 3D joint annotations and is trained only from 2D images such that the pose features can be applied for *novel view* downstream tasks, *e.g.* action recognition and human movement assessment. It also explored whether the learned unsupervised features on action recognition task can be transferred into the action quality assessment task. The experiments showed that not only can the learned pose representations be applied on unseen view videos from the same training data, but they can also be used in *different domains*.

The proposed approach can benefit in decreasing the costs of training for view-invariant action understanding applications. As the learned view-invariant pose features can be transferred into other domains, the existing multi-view datasets can be employed for training. This removes the need for collecting a new multi-view dataset which is expensive and time-consuming. Furthermore, the proposed method allows assessing the quality of human movement from RGB images alone which is particularly helpful in applications where capturing 3D skeletons is challenging, e.g. in healthcare rehabilitation monitoring at home or in the clinic.

As in the pretext stage, the proposed model requires synchronised multi-view frames to learn view-invariant 3D pose representations, future work can 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.

Chapter 7

Conclusion

This thesis has explored *view-invariance* in *human movement assessment* from ambiguous *2D RGB images*. This was investigated in the healthcare domain, for instance, observing subjects synthesising Parkinson's symptoms when they perform specific actions, such as walking or sitting-to-standing, to estimate an objective score for their level of functional mobility. To accomplish this challenging task, this thesis focused predominantly on learning and analysing view-invariant (canonical) spatial or spatio-temporal human pose features. In addition, as all existing human movement quality assessment datasets are single-view, this thesis has introduced two multi-view datasets, SMAD and QMAR, to demonstrate the performance of the proposed approaches, and QMAR is publicly released.

All previous approaches in movement quality assessment are view-specific and cannot operate on data coming from different camera viewpoints from training data. In addition, in healthcare applications, the proposed approaches are based on 3D skeleton data since this data obtain richer information than 2D images. However, capturing accurate skeleton data is challenging in in-the-wild scenarios as it requires specific hardware and environmental settings.

This thesis began dealing with view-invariance by training a model on multiple views. This was facilitated by developing a pose estimation approach that learns 3D human pose features in the same canonical manifold space for different viewpoints such that the extracted pose features can be employed directly for human movement assessment without requiring normalization and dimensionality reduction intermediate steps.

The thesis then explored a human movement assessment approach that can operate on arbitrary *novel viewpoints* at inference time without requiring to be trained on many

viewpoints. In the video understanding domain (*e.g.* action recognition), most state-of-the-art view-invariant approaches need at least two viewpoints for training. However, this thesis introduced a view-invariant model that not only can be trained on multiple views but is also able to tackle the view-invariance by training on *only* a *single viewpoint*.

Finally, this thesis investigated an *unsupervised* solution for the *novel view human movement assessment task*. It proposed a self-supervised approach that learns view-invariant 3D pose representation where the proposed method does not rely on 3D joint annotation, action labels, and camera viewpoint parameters for training. It showed that the view-invariant unsupervised features can be applied *directly* for unseen views downstream tasks such as human movement analysis and action recognition and they are capable to be *transferred* into the *other domains*. However, most current unsupervised 3D pose estimation methods are view-specific and camera information and further processing steps are required to extract view-invariant features from their view-specific output.

7.1 Findings and Limitations

This section reviews briefly the findings and limitations of each chapter of this thesis.

Chapter 3 described recording and details of two multi-view human movement assessment datasets, SMAD and QMAR. It first presented SMAD a multi-modal dataset including, RGB, depth, motion capture, and skeleton data. SMAD has been collected from 4 different views and through 19 healthy subjects that were trained to perform a turn-walk action in both normally and with three types of abnormalities comprising Stroke, Parkinson, and short-limp. Then, to explore further the view-invariance in action analysis, it introduced QMAR, which provides more participants, action types, and viewpoints than SMAD. QMAR has RGB, depth, and skeleton data, and has been recorded through 38 subjects and from six distinct views. The participants were trained to perform turn-walk and standing up and sitting down action types while simulating Parkinson's and Stroke ailments. Additionally, in QMAR, the movements have label that reflects the severity of the abnormality. However, both datasets have been recorded in a stationary environment which does not allow to examine the generalisation of the proposed methods under different environments, so in all experiments, both training and testing sets were selected from the same environment.

Chapter 4 developed a 3D pose estimation method to facilitate assessing the quality of movements from non-skeleton data in the multi-view training scenarios. The proposed method was trained to map all the simultaneous frames of the same scene captured

from different views into the same pose vector in a manifold space. The inputs of the network were body joint heatmaps and body-limb maps generated from 2D RGB images, and the ground truth poses were produced by applying Diffusion Maps, which is a manifold learning technique, on the motion capture data. The accuracy of the proposed method and the efficiency of the estimated pose features for human movement assessment both were examined on the SMAD dataset. The experiments showed that the proposed method can extract the pose features well on single view data even though multiple views are employed for training. However, it cannot be applied to novel viewpoints as it has not been designed explicitly to tolerate the viewpoint variations in unseen view data. Furthermore, although the method removes the need for 3D skeleton data during inference, it still requires motion capture data for training.

To overcome the limitations of the proposed approach in Chapter 4, Chapter 5 explored a method to be capable to assess the quality of human movements from novel viewpoints while it is trained on only one or two views from RGB images. In the end-to-end proposed network, first, a view-invariant spatio-temporal trajectory map is extracted for each body joint, and then the relationship amongst the joint trajectories is exploited to estimate a score for the quality of movement. The performance of the proposed view-invariant network was examined on all four action types of QMAR (W-P, W-S, SS-P, and SS-S) under cross-view and cross-subject protocols, and on all five action types of KIMORE (Ex #1 to Ex #5) under a single-view training scenario. The experiments demonstrated that the proposed method outperforms the baselines in all scenarios on average rank correlation results. However, the method's performance decreases significantly when long-term occlusion happens since its accuracy heavily relies on the quality of the joint heatmaps employed as input.

Finally, Chapter 6 shifted the focus on unsupervised and transfer learning. It developed a self-supervised network that learns view-invariant 3D pose representation without relying on any 3D joint annotations and is trained by exploiting the intrinsic view-invariant properties between simultaneous frames from distinct views, and equivariant properties of augmented frames from the same viewpoint. The proposed method was trained on NTU RGB+D, and evaluated for cross-view and cross-subject action recognition accuracy. Then, the view-invariant 3D pose features learned from NTU RGB+D and action recognition, were transferred into the human movement assessment domain where their efficiency was examined for cross-view, cross-subject and single-view movement quality assessment on QMAR and KIMORE. The experiments showed the benefits of self-supervised learning. However, the proposed method requires synchronized multi-view data in the pretext stage which limits the potential multi-view datasets that can be

applied for training.

7.2 Directions for Future Works

This thesis focused on view-invariance in human movement assessment. The findings of this thesis are initial explorations in this area. As such, there are several interesting avenues for future work in view-invariance in movement quality assessment and the other areas of action understanding field. Four of these directions are outlined bellow.

General Methods for Movement Quality Assessment – In Chapters 4 to 6, all the proposed approaches are trained separately for different action types to learn features that are related to a specific movement type. Future work could investigate the development of a model to be trained on all movement types at the same time. Then, it examines whether the model can still maintain a high accuracy on individual tasks, or explores how the shared representation between different tasks can improve the method's performance on specific tasks.

Learning from Unsynchronized Multi-View Images – Chapter 6 benefits synchronized multi-view frames to learn 3D view-invariant features. However, capturing multi-view datasets recorded by multiple synchronized cameras is challenging and costly. This also limits the number of potential existing multi-view datasets for training. For instance, Ji et al. [59] introduce a large-scale multi-view dataset (UESTC) which has recorded RGB-D videos with entire 360° view angles, but the multi-view sequences have not been captured simultaneously. The dataset has been recorded by two cameras such that one of them is fixed (frontal view) and the other moves to record the rest of the view angels. Another potential venue is to develop an approach that while benefiting multi-view training to learn the 3D view-invariant features, removes the need for synchronized multi-view data.

Unsupervised Synthetic Multi-View Video Generation – To tackle view-invariance, one solution is to generate synthetic multi-view videos. The current approaches, such as [87, 141], require 3D pose and use graphical techniques to generate the synthetic data. However, recent advances in Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs), and Neural Radiance Fields (NeRFs) can be utilized to produce realistic and diverse synthetic multi-view data with less or no supervision.

View-Invariance Beyond Movement Quality Assessment – The issue of view-invariance, which causes the method's performance to drop significantly when applied to unseen view data, is not unique to area of human movement assessment. This issue is also

significant in many other domains of the computer vision field, such as human action localization, prediction, and human-object interaction. To tackle the view-invariance in movement quality assessment, Chapters 4, 5, and 6 proposed approaches to extract view-invariant spatial or spatio-temporal features for unseen view human movement assessment. As the proposed methods are general and not task-related, they can be adapted and applied to different applications. Therefore, another future direction could be to explore adapting the proposed approaches in this thesis for other computer vision tasks and examine their performance in the new domains.

References

[1] J. Albert, P. Glöckner, B. Pfitzner, and B. Arnrich. Data Augmentation of Kinematic Time-Series from Rehabilitation Exercises Using GANs. In *Proceedings of the IEEE International Conference on Omni-Layer Intelligent Systems*, pages 1–6, 2021. 18, 19

[2] B. Bah. Diffusion Maps: Analysis and Applications. 2008. 53

[3] R. Baptista, E. Ghorbel, F. Moissenet, D. Aouada, A. Douchet, M. André, J. Pager, S. Bouilland, et al. Home Self-Training: Visual Feedback for Assisting Physical Activity for Stroke Survivors. *Computer Methods and Programs in Biomedicine*, 176:111–120, 2019. 1, 2, 49

[4] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. In *International Conference on Medical Image Computing and Computer Assisted Intervention Brainlesion Workshop*, pages 161–169. Springer, 2018. 95

[5] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003. 50

[6] G. Bertasius, H. Soo Park, S. X. Yu, and J. Shi. Am I a Baller? Basketball Performance Assessment from First-Person Videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2177–2185, 2017. 8, 15

[7] V. Bijalwan, V. B. Semwal, and T. Mandal. Fusion of Multi-Sensor-based Biomechanical Gait Analysis Using Vision and Wearable Sensor. *IEEE Sensors Journal*, 21(13):14213–14220, 2021. 1

[8] C. M. Bishop et al. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995. 56

[9] X. Bruce, Y. Liu, and K. C. Chan. Skeleton-based Detection of Abnormalities in Human Actions Using Graph Convolutional Networks. In *International Conference on Transdisciplinary AI*, pages 131–137, 2020. 1, 2, 19, 49

[10] X. Bruce, Y. Liu, K. C. Chan, Q. Yang, and X. Wang. Skeleton-based Human Action Evaluation Using Graph Convolutional Network for Monitoring Alzheimer's Progression. *Pattern Recognition*, page 108095, 2021. 19, 36, 37

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 54, 55, 73, 84

[12] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriù, L. Romeo, and F. Verdini. The KIMORE Dataset: Kinematic Assessment of Movement and Clinical Scores for Remote Monitoring of Physical Rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7):1436–1448, 2019. 6, 17, 18, 19, 36, 46, 66

[13] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 20, 73, 75, 77, 78, 81, 83, 106, 107, 108, 109, 110, 111

[14] C. Chambers, N. Seethapathi, R. Saluja, H. Loeb, S. R. Pierce, D. K. Bogen, L. Prosser, M. J. Johnson, and K. P. Kording. Computer Vision to Automatically Assess Infant Neuromotor Risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442, 2020. 19

[15] N. Chatlani and J. J. Soraghan. Local Binary Patterns for 1-D Signal Processing. In *Proceedings of the European Signal Processing Conference*, pages 95–99, 2010. 18

[16] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, S. Stojanov, and J. Rehg. Unsupervised 3D Pose Estimation With Geometric Self-Supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019. 28, 33, 87

[17] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to Scale: Scale-Aware Semantic Image Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016. 9

[18] X. Chen, K.-Y. Lin, W. Liu, C. Qian, and L. 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*, pages 10895–10904, 2019. 28, 87

[19] Y.-B. Cheng, X. Chen, J. Chen, P. Wei, D. Zhang, and L. Lin. Hierarchical Transformer: Unsupervised Representation Learning for Skeleton-based Human Action Recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021. 100

[20] S. H. Chowdhury, M. Al Amin, A. M. Rahman, M. A. Amin, and A. A. Ali. Assessment of Rehabilitation Exercises from Depth Sensor Data. In *International Conference on Computer and Information Technology*, pages 1–7, 2021. 17, 19

[21] R. R. Coifman and S. Lafon. Diffusion Maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. 16, 50, 53

[22] B. Crabbe, A. Paiement, S. Hannuna, and M. Mirmehdi. Skeleton-Free Body Pose Estimation from Depth Images for Movement Analysis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 70–78, 2015. 1, 2, 16, 19, 40, 45, 53, 56, 57, 58, 59, 60, 62, 63, 64

[23] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 70

[24] S. Deb, M. F. Islam, S. Rahman, and S. Rahman. Graph Convolutional Networks for Assessment of Physical Rehabilitation Exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022. 19

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 56, 73

[26] C. Dhiman and D. K. 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. 25, 26, 29

[27] L.-J. Dong, H.-B. Zhang, Q. Shi, Q. Lei, J.-X. Du, and S. Gao. Learning and Fusing Multiple Hidden Substages for Action Quality Assessment. *Knowledge-based Systems*, 229:107388, 2021. 15

[28] M. Dorkenwald, U. Buchler, and B. Ommer. Unsupervised Magnification of Posture Deviations Across Subjects. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 8256–8266, 2020. 95

[29] A. D. P. dos Santos, L. Loke, K. Yacef, and R. Martinez-Maldonado. Enriching Teachers' Assessments of Rhythmic Forró Dance Skills by Modelling Motion Sensor Data. *International Journal of Human-Computer Studies*, 161:102776, 2022. 1

[30] H. Doughty, D. Damen, and W. Mayol-Cuevas. Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6057–6066, 2018. 15

[31] H. Doughty, W. Mayol-Cuevas, and D. 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*, pages 7862–7871, 2019. 1, 15, 75

[32] H. Doughty, I. Laptev, W. Mayol-Cuevas, and D. Damen. Action Modifiers: Learning from Adverbs in Instructional Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 868–878, 2020. 75

[33] C. Du, S. Graham, C. Depp, and T. Nguyen. Assessing Physical Rehabilitation Exercises Using Graph Convolutional Network with Self-Supervised Regularization. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 281–285, 2021. 19

[34] A. Dundar, K. Shih, A. Garg, R. Pottorff, A. Tao, and B. Catanzaro. Unsupervised Disentanglement of Pose, Appearance and Background from Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3883–3894, 2022. 28, 87

[35] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Temporal Cycle-Consistency Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 13

[36] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba. A General Descriptor for Detecting Abnormal Action Performance from Skeletal Data. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1401–1404, 2017. 19, 50

[37] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen, and E. Saba. Efficient and Robust Skeleton-based Quality Assessment and Abnormality Detection in Human Action Performance. *IEEE Journal of Biomedical and Health Informatics*, 24(1):280–291, 2019. 3, 16, 19, 36, 37, 50

[38] P. Esser, E. Sutter, and B. Ommer. A Variational U-Net for Conditional Appearance and Shape Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 95

[39] A. Farhadi and M. K. Tabrizi. Learning to Recognize Activities from the Wrong View Point. In *Proceedings of the European Conference on Computer Vision*, pages 154–166. Springer, 2008. 21

[40] C. Feichtenhofer, H. Fan, J. Malik, and K. He. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6202–6211, 2019. 1, 20

[41] L. Gao, Y. Ji, G. A. Kumie, X. Xu, X. Zhu, and H. T. Shen. View-Invariant Human Action Recognition via View Transformation Network. *IEEE Transactions on Multimedia Journal*, page 1, 2021. 24

[42] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling. In *International Conference on Medical Image Computing and Computer Assisted Intervention Workshop*, volume 3, page 3, 2014. 14, 15

[43] K. Gedamu, Y. Ji, Y. Yang, L. Gao, and H. T. Shen. Arbitrary-View Human Action Recognition via Novel-View Action Generation. *Pattern Recognition*, 118:108043, 2021. 24

[44] E. Ghorbel, K. Papadopoulos, R. Baptista, H. Pathak, G. Demisse, D. Aouada, and B. Ottersten. A View-Invariant Framework for Fast Skeleton-based Action Recognition Using a Single RGB Camera. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2019. 25, 26, 29

[45] K. Gong, J. Zhang, and J. Feng. PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8575–8584, 2021. 28

[46] A. S. Gordon. Automated Video Assessment of Human Performance. In *Proceedings of the International Conference on Artificial Intelligence in Education*, volume 2, page 16–19, 1995. 8, 15

[47] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei. Towards Viewpoint Invariant 3D Human Pose Estimation. In *Proceedings of the European Conference on Computer Vision*, pages 160–177. Springer, 2016. 33

[48] K. Hara, H. Kataoka, and Y. Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 28

[49] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 56

[50] S. Honari, V. Constantin, H. Rhodin, M. Salzmann, and P. Fua. Unsupervised Learning on Monocular Videos for 3D Human Pose Estimation. *arXiv preprint arXiv:2012.01511*, 2021. 28, 32, 33, 87

[51] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5344–5352, 2015. 24

[52] G. Hua, W. Li, Q. Zhang, R. Ding, and H. Liu. Weakly-Supervised Cross-View 3D Human Pose Estimation. *arXiv preprint arXiv:2105.10882*, 2021. 28

[53] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1661–1668, 2014. 33

[54] U. Iqbal, P. Molchanov, T. B. J. Gall, and J. Kautz. Hand Pose Estimation via Latent 2.5D Heatmap Regression. In *Proceedings of the European Conference on Computer Vision*, pages 118–134. Springer, 2018. 30

[55] U. Iqbal, P. Molchanov, and J. Kautz. Weakly-Supervised 3D Human Pose Learning via Multi-View Images in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5243–5252, 2020. 28, 30, 33

[56] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 70, 71, 73

[57] H. Jain and G. Harit. An Unsupervised Sequence-to-Sequence Autoencoder based Human Action Scoring Model. In *IEEE Global Conference on Signal and Information Processing*, pages 1–5, 2019. 13, 15

[58] H. Jain, G. Harit, and A. Sharma. Action Quality Assessment Using Siamese Network-based Deep Metric Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2260–2273, 2020. 15

[59] Y. Ji, F. Xu, Y. Yang, F. Shen, H. T. Shen, and W.-S. Zheng. A Large-Scale RGB-D Database for Arbitrary-View Human Action Recognition. In *Proceedings of the ACM international Conference on Multimedia*, pages 1510–1518, 2018. 21, 24, 29, 116

[60] Y. Ji, Y. Yang, H. T. Shen, and T. Harada. View-Invariant Action Recognition via Unsupervised Attention Transfer (UANT). *Pattern Recognition*, 113:107807, 2021. 22, 24, 67, 77

[61] S. Johnson and M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *British Machine Vision Conference*, volume 2, page 5. Citeseer, 2010. 33

[62] M. Jug, J. Perš, B. Dežman, and S. Kovačič. Trajectory based Assessment of Coordinated Human Activity. In *International Conference on Computer Vision Systems*, pages 534–543. Springer, 2003. 8

[63] I. N. Junejo, E. Dexter, I. Laptev, and P. Púrez. Cross-View Action Recognition from Temporal Self-Similarities. In *Proceedings of the European Conference on Computer Vision*, pages 293–306. Springer, 2008. 21

[64] M. E. Kalfaoglu, S. Kalkan, and A. A. Alatan. Late Temporal Modeling in 3D CNN Architectures with Bert for Action Recognition. In *Proceedings of the European Conference on Computer Vision*, pages 731–747. Springer, 2020. 1, 20

[65] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3D Human Dynamics from Video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 28

[66] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. 75, 78, 81, 83, 105, 106, 107, 108, 109, 110, 111

[67] M. Khokhlova, C. Migniot, A. Morozov, O. Sushkova, and A. Dipanda. Normal and Pathological Gait Classification LSTM Model. *Artificial Intelligence in Medicine*, 94:54–66, 2019. 49

[68] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 96

[69] M. Kocabas, S. Karagoz, and E. Akbas. Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019. 73

[70] M. Kocabas, N. Athanasiou, and M. J. Black. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 28

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012. 16, 56

[72] J. N. Kundu, S. Seth, M. Rahul, M. Rakesh, V. B. Radhakrishnan, and A. Chakraborty. Kinematic-Structure-Preserved Representation for Unsupervised 3D Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11312–11319, 2020. 28

[73] M. I. Lakhal, O. Lanz, and A. Cavallaro. View-LSTM: Novel-View Video Synthesis Through View Decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7577–7587, 2019. 67

[74] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal Convolutional Networks for Action Segmentation and Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 9

[75] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen. A Survey of Vision-based Human Action Evaluation Methods. *Sensors*, 19(19):4129, 2019. 1

[76] Q. Lei, H.-B. Zhang, J.-X. Du, T.-C. Hsiao, and C.-C. Chen. Learning Effective Skeletal Representations on RGB Video for Fine-Grained Human Action Quality Assessment. *Electronics*, 9(4):568, 2020. 1

[77] C. Li and G. H. Lee. Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9887–9895, 2019. 28

[78] J. Li, Y. Wong, Q. Zhao, and M. Kankanhalli. Unsupervised Learning of View-Invariant Action Representations. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1254–1264, 2018. 25, 26, 29, 34, 67, 97, 98, 99, 100, 105, 106, 107, 108, 109, 110, 111

[79] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4750, 2021. 23, 24

[80] W. Li, Z. Zhang, and Z. Liu. Action Recognition based on a Bag of 3D Points. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 9–14, 2010. 24

[81] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool. Domain Generalization and Adaptation Using Low Rank Exemplar SVMs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1114–1127, 2017. 67

[82] Y. Li, X. Chai, and X. Chen. End-to-End Learning for Action Quality Assessment. In *Proceedings of the Pacific Rim Conference on Multimedia*, pages 125–134, 2018. 15, 74, 97

[83] Y. Liao, A. Vakanski, and M. Xian. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28:468–477, 2019. 1, 2

[84] Y. Liao, A. Vakanski, and M. Xian. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):468–477, 2020. 1, 3, 17, 18, 19

[85] L. Lin, S. Song, W. Yang, and J. Liu. MS2L: Multi-Task Self-Supervised Learning for Skeleton based Action Recognition. In *Proceedings of the ACM International Conference on Multimedia*, pages 2490–2498, 2020. 100

[86] D. Liu, Q. Li, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li. Towards Unified Surgical Skill Assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2021. 1

[87] J. Liu, H. Rahmani, N. Akhtar, and A. Mian. Learning Human Pose Models from Synthesized Data for Robust RGB-D Action Recognition. *International Journal of Computer Vision*, 127(10):1545–1564, 2019. 25, 27, 28, 29, 67, 116

[88] M. Liu, H. Liu, and C. Chen. Enhanced Skeleton Visualization for View Invariant Human Action Recognition. *Pattern Recognition*, 68:346–362, 2017. 21, 24, 67, 77

[89] S. Liu, G. Ren, Y. Sun, J. Wang, C. Wang, B. Li, and S. Yan. Fine-Grained Human-Centric Tracklet Segmentation with Single Frame Supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 24

[90] Y. Liu, Z. Lu, J. Li, and T. Yang. Hierarchically Learned View-Invariant Representations for Cross-View Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2416–2430, 2018. 29

[91] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei. Unsupervised Learning of Long-Term Motion Dynamics for Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2212, 2017. 98, 99

[92] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In *International Conference on 3D Vision*, pages 506–516, 2017. 33

[93] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-Time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics*, 36(4):1–14, 2017. 26

[94] I. Misra, L. Zitnick, and M. Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *Proceedings of the European Conference on Computer Vision*, pages 527–544. Springer, 2016. 98, 99

[95] M. Nekoui and L. Cheng. Enhancing Human Motion Assessment by Self-supervised Representation Learning. *British Machine Vision Conference*, 2021. 19, 20, 105, 111, 112

[96] M. Nekoui, F. O. T. Cruz, and L. Cheng. FALCONS: Fast Learner-Grader for Contorted Poses in Sports. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 900–901, 2020. 1

[97] N. Neverova, D. Novotny, and A. Vedaldi. Correlated Uncertainty for Learning Dense Correspondences from Noisy labels. *Proceedings of the Advances in Neural Information Processing Systems*, 32, 2019. 89

[98] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, and M. Mirmehdi. Online Quality Assessment of Human Movement from Skeleton Data. In *British Machine Vision Conference*, pages 153–166, 2014. 1, 3, 16, 17, 19, 35, 36, 49, 50, 52, 53, 61, 63, 73, 134

[99] J.-H. Pan, J. Gao, and W.-S. Zheng. Action Assessment by Joint Relation Graphs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6331–6340, 2019. 1, 2, 74, 97

[100] J.-H. Pan, J. Gao, and W.-S. Zheng. Adaptive Action Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 15

[101] G. Paoletti, J. Cavazza, C. Beyan, and A. Del Bue. Unsupervised Human Action Recognition with Skeletal Graph Laplacian and Self-Supervised Viewpoints Invariance. *British Machine Vision Conference*, 2021. 23, 24

[102] P. Parmar and B. Morris. Action Quality Assessment Across Multiple Actions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1468–1476, 2019. 1, 2, 6, 10, 15, 73

[103] P. Parmar and B. T. Morris. Learning to Score Olympic Events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2017. 2, 9, 15

[104] P. Parmar and B. T. 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*, pages 304–313, 2019. 1, 2, 6, 11, 14, 15, 73, 74, 75, 77, 78, 81, 83, 97, 106, 107, 108, 109, 110, 111

[105] P. Parmar, J. Reddy, and B. Morris. Piano Skills Assessment. *In Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, pages 1–5, 2021. 1, 2

[106] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6988–6997, 2017. 28

[107] A. Piergiovanni and M. S. Ryoo. Recognizing Actions in Videos from Unseen Viewpoints. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4124–4132, 2021. 2

[108] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the Quality of Actions. In *Proceedings of the European Conference on Computer Vision*, pages 556–571. Springer, 2014. 1, 2, 8, 9, 15

[109] Z. Qiu, T. Yao, and T. Mei. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 9

[110] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Histogram of Oriented Principal Components for Cross-View Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2430–2443, 2016. 19, 21, 24, 29, 47

[111] M. J. Raihan, M. A. R. Ahad, and A.-A. Nahid. Automated Rehabilitation Exercise Assessment by Genetic Algorithm-optimized CNN. In *Joint International Conference on Informatics, Electronics & Vision and International Conference on Imaging, Vision & Pattern Recognition*, pages 1–6, 2021. 2, 18, 19

[112] S. Ramagiri, R. Kavi, and V. Kulathumani. Real-Time Multi-View Human Action Recognition Using a Wireless Camera Network. In *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–6, 2011. 29

[113] C. Rao, A. Yilmaz, and M. Shah. View-Invariant Representation and Recognition of Actions. *International Journal of Computer Vision*, 50(2):203–226, 2002. 21

[114] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu. Augmented Skeleton based Contrastive Action Learning with Momentum LSTM for Unsupervised Action Recognition. *Information Sciences*, 569:90–109, 2021. 23, 24, 100

[115] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation. In *Proceedings of the European Conference on Computer Vision*, pages 750–767, 2018. 28, 31, 33, 87, 95, 96

[116] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. 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. 30, 33, 34, 87

[117] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 2007. 56

[118] K. Roditakis, A. Makris, and A. Argyros. Towards Improved and Interpretable Action Quality Assessment with Self-Supervised Alignment. In *The Pervasive Technologies Related to Assistive Environments Conference*, pages 507–513, 2021. 1, 2, 13, 14, 15

[119] O. Ronneberger, P. Fischer, and T. 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. 95

[120] F. Sardari, A. Paiement, and M. 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. 36, 37, 49

[121] F. Sardari, A. Paiement, S. Hannuna, and M. Mirmehdi. VI-Net—View-Invariant Quality of Human Movement Assessment. *Sensors*, 20(18):5258, 2020. 36, 42, 66

[122] F. Sardari, B. Ommer, and M. Mirmehdi. Unsupervised View-Invariant Human Posture Representation. *British Machine Vision Conference*, 2021. 86

[123] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala. Recognizing Actions from Depth Cameras as Weakly Aligned Multi-Part Bag-of-Poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 479–485, 2013. 24

[124] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. Time-contrastive networks: Self-supervised learning from video. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1134–1141, 2018. 33

[125] A. Shafaei and J. J. Little. Real-Time Human Motion Capture with Multiple Depth Cameras. In *Proceedings of the IEEE Conference on Computer and Robot Vision*, pages 24–31, 2016. 33

[126] A. Shahroudy, J. Liu, T.-T. Ng, and G. 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*, pages 1010–1019, 2016. 21, 24, 29, 34, 47, 85, 95, 96

[127] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, 28, 2015. 97

[128] B. K. Shukla, H. Jain, V. Vijay, S. K. Yadav, A. Mathur, and D. J. Hewson. A Comparison of Four Approaches to Evaluate the Sit-to-Stand Movement. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(6):1317–1324, 2020. 1, 2, 49

[129] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014. 56, 73

[130] S. Singh, S. A. Velastin, and H. Ragheb. MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods. In *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance*, pages 48–55, 2010. 29

[131] K. Su, X. Liu, and E. Shlizerman. Predict and Cluster: Unsupervised Skeleton based Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 100

[132] J. J. Sun, J. Zhao, L.-C. Chen, F. Schroff, H. Adam, and T. Liu. View-Invariant Probabilistic Embedding for Human Pose. In *Proceedings of the European Conference on Computer Vision*, pages 53–70. Springer, 2020. 87

[133] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 26

[134] K. S. Tai, P. Bailis, and G. Valiant. Equivariant Transformer Networks. In *Proceedings of the International Conference on Machine Learning*, pages 6086–6095, 2019. 70

[135] S. A. W. Talha, M. Hammouche, E. Ghorbel, A. Fleury, and S. Ambellouis. Features and Classification Schemes for View-Invariant and Real-Time Human Action Recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4): 894–902, 2018. 24

[136] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. 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. 1, 2, 11, 13, 14, 15

[137] L. Tao, A. Paiement, D. Damen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt, and I. Craddock. A Comparative Study of Pose Representation and Dynamics Modelling for Online Motion Quality Assessment. *Computer Vision and Image Understanding*, 148:136–152, 2016. 1, 3, 16, 19, 35, 36, 52, 53, 63

[138] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *science*, 290(5500):2319–2323, 2000. 50

[139] S. Tripathi, S. Ranade, A. Tyagi, and A. Agrawal. PoseNet3D: Learning Temporally Consistent 3D Human Pose via Knowledge Distillation. In *International Conference on 3D Vision*, pages 311–321, 2020. 28, 29, 30, 33, 87

[140] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker. A Data Set of Human Body Movements for Physical Rehabilitation Exercises. *Data*, 3(1):2, 2018. 17, 19, 36

[141] G. Varol, I. Laptev, C. Schmid, and A. Zisserman. Synthetic Humans for Action Recognition from Unseen Viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021. 20, 21, 25, 28, 29, 67, 77, 116

[142] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In *Proceedings of the European Conference on Computer Vision*, pages 601–617, 2018. 33

[143] S. Vyas, Y. S. Rawat, and M. Shah. Multiview Action Recognition Using Cross-View Video Prediction. In *Proceedings of the European Conference on Computer Vision*, pages 427–444. Springer, 2020. 25, 26, 27, 29, 34, 77, 97, 98, 99, 100

[144] D. Wang, W. Ouyang, W. Li, and D. Xu. Dividing and Aggregating Network for Multi-View Action Recognition. In *Proceedings of the European Conference on Computer Vision*, pages 451–467. Springer, 2018. 25, 29, 34, 67

[145] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012. 26, 29

[146] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-View Action Modeling, Learning and Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014. 21, 24, 26, 29, 47

[147] J. Wang, Z. Du, A. Li, and Y. Wang. Assessing Action Quality via Attentive Spatio-Temporal Convolutional Networks. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 3–16. Springer, 2020. 15

[148] J. Wang, J. Jiao, and Y.-H. Liu. Self-Supervised Video Representation Learning by Pace Prediction. In *Proceedings of the European Conference on Computer Vision*, pages 504–521. Springer, 2020. 20

[149] J. Wang, S. Yan, Y. Xiong, and D. Lin. Motion Guided 3D pose Estimation from Videos. In *Proceedings of the European Conference on Computer Vision*, pages 764–780. Springer, 2020. 28

[150] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Proceedings of the European Conference on Computer Vision*, pages 20–36. Springer, 2016. 29

[151] T. Wang, M. Jin, and M. Li. Towards Accurate and Interpretable Surgical Skill Assessment: A Video-based Method for Skill Score Prediction and Guiding Feedback Generation. *International Journal of Computer Assisted Radiology and Surgery*, 16(9):1595–1605, 2021. 1

[152] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 26

[153] D. Weinland, E. Boyer, and R. Ronfard. Action Recognition from Arbitrary Views Using 3D Exemplars. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–7, 2007. 21, 29, 47

[154] C.-Y. Wu and P. Krahenbuhl. Towards Long-Form Video Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 1, 20

[155] L. Xia, C.-C. Chen, and J. K. Aggarwal. View Invariant Human Action Recognition Using Histograms of 3D Joints. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27, 2012. 19, 24

[156] X. Xiang, Y. Tian, A. Reiter, G. D. Hager, and T. D. Tran. S3D: Stacking Segmental P3D for Action Quality Assessment. In *Proceedings of the IEEE International Conference on Image Processing*, pages 928–932, 2018. 9, 15

[157] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 73

[158] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810, 2015. 9

[159] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue. Learning to Score Figure Skating Sport Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4578–4590, 2019. 15

[160] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang. Deep Kinematics Analysis for Monocular 3D Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2020. 28

[161] S. Yan, Y. Xiong, and D. Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition. In *AAAI Conference on Artificial Intelligence*, page 7444–7452, 2018. 22

[162] H. Yao, S. Zhao, C. Xie, K. Ye, and S. Liang. Recurrent Graph Convolutional Autoencoder for Unsupervised Skeleton-based Action Recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021. 100

[163] A. Yilmaz and M. Shah. Actions Sketch: A Novel Action Representation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 984–989, 2005. 21

[164] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou. Group-Aware Contrastive Regression for Action Quality Assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7919–7928, 2021. 12, 13

[165] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-Person Interaction Detection Using Body-Pose Features and Multiple Instance Learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2012. 23, 24

[166] L.-A. Zeng, F.-T. Hong, W.-S. Zheng, Q.-Z. Yu, W. Zeng, Y.-W. Wang, and J.-H. Lai. Hybrid Dynamic-Static Context-Aware Attention Network for Action Assessment in Long Videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 2526–2534, 2020. 14

[167] B. Zhang, Q. Qiang, F. Wang, and F. Nie. Flexible Multi-View Unsupervised Graph Embedding. *IEEE Transactions on Image Processing*, 30:4143–4156, 2021. 50

[168] C. Zhang, H. Zheng, and J. Lai. Cross-View Action Recognition based on Hierarchical View-Shared Dictionary Learning. *IEEE Access*, 6:16855–16868, 2018. 29

[169] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126, 2017. 22, 24, 67, 77

[170] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1963–1978, 2019. 22, 24, 67, 77

[171] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng. Semi-Supervised Action Quality Assessment with Self-Supervised Segment Feature Recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2022. 2, 14, 15

[172] X. Zhang, Y. Wong, M. S. Kankanhalli, and W. Geng. Unsupervised Domain Adaptation for 3D Human Pose Estimation. In *Proceedings of the ACM International Conference on Multimedia*, pages 926–934, 2019. 31, 32, 33

[173] Z. Zhang and H. Zha. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004. 50

[174] L. Zhao, Y. Wang, J. Zhao, L. Yuan, J. J. Sun, F. Schroff, H. Adam, X. Peng, D. Metaxas, and T. Liu. Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12793–12802, 2021. 23, 87

[175] Y. Zhao, X. You, S. Yu, C. Xu, W. Yuan, X.-Y. Jing, T. Zhang, and D. Tao. Multi-View Manifold Learning with Locality Alignment. *Pattern Recognition*, 78:154–166, 2018. 50

[176] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong. Unsupervised Representation Learning with Long-term Dynamics for Skeleton based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, page 2644–2651, 2018. 20

[177] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable ConvNets v2: More Deformable, Better Results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 70

[178] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa. Video and Accelerometer-based Motion Analysis for Automated Surgical Skills Assessment. *International Journal of Computer Assisted Radiology and Surgery*, 13(3):443–455, 2018. 1

[179] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. 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. 94

Appendix A

Deviation of Equation for $f_{P_t}(p_t|p_1,...,p_{t-1})$ – It may be computed as

$$f_{P_t}(p_t|p_1,...,p_{t-1}) = \frac{f_{\mathbb{P}^t}(p_1,...,p_t)}{f_{\mathbb{P}^{t-1}}(p_1,...,p_{t-1})}, \tag{1}$$

where $f_{\mathbb{P}^t}(p_1,...,p_t) = \int_{\Omega_{\mathbb{S}^t}} f_{\mathbb{P}^t,\mathbb{S}^t}(p_1,...,p_t,s_0,...,s_t)$, $\mathbb{P}^t$ denotes $\{p_1,...,p_t\}$, $\mathbb{S}^t$ denotes $\{s_0,...,s_t\}$, and $\Omega_{\mathbb{S}^t}$ refers to the possible values for $\mathbb{S}^t$.

If the following Markovian assumptions are used

$$f_{P_t}(p_t|p_1,...,p_{t-1},s_0,...,s_t) = f_{P_t}(p_t|s_t), \tag{2}$$
$$f_{S_t}(s_t|s_0,...,s_{t-1}) = f_{S_t}(s_t|s_{t-1}), \tag{3}$$

then,

$$\begin{aligned} f_{\mathbb{P}^t,\mathbb{S}^t}(p_1,...,p_t,s_0,...,s_t) &= f_{P_t}(p_t|p_1,...,p_{t-1},s_0,...,s_t) f_{\mathbb{P}^{t-1},\mathbb{S}^t}(p_1,...,p_{t-1},s_0,...,s_t) \\ &= f_{P_t}(p_t|s_t) f_{S_t}(s_t|p_1,...,p_{t-1},s_0,...,s_{t-1}) \\ &\quad f_{\mathbb{P}^{t-1},\mathbb{S}^{t-1}}(p_1,...,p_{t-1},s_0,...,s_{t-1}) \\ &= f_{P_t}(p_t|s_t) f_{S_t}(s_t|s_{t-1}) f_{\mathbb{P}^{t-1},\mathbb{S}^{t-1}}(p_1,...,p_{t-1},s_0,...,s_{t-1}) \\ &\vdots \\ &= f_{S_0}(s_0) \prod_{i=1}^{t} f_{P_i}(p_i|s_i) f_{S_i}(s_i|s_{i-1}), \end{aligned} \tag{4}$$

and Equation 1 becomes

$$f_{P_t}(p_t|p_1,...,p_{t_1}) = \frac{\int_{\Omega_{\mathbb{S}^t}} f_{S_0}(s_0) \prod_{i=1}^{t} f_{P_i}(p_i|s_i) f_{S_i}(s_i|s_{i-1})}{\int_{\Omega_{\mathbb{S}^{t-1}}} f_{S_0}(s_0) \prod_{i=1}^{t-1} f_{P_i}(p_i|s_i) f_{S_i}(s_i|s_{i-1})}. \tag{5}$$

In [98], it is assumed that there is only one acceptable value for S_t that satisfies the constraints on the hidden state, *i.e.* the hidden state linearly increases during a normal

sequence, and it is computed as

$$\begin{aligned}\hat{\mathbb{S}^t} = \{\hat{s_0}, ..., \hat{s_t}\} &= \underset{\{s_0,...,s_t\}}{\text{argmax}}\, f_{\mathbb{S}^t}(s_0, ..., s_t, p_1, ..., p_t) \\ &= \underset{\{s_0,...,s_t\}}{\text{argmax}}\, \frac{f_{\mathbb{P}^t,\mathbb{S}^t}(p_1, ..., p_t, s_0, ..., s_t)}{f_{\mathbb{P}^t}(p_1, ..., p_t)} \\ &= \underset{\{s_0,...,s_t\}}{\text{argmax}}\, f_{S_0}(s_0) \prod_{i=1}^{t} f_{P_i}(p_i|s_i) f_{S_i}(s_i|s_{i-1}).\end{aligned} \tag{6}$$

Therefor, Equation 5 may be simplified as

$$\begin{aligned}f_{P_t}(p_t|p_1, ..., p_{t_1}) &\approx \frac{f_{S_0}(\hat{s}_0) \prod_{i=1}^{t} f_{P_i}(p_i|\hat{s}_i) f_{S_i}(\hat{s}_i|\hat{s}_{i-1})}{f_{S_0}(\hat{s}_0) \prod_{i=1}^{t-1} f_{P_i}(p_i|\hat{s}_i) f_{S_i}(\hat{s}_i|\hat{s}_{i-1})} \\ &\approx f_{P_t}(p_t|\hat{s}_t) f_{S_t}(\hat{s}_t|\hat{s}_{t-1}).\end{aligned} \tag{7}$$