Intelligent Video-based Quality Assessment of Human Activities

by

Mahdiar Nekoui

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science
in
Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering

University of Alberta

Abstract

Can AI-driven robots replace sports officials and rehabilitation physicians in assessing the quality of human activities? An AI judging panel can attend to every nuance of an athlete's performance and bring more just and agility in scoring. The subjectivity of the judging will be removed and no one will heckle the referee anymore. A trained system can monitor the activities of movement disorder patients and assess their progress in the rehabilitation exercise program. There would no longer be any need for the patients to travel to specialized urban health centers and get feedback on their progress. A computer constantly tracks their movements at home and evaluates their performance.

The goal of this thesis is to develop some approaches to automatically evaluate the quality of humans' activities from just the video of their performances. The proposed approaches are mostly based on convolutional neural networks that have demonstrated their effectiveness in analyzing visual imagery.

We first start with grading a diving routine. A human judge keeps track of the coordination among the joints throughout the performance as well as appearance features like amount of splash and smoothness of the flight. The execution score that the referee provides is then multiplied by the routine's degree of difficulty which is obtained from the official benchmark based on the components of the routine. Inspired by this grading schema, we propose a virtual refereeing system that involves both pose and appearance features of a routine in assessing its execution. On the other hand, a difficulty assessor extracts the components of the flight based on the evolution of the pose throughout the routine. Finally, the overall score is reported by multiplying the difficulty and execution scores.

We then extend the AI-driven assessor to be applicable to not only other short-term sports like gym-vault and skiing but also minute-long activities like figure-skating. We propose a modular two-stream network that attends to the hierarchical temporal structure of sports routines and can be easily adapted to score minute-long activities. Both fine and coarse-grained temporal dependencies of pose and appearance features are involved in the assessment procedure. In such contortive sports, the athletes usually look for new angles and turns to configure their bodies in some unusual poses. Such pose configurations are not covered by existing pose datasets. Therefore, we introduce a pose-annotated dataset of extreme poses to support experiments in estimating human body pose with extreme contortions and involving pose features in action quality assessment (AQA).

We finally explore the application of AQA in rehabilitation assessment. We develop a self-supervised method that learns the symmetricity and pace of a normal action from off-the-shelf datasets of healthy people activities. These transferable features are then used to assess the activities of a movement disorder patient based on how impaired and slow they are. The proposed method not only demonstrated superior performance in rehabilitation progress evaluation but also showed a good generalization to infants' general movements assessment.

Preface

Below is the list of our publications based on this dissertation.

- **Mahdiar Nekoui**, Fidel Omar Tito Cruz, and Li Cheng. FALCONS: Fast learner-grader for contorted poses in sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 900–901, 2020.

- **Mahdiar Nekoui**, Fidel Omar Tito Cruz, and Li Cheng. EAGLE-Eye: Extreme-pose action grader using detail bird's-eye view. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2021.

- **Mahdiar Nekoui** and Li Cheng. Enhancing Human Motion Assessment by Self-supervised Representation Learning. In *British Machine Vision Conference (BMVC)* 2021.

To my family

Intelligence is the ability to adapt to change.

– Stephen Hawking.

Acknowledgements

First and foremost, I would like to thank my supervisor Li Cheng for his continuous support and encouragement throughout my M.Sc. I want to also thank my committee members Herb Yang and Xingyu Li for reviewing my thesis.

During my Master's, I have had the opportunity to have the help of two undergraduate students. Here, I would like to thank Fidel Omar Tito Cruz and Zetong (Emma) Zhao for their hard work in the projects we undertook. Collection and annotation of the introduced datasets of this thesis couldn't be possible without their help.

Finally, this thesis is dedicated to my family, specially my parents Masoumeh and Mohammadali, for all the years of their love and support.

Contents

List of Tables

List of Figures

Chapter 1

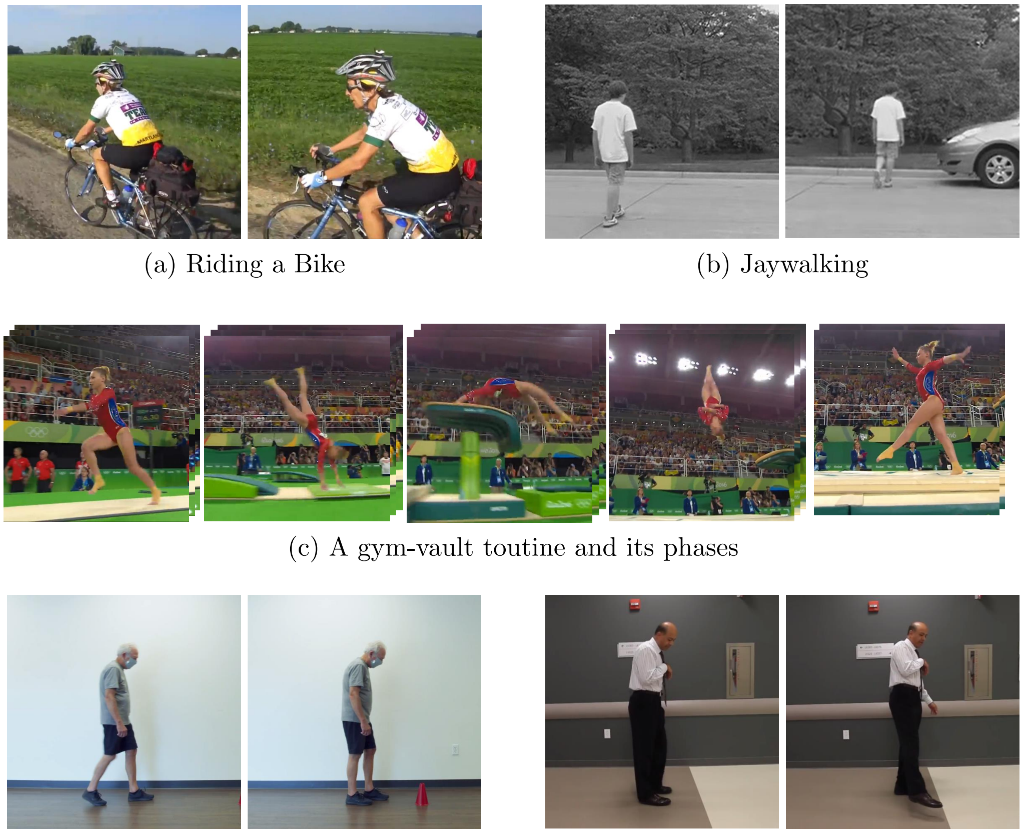
Introduction

Endowing machines with the ability of automatically deriving high-level understanding from the visual world has been a persistent research goal for decades. Designing such intelligent systems would pave the way for automating many human-centered tasks. In the field of medical image analysis, a machine is now able to automatically spot the abnormal cancerous cells in x-ray and MRI scans. The agriculture sector has witnessed a great progress with the help of UAVs that autonomously monitor the farmland and detect plant diseases and insects. Self-driving cars remove the need for having a human driver with the help of image recognition methods that provide a sense of the car's surroundings.

In this thesis, we study the automation of a relatively new task, called human Action Quality Assessment (AQA). The fundamental goal of AQA is to assess how well an action was done and assign a grade to the performer, given the video of the performance. This task would have a broad range of applications in future from automatic scoring of Olympic events to stage detection and rehabilitation progress assessment of the people with physical impairments or disabilities. By automating the AQA task, a computer can then interpret the video of a sports routine and determine the score of the performer. Deploying such a intelligent system not only shortens the events' times by accelerating the grading procedure but also safeguards the scores against the human judges potential misjudgments. Automating AQA can benefit the rehabilitation services as well. Movement disorder is one of the most common neurological diseases affecting one out of ten Canadians [54]. Being monitored for early diagnosis and rehabilitation progress assessment by a clinician opens the door for getting a better treatment and medication. However, getting access to a specialist is not easy for these patients, especially for the ones who live in rural parts considering the Covid-19 barriers. To address this problem, some recent studies [46], [50], [75] have explored the collection of data from tracking the patients' movements by attaching some sensors to their bodies to be later assessed by a virtual clinician. Nevertheless, the sensory system is not only expensive to acquire but also impedes the patients' movements, affecting the assessment accuracy. Nowadays, everyone has a smartphone which is equipped with a camera to take videos of daily living activities. So why not putting one step further and resort to just a video of the movements? The developed action quality assessment approaches of this thesis, in both sports analysis and tele-rehabilitation applications, just take a video of the action to assess the performance.

Although it may seem a trivial task for a human to interpret the visual data, learning to understand and analyze the content of an image or a video is quite hard for a machine. The visual data is just a set of numbers to a computer. A color image is usually represented using a three-dimensional matrix made up of rows and columns of pixels. The third dimension is used for determining the intensity of each shade of primary color channels (red, green, and blue) to form the color of each pixel. A video can be considered as a sequence of these image frames, containing an additional temporal dimension. In order for an AQA machine to have a good evaluation of a performance, both spatial and temporal dependencies throughout the video should be captured.

The AQA task is closely related to a long-standing subfield within the computer vision called Human Action Recognition (HAR). In HAR, the goal is to label a video with the action that is being carried out in it. The early efforts to solve the HAR problem develop handcrafted features like Histogram of Gradients (HOG) [25], Histogram of Motion Boundary (MBH) [64], and Extended SURF (ESURF) [69] that embody the appearance and motion features of an action. An encoder method like Bag of Visual Words (BoVW)



(a) Riding a Bike

(b) Jaywalking

(c) A gym-vault toutine and its phases

(d) Gait freezing in a part of a Parkinson's walking gait (small steps)

(e) An impaired hemiplegic walking gait

Figure 1.1: HAR video samples vs an AQA in sports and health-care

[11] then aggregates these features to recognize the action in a video. With the advent of deep neural networks, computers were able to extract high-level semantic features from a video, leading to higher accuracies in performing the HAR task. In order to get such features in an image, 2D Convolutional Neural Networks (CNNs) apply a succession of convolutional layers and nonlinearities on the image. A simple approach for HAR is to treat it as the task of frame-wise classification of the action using 2D CNNs [22]. 3D CNNs [8], [58] slide a three-dimensional kernel over the video to extract the temporal and spatial feature representations of the action simultaneously. Although these networks perform well in HAR, that won't be the case for AQA. Despite sharing some similarities, there are some differences between HAR and AQA that may challenge the existing works in HAR to be used for the AQA task.

Take a look at Fig.1.1. In HAR, one may recognize the action which is being carried out in a video by looking into a single frame or a set of key frames of the action. On the other hand, to score the performance of an action, all frames should be taken into account. For example, a diver may do a great job throughout the routine but not have a vertical entry to the water. A Parkinson's patient may get tired after some time or experience the freeze of gait and not be able to perform the exercises as good as the beginning. An AQA agent has to keep track of all of the frames to detect abnormalities and assess the whole performance. Besides, there is a specific temporal hierarchical structure for a sports routine. For example, a gym-vault routine is constructed of some medium term phases like approach, first flight, repulsion ,second flight , and landing. Each phase itself is comprised of some short-term elements like twists and saltos. In order for a judge to have a perfect assessment, the fine-grained temporal dependencies between consecutive frames that form a short-term element as well as the coarse-grained dependencies that evaluate the overall smooth performance in medium-term phases should be considered.

As it can be seen in Fig.1.1, there might be a large difference between the samples of two different classes in HAR. However, there is a subtle difference between two sports performances. Considering the competitive nature of these events, specially in final rounds, a non-expert is not even able to discern the worst performance from the best one in most cases. Rehabilitation assessment lies in the middle of these two ends of the spectrum. Non-experts are able to detect the severity of the disease to some extent based on their prior experience. We have seen lots of normal activity samples like walking by healthy people throughout our life. Given this information, we are able to have a guess on the severity of a neuromotor disease based on how impaired and slow the patient is doing the exercises (see the last row of Fig.1.1).

Another difference between the HAR and AQA tasks is the size of their correspondent existing datasets. For example, YouTube-8M [2] and Kinetics-700 [7] datasets provide over 8 million and 600000 annotated video samples respectively to facilitate the use of DNNs in the HAR task. On the other hand, there are only 370 video samples of diving videos in AQA-7 dataset [40]. The

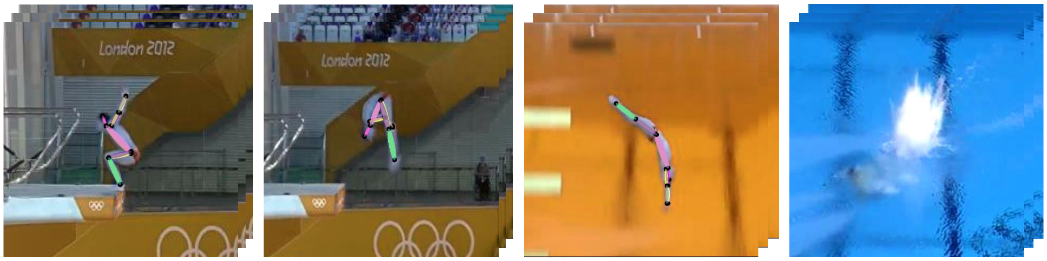


Figure 1.2: To assess a diving routine both pose and appearance features should be taken into account. The athletes may contort their body in some unusual configurations

same thing goes with the KIMORE rehabilitation assessment dataset [6] that only covers about 400 video samples. The small size of the AQA datasets is because of the need for having an expert to annotate their samples with the ground-truth score of each performance. Considering the limited size of the existing AQA datasets, deploying shallow networks or transferring knowledge from the related tasks like HAR becomes vital to avoid the well-known problem of overfitting.

Moreover, unlike the case of HAR in which the appearance-based features are informative enough for most cases to classify an action, integrating pose features in the process of assessing a performance is of great significance. For example, in a diving contest, a judge grades the performances based on both pose features like coordination among the joints and appearance features like amount of splash and smoothness of the flight (see Fig.1.2). In rehabilitation exercise performance assessment, the poor posture of a patient can't be detected by resorting to the pose features evaluation. The dependencies between the joints should be accompanied with appearance features to have a good evaluation over the posture of the body and smoothness of the movements. To this end, unlike previous AQA studies, our developed AQA methods take into account both pose and appearance features to grade a performance.

Estimating the pose of the athlete throughout the sports routine has its own challenges. As evident in Figures 1.1,1.2, the athletes usually contort their bodies in some unusual configurations to demonstrate the flexibility of their body and add to the aesthetic aspect of the performance. Current pose estimation datasets mostly cover daily living activities like walking, sitting, etc. The pose estimators that are trained on these datasets don't show a good performance in estimating the extreme pose configurations of contortive sports. That's why getting access to a dataset that covers the pose annotation of such extreme cases is felt missing in the literature of AQA.

The goal of this thesis is to address the above challenges and propose some video-based approaches to evaluate the performance of an action in both sports analysis and rehabilitation progress assessment applications. Due to availability and larger size of the sports datasets, the most of the AQA works have focused on the sports routine evaluation application. These works rely exclusively on either the appearance or pose features to grade the performance. The pose-based approaches have built their network upon joint location results that a pose estimator provides. Since the estimator is trained on normal activities datasets, the accuracy of the pose estimation in contortive sports like diving, gym-vault, etc. is poor, affecting the grading accuracy adversely as well. Moreover, the current AQA approaches treat a sports routine as a simple atomic action, completely neglecting its complex hierarchical temporal structure. This fact has led to under-performance of the previous works especially in minute-long activities scoring that requires capturing coarse-grained temporal dependencies between distant frames. On the other hand, there are only a few works that explored the rehabilitation assessment application. The privacy concerns have limited the access to video samples of patients' actions. As a result, the current tele-rehabilitation datasets are small, affecting the performance of deep neural networks in assessing the performance videos. It should be noted that the detailed review of the recent works has been brought in **chapter 2**.

In what follows, we present the outline of each chapter and explain how we address the aforementioned issues of the previous works.

In **chapter 3**, we focus more on grading a diving routine. In a diving contest, each athlete is assessed based on how difficult the routine was and how well the diver executed it. The difficulty score is determined based on the components of the performance like the number of somersaults and twists and the position of the flight. To get these components the network tracks the configurations of the joints throughout the performance. In order to have a more accurate pose estimation results in extreme contortions of the body we introduce *ExPose*: annotated dataset of Extreme Poses that covers 3000 annotated images of diving. On the other hand, inspired by human judges' grading schema, a virtual refereeing network evaluates the execution of a diving performance. This assessment would be based on visual clues as well as the body joints sequence of the action video. Finally, the overall score of the performance would be reported as the multiplication of the execution and difficulty scores. The experiments demonstrate that our proposed lightweight network achieves state-of-the-art results compared to previous studies in diving grading.

In **chapter 4**, we aim at proposing an AQA network that is not only applicable to other short-term sports actions like snowboarding, gymnastic vault, and skiing but also can be used to grade a minute-long sports activity like figure-skating. As discussed before, there is a hierarchical temporal structure for most sports routines. Measuring the quality of a sports action entails attending to the execution of the short-term components as well as the overall impression of the whole program. In this chapter, we present Joints Coordination Assessment (JCA) and Appearance Dynamics Assessment (ADA) blocks that are responsible for reasoning about the coordination among the joints and appearance dynamics throughout the performance. We build our two-stream network (pose and appearance) upon the separate stack of these blocks. The early blocks capture the fine-grained temporal dependencies while the last ones reason about the long-term coarse-grained relations. In order to have a minute-long activity assessor, we only need to stack more of these JCA and ADA blocks to involve distant frames temporal relations in the assessment process. We further extended the *ExPose* dataset to cover other contortive sports like synchronized diving, snowboarding, and skiing and we call the new dataset *Generalized ExPose* or shortly *G-ExPose*. We show that the proposed method not only outperforms the previous works in short-term action assessment but also is the first to generalize well to minute-long figure-skating scoring.

Finally, in **chapter 5**, we explore the application of AQA in the rehabilitation assessment task. As discussed before, unlike the case of sports actions evaluation, non-expert humans are able to assess the severity of a patient's disease to some extent based on their prior observations of healthy samples. This assessment would be based on how slow and impaired the action looks to our eyes. Likewise, in order for a healthcare professional to assess the severity of a patient's action properly, he/she must be able to evaluate the impairment and slowness of the performance. These features would be accompanied with detailed analysis of appearance and pose features that the non-expert is not capable to capture. Currently, the limited size of the existing AQA datasets has impeded the previous methods to employ a deep network and capture the pace and impairment features of an action. In this chapter, we propose a method that takes advantage of the large set of healthy people actions representation from HAR datasets to improve the assessment of patients' abnormal movements. A non-expert network first learns the representation of the normal sequences by estimating the pace and a set of manually inpainted joints of the pose sequence in a self-supervised manner. In the next step, an expert network takes these representations as well as the appearance features of the abnormal sequence to assess the quality of the action performed by a patient. The results demonstrate that the proposed method not only outperforms the previous works on Parkinson's and Stroke patients' action assessment but also generalizes well to the new application of infants' general movement assessment.

Chapter 2

Related Works

The AQA is a relatively new task that has recently attracted researches to work on. Most of the existing related works focus on AQA for sports analysis. The availability of sports video footage on media platforms like YouTube has encouraged the researchers to collect datasets and study the video-based automatic evaluation of athletes' performances. Here, we survey the existing works in both sports analysis and health-care applications.

2.1 Pose-based AQA

Due to the challenges of estimating pose in extreme configurations of athletes' body and blurriness of image frames, pose-based AQA has been largely unexplored. Pirsiavash *et al*. [44] presented the first model for assessing the quality of a sports action based on pose features. They first extracted the joints location of the video frames using the Flexible Parts Model [73]. At the next step, they trained a linear SVR on DCT frequency coefficients of the action pose features to regress the final score. Recently, Pan *et al*. [39] decoupled pose features to body parts kinetics and joints coordination, and provided the score by capturing graph-based joint relations. They proposed joint difference and commonality modules to represent the coordination among the joints and motion of certain body parts for consecutive timesteps. The pose features can be extracted using a pose estimator like AlphaPose [12] or Mask-RCNN [17].

These pose-based networks neglect the collaborative role of appearance-based with pose-based clues in score awarding. Furthermore, they rely on the overall score as the only label that needs to be predicted. In addition, their pose estimation modules have been trained on regular existing pose datasets that don't cover unusual contorted pose configurations of Olympics divers. The above reasons generally lead to the underperformance of such networks.

2.2 Appearance-based AQA

On the other end of the spectrum, most studies in the realm of AQA rely exclusively on visual clues of the action in order to regress the score [28], [41], [42], [71]. Parmar and Morris [42] use C3Ds [58] (inflated counterpart of 2D ConvNet) to capture visual spatio-temporal features of the action and feed them to a regression framework (SVR, LSTM, and LSTM-SVR) to provide the final score. Li *et al*. [28] segment videos to multiple fragments and feed them to parallel C3Ds, followed by some 2D convolution layers. They discuss that these fragmented features would have more distinctive power to differentiate a good performance from a bad performance. Tang *et al*. [56] proposed a score distribution learning network that sits on top of I3D appearance features of the action [8] and models the judges' disagreement in grading a routine. However, the heavy computation cost of using 3D CNNs, as well as neglecting the pose features and predicting only the overall score has affected the performance of these models. To address the last issue, Parmar and Morris [41] recently proposed a multi-task approach that jointly learns commentary, action class details, and AQA overall score based on the appearance features extracted from C3Ds. Nevertheless, the first two issues still persist in the latter study.

2.3 Temporal Structure Modeling

Extracting the temporal structure of a video has been extensively studied for action recognition. Niebles *et al*. [37] model each complex action as a composition of some motion segments and use Latent SVM to get the parameters of the model. To capture the long-term temporal relations, TSN [67] evenly divides the video into multiple segments and fuses the class score of sampled snippets from each segment. TRN [77] learns the pair-wise long-term temporal relation between sampled frames at different scales and fuses the resulted features. Recently, Hussein *et al*. [19] proposed a modular layer that sits on top of appearance features of a complex action video to learn its long-term temporal dependencies. All above appearance-based methods have explored the significance of temporal modeling in action recognition. Here we investigate the temporal structure of an action to assess it based on both appearance and pose features.

The significance of long-term temporal modeling in a minute-long activity has made all of the previous sports action assessors only applicable to short-term routines like diving. A simple way to score a minute-long activity like figure-skating is to treat it as a short-term one and simply feed the C3D features of the routine to a SVR [42]. Recently, Xu *et al*. [71] proposed multi-scale skip LSTMs to cover a more broad receptive field. The resulted features are further fused with compacted local feature representation provided by a self-attentive LSTM to regress the score of a figure-skating routine. However, not only the pose features are completely ignored but also the rigid structure of the proposed method is not applicable to short-term activities. Besides, skipping some frames/clips to have a broad receptive field would result in ignoring some useful visual clues.

2.4 Pose Estimation Approaches and Datasets

Human pose estimation is the task of localizing human body's anatomical key-points or simply put joints. In marker-based pose estimation, a set of sensors are attached to the body joints to track their location. However, acquiring the sensors and setting up a motion capture system are both expensive and tedious. On the other hand, marker-less approaches aim to estimate the pose from visual modalities such as images and videos. The classical marker-less approaches for pose estimation use pictorial structure models for the body skeleton. These methods either represent the skeleton in a tree-structured model and encode the spatial dependencies between the connected parts as some templates [45], [49], or form a non-tree model by adding some additional constraints to the tree in order to extract occlusion and long-range dependencies [21], [24]. With the introduction of ”DeepPose” [57], the ability of deep neural networks (DNNs) to extract high-level features encouraged more researches to deploy them for pose estimation and get higher accuracies. DeepPose uses a refined AlexNet and treats the pose estimation problem as a CNN-based regression task. Newel *et al*. [35] hypothesize that both local and global contexts should be involved in the joints localization task and propose a stacked network to perform bottom-up and top-down processing constantly. Sun *et al*. [55] emphasize on the importance of maintaining the high-resolution information throughout a network to get a high estimation accuracy. Their network called HRNet is constructed of a main stream of high-resolution representation which is fused by high-to-low resolution subnetworks' features. Due to the superior performance of HRNet in comparison with the other approaches, here we use it as the pose estimator network to get the joints location.

Among the pose datasets, ”Parse” [48] is one of the earliest containing only 305 in-the-wild annotated images which makes it unsuitable to be used for training DNNs. To address this issue, Lin *et al*. [30] introduce a large-scale dataset (COCO) that contains more than 200,000 labeled images with a focus on multi-person occluded scenarios. Andriluka *et al*. [3] collect about 29000 annotated images covering more rare human poses (MPII). However, the unusual configuration of the athletes' poses in a sports routine is not covered by none of the existing datasets. Here, we introduce ExPose and G-ExPose datasets of blurry images with annotated joints in extreme contortions to facilitate extreme pose estimation in sports analytic.

2.5 AQA for Tele-rehabilitation

A common way to monitor and evaluate the actions of movement disorder patients' is to attach some sensors like gyroscopes or accelerometers to their bodies. These devices can provide accurate kinematic and dynamic analysis of different body parts gait patterns. Zach *et al*. [74] used a lumbar accelerometer to detect Freezing of gait in Parkinson's patients. Trojaniello *et al.* [59] estimated the gait temporal parameters of the hemiparetic patients by attaching an inertial measurement unit (IMU) to their lower back. Using these sensors improves the accuracy in the gait assessment by minimizing the subjectivity. However, acquiring the sensory system in the lab environment would be both expensive and tedious. Besides, the attached sensors may impede the patients from having their normal movements. Therefore, a new set of studies focused on assessing the patients' disease severity from just the video of their movements. Capecci *et al.* [6] introduced the first freely available video dataset of Parkinson, stroke, and back pain patients doing a set of daily exercises (KIMORE dataset). The samples were annotated with quantitative scores of the disease severity. Recently, Sardari *et al.* [51] proposed a view-invariant method to assess the performance of patients and achieved state-of-the-art results on this dataset.

Unlike sports action scoring, the samples of rehabilitation assessment have a lot in common with daily activities like walking and sit-standing covered by large-scale off-the-shelf datasets. This fact inspired us to learn some representations from these healthy samples to be used in assessing the patients' actions.

2.6 Self-supervised Representation Learning

The goal in self-supervised representation learning is to remove the need for having an explicit label to learn representations from the data. To this end, an auxiliary *pretext task* is proposed to learn transferable features inherently from the data itself. In each pretext task, a part of the data is kept visible and the rest is manually hidden. The agent of the pretext task then tries to estimate either the hidden part itself or some properties of it. At the next step, the latent representations that were extracted during the pretext task completion are used to perform the *downstream task* which is the task that we actually want to solve. In image-based pretext tasks, masking a part of an image and trying to predict it [43], estimating the rotation transformation [20], and color channel prediction [26] are some of the surrogate objectives that the model tries to fulfill. The learned representations can then be utilized to solve a downstream task like semantic segmentation and image classification. In comparison to images, videos have an extra temporal dimension that may provide useful information to be adapted to video-based downstream tasks like action recognition. Video clip order prediction [32], motion estimation [29], pose sequence inpainting [76], and video pace prediction [66] are some of these pretext tasks. The non-expert network in **chapter 5** is inspired by the last two to estimate the slowness and impairment of an a patient's movements.

Chapter 3

FALCONS: FAst Learner-grader for CONtorted poses in Sports

3.1 Introduction

Sports is the language of joy and unity. Sporting events are usually among the top most-watched televised broadcasts [10]. Fairness in evaluation is of utmost importance to both the competitors and spectators, hence the need for a structured means to evaluate the athletes and determine the winner. In recent years, the advent of technology has brought more just and agile refereeing to soccer games by the introduction of video assistant referees (VAR). However, some popular fields like diving, gymnastics, etc., are still suffering from the inefficiency of human-based judging systems. For example, in a typical diving contest, 7 judges score the performance of each athlete. These judges should be from different nationalities to that of the contestant, limiting the qualified choices. Furthermore, although the international swimming federation (FINA) has prohibited the judges from looking at the replays to make the grading procedure faster, it takes about 40 seconds to report the score for each performance that itself takes only about 4 seconds. Considering all these issues, it would be a great help to introduce an automatic grading system to score the athletes in a faster and more accurate manner by reducing human intervention. In this chapter we focus on evaluating how a diving routine has been performed and assigning a grade to the performer

To date, the literature of AQA has been dominated by networks that have

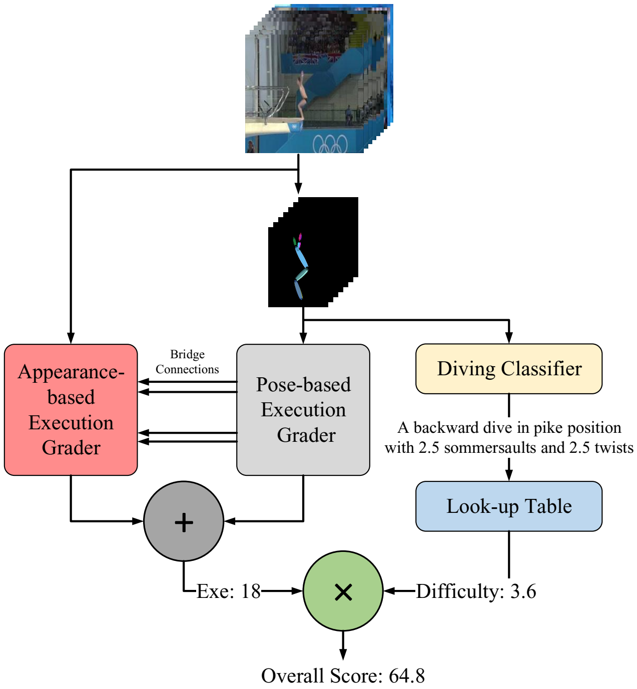

Bridge Connections
Appearance-based Execution Grader
Pose-based Execution Grader
Diving Classifier
A backward dive in pike position with 2.5 sommersaults and 2.5 twists
Look-up Table
+
Exe: 18
×
Difficulty: 3.6
Overall Score: 64.8

Figure 3.1: Overall pipeline of FALCONS.

tried to regress the overall score of each athlete as the only label that needs to be predicted [39], [42], [44], [63]. However, each performance should be considered as a complex action in which not only the quality of the execution but also the difficulty of the task contributes to the final score. Nevertheless, the judges are only responsible for awarding the execution score and the difficulty score would be determined based on a predefined official benchmark released by the corresponding federation of that sport. In a diving contest, the difficulty score of each performance would be awarded based on the type of rotation, position of diving, number of sommersaults, and the number of twists according to FINA difficulty look-up table [13]. On the other hand, the judges award the execution score based on the appearance-based features of the flight (*e.g.* smoothness and aesthetic pleasure of the flight, and amount of splash), and also pose-based features (*e.g.* angle of entry to the water).

As discussed before, existing literature regarding automatic graders take into account either the pose-based or appearance-based features to regress the final score of an action, hence suffering from the limited performance [28], [41], [42], [44], [63]. Here, we decompose the overall score into execution and difficulty. For the former, inspired by what a human judge does, we propose a virtual refereeing system that considers both the pose-based and appearance-based features as the contributors to the execution score. As for the latter, we introduce a difficulty extractor module that classifies the task based on the sequence of body joint arrangements throughout the performance. Consequently, the difficulty score would be determined by feeding the classes (*e.g.* type of rotation and etc.) into the difficulty look-up table. The overall pipeline of our work is depicted in Fig.3.1.

As mentioned in **chapter 1**, the extraction of pose features has its own challenges. In most of the sports video footage, not only the athlete but also the camera is moving fast to track the athlete during the performance, causing motion blurriness of the image frames. Furthermore, the unusual configuration of the athletes' poses is not covered by the existing datasets [3], [30]. These facts have limited the performance of the pose estimation networks in such cases. To address this problem we introduce ExPose, a dataset of extreme poses which includes 4000 annotated blurry images in extreme body contortion scenarios. In order to regress the scores based on the extracted joint sequences, we develop the well-known Spatial Temporal Graph Convolutional Networks (ST-GCN) [72] to also learn dependencies between unconnected joints. ST-GCN extends graph convolutional networks to simultaneously capture spatial and temporal features for action recognition. However, it is only able to learn correlations between the directly connected joints. This fact may affect the performance of the automatic grader in which the symmetry of different parts of the body should contribute to the execution score. To address this issue, we introduce the idea of virtual super-joints. Each super-joint is simply the average of its constituting joints' location and may have connections with other super-joints.

For spatio-temporal appearance features extraction, we divide the whole task video into T subtasks, each consisting of N consecutive frames. In order to learn global spatio-temporal features, we follow [16] that applies 2D spatial convolution filters followed by 1D temporal ones on each subtask. Finally, the temporal dynamics between the subtasks should be encoded to an execution score. However, as the amount of splash plays a greater role than other appearance features in the execution assessment, not all of the subtasks should have the same contribution to the score. To this end, we propose a bridge-connection module that fuses the feature sequence of each subtask with a contribution weight. This module takes the average of confidence score of pose estimation in each subtask and maps it to a weight for each. Simply put, as the performer is under the water in splash capturing frames, the lower the average confidence score of the subtask, the higher its contribution to the Exe score.

Finally, the weighted sum of the score of the joint-based and appearance-based graders generates the execution score (see Fig.3.1). The overall score would then be calculated by multiplication of the extracted execution and difficulty scores. To validate the effectiveness of our method we conducted experiments on existing datasets and demonstrate that our method not only achieves state-of-the-art results in diving grading but also shows acceptable generalization to other fields. The main contributions of this chapter are summarized as follows:

- We introduce *ExPose*, a dataset of blurry images with annotated joints in extreme contortions to facilitate extreme pose estimation in sports analytics.

- We propose a difficulty extractor module that leverages the pose sequence to determine the type and details of the dive and award a difficulty score based on the FINA look-up table. This module doesn't need to be trained in advance and achieves state-of-the-art results.

- Inspired by human judges grading schema, we present a novel virtual refereeing system that regresses the execution score by leveraging both appearance and pose-based features. The results demonstrate the superior performance of the proposed network over previous works with promising generalization to other sports.

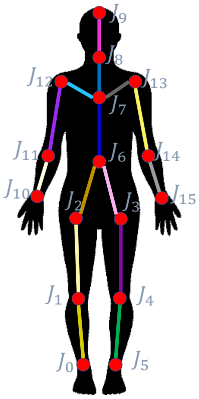


Figure 3.2: The human skeleton structure in *ExPose*

3.2 Dataset

ExPose dataset contains 3000 diving and 1000 gymnastic vault images together with their annotations. The diving images are obtained from four different individual diving events recorded from side-view on two types of boards. The springboard images are obtained from men's 3m final of the 2019 world series and the platform images are taken from men's 10m platform finals of the 2016 European aquatics championships in London, 2018 youth Olympic games in Buenos Aires, and 2019 world series in Sagamihara. The gym vault images are obtained from three different events; men's and women's final of the 2018 Doha world championships and women's final of the 2018 Glasgow European competitions.

In order to collect the dataset, we first queried YouTube to get the original video of each event. Secondly, we filtered out the irrelevant parts of the video (like the opening ceremony and medal presentation) and extracted the frames of each video with the frame rate of 10 fps. Finally, we mostly held out the normal poses of each routine (like approaching to the springboard or after-landing in a gym-vault routine) to focus more on the main part of the execution in which the performer contorts his body in some unusual configurations.

For consistency with existing datasets, the annotations have been provided in MPII dataset format. As depicted in Fig.3.2, this skeleton structure considers 16 joints and 15 bones for the human body.

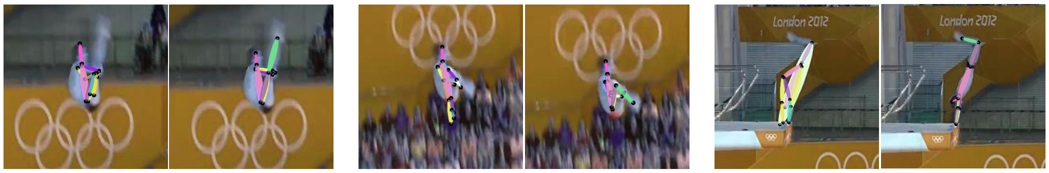


Figure 3.3: The pose estimation network should be trained on a dataset which covers blurry images with extreme configuration of joints. The left side picture of each column is the result of training HRNet on MPII [3] and the right side one corresponds to the results of training the network on our ExPose dataset. As evident, the network trained on our data set better identifies body's true joint locations.

As evident in Fig. 3.3, the HRNet pose estimator does a much better job in estimating the joints location of a contorted body blurry image when it is trained on ExPose.

3.3 Approach

In the following, we present our proposed network consisting of two modules. The first module, virtual referee, evaluates the performance execution based on both of the appearance and pose features. The second module, difficulty assessor, evaluates the difficulty based on the joints sequence of the action.

3.3.1 Virtual Referee

Fig. 3.4 demonstrates our virtual referee pipeline which is comprised of pose-based and appearance-based assessors. In what follows we elloborate more on the details of the two assessors.

Pose-based execution assessor

We use HRNet pose estimation network [55] to extract the pose features of extreme actions in each frame. The network has been trained on our ExPose dataset to be able to handle blurry images with extreme body contortions. It should be noted that the HRNet pose estimator requires the bounding box of human to estimate the joints location. With the help of DiMP visual object tracker [4] we would be able to get the bounding box of the athlete in each

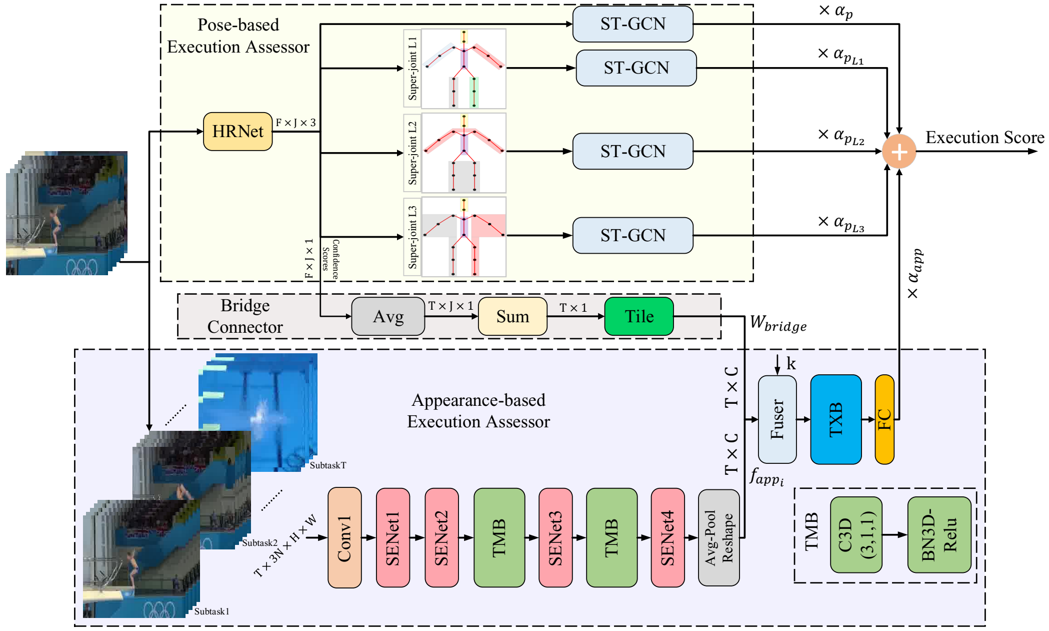


Figure 3.4: Overview of our virtual referee pipeline. This network generates the execution score of the performer based on appearance features as well as pose ones. The bridge connector module links these two to increase the contribution weight for the appearance features with the most importance (splash).

frame. HRNet takes these bounding boxes and estimates the pose of the athlete. So the entanglement of DiMP and HRNet can be seen as a pose tracker.

The extracted joint sequence should be fed into an action regressor that is able to learn the spatio-temporal skeleton-based features of the action. To this end, we use ST-GCN [72] that considers the joints of a skeleton as the nodes of a graph. Spatial edges of the graph connect the structurally neighboring joints and temporal edges connect the same node in consecutive frames. However, this network is not able to capture the dependencies between joints that are unconnected in the predefined skeleton graph.

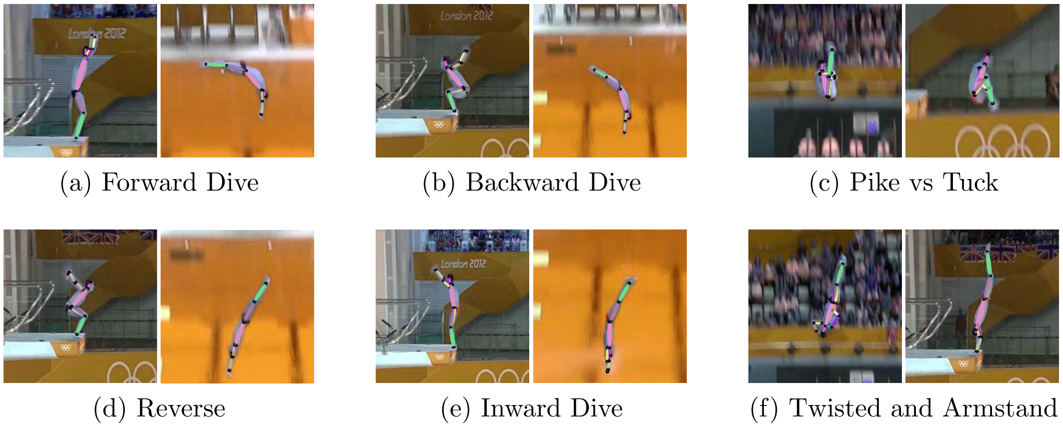
In order to capture the non-local information (for unconnected joints), we introduce three levels of virtual super-joints. Each super-joint represents the average location of its constituting real joints. In the first level, the super-joints consist of the right leg, left leg, right hand, left hand, waist, and head.

As a result, we would be able to capture wrists-shoulders and hips-ankles relationships. The second super-joint level captures dependencies between the upper body and the lower body. Finally, the third super-joint level extracts the symmetricity between the left and the right side of the body. The composition of super-joints is visualized in the upper part of Fig.3.4.

It goes without saying that not only the local relations between connected joints but also richer dependencies like symmetricity between different parts of the body should verify how well the action was performed. Thus, the features of the fixed predefined skeleton body (consisting of all 16 body joints) as well as non-local features of virtual super-joints levels are fed into four parallel ST-GCNs to assess the coordination between the joints and between the body parts (see the upper part of Fig.3.4). The output value of each ST-GCN module is multiplied by its own contribution weight to account for the superior importance of some features over others. For example, intuitively, the contribution weight of the third level of super-joints score should be larger as it is responsible for the balance of the body during the flight.

Appearance-based execution assessor

The proposed network should not only be effective but also have a lightweight configuration to act as fast as possible in both training and testing phases. To this end, we follow StNet [16] structure that instead of using stacked heavy-weight C3Ds, decomposes 3D convolutions into 2D spatial convolutions followed by 1D temporal ones. First, we divide the whole task into T subtasks, each consisting of N frames (in this case T=17, and N=6). Each subtask is fed to 2D convolution layers to extract the local spatio-temporal information. The Conv1 module followed by the 2 SENet [18] layers are responsible to do so (see the lower part of Fig.3.4). The extracted local features are fed to a Temporal Modelling Block (TMB), introduced by [16], to get the temporal features across each subtask. The TMB block is consisted of a 3D convolutional layer followed by a BN3d-ReLU. As the spatial information is already captured by SENets, the spatial filter size of the 3D convolution for TMB is set to 1. In order to get deeper correlations, the result of the TMB module is



(a) Forward Dive (b) Backward Dive (c) Pike vs Tuck

(d) Reverse (e) Inward Dive (f) Twisted and Armstand

Figure 3.5: The type of the rotation, the position, and #twists can be determined based on joint sequences. In addition, based on the the configuration of pelvis and thorax, the #sommersaults and first stand position can be determined. The performance of pose estimation should be acceptable to have a better diving classifier and respectively a better difficulty assessor.

further passed to a stack of SENet-TMB-SENet.

The extracted features (f_{app_i}) would contain local spatio-temporal information of each subtask as well as temporal dynamics across N frames of each subtask. In the next step, we should capture the temporal information between the T subtasks. However, there are some visual clues that should contribute to the execution score more than the others. Performing a rip entry into the water making the least amount of splash is more important than other appearance-based clues like the smoothness of the flight. In order to address the aforementioned concern, we introduce a bridge connection module to increase the contribution weight of the final subtasks which is where the diver makes an entry into the water. As discussed before we use HRNet module to detect the joints in each frame. This module not only provides the joints' locations but also gives the overall confidence of estimation for each joint. For example, in the last frames in which the athlete is under the water, the very distorted visibility of the human body in such frames causes a drastic drop in the confidence score of its joints estimation. Inspired by this fact, we introduce a link module to set the weights of subtasks based on the confidence score of pose estimation in their frames (see the middle part of Fig. 3.4). Given the confidence scores of all frames, we first take the average between the confidence scores of each subtask. Thus, considering the output of HRNet as a $F \times J \times 1$ tensor (where $F = T \times N$ is the number of frames and J is the number of joints), the output of the average module would have size $T \times J$. In the next step, the confidence scores of all joints in each subtask have been summed to form a $T \times 1$ vector and normalized to $(0 - 1)$ interval. Finally, the vector is tiled to have the same size as that of the extracted appearance features $(T \times C)$. The Fuser module takes these two (f_{app_i} and W_{bridge}) as well a scale value (k) to set the contribution of each subtask using the below formula.

$$f_{app_o} = f_{app_i} \odot (1 - W_{bridge}(1 - k)) \tag{3.1}$$

As a result, the lower W_{bridge} a subtask has, the higher contribution to the score regression it would have. Finally, to capture the temporal information between the subtasks, the resulting f_{app_o} is fed to a Temporal Xception Block (TXB). This block, introduced by [16], decomposes the temporal dynamics of the extracted feature sequence into a 1D temporal-wise and a 1D channel-wise convolutions. Deploying this strategy instead of averaging between the features of the subtasks has boosted the results in action classification tasks. Finally, the resulting tensor of the TXB module has been fed to a fully connected layer to regress a number as the appearance based execution score.

In the end, as evident in Fig. 3.4, the virtual referee calculates the weighted sum of the appearance-based execution score and the pose-based ones (including regressed scores of virtual pose levels) to award the final execution score:

$$\begin{aligned} S_{Exe_f} = \alpha_{app} S_{Exe_{app}} + \alpha_p S_{Exe_p} + \alpha_{p_{L1}} S_{Exe_{p_{L1}}} \\ + \alpha_{p_{L2}} S_{Exe_{p_{L2}}} + \alpha_{p_{L3}} S_{Exe_{p_{L3}}} \end{aligned} \tag{3.2}$$

where α_{app}, α_p, and $\alpha_{p_{Li}}$ represent the contribution weight of appearance-based, predefined skeleton pose-based, and virtual-pose levels execution assessment.

3.3.2 Difficulty assessor

Given the joint sequences extracted from the HRNet module and the direction of filming (west-side or east-side camera), our proposed difficulty assessor classifies the performed dive and provides the difficulty score based on the FINA look-up table.

In terms of the rotation type, a performance can be classified into four different groups: forward, reverse, backward, and inward. The group can be determined based on the joints position during frames that capture take-off and entry into the water. For example, when the camera is located in the east-side of the platform and the athlete faces the front of the board (forward or reverse; Fig. 3.5a,3.5d), x_{knee} would be greater than both x_{hip} and x_{ankle} (the origin of the coordinates is located at the top left of the picture). On the other hand, when the athlete takes off with his (her) body back to the water (backward or inward; Fig. 3.5b,3.5e), x_{knee} would be less than both x_{hip} and x_{ankle}. Another metric that helps determine the group is x_{ankle}'s value with respect to other joints' positions during frames that capture entry into the water. As it can be seen in Fig. 3.5a,3.5b,3.5d,3.5e, the joints position in the take-off frame and entry one together distinct a rotation type from another.

For getting the position of the dive, the module looks for a specific pattern among all of the frames. In a pike position, the body is bent at the waist but the legs should remain straight (see Fig. 3.5c). On the other hand, in a tuck position the knees should be pulled tightly to the chest. If a position is neither a pike nor a tuck it would be considered as a free dive. Thus, based on angle between the lower leg and thigh we would be able to determine the position.

In a regular dive we only have the sideview profile of the performance. However, in a twisted dive the performer rotates its body around the vertical axis, exposing the full-body profile during the performance (see Fig. 3.5f). In order to discern a sideview profile from a frontview one, we have set a threshold for captured shoulder-width of the athlete during the performance. The #twists can be determined based on number of the switches from a sideview to a front-view profile and vice versa.

In order to get the #sommersaults, the module monitors the relative positions of the thorax and the pelvis joints along the y-axis. In a normal configuration of the body, the thorax should be located in a higher position than the pelvis. However, this configuration changes as the body is rotated around the horizontal axis by performing a sommersault. The module counts the number of configuration switches to provide the #sommersaults. Furthermore, as evident in Fig. 3.5f, we can also distinguish an armstand dive from the regular one by getting the configuration of the pelvis and the thorax during the take-off frame.

Finally, the difficulty of the execution would be assessed based on the type of rotation, position, #twists, #sommersaults, and whether it was performed in an armstand position or not, according to FINA difficulty look-up table. The resulting difficulty score would be multiplied by the execution score provided by the virtual referee and the overall score determines the ranking of the competitor.

3.4 Experiments

Implementation Details

The weights of the all SENet blocks of the appearance-based assessor module are initialized using the ImageNet pretraied weights. As the first convolution layer takes $3N$ channels instead of 3, the same procedure as [8] have been taken to inflate the ImageNet weights.

The network has been trained for 200 epochs with a learning rate of 0.000075 using the SGD optimizer with the momentum of 0.95 and a batch size of 10. Mean squared error has been used as the loss function of the regression. The network is trained and tested on UNLV-Diving dataset [42]. The videos of the dataset are originally normalized to 103 frames. We ignore the last frame and feed the rest 102 to the network. The number of subtasks (T) is set to 17 and each one is consisted of 6 frames (N). 300 videos of the dataset are dedicated to training and the rest 70 are utilized for testing.

	MTL-AQA	**Ours (ExPose)**	Ours (MPII)
Armstand?	98.65	**100.00**	79.46
Rotation Type	97.30	**97.57**	69.19
Position	95.14	**97.84**	53.78
#Sommersaults	95.68	**98.92**	24.59
#Twists	94.05	**94.32**	28.65

Table 3.1: The results of diving detailed classification on UNLV-Diving dataset [42].

3.4.1 Quantitative Results

Diving Classification

To the best of our knowledge, there are only two studies in detailed diving classification [36], [41]. Table 3.1 shows our results compared to MTL-AQA [41] which performs better than the other one. Aside from outperforming the baseline study, the proposed diving classifier runs online without the need of being trained in advance. It takes the extracted pose sequence from the HRNet module and outputs the detailed classification of the dive as well as the difficulty score. On the other end of the spectrum, the MTL-AQA network uses C3Ds that leads to high computational cost and a huge impact on both training and testing phases speed.

Although this chapter is mostly geared towards diving routines, a similar procedure can be taken to assess the performance difficulty of other sports. For example, the difficulty score of a trampoline routine would be awarded based on the #twists and #sommersaults performed in the whole task. As another example, the difficulty of a gymnastic vault performance is assessed based on the number of turns (twists), the position of flight phase (tuck, pike, or stretched), and type of the approach towards the handspring (backward or inward).

Overall Score Assessment

In order to be consistent with existing literature, we have used Spearman's Rank correlation as the evaluation metric of our network. We evaluate our model on UNLV-Diving dataset [42]. As evident in Table 3.2, our proposed

Method	Spearman's Corr.
Pose-DCT-SVR [44]	53.00
Joint Relation Graphs [39]	76.30
C3D-SVR (best performing in [42])	78.00
MSCADC-STL [41]	79.79
Li *et al.* [28]	80.09
C3D-AVG-STL [41]	83.83
Ours (FALCONS)	**84.53**

Table 3.2: The results of predicting the overall score of the athletes. Our network outperforms the previous baselines. All networks are trained and tested on UNLV-Diving dataset [42].

Ablated Model	Spearman's Corr.
Bridge Blocking	69.11
Virtual Pose Levels Removal	76.68
Only Appearance-based	81.75
Only Pose-based	50.04

Table 3.3: The results of systematically removing the components of the network to study their effectiveness in the final results.

model achieves superior performance than prior methods. The parameters α_{app}, α_p, $\alpha_{p_{L1}}$, $\alpha_{p_{L2}}$, $\alpha_{p_{L3}}$, and k have been set to 0.9, 0.01, 0.01, 0.01, 0.07, and 0.9 respectively. It should be noted that [41] also proposes a multi-task approach in which they augmented the original dataset with captioning to jointly learn the overall AQA score, detailed diving classification, and commentary. This implementation resulted in Spearman's Rank correlation coefficient of 88.08. However, as our method is trained on the original dataset without using excessive information of captioning we have compared our results with their single-task approach.

We also conducted an extensive ablative study to evaluate the components of our network. In the first set of experiments, we blocked the bridge connector by setting the k scale of the Fuser to 1. As a result, all appearance features would have the same contribution to the execution assessment procedure. In order to study the importance of non-local pose features, we removed the contribution of the virtual pose levels by setting their contribution weights ($\alpha_{p_{L1}}$, $\alpha_{p_{L2}}$, $\alpha_{p_{L3}}$) to 0. We further set α_p to 0, relying only on appearance features for execution scoring. Finally, we set the α_{app} to 0 to evaluate the Only Pose-based execution assessor. In this experiment, α_p, $\alpha_{p_{L1}}$, $\alpha_{p_{L2}}$, and $\alpha_{p_{L3}}$ have been set to 0.1, 0.1, 0.1, 0.7 respectively to have the same scale as the original method. The results of Table 3.3 show how each component contributes to the effectiveness of our final model.

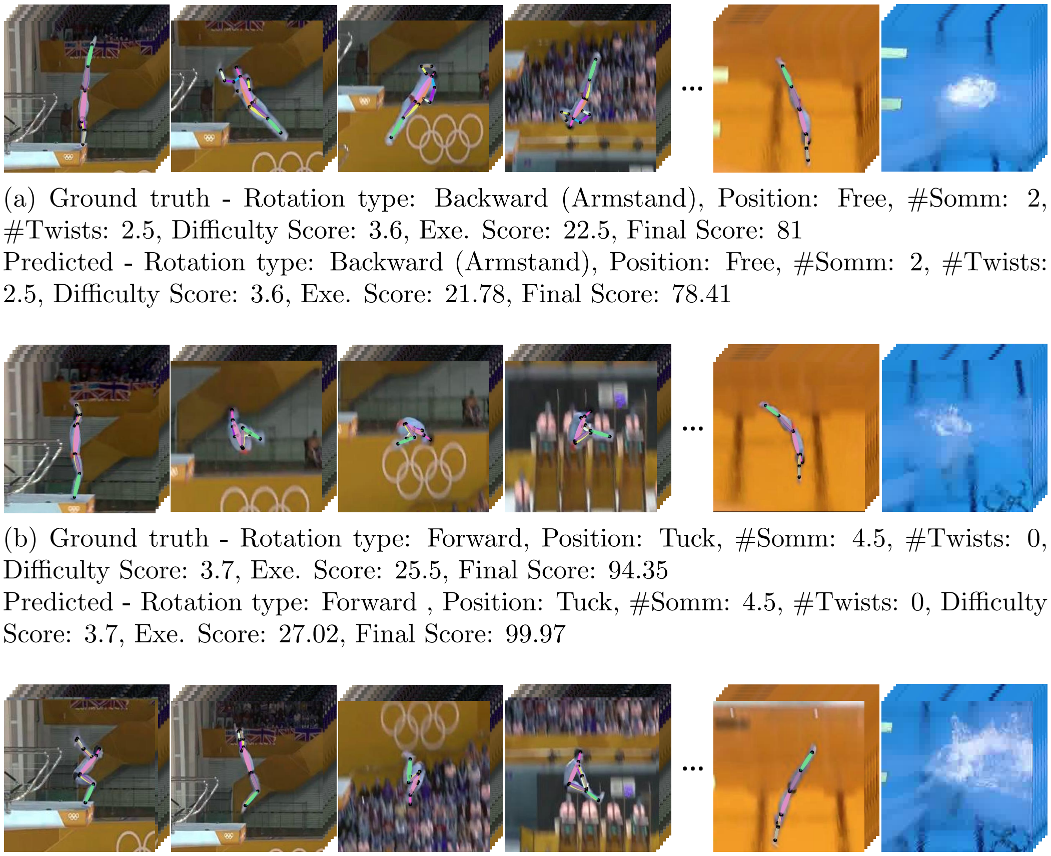
Generalization to Other Fields

Here we present the effectiveness of our method tested on an unseen event from another sport. To this end, we use our pretrained network on the UNLV-Diving dataset and fit a linear regressor on top of the resulting tensor of Fuser module to test unseen gymnastic vault routines. We freeze all other weights of the network to be unbiased around the gymnastic vault performances. As the weight contribution of the appearance features of each subtask is equally important, we blocked the bridge connector by setting the scale (k) to 1. All other parameters have remained the same as before. As we have not provided the framework for assessing the difficulty of the routine in a gymnastic vault performance (classifier + lookup table), we resort to the evaluation of our virtual refereeing system that provides the execution score.

We trained and tested the modified network on the gymnastic vault videos of AQA dataset [42]. The resulted Spearmann's Rank correlation for this task is 27.11. At the first glance it seems that the network is underperforming. However, it should be noted that the ground truth execution scores awarded by the judges are close to each other. In such condition, the difficulty score plays a distinctive role in ranking the athletes. Thus, having an effective difficulty assessor contributes to higher Spearmann's Rank correlation. The promising results of our network on the challenging task of assessing an unseen event from gymnastic vault show that the proposed network can be generalized to be used in other sports.

3.4.2 Qualitative Results

Fig. 3.6 presents some qualitative results for classifying the dive as well as assessing it in terms of execution and difficulty. The visual clues like amount



(a) Ground truth - Rotation type: Backward (Armstand), Position: Free, #Somm: 2, #Twists: 2.5, Difficulty Score: 3.6, Exe. Score: 22.5, Final Score: 81
Predicted - Rotation type: Backward (Armstand), Position: Free, #Somm: 2, #Twists: 2.5, Difficulty Score: 3.6, Exe. Score: 21.78, Final Score: 78.41

(b) Ground truth - Rotation type: Forward, Position: Tuck, #Somm: 4.5, #Twists: 0, Difficulty Score: 3.7, Exe. Score: 25.5, Final Score: 94.35
Predicted - Rotation type: Forward , Position: Tuck, #Somm: 4.5, #Twists: 0, Difficulty Score: 3.7, Exe. Score: 27.02, Final Score: 99.97

(c) Ground truth - Rotation type: Reverse, Position: Tuck, #Somm: 3.5, #Twists: 0, Difficulty Score: 3.3, Exe. Score: 17, Final Score: 56.1
Predicted - Rotation type: Reverse , Position: Tuck, #Somm: 3.5, #Twists: 0, Difficulty Score: 3.3, Exe. Score: 15.21, Final Score: 50.19

Figure 3.6: Qualitative results of our proposed method.

of splash as well as joint configuration during the performance, contribute to the execution score.

3.5 Discussion

In this chapter we present FALCONS, an engine of grading Olympic diving athletes, based on execution and difficulty assessors. Similar to what human judges do, the execution evaluation is based on both visual and pose features of the action. By introducing the notion of virtual super-joints, we augment the local correlations between connected joints with non-local joint dependencies of the action. The extracted pose sequences are also utilized by the bridge connector module to increase the contribution of the splash scene among other appearance clues. For extracting the difficulty of the action we propose a simple assessor that works on the basis of pose features. Finally, the overall score is provided by the multiplication of the execution and difficulty scores. The results show state-of-the-art performance compared to previous studies as well as acceptable generalization to unseen scenes from other sports.

Chapter 4

EAGLE-Eye: Extreme-pose Action Grader using detaiL bird's-Eye view

From just a glance at a video or by looking into its key frames, you may infer what action is being carried out in the video. But what if you are asked to assess the quality of the action? Judging a competitive sporting event and awarding scores to the performers needs a keen eye for details while still looking from the bird's eye view on the whole routine.

In this chapter we address a few other factors that the previous action assessment methods have neglected, causing the under-performance of their networks.

The first factor regards how to model an activity. A sports routine can be considered as a long-term activity, comprising some medium-term phases. For example, in a figure-skating contest the athlete phrases his/her program into intro, verse, chorus, and bridge in unison with the music being played (see Fig.4.1). Each of these phases is composed of some short-term elements like jumps, spins, and footwork sequences. The judge should keep track of the execution of each short-term element to award the Grade of Execution (GOE) [62]. On the other hand, the athlete's overall skating skill and composition of the elements through each medium-term phase are assessed to provide the Program Component Score (PCS) [61]. The same thing can be applied to other sports fields. A gym-vault routine is constructed of the approach,

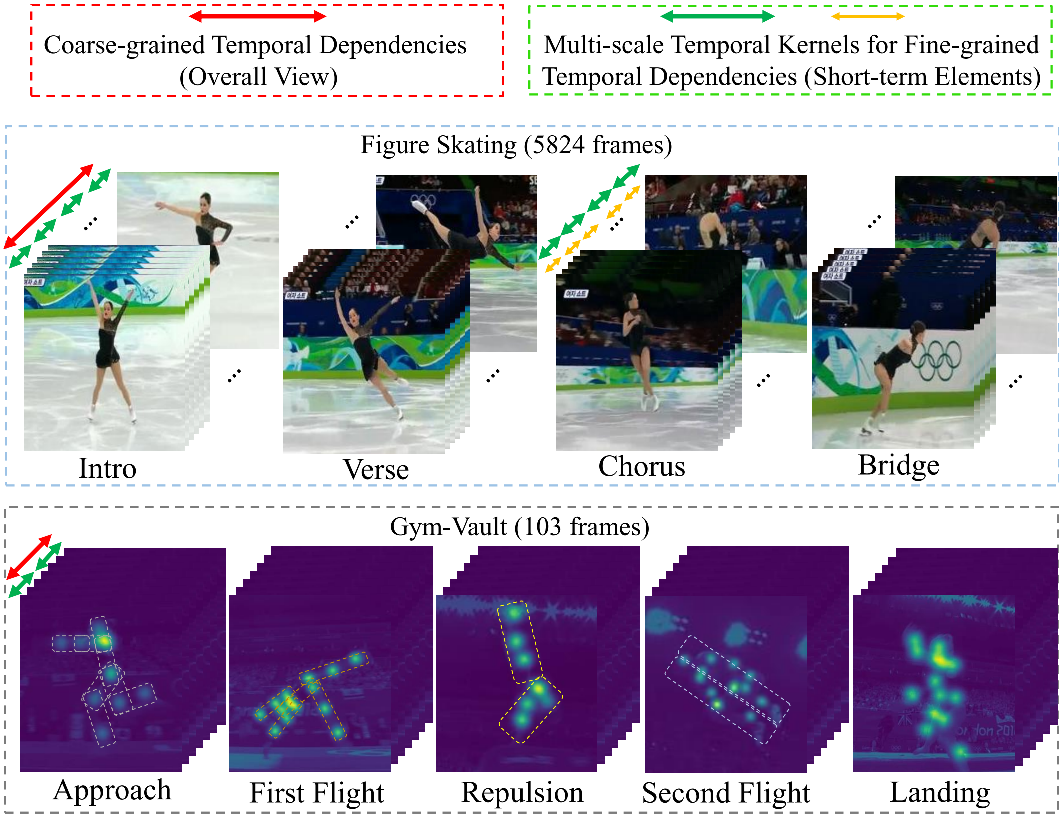

Coarse-grained Temporal Dependencies (Overall View)
Multi-scale Temporal Kernels for Fine-grained Temporal Dependencies (Short-term Elements)
Figure Skating (5824 frames)
Intro
Verse
Chorus
Bridge
Gym-Vault (103 frames)
Approach
First Flight
Repulsion
Second Flight
Landing

Figure 4.1: To assess the performance of an athlete both fine and coarse-grained temporal dependencies should be captured. The assessment is based on the coordination among the joints/body parts as well as appearance dynamics features.

first flight (a half twist on the ground with fully straight knees and body), repulsion (pushing off from the table with straight legs), second flight (airborne performance of saltos and twists in a tuck, pike, or free position), and landing phases (see Fig.4.1). The well-execution of each short-term element like a salto or a twist as well as having an overall smooth performance in medium-term phases would result in a high score. Existing methods [39], [40], [71] mainly rely on short-term temporal relations of few consecutive frames and neglect the long-term temporal relations that create a holistic view over phases. This problem escalates in the case of longer activities like figure skating in which each medium-term phase may take a few minutes to unfold. This fact has led the most of the previous approaches to be only applicable to short-lasting activities like gym-vault [33], [39], [40].

The second factor regards where and when to attend more to award the score. In typical individual sports footage the cameraman tracks the athlete to locate him at the center of the video. Thus, there should be a higher attention on the middle pixels of each frame to assess the performance. In the temporal domain, the judge may deliberately place some components over some others. For example, in a figure skating contest as the time goes on the athlete gets more tired and performing each element becomes more difficult. In order to acknowledge the skill and stamina of the skater, the judge gives bonus marks to well-execution of short-term elements in later phases [60]. Thus, the later parts of the performances are most likely to make the differences between the athletes' scores. Another example is gym-vault in which having a perfect landing in the last frames has a higher contribution than other phases in the scoring schema. On the other hand, there are some parts in long-term activities on which the judge may unintentionally focus more on. The PCS score of each figure-skater is awarded after the whole program which takes about three minutes to complete. As a result, the composition of the components and the holistic view of the last medium-phases are remembered most. Overall, there has been a lack of exploration of these factors in awarding the final score.

In this chapter, we propose EAGLE-Eye a modular two-stream network that sits on top of extracted appearance-based and pose-based features of a sports activity and evaluates the quality of the performance. The first stream is responsible for assessing the coordination among the joints and a variety of body parts with the help of a stack of JCA blocks (see the upper part of Fig.4.2). The first blocks capture the short-term temporal dependencies of the individual joints at different tempos with the help of multi-scale temporal kernels and temporal-wise channel convolutions. By stacking more of these JCAs both the temporal and semantic receptive field gets more broad, contributing to capture the holistic view of the performance as well as dependencies of different body parts/super-joints. Likewise, the second stream captures the both fine and coarse-grained appearance dynamics with the help of its stacked ADA blocks (see the lower part of Fig.4.2). The network is also supplied with some spatial and temporal attention blocks to increase the contribution of the frames' middle pixels and last medium-term phases in assessing the sports action.

We summarize the contributions of this chapter as follows:

- In order to handle estimation of the pose in the extreme contortions of the body in different sports, we have extended the *ExPose* dataset [33] to cover other sports than diving like synchronized diving, snow-boarding, and skiing. It is demonstrated that training the pose estimator on this dataset improves its performance in the extreme pose configurations of such sports.

- We propose a modular network that quantifies how well an action has been performed based on both fine and coarse-grained temporal dependencies. Same as the case of human judges grading schema, both visual and pose clues have been involved in the assessment.

- The proposed network not only outperforms the previous works in short-term actions assessment but also is the first to demonstrate a good generalization to the case of long-term sports activities like figure-skating. We further provide a thorough ablation study to evaluate the effectiveness of each block of the network.

4.1 Approach

This section outlines our proposed action quality assessor. The overview of our pipeline is depicted in Fig.4.2. Here we delve into each block of the network and describe the intuition behind them.

4.1.1 Explicit Spatio-temporal Attention

Given the small size of existing datasets in AQA, proposing a simple attention mechanism that does not make the network deeper is of great importance. Here, we introduce our simple yet effective AQA-specific attention block to capture the most important parts of a routine. The first objective of this block is to model the deliberate higher attention of a judge to the last parts' short-term elements of a routine in the temporal domain. To this end, we

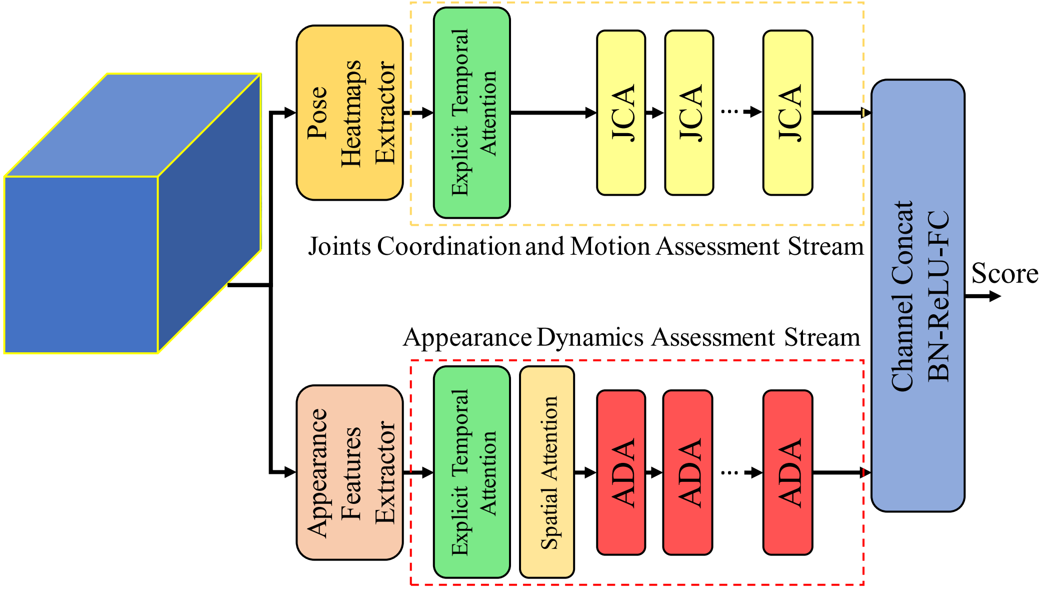


Figure 4.2: Overview of our pipeline. The network regresses the score of each performance based on both pose and appearance clues with attending to short-term elements as well as holistic view of the performance.

propose an explicit temporal attention block that gradually attenuates the contribution of the first phases of a performance. Let's consider the input of this block as X with the dimension of $T \times H \times W \times C$ in which T is the number of timestamps, H and W are the spatial size of each frame, and C is the number of channels. Then the block produces $\tilde{X}$ by element-wise multiplication of the input feature with explicit temporal importance mask (M^e):

$$\tilde{X}_{h,w,c} = X_{h,w,c} \odot M^e \tag{4.1}$$

$$M^e_t = a + (1 - a)\frac{t}{T} \tag{4.2}$$

In which $1 \leq c \leq C$, $1 \leq t \leq T$, $1 \leq h \leq H$, $1 \leq w \leq W$ and a is a constant coefficient $0 \leq a \leq 1$. As a result, $a \leq M^e_t \leq 1$.

The other objective of this module is to increase the spatial attention on the center of each frame. We propose a spatial attention block that applies a gaussian importance mask to its feature map input. As a result, the middle pixels of the appearance features which are more likely to represent the

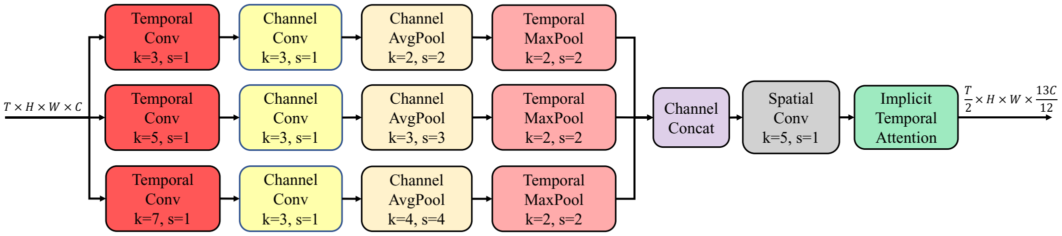


Figure 4.3: The overview of a JCA block.

athlete's body would have a higher contribution in awarding the score.

$$\tilde{X}_{t,c} = X_{t,c} \odot M^s \tag{4.3}$$

$$M^s_{x,y} = e^{-\frac{(x-\mu)^2+(y-\mu)^2}{2\sigma^2}} \tag{4.4}$$

4.1.2 Joints Coordination Assessor (JCA)

Our proposed JCA block is responsible for extracting the temporal pattern of body joints and parts which is depicted in Fig.4.3. Firstly it takes the $T \times H \times W \times C$ pose heatmaps of the routine in which C represents the number of joints. This input is fed to a set of multi-scale channel-wise separable temporal convolutions to capture temporal dependencies of joints at different scales. Utilizing fixed-size temporal kernels seem to be too rigid to model the complex temporal structure of short-term elements in a routine. A diver may perform a somersault and a twist together. These two elements may have different tempos and it is important to capture temporal dependencies throughout each element.

The next step is to capture the coordination among the joints. To this end, a temporal-wise separable channel convolution extracts the dependencies in the semantic subspace in which each channel represents an individual joint.

In assessing an action, the motion and coordination of different body parts is also monitored consistently. The symmetry of different body parts during the performance makes it aesthetically pleasant. In order to systematically capture such features we employ a set of average pooling with different kernel and stride sizes along channels. As a result, a variety of body parts/super-joints at multiple scales are formed. Consequently, the convolution filters of the next JCA block would capture the motion and coordination of these super-joints.

As discussed before, having a holistic view over the performance is of great importance for assessment purposes. Therefore, a temporal max pooling block at the end of each branch increases the temporal receptive field. Thanks to this, the next JCA block would be able to capture longer temporal dependencies. Obviously, minute-long activities like figure-skating require stacking more of these JCA blocks to capture the dependencies between distant frames (see Fig.4.2).

In the case of group activities, the pose heatmap contains more than one instance for each joint. In such cases the dependence between the joints of each performer with another should be extracted. For example, in a synchronized diving contest capturing the symmetry between the performers' joints is an important criterion of assessment. The spatial convolution block of JCA is responsible for catching such dependencies.

Finally, an implicit temporal attention block models the automatic fade of the first parts of a performance in the judge's short-term memory. A judge awards the PCS score of each figure-skating routine after the completion of the whole performance. Each routine takes two minutes and forty seconds on average. With the passage of time, the first parts of the performance become attenuated in the judges memory. To model this effect we propose a sigmoidal implicit temporal importance mask (M^i):

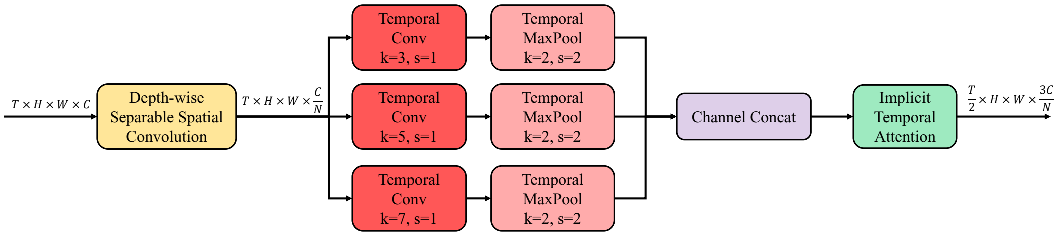
$$\tilde{X}_{h,w,c} = X_{h,w,c} \odot M^i \tag{4.5}$$

$$M^i_t = \frac{1 + be^{\frac{-T}{d}}}{1 + be^{\frac{-t}{d}}} \tag{4.6}$$

In which b and d are constant coefficients ($0 \leq b \leq 1$, $1 \ll d$). As a result $\frac{1+be^{\frac{-T}{d}}}{1+b} \leq M^i_t \leq 1$.

Consequently, the next JCA blocks which are responsible for capturing a holistic view of the performance would perceive a higher attention on the last parts. It should be noted that the explicit temporal attention block impacts the assessment of fine-grained dependencies in both short and long-term actions.

Figure 4.4: The overview of an ADA block.


$T \times H \times W \times C$
Depth-wise Separable Spatial Convolution
$T \times H \times W \times \frac{C}{N}$
Temporal Conv k=3, s=1
Temporal MaxPool k=2, s=2
Temporal Conv k=5, s=1
Temporal MaxPool k=2, s=2
Temporal Conv k=7, s=1
Temporal MaxPool k=2, s=2
Channel Concat
Implicit Temporal Attention
$\frac{T}{2} \times H \times W \times \frac{3C}{N}$

However, the implicit temporal attention block contributes to attending more on last phases coarse-grained dependencies in a long-term action assessment.

4.1.3 Appearance Dynamics Assessor (ADA)

We further propose ADA blocks to capture the dynamics of appearance features. The architecture of an ADA block is depicted in Fig.4.4.

The input to each ADA block is either the output of its previous ADA block or the output of appearance features extractor backbone (like I3D[8]). At the first step, a depth-wise separable spatial convolution captures the cross-channel correlations of the input feature map. It also shrinks the semantic subspace by a factor of N. As a result, stacking ADAs wouldn't lead to the explosion of the number of channels. The proposed AQA head relies heavily on its appearance features extractor backbone to capture the spatial dependencies. Thus, given the small changes that the spatial attention block has made, this lightweight spatial convolution layer suffices in the new setting. Secondly, same as the case of JCA blocks, multi-scale temporal kernels are employed to capture the different visual tempos of appearance clues. Finally, as depicted in Fig.4.2, the resulted feature map of the ADA stream is concatenated with the JCA stream over semantic subspace and fed to a $BN - ReLU - FC$ layer to get the final score.

4.2 Dataset

G-ExPose extends *ExPose*[33] by 7500 annotated images from four different sports.The *G-ExPose* dataset contains 2500 snowboarding, 2000 skiing, 1500 synchronized diving, and 1500 gym-vault 2D annotated images. The snow-

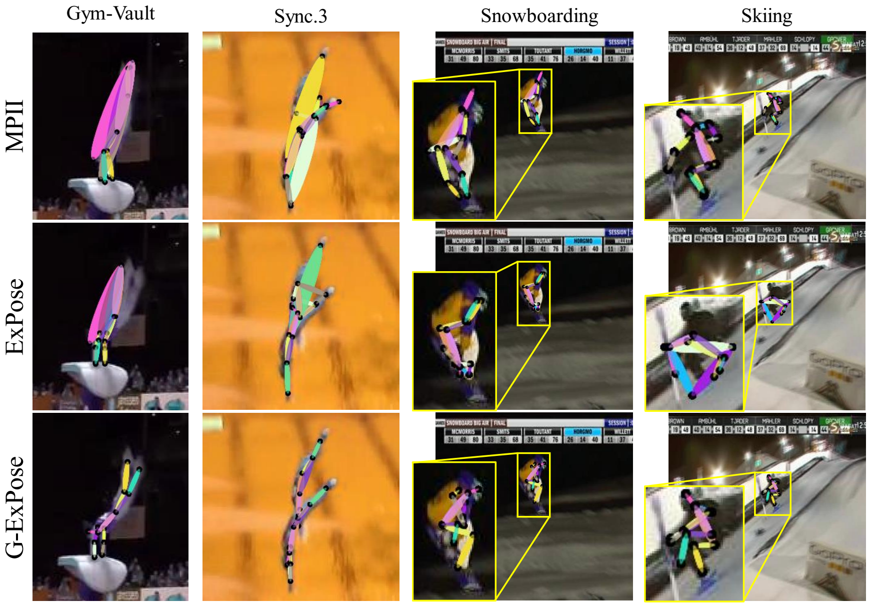


Figure 4.5: Qualitative test-time pose predictions of HRNet[55] model trained on MPII (first row), ExPose (second row), and G-ExPose (third row), respectively.

boarding images are obtained from 2018, 2019, and 2020 X-games competitions at Aspen. The skiing images are taken from X-games 2020 ski big air contests at Aspen and Norway. We extended *ExPose* to also cover highly occluded synchronized diving images by introducing a set of 1500 annotated images from women's 3 meter springboard and men's 10 meter platform synchronized diving finals at 2016 European diving championships in London. We further enlarged gym-vault samples of *ExPose* by annotating 1500 images from Rio 2016 Olympics and Stuttgart 2019 world championships women's vault finals. In G-ExPose, we followed the same image collection and annotation strategy that we used in ExPose.

We first present the qualitative results of training the HRNet pose estimator on the G-ExPose dataset. As evident in Fig. 4.5, MPII dataset is not suitable for estimating the contortive poses of a competitive sports activity which is taken by a moving camera. Besides, although we demonstrated the effectiveness of using ExPose in estimating the pose in a diving routine, utilizing this dataset leads to failure of the estimator in the case of other contortive

Sports Field	**MPII[3]**	**ExPose**	**G-ExPose**
Diving	47.6	**83.7**	—
Sync. 3m	27.2	45.2	**56.7**
Sync. 10m	29.3	53.6	**63.9**
Skiing	18.8	5.5	**30.5**
Snowboarding	20.1	9.1	**31.0**
Gym Vault	24.0	38.8	**53.2**

Table 4.1: The quantitative results of HRNet pose estimator[55] on the 100 annotated images of each extreme sports field when trained on MPII, ExPose and our G-ExPose dataset. The evaluation metric is the standard PCKh@0.5[3], [55]. The position of a joint is correctly estimated if its distance with the ground truth is within 50% of the head segment length.

sports.

Besides the qualitative evaluation, we further quantitatively assessed the effectiveness of the datasets in extreme pose configurations in comparison with *ExPose* and an in-the-wild normal activities pose dataset like *MPII* in Tab. 4.1. To this end, we first picked 10 videos from each field of AQA-7[40] dataset and annotated 10 images from each video with a focus on the main parts of the execution. This dataset contains 1106 sports routine videos as well as their correspondent score from diving, synchronized diving (3 and 10 meters), gym-vault, snowboarding, and skiing. As a result, a set of 600 annotated images from 6 different fields got collected. We then evaluated the performance of the SOTA HRNet [55] pose estimator on the 100 images of each field when it is trained on *G-ExPose*, *ExPose*, and *MPII.* As it can be seen from Tab.4.1, the HRNet which is trained on G-ExPose, outperforms others in the extreme pose estimation task. It should be noted that there is no conflict between G-ExPose and AQA-7 source events.

4.3 Experiments

4.3.1 Datasets and Implementation Details

For short-term AQA we follow recent works [39], [40] and evaluate our approach using the AQA-7 dataset. Each video of the dataset is originally normalized to 103 frames. We follow the same train-test data split as [39], [40].

In order to get the appearance features of each video, we use the output of `mixed-5c` layer of an I3D network pretrained on the Kinetics dataset[8]. Let's consider the input video to the I3D model as a $T\times H\times W\times 3$ matrix. The `mixed-5c` layer of the I3D network outputs a $\frac{T}{8}\times 7\times 7\times 1024$ to be fed into the ADA stream. To make the both ADA and JCA streams output features with the same timesteps, the JCA stream is followed by a temporal max pooling with the stride and kernel size of 8. Besides, an average pooling at the end of JCA stream reduces the spatial size of the pose features output to have the same spatial dimension as the ADA stream output. To stabilize the learning process and capture the complex structure of the data, a BatchNorm-ReLU layer is embedded between two successive ADA (JCA) blocks.

For our pose features extractor backbone we entangled the DiMP [4] visual object tracker with the HRNet pose estimator (trained on *G-ExPose*). The channel shrinkage factor of the ADA blocks (N) is set to 2. The short-term attenuation temporal coefficient (a in Eq.4.2) is set to 0.9. The mean and the standard deviation of the spatial importance mask (see Eq.4.4) are set to 4 and 5 respectively. As discussed before, the implicit temporal attention block should only impact the long-term activities assessing. Thus, we set the long-term temporal attenuation coefficient (b in Eq.4.6) to 0 in short-term action assessment.

For assessing long-term actions we utilize the `fc6` layer of C3D network[58] to get the appearance features of the performance. Since the spatial size of the extracted features is 1, the spatial attention block is removed.

During the training, the backbone networks are frozen. We first train the EAGLE-Eye network on diving samples of AQA-7 dataset. For assessing other short-term sports except skiing, we fine-tune the diving pretrained model on each of the sports separately. The skiing assessor network jump-starts from the model that is fine-tuned on snowboarding samples. In order to assess the long-term figure-skating videos, we first pretrain the EAGLE-Eye on the task of classifying the sports from each other. To this end, we first repeat each short-term sports video (103 frames) to fit 5824 frames of figure-skating samples. We then train the network to classify each sport from the other.

Method	Diving	Vault	Skiing	Snowboard
Pose-DCT-SVR [44]	53.00	—	—	—
ConvISA [27]	—	—	—	—
ST-GCN [72]	32.86	57.70	16.81	12.34
C3D-LSTM [40]	60.47	56.36	45.93	50.29
C3D-SVR [40]	79.02	68.24	52.09	40.06
JR-GCN [39]	76.30	73.58	60.06	54.05
AIM [14]	74.19	72.96	58.90	49.60
C3D-(S+M)LSTM [71]	—	—	—	—
Ours	83.31	74.11	66.35	64.47

Table 4.2: Detailed results on individual actions of AQA-7[40] dataset that contains short-term activities. First and second best are shown in color.

Finally, we use the resulted trained weights to fine-tune the EAGLE-Eye for assessing the figure-skating videos.

In order to assess the effectiveness of the proposed model in long-term sports activities we evaluate it on the extended version of the MIT-Skate [44] by [42]. This dataset contains the awarded scores of 171 single figure-skating videos that take 2.5 minutes on average. In order to normalize all videos to a fixed number of frames we first extract the frames of all videos at 25 fps . We then zero-pad the first frames of each video to fit to 5824 frames which is the longest video's number of frames. We follow [42] and randomly split the dataset into 100 samples for training and 71 for testing. Since a figure skating routine does not involve extreme pose configurations, we train the pose estimator backbone on the COCO+Foot dataset [5]. In the long-term action assessing we set b and d coefficients to 0.5 and 1000 respectively.

We train the model for 500 epochs with the learning rate of 0.005 and the batch size of 20 using Adam optimizer[23]. We use the MSE loss function to train the model and award the scores, following what other regression-based AQA methods[39], [40], [42] have done. To be consistent with the previous works, we use the Spearman's Rank correlation to evaluate the performance of the model and compare the predicted scores with the ground-truth. For further details refer to the supplementary document.

Method	Sync. 3m	Sync. 10m	**Avg. Corr.**	Skating
Pose-DCT-SVR [44]	—	—	—	35.00
ConvISA [27]	—	—	—	45.00
ST-GCN [72] 66.00	64.83	44.33	—	
C3D-LSTM [40]	79.12	69.27	61.65	51.07*
C3D-SVR [40]	59.37	91.20	69.37	53.00
JR-GCN [39]	90.13	92.54	78.49	—
AIM [14]	92.98	90.43	77.89	—
C3D-(S+M)LSTM [71]	—	—	—	57.69*
Ours	91.43	91.58	81.40	60.10

Table 4.3: Detailed results on synchronized short-term actions of AQA-7[40] dataset and long-term figure skating videos of extended MIT-Skate dataset (171 samples)[42]. First and second best are shown in color. Following [39], [40], we use Fisher's z-value to compute the average correlation between all six short-term sports fields of AQA-7 dataset. The results marked with * are obtained by reimplementing the correspondent method. [71] reported 59.00 Sp. Corr. for the old MIT-Skate dataset (150 samples) in the original paper.

4.3.2 Results

We first evaluate the performance of our network on short-term activities of AQA-7 dataset. It should be noted that we have used two ADAs and two JCAs for assessing a short-term action. As it can be seen in Tables 4.2, 4.3, the proposed method outperforms the existing SOTA AQA methods. Its worth mentioning that the JR-GCN [39] and AIM[14] methods have used the excessive optical flow information while our method resorts to the RGB frames input. The largest gaps belong to skiing and snowboarding sports. In these sports' video footage, the size of the athlete is much smaller than the size of the whole frame. Therefore, it is not surprising that methods like [40] which only rely on whole-scene appearance features to regress the score of the performers, are underperforming significantly. Besides, in such fields the position of each individual joint is as important as the symmetry of different body parts during the execution. Failure to grab the board or any insecurity that requires hand movements to remain stable affects the score negatively. Thus, extracting the features of some predefined local patches around the joints (which has been done in[39]) results in neglecting the individual joints position and motion of

Ablated Model	Diving	Vault	Skiing	Snowboard
W/o Explicit Temporal Att. Block	82.66	72.91	62.11	63.23
W/o Spatial Att. Block	82.83	72.11	63.45	61.16
Fixed-size Temporal Kernels	76.71	67.17	55.09	51.99
W/o Channel Avg	81.87	71.24	63.73	62.15
W/o JCA Stream	80.95	70.96	60.46	60.63
W/o ADA Stream	74.30	70.58	57.90	57.90
# JCA and ADA Blocks = $K-1$	79.86	72.24	62.55	59.83
# JCA and ADA Blocks = $K+1$	81.53	71.18	64.94	62.11
Only Whole-Scene Appearance	63.39	68.72	51.79	50.53
W/o Implicit Temporal Att. Block	—	—	—	—
MPII-trained pose estimator	80.19	69.72	62.88	62.95
Ours	**83.31**	**74.11**	**66.35**	**64.47**

Table 4.4: Ablation study results on individual actions of AQA-7 dataset[40]. We systematically removed the components of our network to evaluate their contribution to the full model. K is equal to 2 for short-term actions assessment

some body parts to award the score. On the other hand, our pose-based assessment stream not only works intuitively as an action localizer, it also judges the position of the joints as well as the motion of a variety of body parts.

We further evaluate the performance of our model on the long-term figure-skating sports activity. In order to have a fair comparison with the existing works, we changed the appearance features extractor backbone to C3D to have the same backbone as theirs. Following [71], we feed the output of `fc6` layer of C3D network which is pre-trained on Sports-1M dataset[22] to our ADA stream. As discussed before, long-term temporal reasoning is crucial to model the judge's impression of the overall performance of the figure-skater. To this end, more JCAs and ADAs (here we use 4) are stacked in the long-term activities assessment. As demonstrated in Tab.4.3, our model generalizes well to long-term action assessment task.

We conduct a comprehensive ablation study to evaluate the effectiveness of our models components (see Tables 4.4, 4.5). We first removed the explicit temporal attention block to uniformly attend to all frames in a short-term activity. As a result, we are neglecting the fact that having a clean landing in a snowboarding routine or performing a vertical entry to the water with

Ablated Model	Sync. 3m	Sync. 10m	Skating
W/o Explicit Temporal Att. Block	90.59	89.72	59.66
W/o Spatial Att. Block	**91.50**	90.76	—
Fixed-size Temporal Kernels	87.17	86.81	54.53
W/o Channel Avg	89.28	90.43	57.22
W/o JCA Stream	88.70	90.18	55.49
W/o ADA Stream	85.28	85.64	51.11
# JCA and ADA Blocks = $K-1$	88.16	89.20	57.72
# JCA and ADA Blocks = $K+1$	90.24	88.32	58.41
Only Whole-Scene Appearance	87.83	88.32	44.23
W/o Implicit Temporal Att. Block	—	—	58.73
MPII-trained pose estimator	86.51	87.94	—
Ours	91.43	**91.58**	**60.10**

Table 4.5: Continuation of 4.4. Ablation study results on synchronized short-term sports fields of AQA-7 dataset[40] samples and long-term figure-skating samples of MIT-Skate dataset[42]. K is equal to 2 for short-term and 4 for long-term assessment.

the least amount of splash in a diving performance are the most distinctive features of the execution[1]. In the second set of experiments we removed the spatial attention block of the ADA stream. Therefore, we equally attend to each pixel of the extracted appearance features, no matter whether it belongs to the background or the athlete's body. The third row of Tables 4.4,4.5 refers to the results of using fixed-size temporal kernels instead of multi-scale ones. Consequently, the same temporal kernel size for capturing complex temporal dependencies of the two short-term elements that are performed together would be used. Fourthly, we removed the temporal-wise average pooling of the JCA blocks. As a result, the coordination among the virtual super-joints/body parts would not be captured. In the next set of experiments we removed the whole JCA stream and ADA streams to validate their contribution in the action assessment. The drastic drop of the performance is because of solely relying on either appearance dynamics or joints coordination and motion features. We further changed the number of JCA and ADA blocks to confirm the optimality using two blocks. If we only use one JCA block the dependencies among the formed body parts/virtual super-joints would not be captured. Furthermore, given the small number of frames in short-term activities using

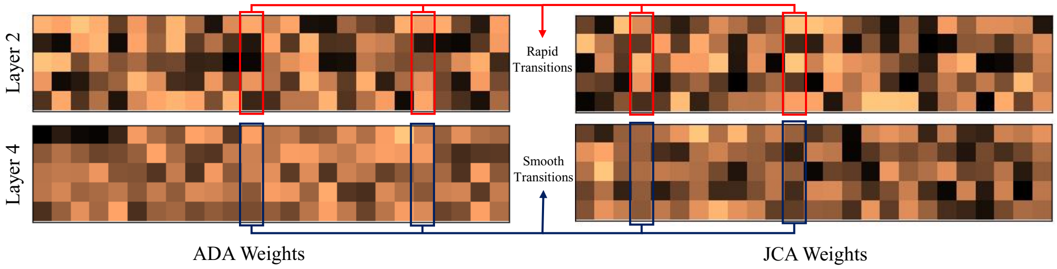
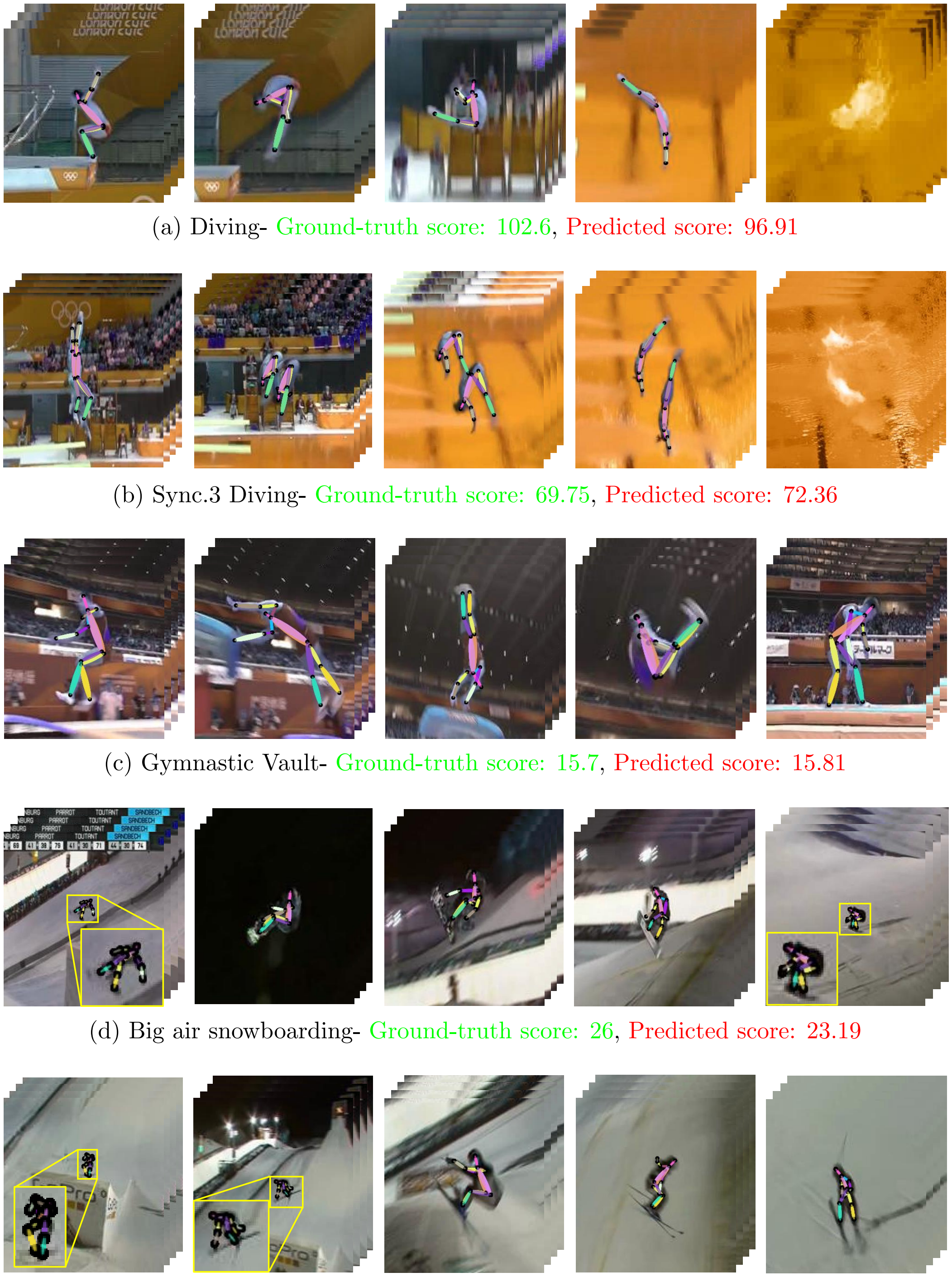


Figure 4.6: Learned weights of the k=5 temporal convolution in ADA (left) and JCA(right) blocks for figure-skating. The upper plot refers to the second ADA(JCA) and the lower one refers to the fourth ADA(JCA). Due to limited space the first 25 channels have been visualized.

three JCA and ADA does not seem beneficial since it leads to the increase of the number of parameters and overfitting problem. Next we evaluated the performance of the network when it only uses the appearance features of the backbone, by removing the JCA and ADA blocks and completely neglecting the pose features. We further evaluated the performance of the network in different numbers of JCAs and ADAs blocks and it turned out that using 4 blocks leads to the best performance. We then removed the implicit temporal attention blocks, assuming that the all medium-term phases of the long-term action have the same contribution in the overall impression. It should be noted that this block is already deactivated in short-term action assessment. Finally, we assessed the performance of the network when the pose estimator is trained on the *MPII* normal activities pose dataset. As listed in Tables 4.4,4.5, the network achieves its full potential when all of its components are utilized.

We then visualize the learned weights of our model in Tab.4.6 following [19]. For brevity, we resorted to the temporal kernel with size of 5 and compared the transitions among the learned weights in each channel between two ADA(JCA) blocks. The upper plot in Fig.4.6 represents the learned weights of the second ADA(JCA) block. The rapid transitions among the weights in each channel demonstrate that this block is capturing the fine-grained temporal dependencies. On the other hand, as depicted in lower plot of Fig.4.6, the transitions among the 4th block learned weights are smoother, confirming the fact that this block is responsible for capturing coarse-grained dependencies.



(a) Diving- Ground-truth score: 102.6, Predicted score: 96.91

(b) Sync.3 Diving- Ground-truth score: 69.75, Predicted score: 72.36

(c) Gymnastic Vault- Ground-truth score: 15.7, Predicted score: 15.81

(d) Big air snowboarding- Ground-truth score: 26, Predicted score: 23.19

(e) Big air skiing- Ground-truth score: 47, Predicted score: 44.20

Figure 4.7: Some qualitative results on short-term actions of AQA-7 dataset[40]

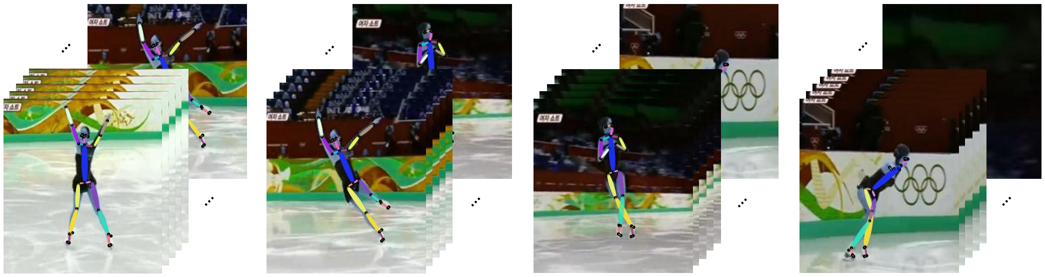


Figure 4.8: The qualitative results of long-term figure-skating sport (MIT-Skate dataset [44])- Ground-truth score: 49.74, Predicted score: 47.18

In the end, we present some qualitative results for assessing the quality of both short and long-term activities as well as their estimated pose sequence. It should be noted that in the short-term actions the pose estimator is trained on our *G-ExPose* while for figure-skating it is trained on the COCO+Foot dataset [5]. Fig.4.7 depicts the qualitative results for five different short-term sports; diving, synchronized diving (3m), gym vault, snowboarding, and skiing. The qualitative results of long-term figure-skating assessment are depicted in Fig.4.8.

4.4 Discussion

In this chapter, we argue that evaluating the quality of an action requires incorporating appearance and pose features of the performance in both fine and coarse-grained temporal scales. To this end, we present a modular two-stream network that sits on top of extracted appearance and pose features of an action to assess it. The first stream is composed of a stack of JCA blocks that are responsible for evaluating the configuration of the joints and body parts throughout the performance. The other stream assesses the appearance dynamics of the action owing to its constituting ADA blocks. Empowering the network with more JCA and ADA blocks leads to capturing long-term coarse-grained temporal dependencies that represent overall impression of the program. Our experimental evaluation demonstrates that our method achieves the state-of-the-art results on short-term action assessment in comparison to prior works. Moreover, the proposed modular network adapts simply to long-term action assessment by stacking more JCA and ADA blocks and outperforms the previous works on this task as well.

Chapter 5

Enhancing Human Motion Assessment by Self-supervised Representation Learning

The problem of labeled data scarcity has marred the effectiveness of using deep neural networks in some fields like medical image analysis. The difficulty of collecting the images and privacy concerns have led to limited access to both healthy and abnormal samples. However, that is not necessarily the case for healthy samples in video-based healthcare monitoring and rehabilitation movement assessment. In such cases, a healthcare professional asks the patient to do some daily activities like walking and sit-standing. Then the performance of the patient is assessed based on posture accuracy of the body parts, motion smoothness, and the speed of the movements. Although getting such video samples of the patients that are labeled by an expert is still an issue, there is a myriad of such daily activities performed by healthy people readily available in datasets like UWA3D [47] and UTKinect [70]. In this chapter, we address the question of: how can we learn a representation from these healthy samples to help develop a more accurate yet shallower network to assess the performance of patients' actions?

We as humans have seen lots of such daily activities in our lives. Our visual system has learned to be sensitive to anomalies it sees like an abnormal walking pace or an impaired posture over time [15], [68]. As a result, given a stroke or Parkinson's patient movement, we would be able to estimate the severity

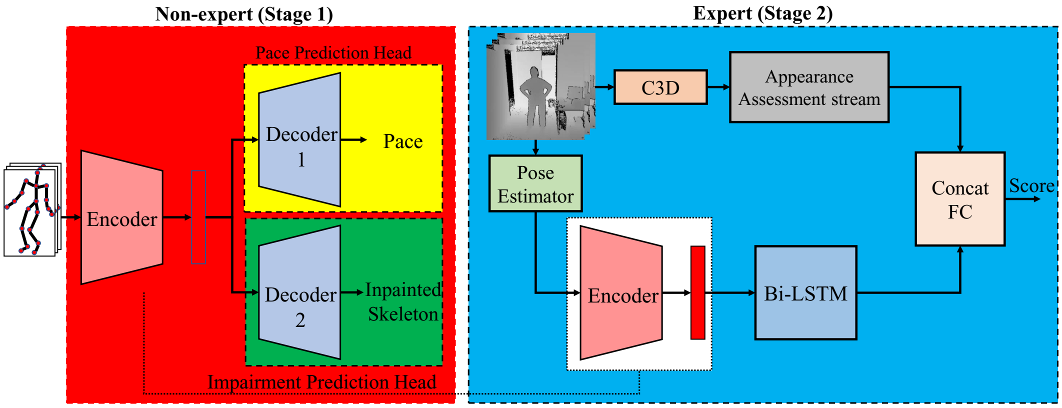


Figure 5.1: An overview of our two-stage pipeline. At stage 1, the non-expert module is trained on the large-scale, well-annotated source benchmarks of healthy samples performing daily living activities. At stage 2, the expert module is further trained on the small-scale and less-annotated target training set for assessing e.g. disease severity, by taking as input both the learned representations from the non-expert module and the motion appearance features of the target dataset. Note the input to expert module is an RGB video. However, since we are not permitted to display the raw color images, their depth images are instead presented here as a substitute.

of the disease to some extent based on how slow or impaired it is. Inspired by this fact, we propose a *Non-expert* network that takes the pose sequence of an activity performed by a healthy person and learns some representations of the action in a self-supervised manner. As it can be seen from the left side of Figure 5.1, this network has a multi-head decoder with a shared encoder. First, some slower pose sequences are generated from the normal paced sample by altering the temporal sampling rate. The goal of the first decoder is to estimate the sampling rate to be as close as possible to the ground-truth one, given the representations of the encoder. Secondly, we manually inpaint some joints of the skeleton in the sequence and the second decoder is employed to estimate those masked joints. After training this deep network on the large dataset of healthy samples we would have a representation that is sensitive to the both pace and impaired posture of the movements.

At the next step, the pose sequences of the target dataset which has a fewer number of samples are fed to the non-expert encoder to get the representations of the action. These representations are fed to a shallower *Expert* network to assess the performance of the patient when doing that specific action (see the right side of Figure 5.1). Since the pose features are not informative enough to assess the smoothness of the movements and overall posture of the body, the expert network is further equipped with another stream of appearance features assessment. These features can come from a backbone network like the well-known C3D [58] or I3D [8]. The whole network can be seen as a collaboration of the deep non-expert and shallow expert networks. The non-expert provides feedback (representations) about the pace and impairment of an action based on lots of healthy samples it has seen beforehand. The expert network takes this feedback as well as the appearance features of the video to quantify how well the action was performed. If the video itself (target dataset) is not available due to privacy issues or any other reason, the expert resorts to the pose representations that come from the non-expert to assess the performance.

Our main contributions in this chapter can be summarized as follows:

- We present a two-stage network to automatically assess the performance of a patient from the RGB video of doing an action. At the first stage, a non-expert network learns representations of a large of-the-shelf dataset pose samples by predicting their pace and impairment in a self-supervised manner. Finally, an expert network leverages the learned representations of the target dataset and appearance features of the video to assess the severity of the disease. To the best of our knowledge, we are the first to make use of self-supervised representation learning in action quality assessment.

- The proposed method not only shows a superior performance in comparison to previous works in rehabilitation progress assessment but also is the first to show a good generalization to the case of infants general movements assessment and their early disease detection.

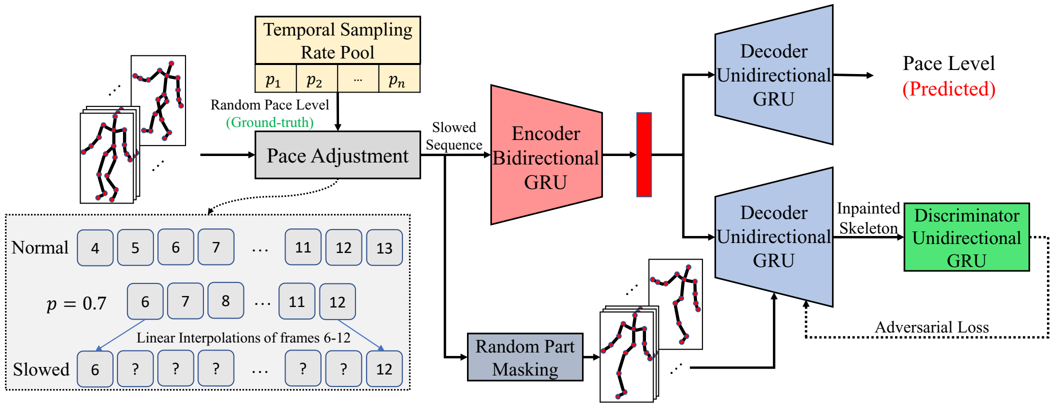


Figure 5.2: The architecture of the Non-expert network. The upper decoder predicts the pace of the sequence and the lower one inpaints the masked joints of the skeleton.

5.1 Approach

This section outlines our two-stage network and gives more details about each block of Figure 5.1. At the first stage, an encoder-decoder based non-expert network learns representations of a pose sequence from an off-the-shelf dataset in a self-supervised manner. Secondly, the expert uses the encoder and the representations from the previous stage to assess the performance of a patient's action sample of the target dataset.

5.1.1 Non-expert

The architecture of our non-expert network is depicted in Figure 5.2. This multi-head network aims at predicting the pace and impairment of a pose sequence without requiring any explicit label and just with the help of the pose sequence itself. Let's denote all the joints of a skeleton sequence as a set $S = \{X_t^k|t = 1, 2, ..., T; k = 1, 2, ..., J\}$, where X_t^k is the position of k^{th} joint in the t^{th} frame. T is the total number of frames and J is the number of the joints in a human body skeleton. At the first step, a pace level is randomly selected from a pool of N temporal sampling rates $P = \{p_i|i = 1, 2, ..., N; p_i \leq 1\}$. This pace level determines how slow the pose sequence should be. Secondly, a pace adjustment module samples $[T \times p_i]$ consecutive frames from S. To make the new sequence have the same length as S one may repeat the sampled frames or use their linear interpolation. Here, we use the second approach to make the new sequence (S') fit T frames. As a result, we would have a pose sequence that is slower than the original one. The smaller p_i gets, the slower S' would be. An overview of how the pace adjustment module works is presented in the lower-left side of Figure 5.2.

The slowed sequence is then fed to our shared encoder (a bidirectional GRU with two hidden layers) to generate the representations of the sample. Each of the decoders then takes these representations to do two different pretext tasks. The first decoder is responsible for estimating the pace of the slowed sequence (upper part of Figure 5.2). We use the multi-class cross-entropy loss for the pace decoder head:

$$L_{pace} = -\sum_{i=1}^{N} p_i \log \hat{p_i} \tag{5.1}$$

Where p_i is the ground truth pace level sampled from the temporal sampling rate pool and $\hat{p_i}$ is the predicted pace by the first decoder.

In parallel, the second decoder takes the representations of the encoder as the initial state of the cells and the manually masked skeleton sequence as the inputs to the cells of the GRU. In order to get the masked skeleton sequence, a random part (b_i) is first sampled from the set of legs and hands of the body's two sides $B = \{b_i | i = 1, 2, 3, 4\}$. The position of each joint that belongs to that part is then set to zero . This would result in a new sequence ($\hat{S'}$). It should be noted that once a part is chosen, all of the frames in the sequence would be masked in the same way. As a result, the decoder wouldn't be able to estimate the masked part of a frame from its neighbors. For the second decoder we use the reconstruction $L2$ loss:

$$L_{rec} = \sum_{t=1}^{T} \sum_{k=1}^{J} (X_t'^k - \hat{X_t'^k})^2 \tag{5.2}$$

Where $X_t'^k$ is the ground-truth unmasked skeleton sequence and $\hat{X_t'^k}$ is the inpainted one. Although the decoder is able to fill in the masked joints, there is no guarantee that the inpainted skeleton is visually plausible. To address this issue, a discriminator sits on top of the inpainted skeleton to adversarially make it more realistic. Thus, the inpainting decoder loss should be revised as follows:

$$L_{inpaint} = L_{rec} + \alpha L_{adv} = \sum_{t=1}^{T} \sum_{k=1}^{J} ({X'}_t^k - \hat{{X'}_t^k})^2 + \\ \alpha \left(\log(Disc(S')) + \log(1 - Disc(\hat{S}')) \right) \tag{5.3}$$

Where α is a constant that adjusts the adversarial loss to make the optimization stable. The parameters of the pace and inpainting decoders are updated w.r.t the L_{pace} and $L_{inpaint}$ respectively. However, since the encoder is shared between these two decoders, its loss function should have a touch of both L_{pace} and L_{rec}. Thus, the encoder's parameters are updated based on the following loss function:

$$L_{enc} = L_{rec} + \beta L_{pace} \tag{5.4}$$

Where β is a constant that controls the weight of two decoders. It should be noted that per Zheng *et al.* [76] suggestion, the encoder should stick to generating the representation regardless of how visually realistic the inpainted skeleton is. This strategy would help the encoder to focus on capturing the motion dynamics for the next stage and not to sacrifice it for style and realisticity of the sequence which can be solely handled by the decoder. Therefore, the L_{adv} shouldn't be involved in the encoder's parameters updating process.

At the end of this stage, we would have an encoder that is going to be used in the next stage to provide representations for samples of the target dataset. In other words, the non-expert learns representations by doing the pretext tasks and the expert uses these representations to perform the downstream task which is the action quality assessment.

5.1.2 Expert

The goal of our expert network is to use the non-expert representations and appearance features of a patient's action to assess it. As depicted in Figure

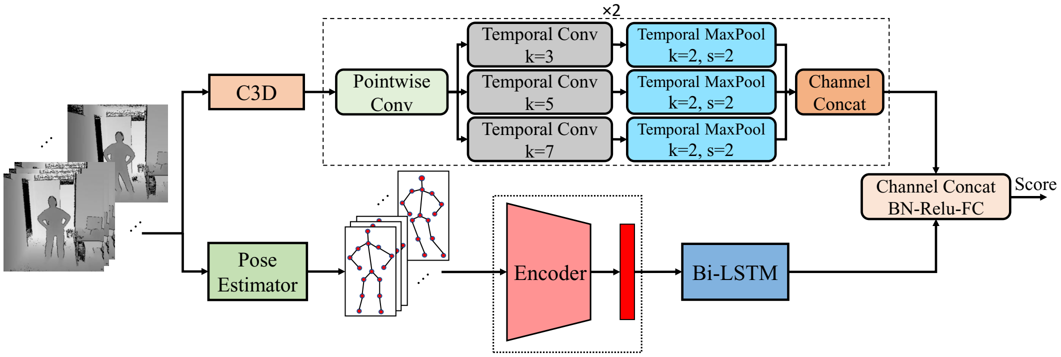


Figure 5.3: The architecture of our two-stream Expert. This network takes the RGB video of a patients action and evaluates it based on both appearance and pose features. In case due to privacy concerns the videos aren't provided and we only have access to the pose sequences, we stick to the lower stream of the network.

5.3, given the RGB video input, this two-stream network assesses the pose and visual clues of the action in parallel.

The upper stream is responsible for evaluating the appearance features. To this end, the video is first fed to a feature extractor backbone like C3D. The resulted features are then fed to a stack of two appearance assessment modules to get more high-level spatio-temporal dependencies inspired by [34]. The first block of this module is a point-wise convolution to reduce the number of channels in the extracted features. As a result, this stream would have a comparable number of channels to the lower one. At the next step, the output goes through a set of depth-wise separable temporal convolutions to capture visual clues with different tempos. Then, a temporal max-pooling layer increases the receptive field in the temporal subspace. As a result, the next module of the stack would observe a broader temporal field to capture long-term temporal dependencies. Finally, the three tensors from the three branches are concatenated over the semantic subspace.

The lower stream first takes the video and extracts the pose sequence of it using an off-the-shelf pose estimator like OpenPose [5]or HRNet [55]. The pose sequence is then fed to the encoder that we trained in the previous stage for the non-expert. At the next step, a shallow bidirectional LSTM with one hidden layer takes the resulted representations to get the dependencies between the frames. Finally, the output of this stream is concatenated with the appearance assessment stream and fed to a stack of batch normalization - ReLU activation - FC to provide the score for the performance.

As a result, our two-stage network not only takes advantage of the large off-the-shelf datasets to learn representations of a movement but also isn't susceptible to be overfitted to the small target dataset. Intuitively, the non-expert function resembles what humans do. It provides some generic representations about the sequence. The expert which could be the healthcare professional in the real world knows these representations and is able to do more detailed analysis about both pose and visual clues to evaluate the patient's performance .

5.2 Experiments

5.2.1 Implementation Details

Dataset: We use the off-the-shelf UWA3D [47] and UTKinect[70] datasets to train the non-expert network on the pretext task. The UWA3D dataset contains 30 daily living actions (e.g. sitting down, bending, etc.) performed by 10 subjects from 4 different views. The dataset has a set of 1075 sequences in total. The UTKinect dataset consists of 200 samples of 10 actions like walking, picking up, etc. performed twice by 10 subjects. The non-expert is first trained on the UWA3D and the learned parameters are then used as the pretrained weights for training the network on the UTKinect dataset.

The expert network is trained on the target KIMORE dataset [6] to do the downstream task of action quality assessment. This dataset collected a set of 78 subjects performing 5 different exercises like squatting and moving a bar. 44 of these subjects are healthy people (29 males, 15 females) and 34 of them (15 males, 19 females) are suffering from motor dysfunctions due to Parkinson's, stroke, or lower back pain. All of the samples are labeled with clinical scores by a healthcare professional. This dataset is the only accessible and publicly available annotated dataset at the time of writing this thesis.

To further assess the performance of the proposed method we study its generalization to the new task of infants' general movement assessment. To this end, we trained the expert network on the dataset of infants' neuromotor risk evaluation released by [9]. This dataset consists of 19 at-risk infants playing with a toy and a set of 85 healthy samples from YouTube. A clinician has annotated the at-risk infants' movements into low, moderate, and high risk for motor dysfunction and getting cerebral palsy (CP) in the future. CP causes stiffness of the joints which affects the normal pace, symmetricity of the movements, and overall balance of the infant. A CP patient may not bring both hands together when playing and weakness of the joints causes delays and slowness in performing the movements.

Training Details: The temporal sampling rate pool in the non-expert network contains five different pace levels: 0.6, 0.7, 0.8, 0.9,1. The slower pace levels represent the severe cases of the disease and $p = 1$ is the normal pace of a healthy person. The reason behind using these levels is to cover all levels of severity based on a recent claim that the average motor frequency of Parkinson patients walking is $\frac{0.94}{1.3}$ slower than that of healthy people [38]. The encoder and both decoders of the non-expert network have two hidden layers. In order to get the appearance features of the target dataset samples, we use the output of `fc6` layer of the C3D network, pretrained on UCF101 dataset [53]. The off-the-shelf pose estimator of the expert is an OpenPose network that is trained on the COCO+Foot dataset [5], [30]. The coordinates of each skeleton sequence are scaled to be in the range $[-1, 1]$. The first frames of each sequence are zero-padded to fit to the longest sequence number of frames (743). The number of units in the layers of the non-expert's encoder, decoder, and discriminator are 1000, 1000, and 200 respectively. The experts single layer encoder consists of 100 units. The learned representation dimension is the same as of the input frames. The adversarial ratio (α) in Eq.5.1.1 is set to 0.01. The pace prediction weight (β) in Eq.5.4 is 0.2.

The non-expert network is trained for 300 epochs on each of UWA3D and UTKinect datasets with the learning rate of 0.0003, decay rate of 0.9, and batch size of 55 using the Adam optimizer [23]. We follow the same data split as [51] and use 70% of the samples for training and the rest for testing.

Method	Ex #1	Ex #2	Ex #3	Ex #4	Ex #5	**Avg.**
C3D [58]	66.00	64.00	63.00	59.00	60.00	62.40
I3D [8]	45.00	56.00	57.00	64.00	58.00	56.00
EAGLE-Eye [34]	70.85	67.34	64.62	65.23	62.42	66.09
VI-Net [51]	79.00	69.00	57.00	59.00	70.00	66.80
Ours	**75.59**	**72.87**	**69.96**	**74.67**	**72.31**	**73.08**
Only Expert	68.42	66.43	67.81	66.97	63.11	66.55
Only Non-expert (U)	60.76	56.98	37.75	59.61	54.52	53.92
Only Non-expert (S)	63.92	60.85	54.44	56.82	57.85	58.78
Only Non-expert (U+S)	66.53	60.09	61.05	59.71	61.11	61.70
C3D as expert	70.19	69.42	67.11	64.59	66.38	67.54
W/o Skeleton inpaint	71.13	68.19	67.60	71.20	66.82	68.99
W/o pace prediction	73.34	70.94	70.84	71.28	69.07	71.09

Table 5.1: Detailed results on the KIMORE dataset [6]. First and second best are shown in color. The lower lines show the ablation study of the network.

The MSE loss function is used to update the parameters of the expert network. The expert is trained for 1000 epochs with batch size of 25. The other training settings of the expert are kept the same as the non-expert's. In order to be consistent with existing AQA studies [39], [42], [51], the Spearman's Rank correlation has been used as the evaluation metric for the results.

5.2.2 Results

The results of our network on the KIMORE dataset are presented in Table 5.1. As it can be seen, our method outperforms the previous works and baselines by a large margin. It should be noted that Sardari *et al.* [51] presented a few architectures with different backbones and here we reported the one with the best performance.

In order to evaluate the effectiveness of the network's components, we conducted a comprehensive ablation study (see the lower rows of Table 5.1). First, we removed the unsupervised representations of the pretext tasks and resorted to the expert network to do the downstream task of action evaluation. In this setting, we initialized the encoder of the expert with random weights. As expected, the performance of the network dropped significantly. That's because we are completely neglecting the first stage of the network that provided help-

Method	Ex #1	Ex #2	Ex #3	Ex #4	Ex #5	**Avg.**
Skl. Recon.	69.17	67.53	70.02	65.39	63.73	67.17
Skl. Recon. + Pace	68.53	69.12	69.53	68.14	64.33	67.93
Mot. Pred.	70.53	70.92	68.17	70.43	68.54	69.72
Mot. Pred. + Pace	72.18	69.92	**70.35**	71.18	70.46	70.82
Order Rec.	65.67	67.29	66.38	67.83	60.52	65.54
Ours	**75.59**	**72.87**	69.96	**74.67**	**72.31**	**73.08**

Table 5.2: Our self-supervised vs baselines on KIMORE and NW-UCLA

ful high-level features about the action. Second, we removed the expert and only used the non-expert network to regress the final score. In this setting (U), we fixed the encoder weights from the previous stage training and put a linear regression layer on top of the representations of the encoder. During the downstream task training, only the regression layer parameters get fine-tuned to study the effectiveness of the learned representations from the pretext tasks. We also studied the effect of initializing the encoder with random weights (S) and using the pretrained weights of unsupervised (U+S), while the expert is completely removed. As it can be seen, using the unsupervised representations of the pretext task as the pretraining weights for the supervised AQA task results in a better performance.

We further evaluated the performance of the model when the non-expert network sticks to one of the pace prediction and skeleton inpainting heads to capture the unsupervised representations. As expected, the model reaches its full potential when both of the heads are utilized. It seems that the inpainting head contributes more to the performance of the network. Intuitively, when trying to inpaint a random masked part of the skeleton, the model tries to analyze the dependencies between the body parts and the neighboring frames. However, in the pace prediction head, the model sees the skeleton as a whole and gets the dependencies between the frames to find the pace of the sequence. Thus, the inpainting head may provide richer representations than the pace prediction one.

In the next set of experiments, we are going to explore the effectiveness of the proposed self-supervised learning approach in human motion assessment.

To this end, we evaluate the performance of the two-stage model when other self-supervised baseline objectives have been set during the non-expert network training (see Table 5.2). For the skeleton reconstruction objective, we mask-out the whole skeleton and let the decoder reconstruct it. In motion learning baseline, the second half of the frames are masked and the decoder tries to predict them given the previous frames (first half). The contribution of adding the pace prediction head to these two have also been studied. Finally, inspired by [29], we shuffled each sequence and asked the decoder to estimate the correct permutation. To this end, we first segment the whole sequence into 25 parts and shuffle these segments. For this baseline, we remove the discriminator and change the decoder to classify the order of the segments by the cross-entropy loss. As can be seen in Table 5.2, our proposed self-supervised approach outperforms the baselines in abnormal movement assessment. Intuitively, to the human eyes, an impaired sequence is the one in which a part of the body moves in an abnormal way compared to the rest of the skeleton joints. The strategy of randomly masking a part of the skeleton and estimating it given the rest of the joints helps to capture local correlations between body parts and gives the representations a sense of symmetry. The complimentary pace prediction head helps to add a sense of slowness of the movement to the representation. As a result, we would have a representation that contains the information that the assessment should be based on. As evident in Fig. 5.4, the patient can not complete the whole cycle of an exercise at the same time of healthy sample. That's when having a representation that has a sense of the movements' slowness and right arm's impairment helps to have an accurate assessment.

It should be noted that we do not claim to provide the best self-supervised approach for human action recognition. The resulted representations of the non-expert is a good fit to the downstream task of abnormal movement assessment which is based on impairment and slowness of a movement. The results of the proposed two-stage network on NW-UCLA action recognition dataset [65] are shown in Tab. 5.3. For action recognition , you may ignore the labels of the target dataset sequences and use them for pretext task train-

Method	NW-UCLA
Skl. Recon.	82.54
Skl. Recon. + Pace	82.71
Mot. Pred.	**84.31**
Mot. Pred. + Pace	84.15
Order Rec.	83.92
Ours	84.02

Table 5.3: Our self-supervised vs baselines on NW-UCLA action recognition dataset

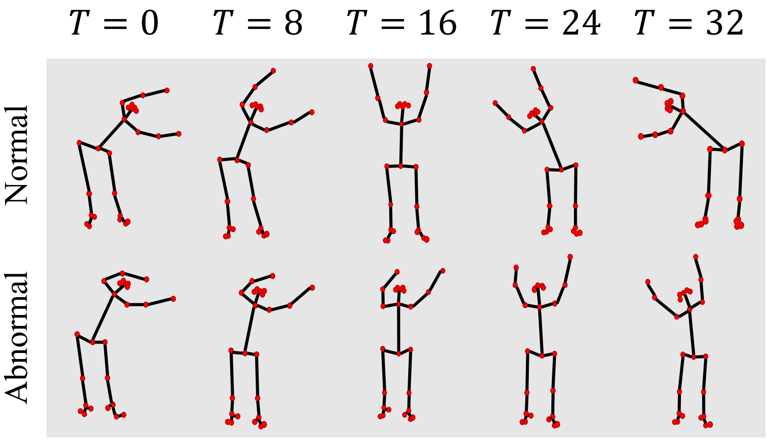


Figure 5.4: Figure4: The cycle of the exercise is not completed in an abnormal sequence of the KIMORE dataset. The ground truth score:23; Predicted scores of the baselines: Skl. Inpaint: 16.71, Skl. Inp. + Pace: 21.38, Motion + Pace: 19.45, Skl. Recon. + Pace: 30.72

ing. On the other hand, due to the scarcity of the target dataset samples in abnormal movement assessment, the downstream and pretext tasks use different datasets. However, it should be noted that these two datasets have to share something in common. Otherwise, the learned representations from the pretext task can not be used for the downstream one. As an example, using the representations of a pretext task on normal daily living activities dataset would not help the case of sports action assessment which involves lots of contorted poses.

We finally evaluated the generalization of our method to the new task of infants' general movement assessment as the downstream task for the expert network. To this end, we used the infants' neuromotor risk dataset [9] to train

Method	**Avg. F1 Score**
EAGLE-Eye	0.83
Only Expert	0.80
Only Non-expert (U)	0.69
Only Non-expert (S)	0.80
Only Non-expert (U+S)	0.86
Full Model	**0.89**

Table 5.4: The results of the model on the infants neuromotor risk dataset [9]. U and S stand for unsupervised and supervised settings. The experiment that is labeled by (U+S) uses the pretrained unsupervised weights to jumsptart the downstream supervised task.

our model. Since this task is relatively new, we compared the performance of our full model with the ablated models as the baselines for this task. As evident from Table 5.4, we get the best results when all of the components of the model are deployed. Since the infants dataset is annotated with 4 risk levels and not a score, we use the cross-entropy loss to update the parameters of the model and micro-averaged F1 score as the evaluation metric to compare our results with the baselines. As the two pretext and downstream datasets have to share a common distribution, the non-expert is fine-tuned using the healthy samples of the infants dataset. Otherwise, there would be a significant difference between what the non-expert tries to encode from the adults dataset and the infants samples that the expert tries to score in the second stage of the network.

5.3 Discussion

In this chapter, we developed a two-stage network to assess the performance of a humans movements. At the first stage, a non-expert network is trained on an off-the-shelf dataset of daily living activities to concurrently perform the pretext tasks of skeleton inpainting and sequence pace prediction in a self-supervised manner. The learned representations by the non-expert as well as appearance features of the target dataset samples are then fed to an expert network to perform the downstream task of action quality assessment. Our experimental evaluation demonstrated that our method not only outperforms the existing works and baselines in rehabilitation progress assessment of patients, but also shows a good generalization to the relatively new task of infants' general movement assessment.

Chapter 6

Conclusion & Future Work

In this thesis, we presented some neural-network based methods to address the problem of action quality assessment in both sports analysis and rehabilitation progress evaluation applications.

In **chapter 3**, we developed a diving routine assessor that regresses the overall score of the performance based on the difficulty of the program and how well it was executed. The proposed assessor takes into account both visual clues(e.g. amount of splash) and pose features (e.g. coordination among the joints) to evaluate the routine. Since the unusual pose configurations of the diver's body is not covered by existing datasets we introduced a dataset of 4000 diving and gym-vault images, accompanied with their pose annotations. The experiments and ablation studies demonstrated the effectiveness of the components of the proposed method in assessing a diving routine.

In **chapter 4**, we put one step further and extended the action quality assessor to be applicable to other second-long sports like snowboarding and skiing as well as minute-long sports activities like figure-skating. To this end, we proposed a modular two-stream network that is capable of extracting both fine and coarse-grained temporal dependencies of a sports routine. The network is comprised of some blocks that are responsible for assessing the coordination among the body parts and appearance dynamics. By stacking more of these blocks, the network would be able to capture distant frames temporal dependencies of the long-term figure-skating routine. We further extended the dataset of **chapter 3** with 7500 annotated images of four different fields (skiing, snowboarding, synchronized diving, gym-vault) to facilitate the assessment of athletes' performances of these sports.

In the end, in **chapter 5**, we explored the application of AQA in rehabilitation assessment and movement disorder stage detection. We developed a two-stage network that evaluates the patients' movements based on how impaired and slow they are. At the first stage, a multi-head non-expert network performed the pretext tasks of pace prediction and skeleton inpainting at the same time. This stage's function is analogous to what non-expert humans have learned about normal pace and symmetricity of the human body while doing daily activities. Then, given a new abnormal action, non-expert humans would be able to assess the severity of disease to some extent. The non-expert network is trained on off-the-self large-scale datasets of healthy people's daily activities like walking. At the next stage, an expert network takes the learned representations of the previous stage to perform the downstream task of AQA. The expert network is trained on the target dataset of movement quality assessment. Our experiments showed that deploying the self-supervised pace and impairment representations of the first stage helps to have a more accurate movement assessment.

For future research, there would be some exciting directions that we delve into here:

- As discussed in **chapter 3**, there is a hierarchical temporal structure for a sports routine. Quite recently, two large-scale hierarchical video datasets for fine-grained action understanding in gym-vault (Finegym [52]) and figure-skating (MCFS [31]) have been introduced. Performing the fine-grained HAR task on a sports routine would result in getting the detailed components of the action which can benefit the grading process. For example, knowing the number of saltos in a gym-vault routine and the temporal localization annotation of each is an extra information that helps the computer to assess the performance based on them. Thus, intuitively, transferring knowledge from fine-grained HAR to AQA should lead to a better grade estimation than using the pretrained weights of coarse-grained HAR or random weights.

- In a figure-skating routine, the athletes try to perform their movements as most coherence as possible to the played song. One interesting idea is to extract features from both audio and video channels and assess the alignment of these two sources of information, as an important element in figure-skating grading schema.

- The other interesting line of work can be the exploration of pose and shape estimation for contortive sports. The beginning of each routine usually starts with normal poses. Therefore, tracking the pose and attending to the temporal dependencies between the frames may contribute to getting a more accurate extreme pose and estimator. Estimating the shape of the athlete, would pave the pay for enlarging the existing AQA datasets by synthesizing other views of a performance. Besides the animated video of the performance can be used for educational purposes.

References

[1] I. S. F. (FIS), *Judges handbook (snowboard and freeski)*, `https://assets.fis-ski.com/image/upload/v1559714418/fis-prod/assets/Draft_SBFK_Judges_Handbook.pdf`.

[2] S. Abu-El-Haija, N. Kothari, J. Lee, *et al.*, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[4] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.

[5] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[6] M. Capecci, M. G. Ceravolo, F. Ferracuti, *et al.*, "The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1436–1448, 2019.

[7] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, *A short note on the kinetics-700 human action dataset*, 2019. arXiv: `1907.06987 [cs.CV]`.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[9] C. Chambers, N. Seethapathi, R. Saluja, *et al.*, "Computer vision to automatically assess infant neuromotor risk," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 11, pp. 2431–2442, 2020.

[10] I. O. Committee, *London 2012 Olympic Games: Global Broadcast Report*, `https://stillmed.olympic.org/Documents/IOC_Marketing/Broadcasting/London_2012_Global_%20Broadcast_Report.pdf`.

[11] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, Prague, vol. 1, 2004, pp. 1–2.

[12] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[13] I. S. Federation, *Fina diving rules*, `http://www.fina.org/sites/default/files/2017-2021_diving_16032018.pdf`.

[14] J. Gao, W.-S. Zheng, J.-H. Pan, *et al.*, "An asymmetric modeling for action assessment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[15] M. A. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements," *Nature Reviews Neuroscience*, vol. 4, no. 3, pp. 179–192, 2003.

[16] D. He, Z. Zhou, C. Gan, *et al.*, "Stnet: Local and global spatial-temporal modeling for action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8401–8408.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[19] N. Hussein, E. Gavves, and A. W. Smeulders, "Timeception for complex action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[20] L. Jing, X. Yang, J. Liu, and Y. Tian, "Self-supervised spatiotemporal feature learning via video rotation prediction," *arXiv preprint arXiv:1811.11387*, 2018.

[21] L. Karlinsky and S. Ullman, "Using linking features in learning non-parametric part models," in *European Conference on Computer Vision*, Springer, 2012, pp. 326–339.

[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: `http://arxiv.org/abs/1412.6980`.

[24] X. Lan and D. P. Huttenlocher, “Beyond trees: Common-factor models for 2d human pose recovery,” in *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, IEEE, vol. 1, 2005, pp. 470–477.

[25] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2008, pp. 1–8.

[26] G. Larsson, M. Maire, and G. Shakhnarovich, “Colorization as a proxy task for visual understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6874–6883.

[27] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in *CVPR 2011*, IEEE, 2011, pp. 3361–3368.

[28] Y. Li, X. Chai, and X. Chen, “End-to-end learning for action quality assessment,” in *Advances in Multimedia Information Processing – PCM 2018*, R. Hong, W.-H. Cheng, T. Yamasaki, M. Wang, and C.-W. Ngo, Eds., Cham: Springer International Publishing, 2018, pp. 125–134, ISBN: 978-3-030-00767-6.

[29] L. Lin, S. Song, W. Yang, and J. Liu, “Ms2l: Multi-task self-supervised learning for skeleton based action recognition,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2490–2498.

[30] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, “Microsoft coco: Common objects in context,” in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[31] S. Liu, A. Zhang, Y. Li, *et al.*, “Temporal segmentation of fine-gained semantic action: A motion-centered figure skating dataset,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, pp. 2163–2171, May 2021. [Online]. Available: `https://ojs.aaai.org/index.php/AAAI/article/view/16314`.

[32] I. Misra, C. L. Zitnick, and M. Hebert, “Shuffle and learn: Unsupervised learning using temporal order verification,” in *European Conference on Computer Vision*, Springer, 2016, pp. 527–544.

[33] M. Nekoui, F. O. T. Cruz, and L. Cheng, “Falcons: Fast learner-grader for contorted poses in sports,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2020.

[34] ——, “Eagle-eye: Extreme-pose action grader using detail bird’s-eye view,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 394–402.

[35] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, Springer, 2016, pp. 483–499.

[36] A. Nibali, Z. He, S. Morgan, and D. Greenwood, "Extraction and classification of diving clips from continuous video footage," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 38–48.

[37] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European conference on computer vision*, Springer, 2010, pp. 392–405.

[38] N. Paker, D. Bugdayci, G. Goksenoglu, D. T. Demircioğlu, N. Kesiktas, and N. Ince, "Gait speed and related factors in parkinson's disease," *Journal of physical therapy science*, vol. 27, no. 12, pp. 3675–3679, 2015.

[39] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019.

[40] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2019, pp. 1468–1476.

[41] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[42] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jul. 2017.

[43] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[44] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 556–571, ISBN: 978-3-319-10599-4.

[45] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 588–595.

[46] B. Qi and S. Banerjee, "Goniosense: A wearable-based range of motion sensing and measurement system for body joints: Poster," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, 2016, pp. 441–442.

[47] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, “Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition,” in *European conference on computer vision*, Springer, 2014, pp. 742–757.

[48] D. Ramanan, “Learning to parse images of articulated bodies,” in *Nips*, Citeseer, vol. 1, 2006, p. 7.

[49] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Strike a pose: Tracking people by finding stylized poses,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, IEEE, vol. 1, 2005, pp. 271–278.

[50] M. W. Rivolta, M. Aktaruzzaman, G. Rizzo, *et al.*, “Evaluation of the tinetti score and fall risk assessment via accelerometry-based movement analysis,” *Artificial intelligence in medicine*, vol. 95, pp. 38–47, 2019.

[51] F. Sardari, A. Paiement, S. Hannuna, and M. Mirmehdi, “Vi-net—view-invariant quality of human movement assessment,” *Sensors*, vol. 20, no. 18, p. 5258, 2020.

[52] D. Shao, Y. Zhao, B. Dai, and D. Lin, “Finegym: A hierarchical video dataset for fine-grained action understanding,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[53] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.

[54] B. Spine and H. Centre, *Movement disorders: From parkinson’s and huntington’s to dystonia and more*, `https://broadviewhealthcentre.com/movement-disorders-from-parkinsons-huntingtons-to-dystonia-more/`.

[55] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *CVPR*, 2019.

[56] Y. Tang, Z. Ni, J. Zhou, *et al.*, “Uncertainty-aware score distribution learning for action quality assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9839–9848.

[57] A. Toshev and C. Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[58] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[59] D. Trojaniello, A. Ravaschio, J. M. Hausdorff, and A. Cereatti, "Comparative assessment of different methods for the estimation of gait temporal parameters using a single inertial sensor: Application to elderly, post-stroke, parkinson's disease and huntington's disease subjects," *Gait & posture*, vol. 42, no. 3, pp. 310–316, 2015.

[60] I. S. Union, *Bonus and deduction rules*, `https://www.isu.org/figure-skating/rules/sandp-handbooks-faq/17823-s-p-who-is-responsible-for-deductions-2019-20/file`.

[61] ——, *Program component chart for single skating*, `https://www.isu.org/isu-statutes-constitution-regulations-technical-rules-2/isu-judging-system/single-and-pair-skating/17596-program-component-chart-id-sp-2019-20/file`.

[62] ——, *Technical panel handbook for single skating*, `https://www.isu.org/isu-statutes-constitution-regulations-technical-rules-2/isu-judging-system/single-and-pair-skating/24781-tphb-single-skating-2020-21-final/file`.

[63] V. Venkataraman, I. Vlachos, and P. Turaga, "Dynamical regularity for action analysis," in *Proceedings of the British Machine Vision Conference (BMVC)*, M. W. J. Xianghua Xie and G. K. L. Tam, Eds., BMVA Press, Sep. 2015, pp. 67.1–67.12, ISBN: 1-901725-53-7. DOI: `10.5244/C.29.67`. [Online]. Available: `https://dx.doi.org/10.5244/C.29.67`.

[64] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *CVPR 2011*, 2011, pp. 3169–3176. DOI: `10.1109/CVPR.2011.5995407`.

[65] 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*, 2014, pp. 2649–2656.

[66] J. Wang, J. Jiao, and Y.-H. Liu, "Self-supervised video representation learning by pace prediction," in *European Conference on Computer Vision*, Springer, 2020, pp. 504–521.

[67] L. Wang, Y. Xiong, Z. Wang, *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*, Springer, 2016, pp. 20–36.

[68] S. N. Watamaniuk and A. Duchon, "The human visual system averages speed information," *Vision research*, vol. 32, no. 5, pp. 931–941, 1992.

[69] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European conference on computer vision*, Springer, 2008, pp. 650–663.

[70] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, IEEE, 2012, pp. 20–27.

[71] 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*, 2019.

[72] S. Yan, Y. Xiong, and D. Lin, *Spatial temporal graph convolutional networks for skeleton-based action recognition*, 2018. [Online]. Available: `https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17135`.

[73] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR 2011*, IEEE, 2011, pp. 1385–1392.

[74] H. Zach, A. M. Janssen, A. H. Snijders, *et al.*, "Identifying freezing of gait in parkinson's disease during freezing provoking tasks using waist-mounted accelerometry," *Parkinsonism & related disorders*, vol. 21, no. 11, pp. 1362–1366, 2015.

[75] W. Zhang, M. Tomizuka, and N. Byl, "A wireless human motion monitoring system for smart rehabilitation," *Journal of Dynamic Systems, Measurement, and Control*, vol. 138, no. 11, p. 111 004, 2016.

[76] 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*, vol. 32, 2018.

[77] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–818.