Enhancing Human Motion Assessment by Self-supervised Representation Learning

Mahdiar Nekoui
nekoui@ualberta.ca
Li Cheng
lcheng5@ualberta.ca
University of Alberta
Edmonton, Canada

Abstract

The space of human motions is vast, ranging from daily behaviors of healthy adults to the slow and stiff motions of Parkinson's patients, or to infant motions. This poses significant challenges when the task is focused on a relatively niche motion subspace such as physical rehabilitation: often the target datasets are limited and less-annotated; meanwhile, there exist large-scale, well-annotated benchmarks, typically consisting of daily activities from healthy adults. This observation inspires us to propose a two-stage pipeline that takes advantage of the best of both worlds: a non-expert network starts to learn the representation of normal motions from source datasets, by estimating the pace and a set of manually inpainted joints of the pose sequence; this is followed by an expert network that takes as input these representations as well as the appearance features of the dedicated motions from the target dataset, to assess the quality of the specific actions. Empirical experiments on two very different motion assessment applications (physical rehabilitation of Parkinson's & stroke patients, and neuromotor behaviors of infants) demonstrate the superior performance of our approach.

1 Introduction

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 [19] and UTKinect [29]. In this paper, 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

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

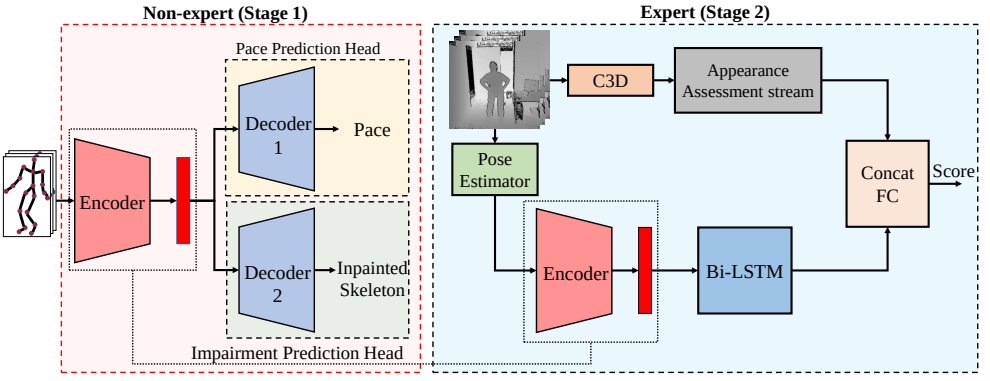

Figure 1: An overview of our two-stage pipeline. At stage 1, the non-expert module is trained on the large-scale, well-annotated source benchmarks of healthy samples performing daily living activities. At stage 2, the expert module is further trained on the small-scale and less-annotated target training 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.

posture over time [6, 28]. As a result, given a stroke or Parkinson's patient movement, we would be able to estimate the severity 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 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 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 [24] or I3D [3]. 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 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.

2 Related Work

Action Quality Assessment (AQA): AQA is the task of evaluating how well an action has been performed and determining a score for the performance. Driven by the vast number of video footage available on Youtube, sports action scoring has attracted the attention of more AQA works than other related tasks. Pirsiavash *et al*. [18] were the first to propose a sports routine action quality assessment network. They first applied DCT transform on the pose sequence of the video to get a set of interpretable high-level features. Finally, by putting a linear SVR on top of these features they regressed the final score of the athlete's performance. Parmar and Morris [16] used C3D network to get the high-level appearance features of the video. These features are then fed to an SVR or LSTM to get the score. Pan *et al*. [15] argued the importance of attending to both individual joints coordination and body parts movements in the assessment of a sports activity. To capture such features they proposed joint relation graphs to assess a routine. Tang *et al*. [23] proposed a score distribution learning method to model the judges' disagreement in grading a routine. Recently, Nekoui *et al*. [13] proposed a two-stream modular network to assess the pose and appearance of an action based on both short and long-term temporal dependencies.

The second research line in AQA focuses on assessing the surgical skills of robotic arms performing some elementary tasks like suturing and knot-tying. To address this problem, Wang *et al*. [27] proposed a multi-task learning approach by employing an auxiliary task of gesture recognition. Gao *et al*. [5] captured the interaction between the robot parts, considering a master-slave relation between them. Quite recently Liu *et al*. [12] argued that the surgical skill should be assessed based on the pattern of movements and field clearness and proposed a multi-path network to do so.

The third task that this paper focuses on is rehabilitation progress monitoring and disease severity assessment. Capecci *et al*. [2] introduced the first freely available 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. Sardari *et al*. [20] proposed a view-invariant method to assess the performance of patients and achieved state-of-the-art results on this dataset.

Unlike the first two tasks, the samples of the third one have a lot in common with daily activities like walking and sit-standing covered by large-scale off-the-shelf datasets. This fact

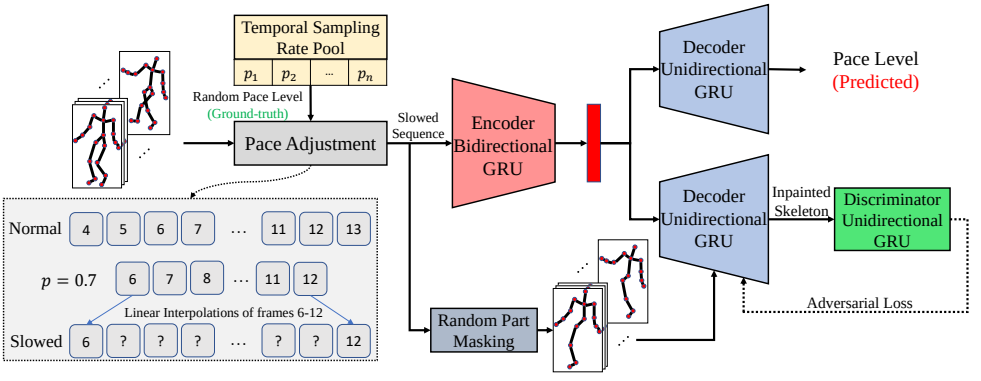

Figure 2: The architecture of the Non-expert network. The upper decoder predicts the pace of the sequence and the lower one inpaints the masked joints of the skeleton.

inspired us to learn some representations from these healthy samples to be used in assessing the patients' actions.

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, a *pretext task* is proposed to learn transferable features inherently from the data itself. Masking a part of an image and trying to predict it [17], estimating the rotation transformation [7], color channel prediction [9], pose sequence inpainting [30], and video pace prediction [26] are some of these pretext tasks. The non-expert network in our work is inspired by the last two to estimate the pace and impairment of an action.

3 Method

This section outlines our two-stage network and gives more details about each block of Figure 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.

3.1 Non-expert

The architecture of our non-expert network is depicted in Figure 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 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 2). We use the multi-class cross-entropy loss for the pace decoder head:

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

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:

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

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:

$$L_{inpaint} = L_{rec} + \alpha L_{adv} = \sum_{t=1}^{T} \sum_{k=1}^{J} (X_t'^k - \hat{X_t'^k})^2 + \alpha \left(\log(Disc(S')) + \log(1 - Disc(\hat{S}'))\right) \tag{3}$$

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:

$$L_{enc} = L_{rec} + \beta L_{pace} \tag{4}$$

Where β is a constant that controls the weight of two decoders. It should be noted that per Zheng *et al.* [30] 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.

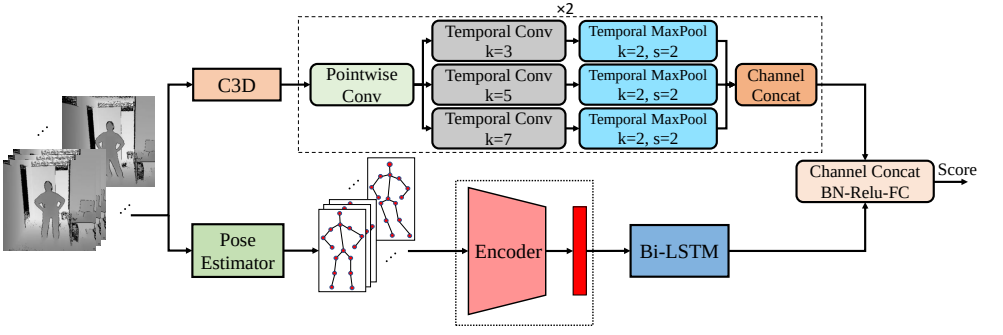

Figure 3: The architecture of our two-stream Expert. This network takes the RGB video of a patients action and evaluates it based on both appearance and pose features. In case due to privacy concerns the videos aren't provided and we only have access to the pose sequences, we stick to the lower stream of the network.

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.

3.2 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 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 [13]. 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 [1]or HRNet [22]. 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 .

4 Experiments

4.1 Implementation Details

Dataset: We use the off-the-shelf UWA3D [19] and UTKinect[29] 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 [2] 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 paper.

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 [4]. 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 [14]. 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 [21]. The off-the-shelf pose estimator of the expert is an OpenPose network that is trained on the COCO+Foot dataset [1, 11]. 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.3 is set to 0.01. The pace prediction weight (β) in Eq.4 is 0.2.

The non-expert network is trained for 300 epochs on each of UWA3D and UTKinect

Method	Ex #1	Ex #2	Ex #3	Ex #4	Ex #5	Avg. Corr.
C3D [24]	66.00	64.00	63.00	59.00	60.00	62.40
I3D [3]	45.00	56.00	57.00	64.00	58.00	56.00
EAGLE-Eye [13]	70.85	67.34	64.62	65.23	62.42	66.09
VI-Net [20]	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 inpainting	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 1: Detailed results on the KIMORE dataset [2]. First and second best are shown in color. The lower lines show the ablation study of the network.

datasets with the learning rate of 0.0003, decay rate of 0.9, and batch size of 55 using the Adam optimizer [8]. We follow the same data split as [20] and use 70% of the samples for training and the rest for testing. 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 [15, 16, 20], the Spearman's Rank correlation has been used as the evaluation metric for the results.

4.2 Results

The results of our network on the KIMORE dataset are presented in Table 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*. [20] 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 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 helpful 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

Method	Ex #1	Ex #2	Ex #3	Ex #4	Ex #5	**Avg.**	NW-UCLA
Skl. Recon.	69.17	67.53	70.02	65.39	63.73	67.17	82.54
Skl. Recon. + Pace	68.53	69.12	69.53	68.14	64.33	67.93	82.71
Mot. Pred.	70.53	70.92	68.17	70.43	68.54	69.72	**84.31**
Mot. Pred. + Pace	72.18	69.92	**70.35**	71.18	70.46	70.82	84.15
Order Rec.	65.67	67.29	66.38	67.83	60.52	65.54	83.92
Ours	**75.59**	**72.87**	69.96	**74.67**	**72.31**	**73.08**	84.02

Table 2: Our self-supervised vs baselines on KIMORE and NW-UCLA

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 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 [10], 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 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 Figure 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 [25] are shown in the last column of Table 2. For action recognition , you may ignore the labels of the target dataset sequences and use them for pretext task training. On the other hand, due to the scarcity of the target dataset samples in abnormal

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**

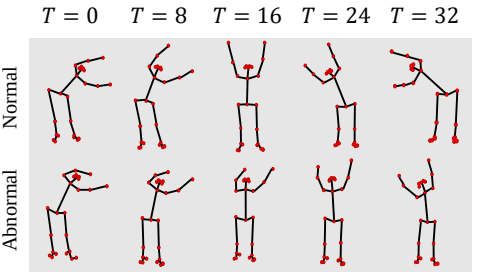

Table 3: The results of the model on the infants neuromotor risk dataset [4]. U and S stand for unsupervised and supervised settings. The experiment that is labeled by (U+S) uses the pretrained unsupervised weights to jumsptart the downstream supervised task.
Figure4: The cycle of the excercise is not 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

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 [4] to train 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 3, 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 Conclusion

In this work, we develop 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 demonstrates 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.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[2] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriù, Luca Romeo, and Federica Verdini. The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7): 1436–1448, 2019.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R Pierce, Daniel K Bogen, Laura Prosser, Michelle J Johnson, and Konrad P Kording. Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442, 2020.

[5] Jibin Gao, Wei-Shi Zheng, Jia-Hui Pan, Chengying Gao, Yaowei Wang, Wei Zeng, and Jianhuang Lai. An asymmetric modeling for action assessment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[6] Martin A Giese and Tomaso Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192, 2003.

[7] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018.

[8] Diederik P. Kingma and Jimmy 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. URL http://arxiv.org/abs/1412.6980.

[9] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017.

[10] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498, 2020.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. *arXiv preprint arXiv:2106.01035*, 2021.

[13] Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. 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*, pages 394–402, 2021.

[14] Nurdan Paker, Derya Bugdayci, Goksen Goksenoglu, Demet Tekdöş Demircioğlu, Nur Kesiktas, and Nurhan Ince. Gait speed and related factors in parkinson's disease. *Journal of physical therapy science*, 27(12):3675–3679, 2015.

[15] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[16] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[17] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[18] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 556–571, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

[19] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European conference on computer vision*, pages 742–757. Springer, 2014.

[20] Faegheh Sardari, Adeline Paiement, Sion Hannuna, and Majid Mirmehdi. Vi-net—view-invariant quality of human movement assessment. *Sensors*, 20(18):5258, 2020.

[21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[23] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9839–9848, 2020.

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[25] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656, 2014.

[26] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020.

[27] Tianyu Wang, Yijie Wang, and Mian Li. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 668–678. Springer, 2020.

[28] Scott NJ Watamaniuk and Andrew Duchon. The human visual system averages speed information. *Vision research*, 32(5):931–941, 1992.

[29] L. Xia, C.C. Chen, and JK 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*, pages 20–27. IEEE, 2012.

[30] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.