PECoP: Parameter Efficient Continual Pretraining for Action Quality Assessment

Amirhossein Dadashzadeh[1], Shuchao Duan[1], Alan Whone[2], Majid Mirmehdi[1]
[1]School of Computer Science [2]Translational Health Sciences
University of Bristol, UK
{a.dadashzadeh, shuchao.duan, alan.whone, m.mirmehdi}@bristol.ac.uk

Abstract

The limited availability of labelled data in Action Quality Assessment (AQA), has forced previous works to fine-tune their models pretrained on large-scale domain-general datasets. This common approach results in weak generalisation, particularly when there is a significant domain shift. We propose a novel, parameter efficient, continual pretraining framework, PECoP, to reduce such domain shift via an additional pretraining stage. In PECoP, we introduce 3D-Adapters, inserted into the pretrained model, to learn spatiotemporal, in-domain information via self-supervised learning where only the adapter modules' parameters are updated. We demonstrate PECoP's ability to enhance the performance of recent state-of-the-art methods (MUSDL, CoRe, and TSA) applied to AQA, leading to considerable improvements on benchmark datasets, JIGSAWS ($\uparrow$ 6.0%), MTL-AQA ($\uparrow$ 0.99%), and FineDiving ($\uparrow$ 2.54%). We also present a new Parkinson's Disease dataset, PD4T, of real patients performing four various actions, where we surpass ($\uparrow$ 3.56%) the state-of-the-art in comparison. Our code, pretrained models, and the PD4T dataset are available at https://github.com/Plrbear/PECoP.

1. Introduction

The analysis and evaluation of the quality of human action performance have a rich application in various real-world scenarios, such as sports events [42, 47, 49], healthcare [9, 32], and skill assessment [35, 38]. Despite recent advances [9, 18, 28, 36, 42, 46, 47, 49], AQA methods are affected by insufficient quantities of annotated data for training deep networks [36]. This is further exacerbated when extra effort is needed to produce very precise labels, e.g. for health-related applications, such as Parkinson's Disease (PD) severity assessment [9, 12, 33, 34]. A common solution to address such problems is to start with a model that is originally pretrained on a large source dataset, commonly

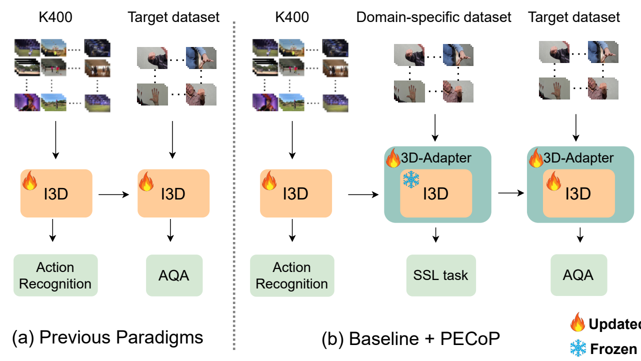

Figure 1. (a) Previous works directly transfer the model pretrained on domain-general data to AQA downstream tasks with target data fine-tuning, (b) in our proposed PECoP framework, the pretrained model continues to learn towards a specific AQA task through an additional pretraining stage, where only a small set of 3D-Adapter parameters are updated on unlabeled domain-specific data in a SSL approach, while the baseline model's weights remain frozen.

Kinetics-400 [27] (K400), and finetune it on one's target AQA dataset [42, 47, 49] (see Fig 1(a)). While better than training from scratch, deploying such pretrained models is sub-optimal for AQA, due to the domain/task discrepancy between action classification and action assessment. For example, in PD action performance scoring, one or two interruptions during the regular rhythm of a patient's movement performing an action can result in a different quality score and this is in contrast to the source pretraining task, where subtle or even more pronounced differences in performing an action should not affect action classification [29].

A promising route to address the shortcomings of this direct jump from classical pretraining to finetuning can be to further pretrain using domain-specific unlabeled data, i.e. continual pretraining – a strategy that has had a remarkable impact in natural language processing (NLP) [19, 20, 45] and recently in image/object classification [1, 40]. When it comes to video-domain tasks (e.g. as in AQA), this additional pretraining stage on in-domain data may be computationally prohibitive or impractical, due to the requirement for updating all parameters and storing pretrained parameter sets for each separate task.

BatchNorm tuning [15, 40] could be used to equip the pretrained model with domain-specific information by only updating the affine parameters of BatchNorm layers, while other pretrained parameters are frozen. Although this technique can greatly reduce the number of trainable parameters, we show that in a continual pretraining framework it can fail on those AQA datasets that are relatively small and more domain-specific, such as JIGSAWS [16].

In this paper, we propose adding a Parameter-Efficient Continual Pretraining (PECoP) adaptation stage to the traditional AQA transfer learning workflow that can efficiently *adapt* the domain-general pretrained model for the downstream AQA task. Inspired by adapter-based methods which have recently achieved strong results with transformer architectures on NLP benchmarks [22, 24, 25, 39], we present 3D-Adapter, a lightweight convolutional bottleneck block which is inserted into a pretrained 3D CNN (e.g. I3D inception modules [3]) and learns domain-specific spatiotemporal knowledge via a self-supervised learning (SSL) approach. During domain-specific pretraining, only the adapter parameters are updated while the original weights of the pretrained model are frozen to allow a high degree of parameter-sharing (see Fig. 1(b)). This greatly reduces the computational and storage costs of conventional continual pretraining, and also prevents overfitting by alleviating catastrophic forgetting [24].

To evaluate our method, we present experiments on three public AQA benchmarks, MTL-AQA [37], JIGSAWS [16] and FineDiving [47]. We also provide comparative results on a new Parkinson's Disease dataset, PD4T, comprising four different PD motor-function tasks by real patients. Our results clearly demonstrate that the proposed continual pretraining approach PECoP can boost the robustness of recent SOTA AQA methods, i.e. TSA [47], CoRe [30] and USDL/MUSDL [42], by a considerable margin.

In summary, the contributions of this work are:

- We propose a parameter-efficient continual pretraining workflow to enhance the efficiency of target tasks.
- We integrate a 3D-Adapter layer for the first time for 3D CNNs for video analysis.
- We introduce a new annotated AQA dataset, PD4T, for the vision community to evaluate various actions performed by actual Parkinson's disease patients.
- We present extensive experiments and ablations to demonstrate that recent state-of-the-art AQA methods can significantly benefit from our proposed continual pretraining approach.

2. Related Work

We address the most recent works that are significantly relevant to our work, including those on action quality assessment, self-supervised learning with respect to AQA tasks, continual pretraining, and adapters.

AQA – Most AQA methods treat the task as a regression problem on various video representations supervised by the scores or labels given by expert judges [32,37,42,46,47,49]. To reduce the inherent ambiguity of such labels, Tang et al. [42] proposed uncertainty-aware score distribution learning (USDL) for AQA. Yu et al. [49] employed contrastive learning [23] and built a contrastive regression (CoRe) framework to learn relative scores by pair-wise comparison. Although these methods have achieved SOTA results on several AQA datasets, they disregard the significant domain gap between their base dataset (i.e. K400) and target AQA dataset. In this paper, we address the pretraining stage towards better targeted learning in AQA methods.

SSL for AQA – Although most of the SOTA works in the AQA literature have focused on supervised learning approaches, a few have recently explored self-supervised learning [30, 32, 41, 50]. In these studies, in addition to the traditional supervised regression loss, the framework is further equipped with an SSL loss during the finetuning stage to improve the performance without the need for additional annotations. For instance, Liu et. al [32] employed a self-supervised contrastive loss to assist their supervised model capture temporal dynamics better in surgical videos. We leverage the advantages of SSL during the *pretraining process* to introduce to the pretrained model a level of domain-specific focus to allow it to handle the downstream AQA tasks more efficiently.

Continual pretraining – In contrast to the traditional approach in transfer learning that follows domain-general pretraining (usually over ImageNet or K400), continual pretraining can enhance learning via in-domain self-supervised pretraining to handle domain shift problems [1,19,39,40,45, 48]. Gururangan et .al [19] showed the importance of an additional pretraining phase with in-domain data to improve their target task performance on text classification. In the image domain, Reed et. al [40] verified that models continually pretrained on datasets that are progressively more similar to the target data can speed up convergence and increase robustness, while being particularly helpful when the target training data is limited. Azizi et. al [1] used a combination of both supervised pretraining on ImageNet, as well as intermediate contrastive SSL [7] on domain-specific medical data, to learn generalisable representations for medical images. In this work, we investigate continual pretraining for the first time in the video domain, in particular on the AQA task. However, instead of continual pretraining of the entire model parameter set, we take advantage of adapter-based tuning to reduce the cost of storage and model pretraining

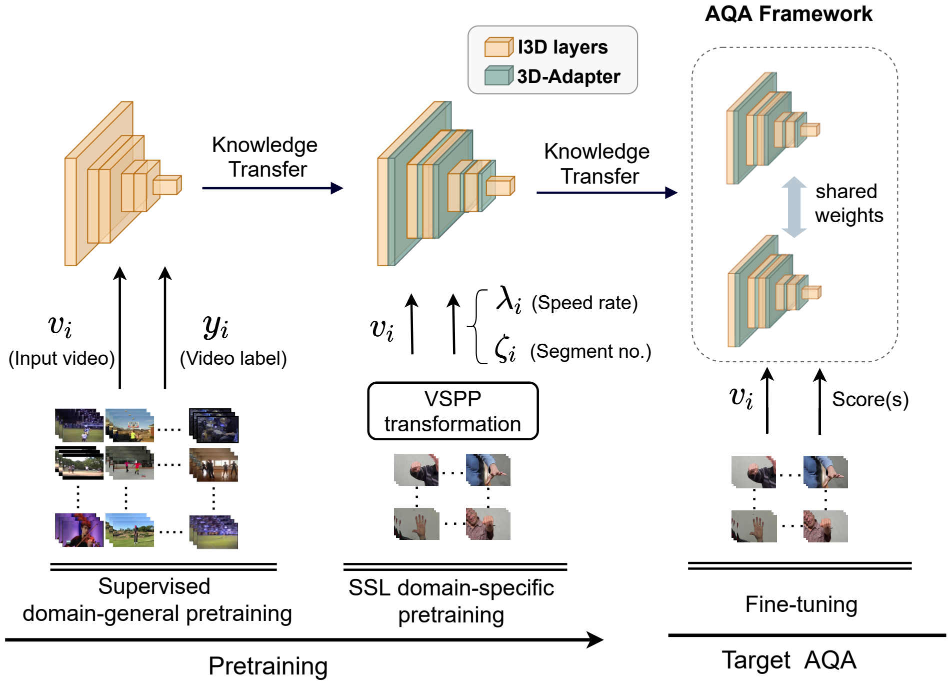

Figure 2. An overview of PECoP – First, a 3D encoder is pretrained on a domain-general dataset (i.e. K400). Then, we equip the pretrained model with 3D-Adapters and update their parameters using VSPP [8], a SSL pretext task, on unlabeled domain-specific data. Finally, we fine-tune the pretrained model on the AQA target task.

on in-domain data while preserving the knowledge obtained through the initial pretraining on domain-general data.

Adapters – Adapters are lightweight bottleneck modules which were designed for Transformer architectures for NLP tasks to conduct parameter-efficient transfer learning [10, 22, 25, 26] for downstream tasks. In computer vision, Chen et al. [6] proposed AdaptFormer to efficiently adapt a pretrained vision transformer model [11] to scalable image and video recognition tasks. In [4], Chen et al. introduced Conv-Adapter with a similar bottleneck architecture proposed for transformers, as in [6, 25], but with convolution layers to enable adapters in 2D CNNs. To the best of our knowledge, there is no work yet to study the effect of Adapters in 3D CNNs, and our proposed approach fills this gap.

3. Proposed Approach

Next, we outline our proposed continual pretraining approach implemented via self-supervised training of 3D-Adapter modules. The pipeline of our framework is shown in Fig. 2.

Let D_g be a large-scale, annotated, domain-general video dataset used for a learning task T_g, and D_t be a target video dataset in the AQA domain for a learning task T_t, with a significant domain discrepancy between T_g and T_t. Then, given an unlabelled video dataset D_q, where $D_q \subseteq D_t$, our aim is to leverage the representations in D_g and D_q to learn a transferable spatiotemporal feature extractor that is able to perform as well as possible on D_t for task T_t.

Domain-general pretraining – A pretrained backbone model, such as one that has been trained on a large video dataset, e.g. an I3D model supervised on K400, serves as our model architecture (see 1st column of Fig. 2).

In-domain SSL continual pretraining – We then equip our K400 pretrained model with randomly initialised 3D-Adapter modules. Our proposed 3D-Adapter has a similar bottleneck architecture as used in Transformers [25, 26] and recently in 2D CNNs [4], however, it requires 3D layers to be applied to 3D CNNs and trained on videos. The architecture of our 3D-Adapter and its integrated design with the inception module of the I3D model is shown in Figure 3. A performance boost can be obtained if a single 3D-Adapter is inserted after the concatenation layer of each inception module.

A 3D-Adapter consists of a downsampling, depth-wise, 3D convolution with learnable weights $\theta_{down} \in \mathbb{R}^{\frac{C_{in}}{\lambda} \times \lambda \times K \times K \times K}$, a non-linear function $f(.)$, e.g. ReLU, followed by an upsampling, point-wise, 3D convolution with learnable weights $\theta_{up} \in \mathbb{R}^{C_{out} \times \frac{C_{in}}{\lambda} \times 1 \times 1 \times 1}$. Here, C_{in} and C_{out} are the channel dimensions of the input and output feature maps, respectively, $K = 3$, and the compression factor λ denotes the bottleneck's dimension. Hence, given an input feature vector $h_{in} \in \mathbb{R}^{C_{in} \times D \times H \times W}$, then the output feature vector $h_{out} \in \mathbb{R}^{C_{out} \times D \times H \times W}$ of our 3D-Adapter is

$$h_{out} = \alpha \odot (\theta_{up} \otimes f(\theta_{down} \bar{\otimes} h_{in})) + h_{in} \ , \quad (1)$$

where $\otimes$ and $\bar{\otimes}$ are point-wise and depth-wise 3D convolution respectively, and α is a tunable scalar hyperparameter in $\mathbb{R}^{C_{out}}$ which is initialised as ones, and $\odot$ denotes element-wise multiplication, following [4, 26].

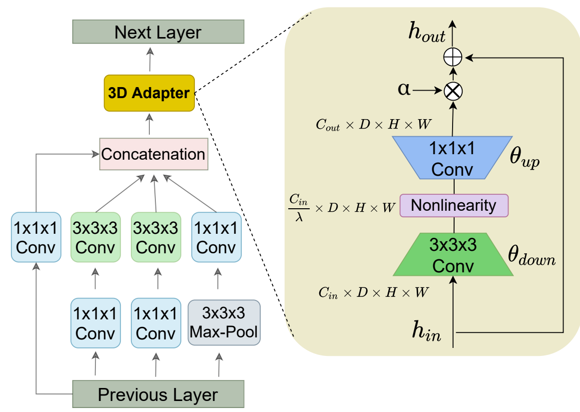

Figure 3. Inception module with adapter used in I3D model.

During the proposed continual pretraining stage, we only allow the 3D-Adapter parameters to be optimised on D_q, while the original model layers' weights stay frozen (middle column in Fig. 2). The training is accomplished through self-supervised learning, with labels automatically generated from unlabeled videos in D_q. Since understanding the quality of action is heavily dependent on movement patterns, we focus on Video Segment Pace Prediction (VSPP) [8] as an SSL pretext task for this stage. VSSP requires our model to temporally explore a video clip and predict the index and speed of a segment within a clip that is sampled at a different speed rate. It was shown in [8] that VSPP is more suited to video motion events than previous video playback prediction tasks [2, 5, 13, 43] by way of much faster convergence. We shall explore the performance of other recent SSL methods (e.g. VideoPace [43] and RSPNet [5]) during our ablations.

Supervised fine-tuning – We select SOTA action quality assessment models CoRe [49], USDL/MUSDL [42], and TSA [47], as example models that can be enhanced with PECoP and then fine-tuned on AQA datasets for direct evaluation. In essence, in this stage, each model's layers, including the original and the introduced adapter layers, are fine-tuned on the target dataset D_t (see rightmost column in Fig. 2).

For **USDL** [42], the features obtained from segments of a video clip pass through our continually pretrained I3D backbone and fused through temporal pooling, and then sent through softmax to generate the predicted quality assessment distribution. The KL loss between the predicted distribution and a Gaussian distribution generated from the ground-truth score is applied for optimisation. **MUSDL** is a multi-path version of USDL which predicts the final score if multiple-judge scores are available, as is the case in the MTL-AQA and JIGSAWS datasets.

For **CoRe** [49], given pairwise query and exemplar videos, their spatiotemporal features are extracted with our continual pretraining I3D and then the two feature sets are combined with the reference score of the exemplar video and passed to the group-aware regression tree to obtain the score difference between the two videos. During inference, the final score is computed by averaging the results from multiple different exemplars.

For **TSA** [47], spatiotemporal features are extracted similarly to CoRe. Then, a temporal segmentation attention module assesses action quality.

4. Experiments

We evaluate our self-supervised continual pretraining adaptation module on the following datasets that span various AQA tasks:

(i) **MTL-AQA** [37] contains 1412 video clips collected from 16 different world events and includes a variety of diving actions, covering both individual and synchronous divers, with videos from different angles. The annotations comprise scores from 7 judges, final scores, difficulty degree and type of diver's action.

(ii) **JIGSAWS** [16] is a collection of 103 surgical activities that cover three distinct tasks: 39 Suturing (S), 28 Needle-Passing (NP), and 36 Knot-Typing (KT). Each task is annotated by multiple sub-scores (that represent e.g., flow of operation, quality of final outcome, and so on), with the final score defined as the sum of these sub-scores.

(iii) **FineDiving** [47] is a fine-grained sports video dataset for AQA which provides 3,000 video of diving with detailed annotations on 52 actions, 29 sub-actions, and 23 difficulty degree types.

(iv) **PD4T** is our new fully annotated dataset offering 2931 videos from 30 PD patients tested longitudinally at 8 week intervals. The videos were captured at 25fps at a resolution of 1920×1080 (reduced to 854x480), using a SONY HXR-NX3 camera. There are a total of 30 subjects, with 22 of these used for training and 8 for testing. The patients, who ranged from 41 to 72 years old, performed PD tasks of gait, finger tapping, hand movement, and leg agility in clinical settings and their Unified Parkinson's Disease Rating Scale (UPDRS) [17] quality scores were assigned by trained clinicians ranging from 0 (normal) to 4 (severe)[1]. Sample frames are shown in Figure 4.

[1]For additional information about the PD4T dataset, see the supplementary materials.

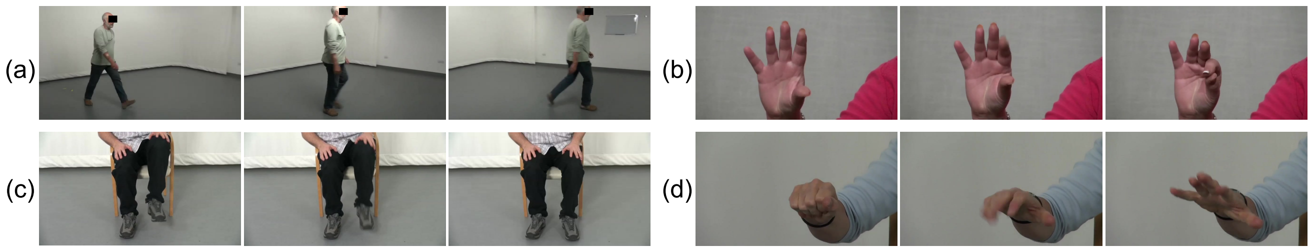

Figure 4. Sample frames from the PD4T dataset: (a) gait, (b) finger tapping, (c) leg agility, and (d) hand movement. All videos are from actual PD patients and were captured in a clinical environment as part of a clinical experiment across several months. For hand movement, finger tapping, and leg agility, data were collected from both the left and right sides for each subject.

Experiment Setup – The experiments were performed on an Nvidia RTX 3090TI GPU under Cuda 11.6 with cuDNN 8.2. We first initialise our I3D model with K400 pretrained weights which then remain frozen throughout the pretraining stage. After adding 3D-Adapters, the classification head is replaced with two randomly initialised FC layers f_λ and f_ζ corresponding to the segment speed and index outcomes of the VSPP [8] pretext task. Different values for f_λ and f_ζ are used for different AQA tasks. We perform SSL pretraining on domain-specific datasets by only updating the 3D-Adapter layers over 8 epochs, with batch size of 16 and SGD with a 1×10^{-3} learning rate. In this stage, we generate 32-frame long video clips, and empirically set the two parameters needed for VSSP to $[\lambda = 4, \zeta = 4]$ or $[\lambda = 4, \zeta = 3]$ which is either at, or close to, those recommended in [8]. Note that the training set videos of the target data is our domain-specific dataset for SSL pretraining. For data augmentation, we randomly crop the video clips to 224×224 followed by horizontal flip and colour jittering of each frame. Following [8], we apply 10x more iterations per epoch for temporal jittering. In all experiments, the input clip length is 32 during pretraining.

Fine-tuning – The pretrained I3D model is then the backbone network of our baselines USDL, MUSDL, CoRe, and TSA, and we evaluate their performance with the Spearman Rank Correlation metric ($\mathcal{S}$), expressed as percentages. For the JIGSAWS dataset, in keeping with other methods [42, 49], we provide values after four-fold cross-validation. In this stage, we adopt similar hyperparameter settings and training/evaluation strategy for each baseline as reported in [42], [49], and [47] respectively.

Comparative Evaluation – We present results on the MTL-AQA [37] and JIGSAWS [16] datasets against SOTA AQA methods MUSDL [42] and CoRe [49] when we enhance them with PECoP, as well as when we enhance them with another recent continual pretraining workflow, HPT [40] (see Table 1). In HPT, which has only been applied to image domain tasks till now, simply additional pretraining steps are introduced on domain-specific datasets with all model parameters updated at every stage. We use the same hyperparameters for training both PECoP and HPT.

While with PECoP improved results are obtained across the board, the improvements on JIGSAWS are very significant, e.g. after adding PECoP, MUSDL's average performance on the three tasks in the JIGSAWS dataset improves to 76% ($\uparrow$ 6%). Similarly, CoRe's average performance on the same tasks increases to 89% ($\uparrow 4\%$). This clearly shows PECoP's effectiveness in narrowing the substantial domain gap between the JIGSAWS dataset and K400.

Further, we note that adding HPT to the baselines results in a performance drop on JIGSAWS. Specifically, CoRe's performance decreases to 80% ($\downarrow$ 5%). This decline can be attributed to overfitting, as HPT requires all model parameters to be pretrained on a relatively small dataset (i.e. $\sim$13M parameters vs. PECoP's 3D-Adapters' $\sim$1M).

In Table 2, we show that not only PECoP dramatically reduces the trainable parameters, it also requires drastically fewer epochs to converge compared with HPT.

Table 3 presents the results on the FineDiving dataset introduced in [47], comparing CoRe [49] and TSA [47], with and without PECoP. Since TSA requires step transition labels for training, we cannot evaluate it on other datasets. As shown, TSA+PECoP improves on TSA by 1.10%. Although TSA outperforms CoRe alone, CoRe+PECoP surpasses TSA and TSA+PECoP to achieve the SOTA performance on FineDiving dataset.

Table 4 presents the results for the PD4T dataset. Given only a single action performance score based on the UPDRS scale [17] is available per clip, we compare PECoP and HPT for USDL instead of MUSDL. We observe the Spearman's rank correlation improves when averaged across the four PD4T tasks for both HPT and PECoP when added to both USDL and CoRe ($\uparrow$ 2.03% and $\uparrow$ 3.56% respectively for PECoP), although HPT performs marginally better ($\uparrow 2.22\%$) when added to USDL. We assume this slight advantage for HPT is likely due to its greater model capacity for handling the complex patterns in the PD4T dataset; how-

Table 1. Spearman Rank Correlation results on MTL-AQA and JIGSAWS, with and without continual pretraining. ⋆ ViSA [31] and MultiPath-VTPE [32] are customised towards surgical skill assessment and not general AQA tasks.

Method	Year	MTL-AQA	JIGSAWS			
		Diving	S	NP	KT	Avg $\mathcal{S}$
USDL [42]	2020	90.66	64	63	61	63
MultiPath-VTPE [32]⋆	2021	-	82	76	83	80
TSA-Net [44]	2021	94.22	-	-	-	-
I3D + MLP [49]	2021	89.21	61	68	66	65
I3D-TA [51]	2022	92.79	-	-	-	-
ViSA [31]⋆	2022	-	84	86	79	83
ResNet34-(2+1)D-WD [14]	2022	93.15	-	-	-	-
MUSDL [42]	2020	92.73	71	69	71	70
MUSDL + HPT [40]	2023	93.49	69	75	72	72
MUSDL + PECoP	2023	**93.72**	**77**	**76**	**76**	**76**
CoRe [49]	2021	95.12	84	86	86	85
CoRe + HPT [40]	2023	94.26	80	81	80	80
CoRe + PECoP	2023	**95.20**	**88**	**90**	**88**	**89**

Table 2. Comparison of PECoP and HPT [40] in terms of storage size and pretraining cost.

Continual Pretraining	#trainble parameters	#epochs	Size
HPT [40]	∼13M	16	∼54MB
PECoP	∼1M	8	∼4MB

Table 3. Results on FineDiving Dataset.

Method	$\mathcal{S}$
CoRe [49]	90.61
CoRe + PECoP	**93.15**
TSA [47]	92.03
TSA + PECoP	93.13

ever, PECoP achieves nearly equivalent performance gains while significantly reducing continual pretraining and storage costs.

5. Ablation Study and Analysis

In this section, we perform ablations on our PECoP framework for AQA tasks, focusing on the influence of different SSL methods employed for domain-specific pretraining, the role of BatchNorm tuning in domain adaptation, and the impact of integrating 3D-Adapters into another 3D CNN, e.g. R3D-18 [21].

Different SSL Methods – We investigate the performance of PECoP when using different SSL methods for domain-specific pretraining, i.e. RSPNet [5], a contrastive learning-based SSL approach (based on MoCo [23]), and VideoPace [43], a transformation-based SSL pretex task (similar to VSPP). In this ablation, CoRe has been used as the AQA baseline and JIGSAWS as the target AQA task. As shown in Table 5, VSSP achieves the best result for domain-specific pretraining.

BatchNorm (BN) Tuning – As mentioned earlier, BN tuning can be used to equip a pretrained model with domain-specific knowledge by only updating the affine parameters of BatchNorm layers. Figure 5 illustrates the comparative performance of BatchNorm tuning (HPT+BN) and other pretraining strategies, such as domain-general pretraining (Dom-G), domain-specific SSL pretraining (Dom-S) from scratch, and PECoP. Again, CoRe is used as the AQA baseline.

Each plot corresponds to an AQA target task taken from our various datasets. On all these tasks, PECoP outperforms HPT+BN, particularly by a large margin on the tasks within PD4T and JIGSAWS datasets. Further, we observe that, HPT+BN performs significantly worse than Dom-G alone on the all three tasks within JIGSAWS dataset. This suggests that the BN affine parameters, β and γ, generally have a negative impact on downstream AQA tasks when facing a significant domain shift. This happens because β and γ are tuned to the feature distribution of the source domain. When these parameters are applied to a significantly different target domain, they fail to correctly adjust feature statistics, resulting in a misalignment between the source and target domains. This issue becomes worse in smaller target datasets, as the limited number of examples makes it harder for the model to learn the true feature distribution, increasing the risk of overfitting.

In addition, the figure also shows that Dom-S performs variably. It fared worse than Dom-G on MTL-AQA, but exceeds Dom-G in nearly all PD4T tasks. This discrepancy suggests that direct transfer from the K400 dataset may not be ideal for PD4T tasks due to a significant domain gap.

Table 4. Spearman Rank Correlation results on the PD4T dataset.

Method	Gait	Finger tapping	Hand movem.	Leg agility	Avg. $\mathcal{S}$
USDL [42]	79.14	42.58	53.93	56.47	58.03
USDL + HPT [40]	**81.93**	46.38	54.15	**58.54**	**60.25**
USDL + PECoP	80.68	**47.44**	**56.19**	58.09	60.06
CoRe [49]	78.87	45.93	54.10	62.34	60.31
CoRe + HPT [40]	81.42	**49.73**	57.06	63.98	63.05
CoRe + PECoP	**82.33**	49.40	**59.46**	**64.27**	**63.87**

Table 5. Determining which SSL pretext task would be better to use - comparing contrastive learning approach to transformation-based ones. The experiment was performed on the JIGSAWS dataset as an example.

Method	S	NP	KT	Avg. $\mathcal{S}$
RSPNet [5]	83	86	84	84
VideoPace [43]	86	87	87	87
VSPP [8]	**88**	**90**	**88**	**89**

On the other hand, Dom-S also performs poorly on all JIGSAWS tasks, implying that self-supervised pretraining from scratch on smaller datasets is sub-optimal. These observations highlight the need for adaptable pretraining strategies, an aim achieved by our PECoP approach, which consistently delivers superior performance in each of the evaluated tasks.

3D-Adapters for ResNet – We evaluate the effectiveness of our 3D-Adapter with another 3D-CNN, i.e. R3D-18 [21] which is a common backbone network for action recognition tasks. We first insert a 3D-Adapter into each 3D residual blocks of R3D-18 backbone (see Fig. 6) and train this model through our continual pretraining framework, PECoP. This model is then used as the backbone network for CoRe to finetune on the target AQA task. We conduct this experiment on PD4T dataset and the results are reported in Table 6. As shown, across all four tasks within PD4T dataset, CoRe+PECoP outperforms CoRe alone (i.e. CoRe with R3D-18 backbone in both cases).

6. Discussion

Learning efficiency of PECoP – PECoP allows the

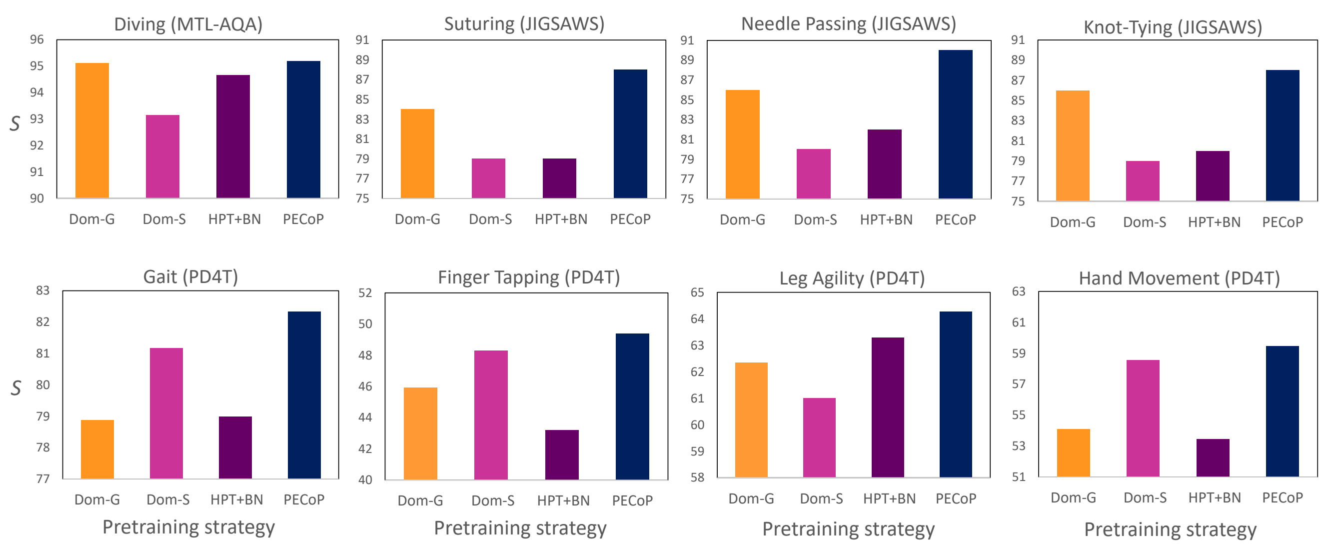

Figure 5. Comparison of PECoP with Domain-Specific SSL Pretraining (Dom-S), Domain-General Pretraining (Dom-G), and BatchNorm Tuning (HPT+BN) across eight different AQA tasks from MTL-AQA, PD4T, and JIGSAWS datasets. Each plot represents a unique AQA task and shows the performance of the four approaches. Dom-G employs pretraining on a domain-general dataset like K400, while Dom-S focuses on domain-specific self-supervised pretraining on target data. HPT+BN fine-tunes only the BatchNorm layers of a pretrained model. PECoP consistently outperforms the other approaches across all tasks, indicating its robustness and adaptability for AQA tasks with different domains and complexities.

Table 6. Spearman Rank Correlation results on the PD4T dataset with R3D-18 backbone used in CoRe.

Method	Gait	Finger tapping	Hand movem.	Leg agility	Avg. $\mathcal{S}$
CoRe	76.16	35.35	50.53	49.96	53.0
CoRe + PECoP	**79.11**	**39.71**	**55.37**	**52.56**	**56.69**

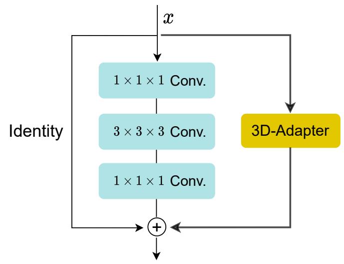

Figure 6. 3D residual block equipped with 3D-Adapter used in the R3D-18 model. We empirically find that such a configuration leads to a better performance.

model to leverage knowledge gained from a previous stage and requires only a small subset of parameters to be specifically learned for the new task. This approach leads to fewer epochs and less training data required for convergence on a new task, resulting in a reduction of both computational resources and pretraining time (refer to Table 2). With fewer trainable parameters, PECoP has a limited capacity to memorise the training data, and as a result, it is forced to learn more generalised patterns, making it less prone to overfitting.

Another benefit to PECoP's efficiency is its ability to avoid the issue of forgetting [24]. Unlike traditional continual pretraining methods [1, 40] that require updating all model parameters, thereby risking forgetting previously learned patterns, PECoP leverages pre-existing domain-general knowledge without modifying this foundational knowledge. Instead, it simply augments it with domain-specific knowledge, thus effectively alleviating the issue of forgetting.

Importance of PECoP for AQA – PECoP is crucial for AQA tasks, offering unique benefits that are valuable in diverse applications, ranging from healthcare settings like PD severity assessment to sports evaluations such as diving. Its scalable design eliminates the need for multiple pretrained models for different assessments, a key advantage where quick and precise evaluations across a range of tasks are essential. Furthermore, PECoP's data efficiency makes it well-suited for healthcare settings where annotated data is often scarce. Its computational efficiency also facilitates faster decisions and saves resources, which is vital in environments with limited computational capabilities.

Limitations and Future Work – One of the key weaknesses in a parameter-efficient continual pretraining approach like PECoP is the potential trade-off between efficiency and complexity in the model. Although using a limited set of adaptable parameters minimises the risk of overfitting, it may unintentionally lead to underfitting, making the model less effective at capturing intricate spatial and temporal features. This limitation becomes especially noticeable in AQA tasks, where accurately capturing subtle details may require a model with greater capacity. Additionally, PECoP has primarily been evaluated using 3D CNN backbones for AQA tasks. The effectiveness of PECoP with other architectures, e.g. transformers, as well as its applicability to other vision tasks (e.g., action recognition and few shot learning), requires further validation. We plan to address these in our future work.

7. Conclusions

We proposed PECoP, a parameter efficient continual pretraining workflow to better transfer the knowledge learned from existing large-scale video datasets (e.g. K400) to AQA target tasks by only updating a small number of additional bottleneck layers (called 3D-Adapters) through self-supervised learning. Alongside the evaluation on benchmark datasets, we also presented results on a new dataset of functional mobility actions performed by actual Parkinson's patients for performance quality assessment, with potential for longitudinal evaluation. Experiments on four AQA datasets (8 different tasks) with three AQA baselines (CoRe, USDL/MUSDL, and TSA) demonstrated the significant advantages of PECoP over the conventional continual pretraining approach with respect to both generalisation ability, storage needs, and training cost.

8. Acknowledgements

The authors would like to gratefully acknowledge the contribution of the Parkinson's study participants. The 1st author is funded by a kind donation made to Southmead Hospital Charity, Bristol and by a donation from Caroline Belcher via the University of Bristol's Development and Alumni Relations. The clinical trial from which the video data of the people with Parkinson's was sourced was funded by Parkinson's UK (Grant J-1102), with support from Cure Parkinson's.

References

[1] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022. 1, 2, 8

[2] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. SpeedNet: learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 4

[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. 2

[4] Hao Chen, Ran Tao, Han Zhang, Yidong Wang, Wei Ye, Jindong Wang, Guosheng Hu, and Marios Savvides. Conv-Adapter: exploring parameter efficient transfer learning for convnets. *arXiv preprint arXiv:2208.07463*, 2022. 3, 4

[5] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. RSPNet: relative speed perception for unsupervised video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1045–1053, 2021. 4, 6, 7

[6] Shoufa Chen, GE Chongjian, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, 2022. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[8] Amirhossein Dadashzadeh, Alan Whone, and Majid Mirmehdi. Auxiliary learning for self-supervised video representation via similarity-based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4231–4240, 2022. 3, 4, 5, 7

[9] Amirhossein Dadashzadeh, Alan Whone, Michal Rolinski, and Majid Mirmehdi. Exploring motion boundaries in an end-to-end network for vision-based parkinson's severity assessment. In *10th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2021*, pages 89–97. SciTePress, 2021. 1

[10] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, pages 1–16, 2023. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[12] Mark Endo, Kathleen L Poston, Edith V Sullivan, Li Fei-Fei, Kilian M Pohl, and Ehsan Adeli. Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 130–139. Springer, 2022. 1

[13] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 919–929, 2020. 4

[14] Shafkat Farabi, Hasibul Himel, Fakhruddin Gazzali, Md Bakhtiar Hasan, Md Hasanul Kabir, and Moshiur Farazi. Improving action quality assessment using weighted aggregation. In *Pattern Recognition and Image Analysis: 10th Iberian Conference, IbPRIA 2022, Aveiro, Portugal, May 4–6, 2022, Proceedings*, pages 576–587. Springer, 2022. 6

[15] Jonathan Frankle, David J Schwab, and Ari S Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. In *International Conference on Learning Representations*. 2

[16] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, 2014. 2, 4, 5

[17] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008. 4, 5

[18] Rui Guo, Hao Li, Chencheng Zhang, and Xiaohua Qian. A tree-structure-guided graph convolutional network with contrastive learning for the assessment of parkinsonian hand movements. *Medical Image Analysis*, 81:102560, 2022. 1

[19] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020. 1, 2

[20] Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. *arXiv preprint arXiv:1904.02817*, 2019. 1

[21] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017. 6, 7

[22] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 2, 3

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual rep-

resentation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 6

[24] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *ACL/IJCNLP (1)*, 2021. 2, 8

[25] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 3

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 4

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[28] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 1

[29] Qing Lei, Ji-Xiang Du, Hong-Bo Zhang, Shuang Ye, and Duan-Sheng Chen. A survey of vision-based human action evaluation methods. *Sensors*, 19(19):4129, 2019. 1

[30] Mingzhe Li, Hong-Bo Zhang, Qing Lei, Zongwen Fan, Jinghua Liu, and Ji-Xiang Du. Pairwise contrastive learning network for action quality assessment. In *Computer Vision–ECCV 2022: 17th European Conference, 2022, Proceedings, Part IV*, pages 457–473. Springer, 2022. 2

[31] Zhenqiang Li, Lin Gu, Weimin Wang, Ryosuke Nakamura, and Yoichi Sato. Surgical skill assessment via video semantic aggregation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII*, pages 410–420. Springer, 2022. 6

[32] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2021. 1, 2, 6

[33] Mandy Lu, Qingyu Zhao, Kathleen L Poston, Edith V Sullivan, Adolf Pfefferbaum, Marian Shahid, Maya Katz, Leila Montaser Kouhsari, Kevin Schulman, Arnold Milstein, et al. Quantifying parkinson's disease motor severity under uncertainty using mds-updrs videos. *Medical Image Analysis*, 73:102179, 2021. 1

[34] C Morgan, A Masullo, H Isotalus, E Tonkin, M Mirmehdi, F Jovan, T Whone, G Oikonomou, R McConville, G Tourte, et al. Real-world sit-to-stand evaluation. In *MOVEMENT DISORDERS*, volume 37, pages S195–S196. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA, 2022. 1

[35] Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *Computer Vision–ECCV 2022: 17th European Conference, 2022, Proceedings, Part XXXVIII*, pages 105–123. Springer, 2022. 1

[36] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1468–1476. IEEE, 2019. 1

[37] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 2, 4, 5

[38] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2021. 1

[39] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: a framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, 2020. 2

[40] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2584–2594, 2022. 1, 2, 5, 6, 7, 8

[41] Konstantinos Roditakis, Alexandros Makris, and Antonis Argyros. Towards improved and interpretable action quality assessment with self-supervised alignment. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 507–513, 2021. 2

[42] 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. 1, 2, 4, 5, 6, 7

[43] 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. 4, 6, 7

[44] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. TSA-Net: tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4902–4910, 2021. 6

[45] Han Wu, Kun Xu, Linfeng Song, Lifeng Jin, Haisong Zhang, and Linqi Song. Domain-adaptive pretraining methods for dialogue understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 665–669, 2021. 1, 2

[46] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure

skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30(12):4578–4590, 2019. 1, 2

[47] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2949–2958, 2022. 1, 2, 4, 5, 6

[48] Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W Oard. C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2512, 2022. 2

[49] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7919–7928, 2021. 1, 2, 4, 5, 6, 7

[50] Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Semi-supervised action quality assessment with self-supervised segment feature recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6017–6028, 2022. 2

[51] Yu Zhang, Wei Xiong, and Siya Mi. Learning time-aware features for action quality assessment. *Pattern Recognition Letters*, 158:104–110, 2022. 6