# Domain Knowledge-Informed Self-Supervised Representations for Workout Form Assessment

Paritosh Parmar[1,2], Amol Gharat[2], and Helge Rhodin[1]

[1] University of British Columbia
[2] FlexAI Inc.

**Abstract.** Maintaining proper form while exercising is important for preventing injuries and maximizing muscle mass gains. Detecting errors in workout form naturally requires estimating human's body pose. However, off-the-shelf pose estimators struggle to perform well on the videos recorded in gym scenarios due to factors such as camera angles, occlusion from gym equipment, illumination, and clothing. To aggravate the problem, the errors to be detected in the workouts are very subtle. To that end, we propose to learn exercise-oriented image and video representations from unlabeled samples such that a small dataset annotated by experts suffices for supervised error detection. In particular, our domain knowledge-informed self-supervised approaches (pose contrastive learning and motion disentangling) exploit the harmonic motion of the exercise actions, and capitalize on the large variances in camera angles, clothes, and illumination to learn powerful representations. To facilitate our self-supervised pretraining, and supervised finetuning, we curated a new exercise dataset, *Fitness-AQA* (https://github.com/ParitoshParmar/Fitness-AQA), comprising of three exercises: BackSquat, BarbellRow, and OverheadPress. It has been annotated by expert trainers for multiple crucial and typically occurring exercise errors. Experimental results show that our self-supervised representations outperform off-the-shelf 2D- and 3D-pose estimators and several other baselines. We also show that our approaches can be applied to other domains/tasks such as pose estimation and dive quality assessment.

## 1 Introduction

Detecting errors in gym exercise execution and providing feedback on it is crucial for preventing injuries and maximizing muscle gain. However, feedback from personal trainers is a costly option and hence used only sparingly—typically only a few days a month, just enough to learn the basic form. We believe that an automated computer vision-based workout form assessment (*e.g.*, in the form of an app) would provide a cheap and viable substitute for personal trainers to continuously monitor users' workout form when their trainers are not around. Such an option would also be helpful to the socio-economically disadvantaged demographic who cannot afford or have access to personal trainers.

While fitness apps have recently become popular, the existing apps only allow the users to make workout plans—they do not provide a functionality to

assess the workout form of the users. To detect errors in the workout videos, it is important to analyze the posture of the person. Academic research in workout form assessment so far has been limited to simple, controlled conditions [3, 27], where posture can be reliably estimated using off-the-shelf (OTS) pose estimators [2, 19, 22]. Ours, on the other hand, is the first work to tackle the problem of workout form assessment distinctly in complex, real-world gym scenarios, where, people generally record themselves using ubiquitous cellphone cameras that they place somewhere in the vicinity; which results in large variances in terms of camera angles, alongside clothing styles, lighting, and occlusions due to gym equipment (barbells, dumbbells, racks). These environmental factors combined with the subtle nature of workout errors (refer to Fig. 1) and the convoluted, uncommon poses that people go through while exercising, cause major challenges for OTS pose estimators (refer to Fig. 1), and consequently, workout form errors cannot be reliably detected from pose. To mitigate this in the absence of workout datasets labeled for human body pose, we propose to replace the error-prone pose estimators with our more robust domain knowledge-informed self-supervised representations that are sensitive to pose and motion, learned from unlabeled videos — helps in avoiding annotation efforts. Towards those ends, our contributions are as follows:

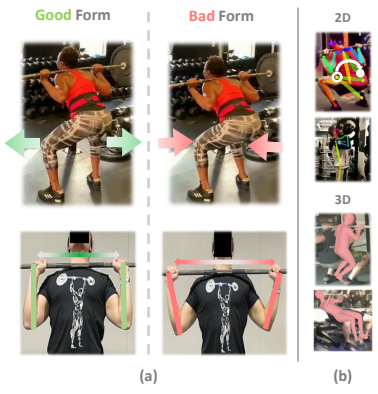


**Fig. 1. Concept.** (a) Errors of small magnitude generally occurring in workout form: Good column shows correct posture/execution (knees should be outwards), while the Bad column shows erroneous form during exercising. (b) Examples of failures of off-the-shelf 2D- and 3D-pose estimators in real-world gym scenarios (compare the discrepancies in pose estimation with the magnitude of the errors to be detected). We tackle the problem of detecting errors in workout form. To do so more accurately, we replace the error-prone pose estimators with our more robust fitness domain-oriented representations learnt using self-supervision.

1. **Novel self-supervised approaches that leverage domain knowledge**. We initiate the work in the direction of domain knowledge-informed self-supervised representation learning by developing two contrastive learning-based approaches that capitalize on the harmonic motion of workout actions and the large variance in unlabeled gym videos to learn robust fitness domain-oriented representations (Sec. 3). Our domain knowledge-informed self-supervised representations outperform 2D- and 3D-pose estimators [2, 22, 27], and various general self-supervised approaches [1, 5, 17, 18] on the task of workout form assessment on existing and our newly introduced datasets. This indicates that future work on representation learning would benefit from us-

ing domain knowledge in designing self-supervised methods, especially when tackling problems involving real-world data.

2. **Workout form assessment dataset**. To facilitate our self-supervised approaches, as well as the subsequent supervised workout form error detection, we collected the largest, first-of-its-kind, in-the-wild, fine-grained fitness assessment dataset, covering three different exercises (Sec. 4) and a small labeled subset for evaluation. We show that this in-the-wild dataset provides a significantly more challenging benchmark than the existing ones recorded in controlled conditions.

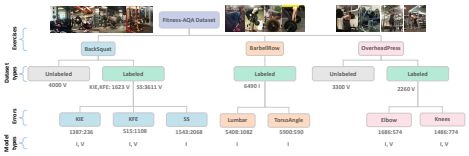


**Fig. 2. Fitness-AQA dataset hierarchy.** Numbers below the dataset type indicate dataset size; and those under the errors indicate the ratio of non-erroneous:erroneous samples. I, V indicate if the error detection is static image- or multiframe (video)-based.

## 2 Related Work

*Action Quality Assessment (AQA)/Skills Assessment (SA).* Our work can be classified under AQA/SA, which involve the computer vision-based quantification of the quality of movements and actions. Works in AQA/SA have mainly been focused on domains like physiotherapy [8, 24, 31, 41, 45], Olympic sports [4, 30, 34, 44, 49, 51], various types of skills [7, 25, 32, 48]. However, workout form assessment, especially, in real-world conditions, has not received much attention.

Approaches in AQA can be organized into 1) human pose features-based [28, 35]; 2) image and video features-based [33, 34]. Pose-based approaches use OTS pose estimators to extract 2D or 3D coordinate positions of various human body joints. These approaches have the disadvantage that poor estimation of the pose can adversely affect the final output. This is especially prevalent in non-daily action classes like fitness and sports domains. This can be mitigated, for example, by annotating domain-specific datasets [4], but that requires a considerable amount of manual annotation efforts, financial resources, and 3D annotations can only be obtained in controlled conditions. Therefore, we propose to learn domain-oriented pose-sensitive representations from unlabeled videos, which can be finetuned using only a small labeled dataset.

Closest to ours is the work on backsquat assessment by Ogata *et al.* [27]. However, a) they used OTS pose estimators, whereas we develop self-supervised approaches to learn more powerful representations; b) being dependent on OTS pose estimators, their approach is limited only to simple, controlled environments, whereas our approach is applicable to complex, real-world scenarios (Sec. 5); and c) their dataset contains only single exercise and was collected

in simpler conditions and a single human, whereas our dataset contains three exercises and was collected in real-world gym scenarios and numerous humans (further differences discussed in Sec. 4).

*SSL.* Earlier work in this area include those of autoencoders [12], which learn low-dimensional representations by reconstructing the input. Le *et al.* [23] propose a way to learn hierarchical representations from unlabeled videos using unsupervised learning, which was also considered as a feature extractor in an earlier AQA work [35], but was found to perform worse than an OTS pose estimator. Recently, Chen *et al.* [5] proposed a simple siamese approach to learn representations that obtain competitive results on various benchmarks. Various general SSL works also propose to leverage properties of videos. Misra *et al.* [26] and Xu *et al.* [50] propose to exploit temporal order of frames and clips. Predicting the amount of rotation in images and videos was used as a pretext task by Gidaris *et al.* [9] and Jing *et al.* [18]. Wang *et al.* [46] leveraged motion and appearance statistics to learn self-supervised video representations. Benaim *et al.* [1] and Wang *et al.* [47] used video speed prediction as the pretext task. In addition to video speed prediction, Jenni *et al.* [17] proposed to use wider range of temporal transformations for pretext task. In contrast, we developed domain knowledge-informed SSL approaches that we show outperform general SSL approaches. A few works propose to leverage time-contrast to learn representations using self-supervision [14,15,42]. However, these temporal models either consider a single-view or a single subject. Our pose contrastive approach, on the other hand, simultaneously exploits cross-view and cross-subject information to learn more meaningful representations.

Another work proposes to disentangle pose and appearance from multiple views with a geometry-aware representation [37]. However, this approach is not tailored for exercise analysis, and requires calibrated multi-view datasets. Inspired by this method, we develop a variant—our pose and appearance disentangling baseline—applicable to our dataset.

## 3 Method

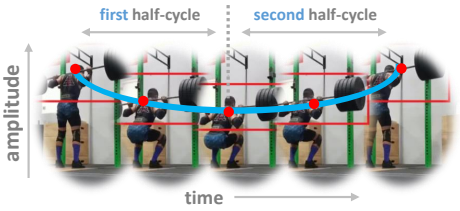


**Fig. 3. Barbell trajectory.** Red bounding boxes (bboxes) - barbell object detected; Red dots: the center of bboxes; Blue curve: the parabolic trajectory of the barbell traced out.

In this section, we present our self-supervised approaches for learning image and video representations. Subsequently, an error detection network is trained

to map these self-supervised representations to workout form error probabilities. Note that in the following, we have presented our approaches using BackSquat as an exemplary exercise, but our methods are applicable to other exercises.

**Preliminary: *Synchronizing videos.*** Our methods build upon quasi-synchronized videos. In some datasets, such as Human3.6M [16], synchronized videos recorded from multiple angles is already available using special setups, which allows unsupervised learning, *e.g.*, as done in [37]. However, we are not using any kind of such special setups. Thus, we quasi-synchronize the videos using the following method. Given a collection of videos of (different) people performing the same exercise, we detect the barbell/weight over time to get a motion trajectory, which when plotted against time traces an approximately parabolic curve as shown in Fig. 3. These trajectories are then amplitude-normalized. Object size, resolutions do not affect the normalization, as we are using the center of the bounding box; and the vertical movement of the barbell can be reliably recorded from various viewpoints (unless extreme, like top-view of the scene—unrealistic, anyway). Now, we leverage the following property to synchronize the videos: for a given elevation of the object (or equivalently, the amplitude of the trajectory), the people doing the same exercise would be in approximately the same pose. This holds across different subjects, different video instances, and across different views/camera angles, which allows us to synchronize videos of different subjects in different environments/scenes.

### 3.1 Self-Supervised Pose Contrastive Learning

**Objective.** Given the synchronized video samples of the same exercise (*e.g.*, BackSquat), in this approach, we aim to learn richer human pose information using self-supervised contrastive learning. In contrastive learning, same or similar samples are pulled together, while dissimilar samples are pushed apart [6]. In our case, we hypothesize that we can extend contrastive learning to learn human pose-sensitive representations. Particularly, we propose a self-supervised pretext task, which aims to pull together images (frames of videos) containing humans in similar poses, while pushing apart images with humans in dissimilar poses as shown in Fig. 4 . Note that, this approach operates on single frame-triplets (not videos or clips) at a time.

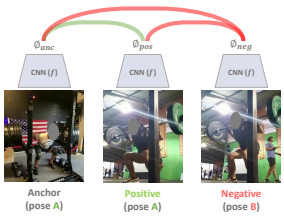


**Fig. 4. Cross-View Cross-Subject Pose Contrastive learning (CVCSPC).** Red lines indicate repulsion, while the green line indicates attraction in the representation space.

**Constructing triplets for contrastive learning.** Once we have the normalized barbell trajectories, for any given anchor input, $I_{\text{anc}}$, we retrieve the corresponding positive input frames with similar object elevation, $I_{\text{pos}}$, and the negative input frames with a difference in object elevation of more than a threshold value ($\delta$), $I_{\text{neg}}$, from across video instances; and subsequently build triplets of $\{I_{\text{anc}}, I_{\text{pos}}, I_{\text{neg}}\}$. Such triplets provide a cross-view, cross-subject, cross-video-instance self-supervisory signal that has not yet been leveraged by the existing computer vision approaches to learn pose sensitive representations. These triplets also offer strong, in-built data augmentations. A recent work [40] observed that background augmentation can help increase the robustness of self-supervised learning. Our method not only provides such background augmentation, but also provides foreground augmentation in terms of appearance (clothing, body type, gender, etc.). We term our approach Cross-View Cross-Subject Pose Contrastive learning (CVCSPC).

**Contrastive learning.** We use the constructed triplet, $\{I_{\text{anc}}, I_{\text{pos}}, I_{\text{neg}}\}$, to learn good representations through self-supervised contrastive learning. Let $f$ represent a 2D-convolutional neural network (CNN) backbone, which when applied to $I_{\text{anc}}, I_{\text{pos}}, I_{\text{neg}}$, yields $\phi_{\text{anc}}, \phi_{\text{pos}}, \phi_{\text{neg}}$, respectively. In contrastive learning, $\phi_{\text{anc}}$ and $\phi_{\text{pos}}$ are forced to be similar, *i.e.*, $\phi_{\text{anc}} \approx \phi_{\text{pos}}$, while $\phi_{\text{anc}}$ and $\phi_{\text{neg}}$ are forced to be dissimilar, *i.e.*, $\phi_{\text{anc}} \neq \phi_{\text{neg}}$, as illustrated in Fig. 4 . Following [43], we optimize the parameters of $f$ during the self-supervised training, by minimizing the distance ratio loss [13],

$$\mathcal{L} = -\log \frac{\mathrm{e}^{-||\phi_{\text{anc}} - \phi_{\text{pos}}||_2}}{\mathrm{e}^{-||\phi_{\text{anc}} - \phi_{\text{pos}}||_2} + \mathrm{e}^{-||\phi_{\text{anc}} - \phi_{\text{neg}}||_2}}. \tag{1}$$

### 3.2 Self-Supervised Motion Disentangling

Motion cues can be useful in detecting many workout form errors. Different from our pose-contrastive approach, this approach uses motion information to detect anomalies in workout form. In the following, we first present the preliminary information, before describing our method.

**Preliminaries**

- **Useful property 1: Harmonic motion**. Workout actions have a desirable property of exhibiting harmonic motion. For example, during benchpress (an exercise targeting the chest muscles), the person would be lifting the barbell above their chest and then bringing it down to the starting point; or during squats, the person would be squatting down (first half-cycle in Fig. 3) and then getting up (second half-cycle in Fig. 3).
- **Useful property 2: Bias in temporal location of form-errors**. People are more likely to make errors (anomalous motions) when lifting up the weights (one half-cycle of the harmonic motion, as in Fig. 3), rather than lowering the weights (another half-cycle of the harmonic motion).
- **Global motion.** The actual, regular motion of the workout action. For example, in Backsquat, the person squatting down and getting up.

– **Local motion.** The small-scale, fine-grained, irregular motion of the body parts (ref. Fig. 5). For example, in Backsquat, the knees abnormally going inward/outward or forward. So, while the global motion refers to regularities in motion patterns, local motion would cover anomalies in motion patterns.

**Objective.** Our goal is to learn self-supervised representations that are sensitive to local (anomalous) motions. The above discussed properties can provide a very useful, freely available signal that has not yet been exploited for this task by the existing computer vision approaches. We design a contrastive learning-based self-supervised approach to disentangle the local motion from the global motion.

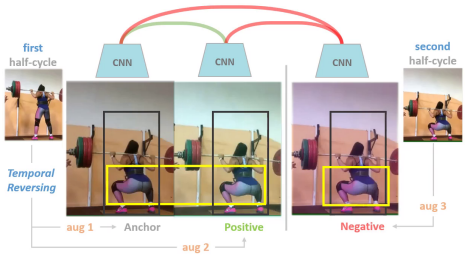


**Fig. 5. Motion Disentangling (MD) approach.** *Please view in AdobeReader to play the embedded animation for better explanation.* **Black** boxes: global motion (getting up, here); **Yellow** boxes: local motions (the knees rotating inwards under the influence of heavy training weight); aug: augmentations. Here we have applied very weak augmentation (only color augmentation) for representative purpose—to better illustrate the concept. However, in practice, we apply much stronger augmentations. Red lines indicate repulsion, while green indicates attraction in the representation space.

**Accentuating the local motion.** Temporally reversing any one of the half-cycles would, in general, make both half-cycles identical in terms of the global motion, while they would still differ in terms of the local motion. In other words, contrasting the two half-cycles after temporally reversing any one of them, helps accentuate the anomalous local motion, as shown in Fig. 5.

**Constructing triplets for contrastive learning.** The first half-cycle serves as the anchor; an augmented copy of the anchor serves as the positive input. The second half-cycle serves as the negative input. As discussed previously, we randomly temporally-reverse either the {anchor, positive} pair or the {negative} input to make the global motion of all three identical. In practice, we randomly and independently applied the following augmentations on the triplets: image horizontal flipping, partial image masking, image translation, image rotation, image blurring, image zooming, color channel swapping, temporal shifting.

**Contrastive learning.** We use a 3DCNN as the backbone for this model, and Eq. 1 as the loss function for this self-supervision task. Through contrastive learning, the 3DCNN learns to capture the previously discussed local, anomalous motions that are accentuated in our specially created triplets.

Anomalous motions maybe harmful or they can be beneficial. For example, knees buckling inwards during squatting is harmful, while knees going outwards is not. Therefore, during the finetuning phase, we aim to calibrate representations learnt using self-supervision to distinguish between harmful irregularities and harmless variations.

## 4 Fitness-AQA Dataset

Since exercise or workout assessment is an emerging field, there is a shortage of dedicated video datasets. To the best of our knowledge, the Waseda backsquat dataset by Ogata *et al.* [27] is the only publicly available such dataset. However, this dataset has shortcomings such as: it contains samples from a single human subject; the human subject is deliberately faking exercise errors; no kind of exercising weights, such as barbells and dumbbells, are used; the videos do not include realistic occlusions.

*Dataset Collection.* To fill the void of real-world datasets, we collected the largest exercise assessment dataset from video sharing sites such as Instagram and YouTube. We considered the following three exercises: 1) BackSquat; 2) BarbellRow; and 3) Overhead (shoulder) Press. In addition to the labeled data, we also collected an unlabeled dataset to learn human pose focused representations in self-supervised ways (discussed in Sec. 3). The purpose of the labeled dataset is to finetune our models to do actual error detection and quantify the performance of our models. We have provided statistics and illustrated the full hierarchy of our Fitness-AQA dataset in Fig. 2 . Illustrations of exercise errors are provided in the supplementary material.

*Annotations by Expert Trainers.* We employed two professional gym trainers to annotate our dataset for error labels. Due to this, even very subtle errors are caught and annotated accordingly. Errors range from very subtle to very severe. Unique properties of our dataset:

- **Real-world videos**. Unlike the existing dataset [27], we collected our dataset from actual real-world videos in actual gyms recorded by the people without any scripts. Due to this, the videos are naturally recorded from a wide range of azimuthal angles, inclination angles, and distances. Our samples were automatically processed to contain a single repetition.
- **People making errors under the impact of actual weights**. In the existing dataset [27], people are instructed to make deliberate exercise mistakes without being under the influence of actual weights. Our dataset, on the other hand, captures cases where people are naturally making mistakes (without any instructions), under the influence of actually heavy weights. Due to this, we believe that there is no bias towards exaggerated errors, and contains natural, subtler error cases.
- **Occlusions**. Having captured in actual gyms, human subjects are partially occluded by barbell weights, weight racks or other equipment like benches.
- **Various types of clothing, background, illumination**. Since we did not hire any specific group of people to collect the dataset, the samples in our

dataset are likely to come from numerous unique individuals, which results in a large number of clothing styles, and colors; different gyms (in terms of the room arrangement, and the background); other people in the background; and lighting conditions.
– **Unusual poses**. Exercise actions result in much more convoluted human body positions than those covered in the existing pose estimation datasets.

## 5 Experiments

To validate our contributions, we compared our features against various baselines and off-the-shelf pose estimators in simple (Case Study 1) and complex conditions (Case Study 2), showing significant improvements in the latter case.

We took a two-step approach towards detecting errors in exercising videos. Our models were first trained on the unlabeled datasets using self-supervision, and then used as feature extractors on the supervised datasets. For imbalanced datasets, we used class weights (in cross-entropy loss) inversely proportional to the class size. Note that the labeled dataset contains only the exercise error as ground-truth annotation and no information related to human pose. As such, our models did not use any pose-related ground-truth.

For the motion disentangling model, since the temporal model is already baked in it, we simply finetuned the model end-to-end on the labeled dataset for error detection. We used 32 frames for all types of errors.

For all 2DCNN-based approaches, we a learnt ResNet1D temporal model [27] that aggregates frame-level features for supervised error prediction on our labeled dataset. We used about 200 frames during error detection. Finetuning end-to-end on such a long sequence is not recommended [33, 49, 52]. Therefore, in this case, the 2DCNN backbone is not finetuned unless specified otherwise.

*Implementation details.* We used ResNet-18 [11] as the backbone CNN unless specified otherwise. We used custom YOLOv3 [36] to detect barbells/weights; and normalized the amplitudes of the trajectories to -180 to 180 (simply for a resemblance to a circle). Specifications regarding each approach are as follows:
– **Pose Contrastive approach (CVCSPC)**. We used a threshold gap of 30 between anchor/positive and negative inputs. We initialized our backbone CNN with ImageNet weights. We used ADAM optimizer [21] with an initial learning rate of 1e-4 and optimized for 100 epochs with a batch size of 25.
– **Motion Disentanglement approach (MD)**. We used R(2+1)D-18 [10] as our backbone CNN. We sampled 16 frames from each half-cycle. We randomly applied strong augmentations. We initialized our backbone CNN with Kinetics [20] pretrained weights. We optimized our models using ADAM optimizer with an initial learning rate of 1e-4 for 20 epochs with a batch size of 5.

Further details provided in the supplementary material.

### 5.1 Case Study 1: Simple Conditions

The Waseda Squat dataset [27] provides an excellent labeled dataset for evaluating exercise errors in controlled conditions. The publicly available portion of

| Feature extraction | Modality | Accuracies ↑ | | | | | |
|---|---|---|---|---|---|---|---|
| | | KIE | CVRB | CCRB | SS | KFE | Avg |
| HMR-TDM [27] | 3D Pose | 89.80 | **98.65** | 93.05 | **87.30** | 83.58 | 89.08 |
| Ours CVCSPC | Image | **95.92** | 91.89 | **94.44** | 77.77 | **89.55** | **89.92** |

**Table 1. Performance comparison on Waseda Squat dataset**.

this dataset contains samples from a single human subject. This dataset was not captured in a gym-like setting, but rather in home, and office-like settings. Each sample contains multiple squat repetitions. Note that the publicly available train/val/test split is different from that used in the original paper. Using this dataset, we experimented detecting the following errors: knees inward error (KIE); convex rounded back (spine) (CVRB); concave rounded back (spine) (CCRB); shallow squat (SS); knees forward error (KFE). To do so, we trained classifiers to distinguish between each of these error classes and good squat class (samples belonging to this class did not contain any errors). In this experiment, we compared features from our CVCSPC method (self-supervisedly trained on our unlabeled BackSquat dataset) against the Temporal Distances Matrices (TDM) derived from HMR pose estimator [19]. HMR-TDM features were made available by Ogata *et al.* [27]. During feature extraction, we resized the input images to 320×320 pixels, and considered the center 224×224 pixel crop. We did not consider our MD model because this dataset has multiple repetitions in each sample, and the sequence length is 300 frames, which is about 9 times longer than our MD model sequence length (32 frames). And, consequently, if we temporally downsample the sequence, it would lose a lot of information.

The results are summarized in Table 1, where we report accuracies. We found that our model outperformed existing methods [27] on three types of errors: KIE, CCRB, and KFE; with the performances being notably better on KIE and KFE errors. Even though not consistently across all the errors, our self-supervisedly learnt features outperformed HMR-TDM features on overall average performance. Note that large performance gap is not expected on this dataset, as OTS pose estimators work quite well in these simpler conditions.

### 5.2 Case Study 2: In-The-Wild Conditions

Next, we considered evaluating our approach on more complex datasets. For that, we considered our labeled datasets, which we introduced in Sec. 4, where we also discussed the reasons that make our new in-the-wild dataset more challenging. Unless mentioned otherwise, we divided the datasets into train-, validation-, and test-splits of 70%, 15%, and 15%, respectively.

*Baselines.* We compared our self-supervised feature extractors with the following models and features:

- ImageNet: ImageNet [39] pretrained ResNet-18 [11]
- Kinetics: Kinetics [20] pretrained R(2+1)D-18 [10]
- SPIN-TDM: Temporal Distance Matrices (TDM) [27] constructed from the output of SPIN [22] (3D joint positions)

- OpenPose-TDM: Temporal Distance Matrices [27] constructed from the output of OpenPose [2] (2D joint positions). Originally, TDM was proposed for 3D joint positions, but we also experiment with constructing TDMs from 2D joint positions.
- SimSiam: ImageNet pretrained model adapted/trained to our dataset using a general self-supervised image representation learning approach: SimSiam [5].
- Ours PAD: Inspired from [29, 37], we developed an autoencoder-based approach that learns to disentangle pose and appearance of the human. Pose vector is then used for error-detection. We term this pose and appearance disentangling approach Ours PAD. We initialized the encoder with ImageNet weights. We have elaborated on this baseline in the supplementary material.
- VideoSpeed-1: Kinetics pretrained model adapted/trained to our dataset using the pretext task of predicting speed of videos [1]. We considered the following speeds: 1x (normal), 2x (faster), 3x, 4x (fastest).
- VideoSpeed-2: same as VideoSpeed-1, but for 1x speed, we sampled frames uniformly from entire sequence. For higher speeds, it would create the effect of repeating the sequences. So, it can equivalently be considered as counting the exercise repetitions.
- VideoRot: Kinetics pretrained model adapted/trained to our dataset using the pretext task of predicting rotation amount of videos [18]. Rotation amount is selected randomly from $\{0°, 90°, 180°, 270°\}$.
- TemporalXform: Kinetics pretrained model adapted to our dataset using the pretext task of predicting various temporal transforms [17].
- Ours TemporalXform-1: We developed a contrastive learning-based approach in which the negative input is more temporally shifted than the positive input relative to the anchor. We initialized with Kinetics pretrained model.
- Ours TemporalXform-2: We developed another contrastive learning-based approach in which the negative is more temporally distorted than the positive input. We initialized with Kinetics pretrained model.

*Performance metric.* Since this dataset is imbalanced, we report the F1-score, instead of the accuracy.

**Dataset: Fitness-AQA BackSquat**

*Knees Inward and Knees Forward Errors.* First, we evaluated all the approaches on knees inward (KIE) and forward (KFE) errors. The results are summarized in Table 2. Additionally, here, we also considered a single-view, single-subject version of our cross-view, cross-subject pose-contrastive approach. In this version, anchor, positive, and negative inputs all belonged to the same video instance. We applied strong augmentations (rotation, translation, masking image regions, color channel order changing, zooming, blurring) during training this model. We refer to this approach as Vanilla-PC. We observed the following. 1) Training both image- and video-based self-supervision methods on our dataset helped in improving over their respective base models (ImageNet pretrained model and Kinetics pretrained model). 2) Our Vanilla Pose Contrastive learning improved the performance even more than our PAD. However, off-the-shelf pose estimator, OpenPose, still worked better than this model. 3) By contrast, our full pose-

| Feature extraction model | Modality | F-score ↑ | |
|---|---|---|---|
| | | KIE | KFE |
| OpenPose-TDM [2, 27] | 2D Pose | 0.4143 | 0.8123 |
| OpenPose-TDM* [2, 27] | 2D Pose | 0.3186 | 0.7968 |
| SPIN-TDM [22, 27] | 3D Pose | 0.2878 | 0.7761 |
| ImageNet [39] | Image | 0.1923 | 0.7725 |
| SimSiam [5] | Image | 0.2270 | 0.7868 |
| Ours PAD | Image | 0.3180 | 0.7784 |
| Ours Vanilla PC | Image | 0.4118 | 0.7965 |
| Ours CVCSPC | Image | **0.5195** | 0.8286 |
| Kinetics [20] | Video | 0.2970 | 0.8184 |
| VideoSpeed-1 [1] | Video | 0.3095 | 0.8155 |
| VideoSpeed-2 | Video | 0.3617 | 0.8000 |
| VideoRot [18] | Video | 0.3333 | 0.8138 |
| TemporalXform [17] | Video | 0.3414 | 0.8319 |
| Ours TemporalXform-1 | Video | 0.3457 | 0.8097 |
| Ours TemporalXform-2 | Video | 0.2286 | 0.8184 |
| Ours MD | Video | 0.4186 | **0.8338** |
| Ours MD + CVCSPC | Image, Video | **0.5263** | **0.8468** |

**Table 2. Performance comparison on Knees Inward and Knees Forward errors on our BackSquat dataset**.

| Feature extraction model | Modality | F-score ↑ |
|---|---|---|
| OpenPose-TDM [2, 27] | 2D Pose | 0.8340 |
| SimSiam [5] | Image | 0.8286 |
| Ours CVCSPC | Image | **0.8694** |

**Table 3. Performance comparison on detecting Shallow Squat error.**

contrastive model, CVCSPC outperformed all the models on KIE; for completeness, we also computed OpenPose baseline with our hyperparameter settings referred to as OpenPose*. 4) CVCSPC performing better than Vanilla PC also reinforced the importance of considering our cross-view and cross-subject conditions during pose contrastive learning. 5) Our MD model performed the best and second best on KFE and KIE, respectively. TemporalXform performed the best among general video self-supervised approaches. 6) Our domain knowledge-informed self-supervised approaches outperformed general self-supervised approaches, indicating the importance of using domain knowledge in designing self-supervised approaches. 7) Our contrastive learning-based approaches (CVCSPC and MD) worked better than our reconstruction-based approach (PAD). Furthermore, ensemble of our contrastive approaches outperformed all the models. Attention visualizations presented in the supplementary material.

Note that in all the subsequent experiments, we selected only the best performing methods for further evaluation.

*Shallow Squat Error.* We further considered evaluating and comparing approaches on another squat error—shallow squat error. Since shallow depth error is a static type of error, image models (2DCNN-based) are more suitable, where errors are detected in singular images, as opposed to in a stack of video frames. Using a 3DCNN for detecting single frame-based errors does not make sense. Therefore, we have not considered our MD approach for single frame-based errors. Single image detection also made end-to-end learning more feasible, so we finetuned our models end-to-end. The results are summarized in Table 3. We observed that OpenPose worked better than SimSiam. Our self-supervised learning performed the best, showing the importance of learning task-oriented representations, and its utility even in end-to-end finetuning scenarios.

**Dataset: Fitness-AQA OverheadPress.** Further, we evaluated and compared approaches on a different exercise—OverheadPress. The results are summarized in Table 4. We observed that video-based approaches worked better than image-based approaches on this exercise. Both of our proposed approaches outperformed the off-the-shelf pose estimator.

| Feature extraction model | Modality | F-score ↑ | |
|---|---|---|---|
| | | Elbow Err. | Knees Err. |
| OpenPose-TDM [2, 27] | 2D Pose | 0.4265 | 0.7131 |
| SimSiam [5] | Image | 0.4145 | 0.5301 |
| Ours CVCSPC | Image | 0.4522 | 0.7203 |
| TemporalXform [17] | Video | 0.4138 | 0.8416 |
| Ours MD | Video | **0.4552** | **0.8452** |

**Table 4. Performance comparison on detecting Elbow and Knees errors in OverheadPress exercise**.

| Feature extraction model | Modality | F-score ↑ | |
|---|---|---|---|
| | | Lumbar Err. | Torso Err. |
| OpenPose-TDM [2, 27] (SQ→BR) | 2D Pose | 0.5422 | 0.4060 |
| SimSiam [5] (SQ→BR) | Image | 0.5934 | 0.4543 |
| Ours CVCSPC (SQ→BR) | Image | **0.6057** | **0.4800** |
| Ours CVCSPC (OHP→BR) | Image | 0.5760 | 0.4675 |
| Ours CVCSPC (SQ+OHP→BR) | Image | **0.6338** | **0.5261** |

**Table 5. Cross-exercise transfer performance. Detecting Lumbar and Torso-Angle errors in BarbellRow exercise.**

### 5.3 Cross-Exercise Transfer

It is common to not have enough labeled data for each exercise. In such cases, it would be useful to transfer models from an exercise with abundant data over to exercises with limited data. So, in this experiment, we first transferred our model trained on BackSquat (SQ) exercise to BarbellRow exercise, where we detected two kinds of errors: Lumbar and TorsoAngle errors. Since these errors are static errors, we considered transferring our CVCSPC model. Note that in this experiment we used only a small amount of training data (details in the Supplementary Material). The results are presented in Table 5. We observed that models pretrained using our proposed self-supervised approach performed better than baselines even when finetuned to a different exercise action. We also transferred from Overhead Press (OHP), & noted improvements. Lastly, we also tried the ensemble of our SQ & OHP transferred models, which worked the best.

### 5.4 Applications to Other Domains

**Pose Estimation.** We conducted a novel pose retrieval experiment where we retrieved images based on query poses using our pose-contrastive embeddings. From the results shown in Fig. 6, it can be seen that compared to SimSiam embeddings, ours are much better at encoding pose information, even with camera angle variation. We believe that our representations can be decoded into actual 2D/3D joint positions, by using a small pose-annotated dataset. We will explore this further in future research.

**Dive Quality Assessment.** *While we use symmetry to simplify problems, our methods are generalizable, e.g.*, we applied our motion disentangling method for

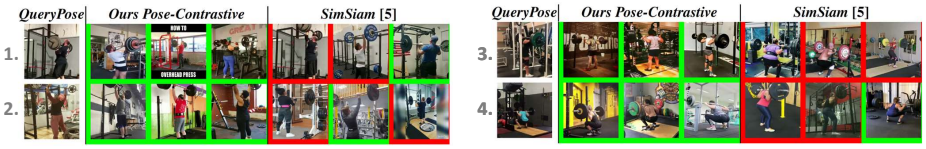


**Fig. 6.** Results of **pose-based retrieval** experiment.

assessing the quality of Olympic dives on MTL-AQA dataset [34]. Global motion & local motions here refer to the motion of the dive-classes & the errors in them, respectively. To disentangle local motion, we match-contrast dives from the same dive-class from the same diving events so that the background remains same. We used supervised dive-classification pretraining as the baseline. Performance metric is Spearman's rank correlation (higher is better). We found significant improvement after incorporating our motion disentangling approach as shown in Table 6, even surpassing previous self-supervised state-of-the-art [38].

| Model | SSL SoTA [38] | Ours baseline | Ours MD |
|---|---|---|---|
| Sp. Corr. | 0.7700 | 0.5665 | **0.7763** |

**Table 6. Motion disentangling for Dive quality assessment.**

## 6 Conclusion

In this paper, we addressed the problem of assessing the workout form in real-world gym scenarios, where we showed that pose-features from off-the-shelf pose estimators cannot be reliably used for detecting subtle errors in workout form, as these pose estimators struggle to perform well due to unusual poses, occlusions, illumination, and clothing styles. We tackled the problem by replacing these noisy pose features with our more robust image and video representations learnt from unlabeled videos using domain knowledge-informed self-supervised approaches. Using self-supervision helped in avoiding the cost of annotating poses. Mapping of our self-supervised representations to workout form error probabilities was learnt using a much smaller labeled dataset. We also introduced a novel dataset, Fitness-AQA, containing actual, unscripted exercise samples from real-world gyms. Experimentally, we found that while our self-supervised features performed comparably in simpler conditions, they outperformed off-the-shelf pose estimators and various baselines in complex real-world conditions on multiple exercises. We also showed that pose information is encoded in our representations; and our motion disentangling approach can be used to assess quality of motion in other domains.

## References


1. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9922–9931 (2020) 2, 4, 11, 12
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019) 2, 11, 12, 13
3. Chen, S., Yang, R.R.: Pose trainer: correcting exercise posture using pose estimation. arXiv preprint arXiv:2006.11718 (2020) 2
4. Chen, X., Pang, A., Yang, W., Ma, Y., Xu, L., Yu, J.: Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. arXiv preprint arXiv:2104.11452 (2021) 3
5. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021) 2, 4, 11, 12, 13
6. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005) 5
7. Doughty, H., Mayol-Cuevas, W., Damen, D.: The pros and cons: Rank-aware temporal attention for skill determination in long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7862–7871 (2019) 3
8. Du, C., Graham, S., Depp, C., Nguyen, T.: Assessing physical rehabilitation exercises using graph convolutional network with self-supervised regularization. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 281–285. IEEE (2021) 3
9. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018) 4
10. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018) 9, 10
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 9, 10
12. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science **313**(5786), 504–507 (2006) 4
13. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015) 6
14. Honari, S., Constantin, V., Rhodin, H., Salzmann, M., Fua, P.: Unsupervised learning on monocular videos for 3d human pose estimation. arXiv preprint arXiv:2012.01511 (2020) 4
15. Hyvarinen, A., Morioka, H.: Unsupervised feature extraction by time-contrastive learning and nonlinear ica. Advances in Neural Information Processing Systems **29**, 3765–3773 (2016) 4
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (jul 2014) 5

17. Jenni, S., Meishvili, G., Favaro, P.: Video representation learning by recognizing temporal transformations. In: European Conference on Computer Vision. pp. 425–442. Springer (2020) 2, 4, 11, 12, 13
18. Jing, L., Yang, X., Liu, J., Tian, Y.: Self-supervised spatiotemporal feature learning via video rotation prediction. arXiv preprint arXiv:1811.11387 (2018) 2, 4, 11, 12
19. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018) 2, 10
20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 9, 10, 12
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
22. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019) 2, 10, 12
23. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR 2011. pp. 3361–3368. IEEE (2011) 4
24. Li, J., Bhat, A., Barmaki, R.: Improving the movement synchrony estimation with action quality assessment in children play therapy. In: Proceedings of the 2021 International Conference on Multimodal Interaction. pp. 397–406 (2021) 3
25. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9522–9531 (2021) 3
26. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. pp. 527–544. Springer (2016) 4
27. Ogata, R., Simo-Serra, E., Iizuka, S., Ishikawa, H.: Temporal distance matrices for squat classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 2, 3, 8, 9, 10, 11, 12, 13
28. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 3
29. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A., Zhang, R.: Swapping autoencoder for deep image manipulation. Advances in Neural Information Processing Systems **33**, 7198–7211 (2020) 11
30. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1468–1476. IEEE (2019) 3
31. Parmar, P., Morris, B.T.: Measuring the quality of exercises. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 2241–2244. IEEE (2016) 3
32. Parmar, P., Reddy, J., Morris, B.: Piano skills assessment. arXiv preprint arXiv:2101.04884 (2021) 3
33. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–28 (2017) 3, 9
34. Parmar, P., Tran Morris, B.: What and how well you performed? a multitask learning approach to action quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 304–313 (2019) 3, 14

35. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: European Conference on Computer Vision. pp. 556–571. Springer (2014) 3, 4
36. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 9
37. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3d human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 750–767 (2018) 4, 5, 11
38. Roditakis, K., Makris, A., Argyros, A.: Towards improved and interpretable action quality assessment with self-supervised alignment. In: The 14th PErvasive Technologies Related to Assistive Environments Conference. pp. 507–513 (2021) 14
39. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 10, 12
40. Ryali, C.K., Schwab, D.J., Morcos, A.S.: Characterizing and improving the robustness of self-supervised learning through background augmentations. arXiv preprint arXiv:2103.12719 (2021) 6
41. Sardari, F., Paiement, A., Hannuna, S., Mirmehdi, M.: Vi-net—view-invariant quality of human movement assessment. Sensors **20**(18), 5258 (2020) 3
42. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 1134–1141. IEEE (2018) 4
43. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7396–7404 (2018) 6
44. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: Uncertainty-aware score distribution learning for action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9839–9848 (2020) 3
45. Tao, L., Paiement, A., Damen, D., Mirmehdi, M., Hannuna, S., Camplani, M., Burghardt, T., Craddock, I.: A comparative study of pose representation and dynamics modelling for online motion quality assessment. Computer vision and image understanding **148**, 136–152 (2016) 3
46. Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W.: Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In: CVPR. pp. 4006–4015 (2019) 4
47. Wang, J., Jiao, J., Liu, Y.H.: Self-supervised video representation learning by pace prediction. In: European conference on computer vision. pp. 504–521. Springer (2020) 4
48. Wang, T., Wang, Y., Li, M.: 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. pp. 668–678. Springer (2020) 3
49. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score figure skating sport videos. IEEE transactions on circuits and systems for video technology **30**(12), 4578–4590 (2019) 3, 9

50. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: Computer Vision and Pattern Recognition (CVPR) (2019) 4
51. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7919–7928 (2021) 3
52. Zeng, L.A., Hong, F.T., Zheng, W.S., Yu, Q.Z., Zeng, W., Wang, Y.W., Lai, J.H.: Hybrid dynamic-static context-aware attention network for action assessment in long videos. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2526–2534 (2020) 9