Abstract

Maintaining proper form while exercising is important for preventing injuries and maximizing muscle mass gains. While fitness apps are becoming popular, they lack the functionality to detect errors in workout form. Detecting such errors naturally requires estimating users’ 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-specific 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 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, 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- & 3D-pose estimators and several other baselines.

1. Introduction

Detecting errors in users’ 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 disadvan-

taged 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 humans. Academic research in workout form assessment so far has been limited to simple, controlled conditions [22], where posture can be reliably estimated using off-the-shelf
(OTS) pose estimators \cite{DBLP:journals/corr/abs-1907-05537}. 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:

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, and one reconstruction-based self-supervised approaches that capitalize on the harmonic motion of workout actions and the large variance in unlabeled gym videos to learn robust fitness domain-specific representations (Sec. 3). Our representations outperform various baselines including 2D- and 3D-pose estimators on the task of workout form assessment on existing and our newly introduced datasets. We believe that, in general, future work on representation learning would benefit from using 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 enclosing a small labeled subset for evaluation. We show that this in-the-wild dataset provides a significantly more challenging benchmark than the existing one recorded in controlled conditions.

Note that our objective is not to improve the performance directly on the task of human pose estimation per se, but rather to improve the performance on exercise error detection.

**Ethics:** We have blurred the faces in our paper. Our dataset is diverse and not limited to any particular demographic.

### 2. Related Work

In the following, we discuss relevant prior work in the areas of action quality assessment (AQA) and SSL.

**AQA and Skills Assessment (SA).** Our work can be classified under AQA and SA, which involves the computer vision-based quantification of the quality of movements and actions. Works in AQA and SA have mainly been focused on domains like physiotherapy \cite{DBLP:journals/corr/abs-1907-05537, DBLP:journals/infocoms/Gonzalez-Torre-Macias-Gonzalez-Morales-Ruiz-Morales-Medina-Medina-19, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/tmometrics/Malik-18, DBLP:journals/tmometrics/Ruiz-Morales-Medina-Medina-18, DBLP:journals/tmometrics/Schreck-Ruiz-Medina-Medina-18}, Olympic sports \cite{DBLP:journals/corr/abs-1907-05537, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/infocom/Li-Yang-Ding-Yu-18}, various types of skills \cite{DBLP:journals/corr/abs-1907-05537, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/infocom/Li-Yang-Ding-Yu-18, DBLP:journals/infocom/Li-Yang-Ding-Yu-18}. 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 \cite{DBLP:journals/corr/abs-1907-05537, DBLP:journals/corr/abs-1907-05537}; 2) image and video features-based \cite{DBLP:journals/corr/abs-1907-05537, DBLP:journals/corr/abs-1907-05537}. 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 [3], 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-specific 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. [22]. 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 [10], which learn low-dimensional representations by reconstructing the input. Le et al. [18] 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 [29], but was found to perform worse than an OTS pose estimator. More recent SSL methods include leveraging properties of video data like: 1) the temporal order [21, 40]; 2) the objects’ upright positions [7]; and 3) motion and appearance statistics [37]. A few works propose to leverage time-contrast to learn representations using self-supervision [12, 13, 34]. 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 [31]. 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 approach—applicable to our dataset.

3. Method

In Secs. 3.1-3.3, we present our self-supervised approaches. 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 - temporally aligning videos. Our methods build upon temporally aligned videos. Given a collection of videos of 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. We leverage the following property to temporally align 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. Note that, alternatively, other methods, e.g. optical flow, can be used to temporally align videos.

3.1. Self-Supervised Pose Contrastive Learning

Objective. Given the temporally aligned videos of the same exercise action, 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 [4]. In our case, we hypothesize that we can extend contrastive learning to learn human pose-sensitive representations. Particularly, we propose a self-supervised pre-text task, which aims to pull together images (frames of videos) containing humans in similar poses, while pushing apart images with humans in dissimilar poses. Note that, this approach operates on single frame-triplets (not videos or clips) at a time.

**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. 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. We optimize the parameters of \( f \) during the self-supervised training, by minimizing the distance ratio loss [11],

\[
\mathcal{L} = -\log \frac{e^{-||\phi_{\text{anc}} - \phi_{\text{pos}}||_2}}{e^{-||\phi_{\text{anc}} - \phi_{\text{pos}}||_2} + e^{-||\phi_{\text{ane}} - \phi_{\text{neg}}||_2}}.
\]

### 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 Backsquats, 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 Backsquats, 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.

**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 identify 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.

### 3.3. Self-Supervised Pose and Appearance Disentangling

Different from our previous two contrastive learning-based approaches, in this image reconstruction-based approach, we aim to disentangle the pose of the humans doing exercises from their appearance (in computer vision sense). We extend the method proposed by Rhodin *et al.* [31], which disentangles camera view and appearance. We extend their concept to a new setting and propose to disentangle the pose and the appearance, while the camera view remains the same. To accomplish this, we leverage an autoencoder setup. This approach involves using two different frames from an exercise-video instance. The person doing the exercise would be in different poses in both frames. We, then, split the encoder embedding into a pose and an appearance vector; and swap the appearance vectors between the two frames before reconstructing the frames using a decoder. The approach is shown in Fig. 6 and additional details are provided in the Supplementary Material. We hypothesize that the swapping operation disentangles appearance from pose by separating time-varying and time-invariant features. To ensure that appearance feature do not collapse to trivial/null solution, we use appearance vector of much larger size than the pose vector. Pose vector is only 32-dim vector. We then use the pose features for error-detection.

### 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.* [22] 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. To that end, we collected the largest exercise assessment dataset from video sharing sites such as Instagram and YouTube.

We considered the following three exercises: 1) Back-Squat; 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. Unique properties of our dataset:

- **Real-world videos.** Unlike the existing dataset [22], 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 [22], 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, and/or 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.
- **Annotated by multiple expert trainers.** Our dataset has been annotated by professional gym trainers. Due to this, even very subtle errors are caught and annotated accordingly. Errors range from very subtle to very severe.

We have provided descriptions of various exercise errors in Table 6 (after the References).

### 5. Experiments

To validate our contributions, in our evaluation, 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. We first trained our models on the unlabeled datasets, and then used our self-supervisedly learnt models 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 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. Therefore, in this case, the 2DCNN backbone is not finetuned unless specified otherwise.

| Feature extraction | Modality | Accuracies |
|---------------------|----------|------------|
|                     |          | KIE    | CVRB  | CCRB  | SS   | KFE | Avg   |
| HMR-TDM [22]        | 3D Pose  | 89.80 | 96.65 | 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.

**Implementation details.** We used ResNet-18 [9] as the backbone CNN unless specified otherwise. We used custom YOLOv3 [30] to detect barbells/weights; and normalized the amplitudes of the trajectories to -180 to 180 (simply for a resemblance to a full 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 pretrained weights. We used ADAM optimizer [16] 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 [8] as our backbone CNN. We sampled 16 frames from each half-cycle. We randomly applied strong augmentations. We initialized our backbone CNN with Kinetics [15] 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.

- **Pose and Appearance Disentanglement approach (PAD).** We initialized our backbone CNN, which was based on ResNet-18, with the ImageNet pretrained weights. We used ADAM optimizer with an initial learning rate of 1e-4 and optimized it for 500 epochs with a batch size of 25.

Further details provided in the supplementary material.

### 5.1. Case Study 1: Simple Conditions

The Waseda Squat dataset [22] provides an excellent labeled dataset for evaluating exercise errors in controlled conditions. The publicly available portion of 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 \[14\]. HMR-TDM features were made available by Ogata et al. \[22\]. 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 \[22\] 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 one does not expect the performance gap to be large on this dataset, as in these simpler conditions, pose estimators work quite well.

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 \[32\] pretrained ResNet-18 \[9\]
- Kinetics \[15\] pretrained R(2+1)D-18 \[8\]
- Temporal Distance Matrices (TDM) \[22\] constructed from the output of SPIN \[17, 22\] (3D joint positions)
- Temporal Distance Matrices \[22\] constructed from the output of OpenPose \[1\] (2D joint positions). Originally, TDM was proposed for 3D joint positions, but we also experiment with constructing from 2D joint positions.
- ImageNet pretrained model adapted to our dataset using a general self-supervised image representation learning approach: SimSiam \[2\]

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

5.2.1 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) Kinetics pretrained 3DCNN baseline outperformed the 2DCNN counterpart; 2) Adaptation using the general SSL methods (SimSiam) improved the performance of regular ImageNet pretrained model; 3) Improvement brought by our PAD model was more than that by general SSL method; 4) Our vanilla pose-contrastive learning improved the performance even more than PAD. However, off-the-shelf pose estimator, OpenPose still worked better than this baseline; 5) By contrast, our full pose-contrastive model, CVCSPC outperformed all the models on KIE; for completeness, we also computed OpenPose baseline with our hyperparameter settings referred to as OpenPose∗; 6) Our MD model performed best on KFE; 7) CVCSPC performing better than Vanilla PC also reinforced the importance of considering our cross-view and cross-subject conditions during pose-contrastive learning; 8) Our contrastive learning-based approaches worked better than our reconstruction-based approach; 9) ensemble of our contrastive approaches outperformed all the models.

In all the subsequent experiments, we considered only the best performing methods for further evaluation.

Attention visualization. To see where our CVCSPC and MD focused after self-supervised training, we visualized using PCA. We also compared the location of attentions
with those before the self-supervised training (refer to Fig. 7). We found that after self-supervised training, the CNN focused more on the human doing the exercise, around the important body parts, for example, legs in case of squats.

**Shallow Squat.** 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. 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 end-to-end finetuning allowed the regular ImageNet pretrained model to outperform OpenPose in complex cases. Our self-supervised learning performed the best, showing how important it is to learn task-specific representations, and that our self-supervised training has utility in end-to-end finetuning scenarios as well.

### 5.2.2 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.

### 5.3. Cross-Exercise Transfer

It is common to not have enough labeled data for each exercise. In such cases, it would be useful if we can transfer models from an exercise with abundant data over to exercises with limited data. So, in this experiment, we transferred our model trained on BackSquat 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.

### 6. Limitations

Our approaches require that the body parts in which the error is to be detected, should be active/moving during the exercise to work well. For example, we do not expect our approach to work well on BarbellRow Lumbar error if trained using our SSL techniques on BarbellRow (lumbar does not move in this exercise). However, a way to overcome this is as we did in Sec. 5.3—by transferring the model from an exercise where the body part of interest was active (backsquat in our case).

---

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

| Feature extraction model | F-score  |
|--------------------------|----------|
| OpenPose-TDM [1, 22]     | 0.8340   |
| ImageNet [32]            | 0.8398   |
| Ours CVCSPC              | **0.8694** |

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

| Feature extraction model | F-score  |
|--------------------------|----------|
| Elbow Err.               | Knees Err. |
| OpenPose-TDM [1, 22]     | 0.4265   | 0.7131   |
| ImageNet [32]            | 0.3687   | 0.5622   |
| Ours CVCSPC              | 0.4522   | 0.7203   |
| Kinetics [15]            | 0.4252   | 0.8382   |
| Ours MD                  | **0.4552** | **0.8452** |

**Table 5. Performance comparison on detecting Lumbar and Torso-Angle errors in BarbellRow exercise.**

| Feature extraction model | F-score  |
|--------------------------|----------|
| Lumbar Err.              | Torso Err. |
| OpenPose-TDM [1, 22]     | 0.5422   | 0.4060   |
| ImageNet [32]            | 0.5960   | 0.4340   |
| Ours CVCSPC              | **0.6057** | **0.4800** |
7. 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, 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.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2, 7, 8

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 7

[3] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *arXiv preprint arXiv:2104.11452*, 2021. 2, 3

[4] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 1, pages 539–546. IEEE, 2005. 4

[5] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. 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*, pages 7862–7871, 2019. 2

[6] Chen Du, Sarah Graham, Colin Depp, and Truong Nguyen. 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)*, pages 281–285. IEEE, 2021. 2

[7] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 3

[8] Kesho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 6, 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[10] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 3

[11] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015. 4

[12] Sina Honari, Victor Constantin, Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised learning on monocular videos for 3d human pose estimation. *arXiv preprint arXiv:2012.01511*, 2020. 3

[13] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016. 3

[14] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6, 7, 8

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2, 7

[18] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011*, pages 3361–3368. IEEE, 2011. 3

[19] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. Improving the movement synchrony estimation with action quality verification. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 397–406, 2021. 2

[20] 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. 2

[21] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3
[22] Ryoji Ogata, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Temporal distance matrices for squat classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019. 1, 3, 5, 6, 7, 8
[23] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019. 2
[24] 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. 2
[25] Paritosh Parmar and Brendan Tran Morris. Measuring the quality of exercises. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 2241–2244. IEEE, 2016. 2
[26] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. arXiv preprint arXiv:2101.04884, 2021. 2
[27] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 20–28, 2017. 2, 6
[28] 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 Conference on Computer Vision and Pattern Recognition, pages 304–313, 2019. 2
[29] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In European Conference on Computer Vision, pages 556–571. Springer, 2014. 2, 3
[30] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. 6
[31] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 750–767, 2018. 3, 5
[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3):211–252, 2015. 7, 8
[33] Faegheh Sardari, Adeline Paiement, Sion Hannuna, and Majid Mirmehdi. Vi-net—view-invariant quality of human movement assessment. Sensors, 20(18):5258, 2020. 2
[34] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In 2018 IEEE international conference on robotics and automation (ICRA), pages 1134–1141. IEEE, 2018. 3
[35] 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. 2
[36] Lili Tao, Adeline Paiement, Dima Damen, Majid Mirmehdi, Sion Hannuna, Massimo Camplani, Tilo Burghardt, and Ian Craddock. A comparative study of pose representation and dynamics modelling for online motion quality assessment. Computer vision and image understanding, 148:136–152, 2016. 2
[37] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In CVPR, pages 4006–4015, 2019. 3
[38] Tianyu Wang, Yijie Wang, and Mian Li. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 668–678. Springer, 2020. 2
[39] 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. 2, 6
[40] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In Computer Vision and Pattern Recognition (CVPR), 2019. 3
[41] 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. 2
[42] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2526–2534, 2020. 6
| Exercise     | Error type      | Correct                              | Incorrect                                           |
|--------------|-----------------|--------------------------------------|-----------------------------------------------------|
| BackSquat    | Knees Inward Error | Knees pointing outwards              | knees buckling in                                    |
| BackSquat    | Knees Forward Error | Knees should be aligned over the toes | knees moving forward excessively                      |
| BackSquat    | Shallow Squat Error | Glutes below knees-line              | glutes above knees-line                              |
| BarbellRow   | Lumbar Error     | Lower back should be neutral         | Exaggerated curvature of the lumbar spine            |
| BarbellRow   | Torso-Angle Error | Back aligned with the hips at a $45^\circ$-$90^\circ$ angle with the core | Too high less than $45^\circ$ torso inclination / Too low, torso below parallel, compared to the floor |
| OverheadPress| Elbow Error      | Elbows underneath the wrists         | Elbows flaring out or caving in                      |
| OverheadPress| Knees Error      | Knees should be locked at all times  | Knees are bent. Knees go from being bent to straight to help jerk the bar up (because the weights overly heavy for the person) |

Table 6. Description of various errors covered in our dataset.