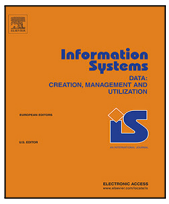
Information Systems 134 (2025) 102579

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is



Pose estimation analysis and fine-tuning on the REHAB24-6 rehabilitation dataset

Andrej Černek [a,*], Jan Sedmidubsky [a], Petra Budikova [b]

[a] Faculty of Informatics, Masaryk University, Botanická 68a, Brno, 602 00, Czechia
[b] VisionCraft, Výstaviště 405/1, Brno, 603 00, Czechia

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.13305825

Keywords:
REHAB24-6 dataset
Pose estimation
Motion capture
Rehabilitation exercise
Skeleton format
Fine-tuning 2D/3D detectors
Similarity of repetitions

ABSTRACT

Human motion analysis is a key enabler for remote healthcare applications, particularly in physical rehabilitation. In this context, mobile devices equipped with RGB cameras seem to be a promising technology for monitoring patients during home-based exercises and providing real-time feedback. This relies on pose estimation algorithms that extract spatio-temporal features of human motion from video data. While state-of-the-art models can estimate body pose from mobile video streams, their effectiveness in rehabilitation scenarios remains underexplored. To address this, we introduce the REHAB24-6 dataset, which includes untrimmed RGB videos, 2D and 3D skeletal ground truth annotations, and temporal segmentation for six common rehabilitation exercises. We also propose an evaluation protocol for assessing different aspects of quality of pose estimation methods, dealing with challenges that arise when different skeleton formats are compared. Additionally, we show how fine-tuning of existing models on our dataset leads to improved quality. Our experimental results compare several state-of-the-art approaches and highlight their key limitations – particularly in depth estimation – offering practical insights for selecting and improving pose estimation systems for rehabilitation monitoring.

1. Introduction

The goal of *human pose estimation* (*HPE*) methods is to identify the 2D or 3D positions of key body joints across consecutive frames of an RGB video stream. These estimated joint positions form a simplified spatio-temporal representation of human motion, referred to as a *skeleton sequence*. Analyzing this skeleton representation unlocks new opportunities for various applications [1], with healthcare being one of the most promising fields. In this domain, computer-assisted motion monitoring can, for instance, enhance the effectiveness of physical therapy treatments [2].

This paper is driven by the practical need for rehabilitation-support software designed to assist patients during their home-exercise phase of physical therapy. In such software, reference exercises would be recorded for individual patients at a physiotherapy clinic and then compared to motions captured during their home exercises. These movements would be represented as skeleton sequences, estimated by applying a suitable HPE method to RGB video streams recorded using a standard smartphone. To ensure reliable monitoring of exercise accuracy, it is critical to select an HPE method that is both efficient and precise. But how do we identify the most appropriate method from the many state-of-the-art research models and commercial solutions [3]?

Unsurprisingly, questions about the quality of HPE methods have been widely explored. Several established benchmarks and datasets exist to evaluate HPE methods [4,5]. However, these evaluations provide only partial insights into the suitability of the underlying deep *neural network* (*NN*) models. There are several reasons for this limitation. First, existing testbeds often focus on basic everyday motions (e.g., sitting, walking, or picking up a phone) that may differ significantly from physical therapy exercises. Second, HPE methods utilize various skeleton formats, but existing benchmarks typically compare only NNs that use the *same* skeleton format (body model) as the dataset. Third, standard benchmarks primarily evaluate skeleton data based on joint-coordinate accuracy and do not consider the usefulness of these data for application-specific tasks, such as evaluating skeleton data similarity in the context of physical therapy.

As discussed, there is a clear need for a specialized testbed focused on rehabilitation exercises, along with a methodology for comparing the quality of HPE methods across different skeleton formats while considering their practical utility. This need is analyzed in our recent work [3]. In this manuscript, we provide an extended version of this work by incorporating additional insights and analyses. As a whole, all the achieved contributions can be summarized as follows.

* Corresponding author.
E-mail addresses: cernek@mail.muni.cz (A. Černek), sedmidubsky@mail.muni.cz (J. Sedmidubsky), budikova@visioncraft.cz (P. Budikova).

https://doi.org/10.1016/j.is.2025.102579
Received 13 February 2025; Received in revised form 12 May 2025; Accepted 23 June 2025
Available online 7 July 2025

- REHAB24-6, a dataset of rehabilitation exercises captured using two synchronized RGB cameras and a motion-capture system that provides ground truth in the form of accurate 3D joint coordinates. The dataset is also equipped with annotations of performed repetitions of six types of rehabilitation exercises, along with descriptions of typical problems that arise during exercise.
- A methodology for evaluating diverse HPE methods on the REHAB24-6 dataset, focusing on both the position error of joint coordinates and the usefulness of obtained skeleton data in a real-world application task of distinguishing between correctly and incorrectly performed exercises.
- Modifications of existing NN architectures along with their fine-tuning to demonstrate how fine-tuning 3D HPE methods and the underlying 2D pose estimators can increase the overall HPE accuracy in the context of introduced cross-subject, cross-camera, and cross-exercise scenarios.
- Extensive experimental evaluation of several state-of-the-art HPE methods and their fine-tuned variants, with a detailed analysis of factors that influence the precision and practical applicability of individual approaches.

The manuscript is further organized as follows. Section 2 describes existing HPE methods, the comparison of publicly available datasets for evaluation of HPE methods, and how the HPE quality is determined. Section 3 describes the introduced REHAB24-6 rehabilitation dataset, including the description of provided data, the description of exercises along with typical errors that arise during exercising, and dataset limitations along with planned extensions. Section 4 introduces the qualitative and quantitative metrics for determining the HPE quality in the context of rehabilitation, including the ways of dealing with varying skeleton formats and improper alignment of estimated skeletons with respect to ground-truth skeletons. Section 5 describes how the underlying architectures of HPE methods can be fine-tuned to increase the estimation quality. Section 6 extensively evaluates the quality of existing HPE methods using the proposed evaluation methodology on the REHAB24-6 dataset and discusses observed issues.

2. State-of-the-art approaches

The development and evaluation of HPE methods are closely linked to the availability of training/test data. In this section, we first provide a brief survey of state-of-the-art HPE approaches and highlight those 2D and 3D HPE methods that are highly ranked for general-purpose scenarios. Then, we focus on available datasets of human motion data and discuss their limitations in terms of scope and available metadata.

2.1. Human pose estimation

HPE entails extracting time series of joints from videos, with the most common representation being a *skeleton* (a set of connected joints or keypoints in general) [6,7]. We differentiate between 2D estimation, which infers the positions of joints in the *pixel space* of video frames, and 3D, where the positions are set in a *virtual coordinate system*, often normalized using the position of one joint in the origin.

State-of-the-art multi-person 2D pose estimation approaches comprise detecting the object location (e.g., using the YOLOv3 object detector), followed by a single-person pose estimation on each object, like in the cases of HRNet [8], AlphaPose [9], and BlazePose [10]. As a result, if the object's bounding box is incorrectly selected, the 2D model cannot detect anything meaningful. YOLOv7 [11,12] avoids this by detecting the bounding boxes and keypoints in a single pass.

Joint-based monocular (from a single camera stream) 3D pose estimation can also be done in a similar vein. However, the low diversity of datasets with 3D ground truth often leads to difficulties in generalization. The state-of-the-art instead lies in 2D-to-3D lifting, which focuses on predicting the depth of 2D poses generated by a 2D pose detector.

Since a single view cannot capture all the potential depth information, monocular models like MotionBERT [13], MHFormer [14] or STCFormer [15] try to recover additional depth clues from the temporal domain by using a sliding window over the 2D pose sequences generated by an off-the-shelf model, resulting in smoothing. Some commercial solutions provide the entire pipeline, e.g., MediaPipe Pose[1] uses modified BlazePose for 2D estimation and then executes lifting to 3D with the help of the mesh-based GHUM [16]. Even if there are multi-view models [4] providing higher estimation accuracy, they often require camera calibrations, which is impractical for a home-exercise environment.

2.2. Human motion datasets and skeleton formats

High-quality datasets for training are vital to the effectiveness of the pose estimation. Table 1 compares the existing datasets with regard to the attributes relevant to the physical therapy correctness discrimination. Overall, the weakness of existing fitness/rehabilitation datasets is a limited range of exercises with available RGB videos and missing information about the correctness of each exercise repetition.

EC3D [23] is one of the few datasets with incorrect exercise repetitions, but it only covers three exercises, and the videos are private. Another such dataset, UI-PRMD [28], also lacks publicly available videos, and the correctness is not evaluated by an expert. Similarly, IRDS [30] only offers depth videos, and the qualification of the annotators is not specified. While KIMORE [24] supplies correctness scores from multiple experts for healthy and disordered participants across five lower-back exercises, the data are not available on the level of individual repetitions. These characteristics are important for deciding which HPE method is suitable for exercise monitoring. This gap is filled by the introduced REHAB24-6 dataset.

From other characteristics, we noted the video format: synthetic (mixed reality or video game) videos offer a larger diversity of environments and views but are less realistic. The skeleton sequences also vary in quality, where mocap ground truth (e.g., REHAB24-6 or Human3.6M) provides more significant accuracy than Kinect or HPE predictions (e.g., KERAAL).

2.3. HPE quality evaluation

The standard measure for evaluating HPE methods is the Mean Per Joint Position Error (MPJPE) [18], which calculates the mean Euclidean distance between the real and predicted position of each joint in all frames of a given benchmark dataset. However, this requires that the HPE method uses the same skeleton format as the benchmark. Unfortunately, there is no standardized skeleton format, and a variety of formats can be found in the datasets, differing not only in the number of joints but also in their exact position on the human body. For example, the most widely used 2D training and benchmarking dataset is MS COCO [17], which features a 17-joint skeleton extended to 33 joints by BlazePose. On the other hand, the primary dataset for 3D pose estimation is Human3.6M [18] with a different skeleton of 17 joints, which differs from COCO in joints on the head and body core. Consequently, only the HPE methods trained on the same dataset are usually compared.

[1] https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/pose.md.

Table 1
Comparison of properties of general-purpose (MS-COCO and Human3.6M) and fitness-based (the rest) human motion datasets. The "N/A" symbol denotes that clear information about a given feature is not provided or the feature is not publicly available. The "Modality" columns represents whether RGB videos, and 2D or 3D skeleton sequences are provided. The "lighting" column indicates whether there are varying lighting conditions across the dataset videos. The "segment." column indicates the presence of temporal segmentation on the level of exercise repetitions, while the "correct." column indicates the presence of annotations of how well the exercises were performed.

Dataset	Modality			RGB properties			Repetitions		Dimensions: # of		
	RGB	2Dj	3Dj	#cams	lighting	Real/Synth	segment.	correct.	subjs	exercs	frames
MS-COCO [17]	✓	✓	✗	1	✗	R	✗	✗	25 K	N/A	330 K
Human3.6M [18]	✓	✓	✓	4	✗	R+S	✗	✗	11	N/A	3.6 M
Fit3D [19]	✓	✓	✓	4	✗	R	✓	✗	11	47	>3 M
Lower body [20]	N/A	N/A	N/A	1	✗	R	N/A	✗	20	31	1.9 M
FLAG3D [21]	✓	N/A	✓	1	✗	S	N/A	✗	24	60	20 M
mRI [22]	✓	✓	✓	2	✗	R	✗	✗	20	12	160 K
EC3D [23]	N/A	✓	✓	4	✗	R	✓	✓	4	3	30 K
KIMORE [24]	N/A	✗	✓	1	✗	R	✗	✗	78	5	N/A
InfiniteForm [25]	✓	✓	✓	N/A	✓	S	N/A	✗	N/A	15	60 K
UCOPhyRehab [26]	✓	✓	✓	5	✗	R	✗	✗	27	16	1.6 M
KERAAL [27]	✓	✓	✓	1	✗	R	✗	✓	21	3	N/A
UI-PRMD [28]	N/A	✗	✓	N/A	✗	R	✓	✓	10	10	N/A
K3Da [29]	N/A	N/A	✓	1	✗	R	✗	✗	54	13	N/A
IRDS [30]	N/A	✓	✓	1	✗	R	✓	✓	29	9	N/A
EmoPain [31]	N/A	N/A	N/A	8	✗	R	✗	✗	50	11	N/A
REHAB24-6	✓	✓	✓	2	✓	R	✓	✓	10	6	370 K

3. REHAB24-6: Dataset Description

To enable the evaluation of HPE models and the development of exercise feedback systems, we introduce the REHAB24-6 rehabilitation dataset, which is made publicly available in the Zenodo repository: https://doi.org/10.5281/zenodo.13305825. The main focus is on a diverse range of exercises, views, body heights, lighting conditions, and exercise mistakes. With the RGB videos, skeleton sequences, repetition segmentation, and exercise correctness labels, this dataset currently offers the most comprehensive testbed for exercise-correctness-related tasks.

3.1. Recording conditions

Our laboratory setup included 18 synchronized sensors (2 RGB video cameras, 16 ultra-wide motion capture cameras) spread around an 8.2×7 m room, as shown in Fig. 1. The RGB cameras were located in the corners of the room, one in a horizontal position (hor.), providing a larger field of view (FoV), and one in a vertical position (ver.), resulting in a narrower FoV. Both types of cameras were synchronized with a sampling frequency of 30 frames per second (*FPS*).

The subjects wore motion capture body suits with 41 markers attached to them, which were detected by optical cameras. The OptiTrack Motive 2.3.0 software inferred the 3D positions of the markers in *virtual centimeters* and converted them into a skeleton with 26 joints, forming our human pose 3D *ground truth* (*GT3D*).

To acquire a 2D version of the ground truth (*GT2D*) in *pixel coordinates*, we applied a projection of the virtual coordinates into the camera using the simplified pin-hole model. We estimated the parameters for this projection as follows. First, the virtual position of the cameras was estimated using a measuring tape and knowledge of the virtual origin. Then, the orientations of the cameras were optimized by matching the virtual marker positions with their position in the videos.

We also simulated changes in lighting conditions: a few videos were shot in the natural evening light, which resulted in worse visibility, while the rest were under artificial lighting.

3.2. Description of exercises

Ten subjects participated in our recording and consented to release the data publicly: six men and four women of different ages (from 25 to 50) and fitness levels. The following six types of exercises, which constitute a representative sample for rehabilitating various body parts, were recorded. A graphical illustration of all the types of exercises is provided in Fig. 2.

Ex1. Arm abduction: sideway raising of a straightened right arm from 0° to about 180°. Torso is upright. Exercising shoulder is not lifted, arm moves in line with the body (from a side view).

Ex2. Arm VW: fluent transition of arms between V (arms straight up) and W (elbows down, hands up) shape. Torso is upright, legs are slightly bent. Arms move in line with the body (from a side view).

Ex3. Push-ups: push-ups with hands on a table. Knees, hips and shoulders are in line. Elbows do not move sideways. Only toes touch the ground.

Ex4. Leg abduction: sideway raising of a straightened leg from 0° to 45–60°. Torso is upright, standing leg slightly bent.

Ex5. Leg lunge: downward movement of the body with one leg in front and one below the body. Front knee should get to 90° position while keeping in line with ankle and not getting in front of the toe.

Ex6. Squats. Torso is upright, with whole feet on the ground.

3.3. Recorded data

A physiotherapist instructed the subjects on how to perform the exercises so that at least ten repetitions were done in a correct way and ten more incorrectly. The participants had a certain degree of freedom, e.g., which leg they used in Ex4 and Ex5. For incorrect exercising, the physiotherapist guided subjects to simulate exercise mistakes, reflecting the most common exercising problems from real rehabilitation treatments. An overview of exercising problems captured in the dataset is provided in Table 2.

Every exercise was also executed in two directions, resulting in different *views* of the subject depending on the camera. Facing the horizontal camera resulted in a *front* view for that camera and a *profile* from the other. Facing the wall between the cameras shows the subject from *half-profile* in both cameras. In Fig. 1, these directions can be seen as filled arrows. A rare direction, only used for push-ups due to the use of the table, was facing the vertical camera, shown as a dashed arrow in Fig. 1, with the views being reversed compared to the first orientation.

In summary, the recorded dataset contains the following data:

- 65 recordings (184,825 frames, 30 FPS):

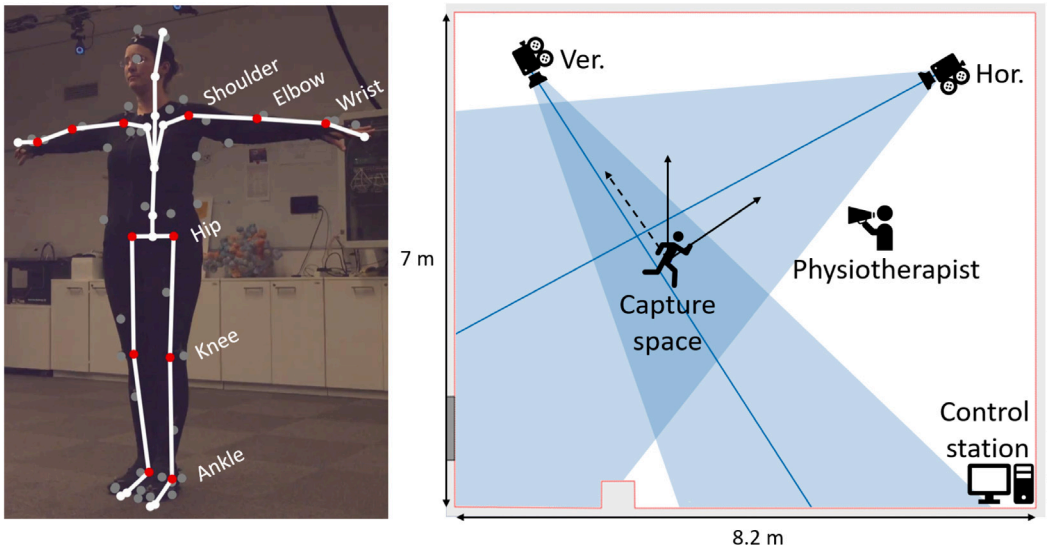


Fig. 1. The floor plan shows the placement of the RGB cameras, the capture space they form, the direction of exercising, and an example of a (cropped) vertical camera frame with the body model (white skeleton) and marker (gray points) projection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Exercising errors captured in the REHAB24-6 dataset.

Exercise	Body posture problems	Exercising problems
Ex1: arm abduction	Anteversion of pelvis; forward head posture	Moving arm not in line with the body; elevation of exercising shoulder; torso tilted to one side; insufficient arm abduction; arm not straight but bending in the elbow
Ex2: arm VW	Anteversion of pelvis; legs straight (should be slightly bent); forward head posture; forward chest posture	Arms not in line with the body; range of movement too small; elbows get behind the body
Ex3: push-ups	Heels touch the ground; anteversion of pelvis; elevated pelvis	Elbows move to the side; push-ups not deep enough; body tilted to one side; head going forward; head going backward
Ex4: leg abduction	Anteversion of pelvis; forward head posture; overly bent standing leg	Torso tilted to side; too large range of motion with hip rotation; loss of balance; knee of the standing leg moves sidewards; pelvis rotation
Ex5: leg lunge	Anteversion of pelvis	Lunge not deep enough; back or front knee not in line with ankle; loss of balance; front knee gets in front of toes; torso tilted to side; whole body moves front–back, not only up–down
Ex6: squats	Anteversion of pelvis	Ankle or knee moves inside; torso tilted to side; torso bent forward; too deep or too low squat; lifted heels; bent back

  - RGB videos from two cameras (hor./ver.), resulting in 370 K frames;
  - 3D and 2D projected positions of 41 motion capture markers;
  - 3D and 2D projected positions of 26 skeleton joints;

- Annotation of 1072 exercise repetitions:
  - Temporal segmentation (start/end frame, most between 2–5 s);
  - Binary correctness label (around 90 from each category in each exercise, except Ex3 with around 50);
  - Exercise direction (around 90 from each direction in each exercise);
  - Lighting conditions label.

All evaluations were conducted by the same experienced physical therapist, specializing in trauma and sports rehabilitation. To ensure accuracy, the assessments were cross-verified by another expert in human motion analysis. A repetition was considered correct if the participant maintained the required posture and executed the dynamic movement in a way that effectively activated the target joints and muscles.

3.4. Dataset limitations and planned extensions

Due to financial constraints, the REHAB24-6 dataset currently includes a limited number of exercises and subjects. We focused on a diverse set of *standing* exercises that are commonly used in real-world physical rehabilitation and are suitable for software-based monitoring — specifically, those with movements that are sufficiently pronounced to be detectable by computer vision methods. Actors were selected to represent patients likely to adopt new technologies for exercise monitoring; early adopters are typically younger individuals, so we emphasized a range of ages and fitness levels within that demographic.

A key limitation of the dataset is that it only includes recordings of healthy individuals. While a few participants had prior injuries or surgeries requiring rehabilitation, none had acute movement impairments at the time of recording. Although simulated movement disorders were performed under the supervision of a physical therapist, the inclusion of real patient data would provide much greater clinical relevance. To address this, we are currently collaborating with several healthcare institutions to expand the dataset with recordings from individuals undergoing actual rehabilitation.

In terms of movement variety, we aim to broaden the dataset to include exercises performed in *sitting* and *lying* positions. Exercises in the lying position, in particular, are both common in rehabilitation and known to pose challenges for HPE systems, making their inclusion especially important for evaluating real-world performance. To further increase the size of the dataset and the visual variety of data, we also plan to employ generative AI models, such as Vid2Vid [32], PyramidFlow [33], or RunwayML [34], for creating artificial variations of recorded RGB data.

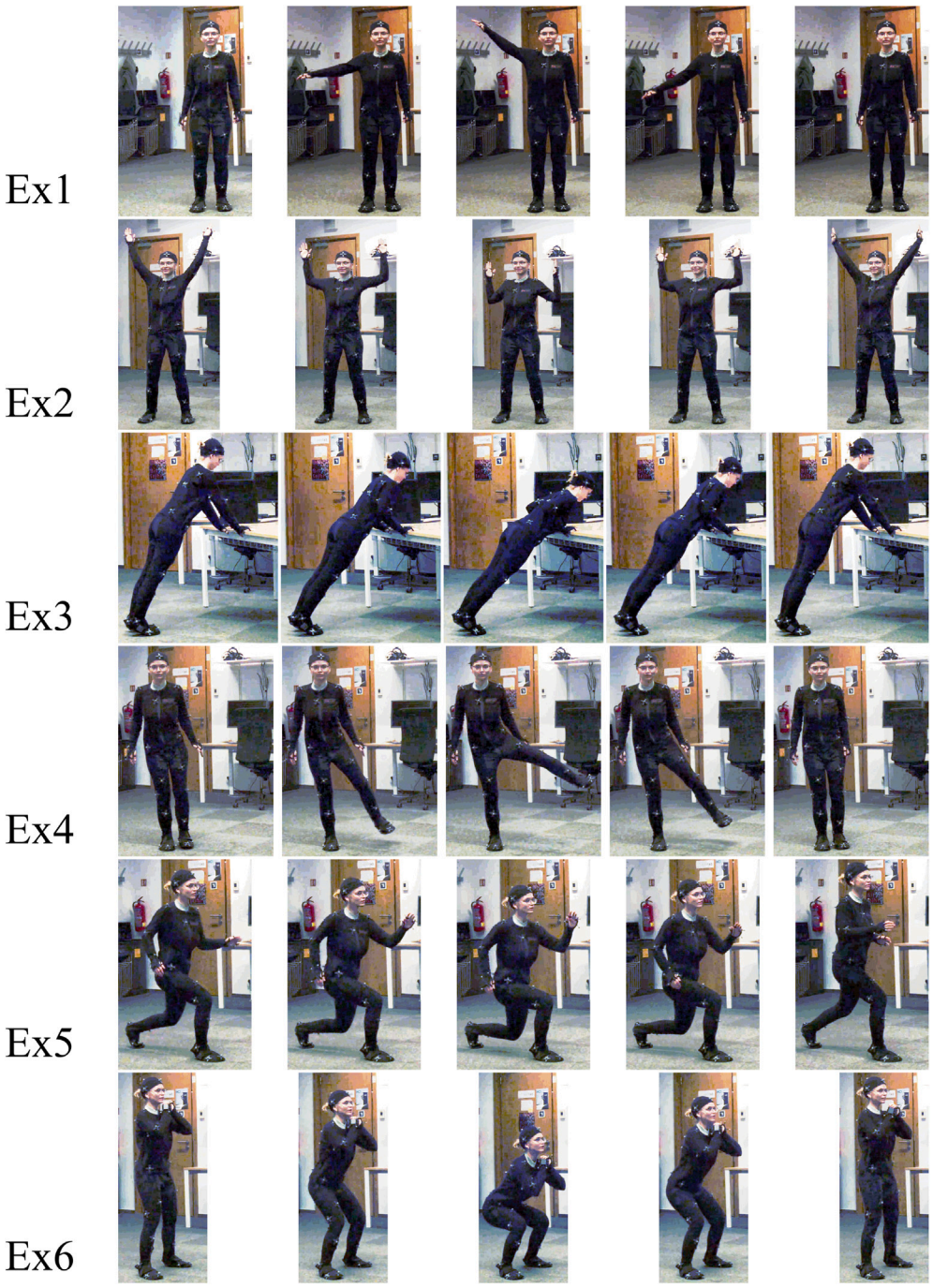


Fig. 2. Illustration of all types of exercises (Ex1–Ex6); each exercise is visualized in one row.

4. REHAB24-6 Evaluation Protocol

One of the primary objectives of the REHAB24-6 dataset is to provide a testbed for evaluating HPE models. In this section, we outline the recommended quality metrics for use with REHAB24-6, and present our solutions to several challenges that emerge when HPE techniques are evaluated on top of skeleton data that have different properties than the data used for HPE model training.

4.1. Quality metrics

A standard criterion of HPE quality is the MPJPE [18], which can be applied to both 2D and 3D skeletons to evaluate the precision of joint position estimates. However, the MPJPE measure may not directly correspond to the usefulness of HPE outputs for real-life applications, such as deciding the correctness of rehabilitation exercises, for which the mutual relationships between estimated positions of main body parts may be more important than the exact coordinates of individual joints. Therefore, we propose to complement the MPJPE metric with two measures of application usefulness: the 1-nearest neighbor (1-NN) classification accuracy and the Silhouette coefficient.

4.1.1. Joint position error

The MPJPE calculates the mean Euclidean distance between the ground-truth and estimated positions of all joints across all poses and video sequences in a given test collection. All joints are considered equally important. In its basic form, the MPJPE measure is only applicable to situations where the estimated skeleton has the same format

Fig. 3. Visualization of GT (yellow points) and predicted (red points) positions of a mostly static left-hip joint from the whole video: a high standard deviation indicates incorrect detections (left), while a low standard deviation along with high error indicates the difference between skeleton formats (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

as the ground truth. Because of this, only comparisons between HPE models that share the same skeleton format are usually performed, as discussed in Section 2.3.

The objective of REHAB24-6 testbed, however, is to allow comparison between a wide range of HPE models that may not share the same skeleton format. Therefore, we propose several techniques for accommodating different body models during HPE evaluation, which are presented in Section 4.2.

4.1.2. Discriminating exercise correctness

Physiotherapy assistance tools should mainly distinguish between correct and incorrect exercise execution, so we also look for metrics that would evaluate the discriminating power of individual HPE results. The correct exercise instructions, as provided by the physiotherapist, typically use angle thresholds rather than absolute joint positions. Accordingly, we calculate a set of main body angles as a baseline representation of each pose. In particular, we use 8 angles defined by limb roots (roughly armpits and groin) and middle points (knees, elbows). The angle is computed as a Cosine distance between the two corresponding bones. Thus, each exercise repetition (motion) is represented as an 8-dimensional time series. To determine the similarity of two motions (i.e., 8D time series) of variable lengths, we apply the Dynamic Time Warping with the Euclidean distance as the internal metric. We are aware of the fact that such multi-dimensional angle-based time series cannot capture all the discrimination patterns of a given rehabilitation exercise. Moreover, each exercise can be used for different purposes (e.g., Ex6 "squat" for rehabilitation after surgery of the hip or surgery of the knee) that need different domain-specific features. Searching for suitable features for a given combination of exercise and purpose is out of the scope of this manuscript.

Our first approach is to model the correctness assessment as a classification problem, assuming that correct exercise repetitions are similar to each other, while incorrect ones are also more similar to each other. We apply the 1-NN classifier on the angle-based representations in the leave-one-out scenario to evaluate the classification accuracy for each subject.

Next, we want the similarity between correct repetitions to be smaller than the similarity within incorrect repetitions. For this purpose, we propose to employ the Silhouette coefficient [35] and calculate the mean of these coefficients for each subject from the perspective of the correct repetitions.

4.2. Addressing key issues in HPE evaluation

As discussed in Section 2.3, the standard benchmarks rely on the evaluation dataset and the HPE detectors using the same skeleton format, and do not provide support for evaluating detectors with different formats. In contrast, the REHAB24-6 dataset aims to allow comparisons between a wide range of techniques, which requires bridging the gap between different skeleton formats. Moreover, it is necessary to deal with the possibility of detecting an incorrect person in the video, and provide a fair alignment of ground-truth and estimated skeleton sequences for evaluation of the 3D position error.

4.2.1. Unification of skeleton formats

HPE methods differ in the set of keypoints (i.e., the skeleton format) they accept as input and return as output. The 3D methods trained on the Human3.6M [18] dataset (e.g., MHFormer, STCFormer, or MotionBERT) use a 17-joint skeleton for both input and output. On the other hand, most 2D HPE methods were trained on the MS-COCO [17] dataset that uses a 17-joint skeleton different from Human3.6M. A 33-joint skeleton, a superset of the MS-COCO model, is used by MediaPipe Pose. On the contrary, the introduced REHAB24-6 dataset provides a different skeleton format with 26 joints.

In theory, the optimal solution for the skeleton format issue would create a precise mapping between individual formats. However, calculating additional keypoints would require their exact specifications, which are not usually available. Even if definitions were available, the transformation of the coordinates based on such definitions would still be problematic in 2D space. Since MHFormer and MotionBERT were trained with Human3.6M inputs for which 2D detectors are not generally available, their public implementations provide some simple conversions to this Human3.6M format to enable compatibility with 2D models. However, they rely on the trained neural networks' ability to correct the joints' exact location during the inference, so they do not guarantee a correct transformation. More advanced learning-based approaches, such as autoencoders [36], also offer an interesting option for transformations between skeleton formats. However, these models cannot ensure that the errors in the original output data are not removed, and therefore, they are also unsuitable for evaluation.

Our solution: Joint subset mapping. A more straightforward solution for the skeleton format discrepancies is to match only keypoints with close-to-identical semantics between different skeleton formats. We identify a greatest common subset of 12 keypoints present in all the used skeleton formats: 3 points for each limb (shoulders, elbows, and wrists on arms; hips, knees, and ankles on legs), as shown in red in Fig. 1. To fairly compare the accuracy of HPE methods, we consider only these 12 joints in experimental evaluations. It would be possible to consider even smaller subsets for each exercise, to fit the rehabilitation needs, but this would prevent error comparisons between the exercises.

Nevertheless, even these 12 joints might not align perfectly across different skeleton formats, manifesting in a systematic error. This error can be analyzed by comparing how the distance between the estimated and ground-truth positions of a given pair of matched joints differs over time. Based on such distances, it is trivial to calculate the mean and standard deviation for each pair of matched joints. If the standard deviation is high, it suggests that the estimation error is random due

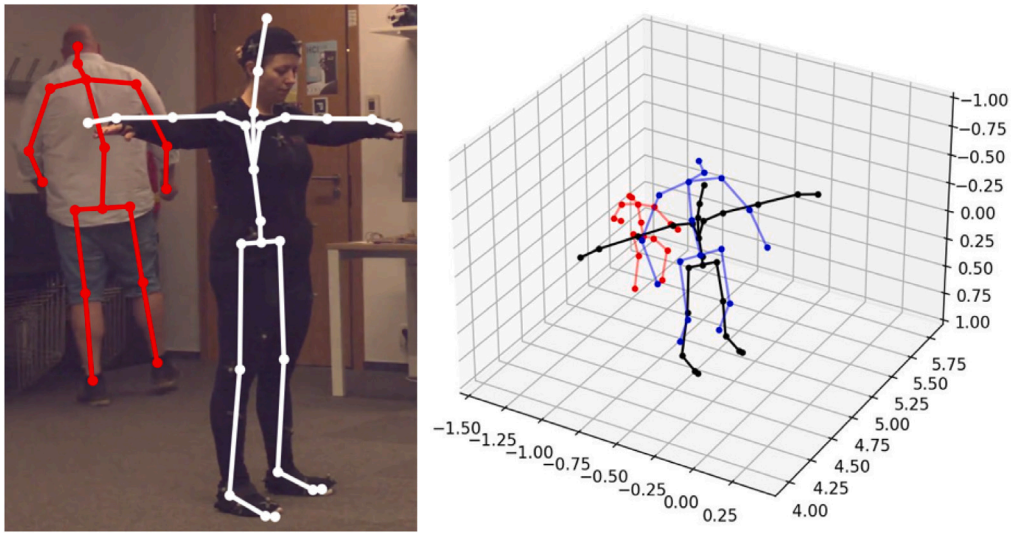


Fig. 4. Scenario in which there would be an overlap of GT bounding box (white) with the prediction (red) if arm joints were used (left). Comparison of the *BestFit* (blue) and *BBox* (red) transformations against GT (black) in such a scenario (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to incorrect pose estimation (see Fig. 3-left). On the other hand, if the standard deviation is low and the error high, this might point in the direction of a systematic shift in the estimation of the joint's location (see Fig. 3-right).

4.2.2. Separation of object detection error in 2D pose detection

2D pose detectors return joint positions in pixels, which are directly comparable to the ground truth projected to 2D during MPJPE evaluation. However, matching the skeletons only makes sense if the skeletons represent the same person. As the object detection parts of the pose detectors might detect another object in the video (e.g., the physiotherapist or a hallucination), we propose to measure these errors in object detection and joint position separately.

Our solution: Removal of frames with incorrect person detections. To determine whether the same person is represented in GT and HPE output, we apply the standard Intersect over Union with a low threshold on the bounding boxes encompassing the GT and predicted skeletons in 2D. We only use the leg and shoulder joints to minimize the influence of differing skeleton formats and not to make the bounding box too sparse, like in Fig. 4-left. Finally, we filter these incorrectly detected frames from the joint position error calculations.

4.2.3. Similarity transformations for 3D error evaluation

With the 3D ground truth in virtual centimeters and predictions in arbitrary units, the evaluation of 3D position error requires that the predictions are first suitably scaled and normalized. Traditional implementations of the 3D MPJPE use poses normalized with one joint in the origin (usually the center of the pelvis). Since this joint is inconsistent across different body models, we analyze the traditionally used *Best Fit* transformation and identify its weaknesses. To remedy them, we propose an alternative *Bounding Box* transformation.

Best fit (BestFit) transformation. To fit the predicted joints onto the ground truth joints, a standard solution is to perform the so-called Procrustes transformation, which scales, translates, and rotates the coordinates to produce the minimal MPJPE. Scaling provides the same units as ground truth, while the translation and rotation move the matched joints close to each other. However, such transformation can mask specific issues, particularly in cases where the poses do not match the direction or are not even overlapping, like in Fig. 4-right. It also might not correctly show which joints contribute most to the error.

Our solution: Bounding box (BBox) transformation. Since we have approximate camera parameters, we can rotate the ground truth to match what NNs try to predict and forgo the rotation step from the previous transformation. Similarly, we can calculate specific translations in the image axes based on the positions of the 2D and 3D bounding boxes, i.e., the distance between the 2D bounding box centers relative to their sizes should match the distance between the 3D bounding boxes. Therefore, the only steps that are taken from the BestFit manner are scaling (to have the correct units) and shifting in depth (there is no universal reference to set the predicted depth correctly).

Among other advantages, the proposed BBox transformation allows us to calculate the separate error in 2D (height and width of the image) and depth more reliably than with the BestFit transformation, providing insight into how effective the monocular 3D pose estimation is compared to 2D. However, the transformation still retains minor issues: it cannot detect distance differences in the depth (if one person occludes the other and the NN tracks the incorrect one), and it does not consider the fact that 2D and 3D predictions may not match perfectly (e.g., if the 3D model smooths the frames). Nevertheless, these issues are rare and, therefore, have minimal impact on the mean errors.

5. Fine-tuning of HPE methods

To adjust a pre-existing neural network to a different skeleton format or improve its accuracy on specific REHAB24-6 motions, we may employ *fine-tuning*, i.e., a form of transfer learning. This involves executing another training phase with the original weights as the initial state of the NN, which was previously shown to improve the fidelity of the HPE methods [37]. The original architectures typically match the neurons in specific layers with the number of joints times the number of spatial dimensions, e.g., the output layer in 2D and both input and output layers in 3D architectures. If skeleton formats are similar (i.e., having the same number of joints with similar semantics), it is sufficient to map the input/output neurons with the corresponding joints, to maximize the potential in the original weights. If more joints are to be added, the architecture must be correspondingly modified by adding new neurons, particularly to the output layer (or, additionally, to the input layer in the case of 3D estimators). For example, if a 2D detector outputs 17 joints, training it on REHAB24-6 data with our full skeleton format of 26 joints would require adding 9 neurons per dimension to the output layer.

When training 3D estimators, there is a choice of whether to use the 2D ground truth projection as the input or predictions from some (potentially also fine-tuned) 2D estimator. It is generally preferred to use the predictions as they more closely follow the real-world scenario and may improve the network's tolerance to outlying input frames (in networks that work over a sliding window).

5.1. Cross-validation for fine-tuning

One of the main challenges in machine learning is avoiding over- and under-fitting, and fine-tuning is no different. The standard solution to this, called cross-validation, is splitting the dataset into training and testing data. For our dataset, we define the following three splits.

- *Cross-subject*: It is the most common split type for pose estimation. In our case, we select two representative subjects (a tall man and a short woman) out of 10 for testing, totaling roughly 22% of the frames.
- *Cross-camera*: We select frames from the vertical camera for testing and frames from the horizontal camera for training, which reduces the training set to half of the dataset size.
- *Cross-exercise*: It is a specific rehabilitation scenario that aims at tracking the estimator's ability to handle unseen exercises. For testing, we select Ex2 as a medium-difficulty exercise, comprising around 25% of the dataset size.

6. Experimental evaluation

The collected REHAB24-6 dataset gives us a unique opportunity to gain insight into the effectiveness and efficiency of 2D and 3D pose detectors on in-the-wild rehabilitation videos and the exercise discrimination task. We also evaluate how such detectors can be fine-tuned to better accommodate the nature of the REHAB24-6 dataset.

6.1. Overview of evaluated HPE methods

The experiments included four 3D pose estimators with various configurations: MotionBERT [13] (two network sizes — Lite/Full, both with a smoothing window of 243 frames), MHFormer [14] (with a window of 351), STCFormer [15] (with a window of 81), and MediaPipe Pose [10,16] (three network sizes — Lite/Full/Heavy). We selected the models trained with the widest available smoothing windows.

Aside from MediaPipe Pose, these models demand 2D joints as input, for which we tested AlphaPose [9] (Fast Pose trained on the Halpe dataset, as recommended by MotionBERT), HRNet [8] (w48, recommended by MHFormer and STCFormer). On top of the combinations of recommended 2D and 3D models, we also tried YOLOv7 [11,12] and even the MediaPipe Pose (heavy) with the MHFormer and MotionBERT-lite, which seemed to us as the most promising ones. We mostly used existing implementations, including object detection and pose tracking, or implemented a naive approach if unavailable. However, we decided not to analyze the influence of these parts of the pose estimation pipeline.

As input for fine-tuning, we always used exercise repetitions only to avoid any interference from invalid frames. All evaluations, including the fine-tuning experiments, are performed on the 12 joint subset, as defined in Section 4.2.1.

6.2. Invalid frames

As discussed in Section 4.2.2, the object detection and pose-tracking methods can cause some frames to have completely missing or incorrect estimations, which would drastically increase the errors. Table 3 shows that the amount of invalid frames is insignificant outside of exercises Ex3 and Ex4. In particular, the drop in AlphaPose coverage was caused by the pose tracking method that provided fragmented pose sequences. Since the analysis of the effects of these methods was not the aim of this work, we filtered out the frames that were not detected and those where the predictions had low bounding box overlap with ground truth for the remaining experiments.

Table 3
The mean percentage of valid frames (in %). **Best** and second best (not rounded) results in each column are highlighted.

2D model	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6
AlphaPose [9]	99.9	97.9	80.9	92.1	96.9	97.7
HRNet [8]	99.9	**100.0**	94.9	97.0	**100.0**	**100.0**
YOLOv7 [11,12]	98.6	99.6	91.4	95.6	97.9	99.6
MPP-heavy [10,16]	**100.0**	99.9	**97.7**	**99.1**	99.9	99.9

Table 4
The 2D MPJPE (in pixels).

2D model	hor. camera						ver. camera					
	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6
AlphaPose	19	21	26	19	19	21	36	40	**29**	35	40	35
HRNet	19	**20**	**24**	**18**	**18**	**19**	**32**	**36**	31	**33**	38	**30**
YOLOv7	**18**	21	27	20	21	21	34	40	33	33	46	32
MPP-heavy	22	22	27	23	20	22	43	38	36	35	**36**	32

Table 5
The 3D MPJPE (in centimeters). The italicized rows represent an artificial scenario with an ideal 2D ground-truth (GT2D) input.

3D model	2D input	BestFit	BBox		
		3D	3D	2D	Depth
MotionBERT-full [13]	AlphaPose	9.2	14.0	6.5	11.2
MotionBERT-lite [13]	AlphaPose	9.4	13.5	6.3	10.7
	HRNet	9.2	13.6	6.0	11.1
	YOLOv7	9.2	13.6	6.2	10.9
	MPP-heavy	9.7	14.0	6.5	11.1
	GT2D	*7.7*	*11.3*	*3.8*	*10.0*
STCFormer-81 [15]	HRNet	10.9	15.8	7.7	12.3
MHFormer-351 [14]	HRNet	9.0	14.6	6.4	12.1
	YOLOv7	9.4	15.0	6.7	12.4
	MPP-heavy	9.4	15.3	7.2	12.3
	GT2D	*7.9*	*12.0*	*4.2*	*10.6*
MPP-heavy [10,16]		8.7	12.0	6.6	8.7
MPP-full [10,16]		**8.4**	**11.7**	6.6	8.4
MPP-lite [10,16]		8.6	12.3	6.8	8.9

6.3. 2D error

With the invalid frames filtered out, we can look at the errors of the 2D pose detectors. Since the 3D detectors use their outputs, we can expect that their choice might also affect the overall 3D performance. Table 4 shows HRNet achieving the best scores and MediaPipe the worst.

It is important to realize that the pixel error is relative to the image and person size: In the videos from the horizontal camera, the standing person had a height of approximately 700–900 pixels, depending on their actual height. On the other hand, in the Vertical videos, the heights ranged around 1000–1300 pixels, roughly 1.5 times more, and as a result, the errors are more significant as well.

Moreover, as discussed in Section 4.2.1, we should still consider the systematic error in estimated coordinates caused by the skeleton format variance. Fig. 3 shows that in some views, we can clearly see and quantify such error: The hip error between the MediaPipe Pose and REHAB24-6 skeleton formats is around 50 pixels in a given video, which we can rely on thanks to a low standard deviation of below 5. However, in other views, this error either differs or cannot be relied on because of the high standard deviation.

6.4. 3D and depth error

Moving from the 2D-to-3D detectors, we are particularly interested in how the added depth affects the error. The BBox transformation allows us to see the error broken down into the error in depth and the error in width and height (this should correlate with the 2D error).

Table 6
The best model on each exercise (the most appropriate camera views – with minimum error – are highlighted in bold for each exercise).

Exercise	3D model	2D model	Camera	View	3D MPJPE	
					BestFit	BBox
Ex1	MHFormer-351	YOLOv7	hor.	front	7.9	11.3
				half-profile	5.7	8.4
			ver.	half-profile	7.4	19.3
				profile	13.5	19.0
Ex2	MHFormer-351	HRNet	hor.	**front**	6.3	9.2
				half-profile	6.5	10.2
			ver.	half-profile	7.6	21.2
				profile	11.3	21.0
Ex3	MPP-heavy		hor.	**profile**	8.1	11.4
			ver.	front	10.7	14.5
Ex4	MotionBERT-lite	AlphaPose	hor.	**front**	5.5	8.1
				half-profile	8.2	11.6
			ver.	half-profile	7.2	13.1
				profile	12.0	16.1
Ex5	MotionBERT-lite	HRNet	hor.	front	11.3	13.2
				half-profile	8.8	9.7
			ver.	half-profile	12.8	14.9
				profile	10.0	12.3
Ex6	MotionBERT-lite	AlphaPose	hor.	front	10.4	12.4
				half-profile	7.3	9.0
			ver.	half-profile	10.4	12.4
				profile	12.1	14.5

This separation cannot be correctly achieved with the traditional BestFit transformation due to the rotations, but the BBox transformation also causes greater reported errors and, unlike BestFit, is not directly comparable to existing literature.

Looking at Table 5 with the errors aggregated across all videos (i.e., all exercises and cameras), the MediaPipe Pose overcame the other models thanks to its lower depth error. Still, the higher depth error than the 2D error points to the main bottleneck of monocular (single camera view) pose estimation methods.

We can also observe that changing the input 2D model changes the 3D error very little, by less than a centimeter, so the percentage of valid frames and the efficiency matter more for the correct choice. Even with the ideal ground-truth input ("GT2D" lines), the depth error remains similar to the realistic inputs.

However, we can see more significant differences in the best views for each exercise, like in Table 6. Here, MediaPipe Pose tops the other models only in a single exercise, and the best 2D models also vary. It is still worth noting that even at this level of detail, the choice of 2D detectors influences the errors less than that of the 3D model.

6.5. Fine-tuning 2D HPE methods

The 3D BestFit errors are double what the models report on their trained dataset, which could suggest overfitting to the dataset's camera views and input skeleton formats. In Table 7, we show how the pixel error of the best-performing HRNet 2D estimator can be significantly decreased by fine-tuning the corresponding NN model with 2D ground truth data of the REHAB24-6 dataset. The reported results (HRNet-FT) demonstrate that the error decreased more than two times across all exercises. This error is also low on all joints; only a right-hand fingertip joint is slightly elevated due to worse visibility in some views. This also proves the ability to add more joints to a pre-trained model.

However, we cannot rule out overfitting to the dataset's recording conditions, e.g., the same black body suit worn by all the subjects. This can be observed in Table 8 as the fine-tuned model underperforming the original model in the cross-camera scenario. Possible solutions include the use of data augmentation, particularly synthetic videos [38], which could enable greater diversity of clothing, environment, and even camera views, but is outside the scope of this paper.

Table 7
The 2D MPJPE (in pixels) on the cross-subject scenario.

2D model	hor. camera						ver. camera					
	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6
AlphaPose	22	24	25	22	20	21	40	42	32	45	44	36
HRNet	16	22	23	21	19	21	37	40	33	39	38	30
YOLOv7	22	24	27	23	22	21	38	46	36	41	46	33
MPP-heavy	27	25	26	25	20	22	47	40	38	40	39	32
HRNet-FT	7	8	12	9	8	9	18	15	15	19	16	14

Table 8
The 2D MPJPE (in pixels) on cross-exercise and cross-camera scenarios.

2D model	Cross-ex. (Ex2)		Cross-camera (ver.)					
	hor.	ver.	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6
AlphaPose	21	40	36	40	29	35	40	35
HRNet	20	36	32	36	31	33	38	30
YOLOv7	21	40	34	40	33	33	46	32
MPP-heavy	22	38	43	38	36	35	36	32
HRNet-FT	13	19	39	46	56	32	41	35

6.6. Fine-tuning 3D HPE methods

For fine-tuning of 3D estimators, we explored three options for the estimator's input: the ideal input (GT2D), and the original (HRNet) and the fine-tuned (HRNet-FT) predictions. In Table 9, we demonstrate that both prediction-based cases show a particular improvement in depth estimation, exceeding the best non-fine-tuned model (MPP-full). Nevertheless, these still fall behind the model trained on the ground truth (GT2D) with the ideal inputs (GT2D). Furthermore, the cross-camera scenario shows greater robustness of 3D models from over-fitting when learned on original predictions, as it does not access any of the potential visual clues in our dataset.

What the HRNet case does not achieve is learning new joints from our dataset. Additional experiments show that only joints inside the boundaries of the original skeleton format (e.g., spine) were learned, unlike the rest (e.g., fingertips and toes).

Table 9
The 3D MPJPE (in centimeters) of fine-tuning the 3D MHFormer detector (blank Fine-Tune cells represent the results from the original pre-trained model). The italicized rows represent an artificial scenario with an ideal 2D ground-truth evaluation input (GT2D).

Scenario	Fine-Tune		Evaluation	BestFit	BBox		
	Input	Output	Input	3D	3D	2D	Depth
Cross-subject			*GT2D*	*7.9*	*11.6*	*4.2*	*10.2*
			HRNet	9.0	14.5	6.3	12.1
	GT2D	*GT3D*	*GT2D*	*3.1*	*4.0*	*2.1*	*3.0*
	GT2D	GT3D	HRNet	7.2	9.4	5.5	6.5
	HRNet	GT3D	HRNet	5.4	7.5	5.4	4.1
	HRNet-FT	GT3D	HRNet-FT	3.9	5.1	3.1	3.4
Cross-exercise			*GT2D*	*7.5*	*13.8*	*4.4*	*12.6*
			HRNet	7.9	15.1	6.1	13.0
	GT2D	*GT3D*	*GT2D*	*5.4*	*6.8*	*3.3*	*5.2*
	GT2D	GT3D	HRNet	7.6	9.8	5.5	7.0
	HRNet	GT3D	HRNet	7.2	9.2	5.6	6.2
	HRNet-FT	GT3D	HRNet-FT	6.2	8.2	4.6	5.7
Cross-camera			*GT2D*	*9.4*	*14.8*	*5.2*	*13.1*
			HRNet	10.2	18.3	7.7	15.3
	GT2D	*GT3D*	*GT2D*	*8.4*	*13.1*	*5.7*	*10.6*
	GT2D	GT3D	HRNet	10.1	14.9	7.5	11.3
	HRNet	GT3D	HRNet	9.3	14.5	8.9	9.4
	HRNet-FT	GT3D	HRNet-FT	10.0	14.6	8.3	10.1

Table 10
1-NN accuracy [%] (all from hor. camera with front view on Ex1, profile on Ex3, and half-profile on the rest).

3D model	2D model	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6
Ground truth		97	84	**99**	**97**	**96**	93
MotionBERT-lite	AlphaPose	94	89	88	89	95	89
	HRNet	97	91	96	84	**96**	93
MHFormer-351	HRNet	98	**92**	97	91	**96**	94
	YOLOv7	**98**	90	80	89	93	**95**
MPP-heavy		95	89	97	92	95	92

Table 11
Silhouette coefficient ($\in [-1, 1]$, all from hor. camera with front view on Ex1, profile on Ex3 and half-profile on the rest).

3D model	2D model	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6
Ground truth		0.47	−0.03	**0.71**	**0.48**	**0.47**	**0.51**
MotionBERT-lite	AlphaPose	**0.54**	0.05	0.56	0.36	0.37	0.40
	HRNet	0.53	0.00	0.68	0.35	0.39	0.40
MHFormer-351	HRNet	0.47	0.09	0.66	0.38	0.44	0.47
	YOLOv7	0.47	**0.13**	0.31	0.30	0.45	0.45
MPP-heavy		0.48	−0.01	0.57	0.30	0.39	0.38

6.7. Discriminating exercise correctness

While joint position errors give us some comparison between the models, they do not show how well the models generally behave. Conversely, the 1-NN classification and Silhouette coefficients can also be calculated on ground truth, offering additional insight on the best 2D–3D model combinations (based on Table 6).

Note that the results on ground truth do not achieve 100% 1-NN accuracy as well as the Silhouette coefficient of 1.0 since the simple sequence of 8-dimensional angle features might not be the most suitable representation for the discrimination task. Similarly, in the case of Ex2, both ground truth and predictions offer worse results, further suggesting the limitations of this simplified representation. Sometimes, the results of HPE models are slightly better compared to GT — this is mainly caused by improper estimation of joint coordinates on incorrectly performed repetitions, which better contributes to the discrimination.

The 1-NN classification (Table 10) shows minimal variation between the models and GT, which is promising for exercise correctness detection but provides no clear idea of which model to select. Interestingly, the MotionBERT with AlphaPose 2D inputs is the worst, even though it is among the best in 3D MPJPE position error.

On the other hand, Silhouette coefficients (Table 11) support MHFormer as the best model across several exercises, despite its MPJPE only being the best on exercise Ex1. It is essential to mention that the coefficients varied between the subjects and, in rare cases, dropped the scores below zero (this indicates that correctly and incorrectly performed repetitions are not well discriminated).

Nevertheless, both metrics lack correlation with the 3D errors, which gives credence to the idea that positional error is not necessarily the only sufficient metric to judge the pose detector quality. Note that, unlike 1-NN, the Silhouette coefficient considers all distances and ideally expects a single cluster, so these two metrics also do not correlate.

Table 12
The runtime speed of HPE NNs in the frame-per-second rate (FPS).

Model	Type	GPU	FPS
AlphaPose	2D	✓	5.7
HRNet	2D	✓	6.2
YOLOv7	2D	✓	3.3
MotionBERT-full	3D	✓	212.6
MotionBERT-lite	3D	✓	308.1
STCFormer-81	3D	✓	20.4
MHFormer-351	3D	✓	13.2
MPP-heavy	2D & 3D	✗	5.0
MPP-full	2D & 3D	✗	14.2
MPP-lite	2D & 3D	✗	18.3

6.8. Efficiency

All models were executed on a Windows 10 machine (Intel i5-2400 CPU, 8 GB RAM) with Nvidia GeForce GTX 960 GPU with 4 GB VRAM. The only models that were tested on CPU only were the MediaPipe Pose ones, as the TensorFlow Lite 2 focuses primarily on mobile devices, including GPU support (see Table 12).

Nonetheless, MediaPipe Pose finished the computation fastest if we consider that the 3D models require the 2D inputs first. MotionBERT models achieved high speed thanks to the official implementation supporting parallelization, unlike the other networks, and theoretically, it could compete with MPP, given a fast 2D model. Overall, the 2D detectors create a bottleneck that limits the 2D-to-3D lifting strategy in efficiency.

From our experience, the 10 FPS rate constitutes the minimal requirement for real-time rehabilitation feedback. This is met by the MediaPipe Pose tool only, reaching the 18 FPS rate with the most compact model. For other 3D detectors, the underlying 2D models would have to be further optimized.

7. Conclusions

This paper gives deeper insight into the new rehabilitation REHAB24-6 dataset, which provides a variety of exercises, views, body heights, lighting conditions, and exercise mistakes. The dataset, together with the proposed BBox transformation, also serves as a testbed for evaluating 2D and 3D pose estimation methods trained on other datasets with distinct body models to get more realistic performance on in-the-wild videos. With the three newly introduced cross-validation splits (cross-subject, cross-camera, and cross-exercise), the dataset is also suitable for training and fine-tuning HPE methods.

The experiments on a selection of state-of-the-art 3D HPE models show higher joint position errors than in the standard benchmarks, partially due to the mismatch of body models but also due to the current limitations of HPE methods. We demonstrate that depth estimation is truly the main challenge of HPE, so we recommend preferring 2D pose estimators over 3D if possible. Otherwise, the choice of 2D detectors influences the errors less than that of the 3D model. The MediaPipe Pose method has the lowest 3D position error in general, but different models provide the best results on individual exercises. Fine-tuning also shows potential for surpassing MediaPipe Pose, particularly thanks to the depth estimation improvements. Nonetheless, from the efficiency point of view, MediaPipe Pose provides the best results, working fast even on CPUs.

Further experiments on the specific application of exercising correctness discrimination show no clear winner among the HPE methods either. However, they give us the intuition that 3D skeletons produced by HPE methods could work as well as the GT skeleton sequences in exercise monitoring applications.

In the future, it would be interesting to investigate the adaptability of HPE models to different body postures (e.g., a lying person) or to the specific requirements of patients with varying levels of mobility. There is also a room for exploration of smoothing methods for both 2D and 3D outputs beyond the model's native capabilities, but this is limited by the real-time requirements of our application focus. In addition, it is possible to increase the size of the dataset by employing generative AI models to create artificial alternatives to recorded RGB data.

CRediT authorship contribution statement

Andrej Černek: Writing – original draft, Methodology, Writing – review & editing, Software, Conceptualization. **Jan Sedmidubsky:** Writing – original draft, Conceptualization, Writing – review & editing, Supervision. **Petra Budikova:** Writing – original draft, Conceptualization, Writing – review & editing, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is co-financed from the state budget by the Technology Agency of the Czech Republic under the TREND Programme; project "VisioTherapy: Supporting physiotherapy treatments using computer-based movement analysis" (No. FW09020055).

Data availability

All data used in our experiments are publicly available in the Zenodo repository: https://doi.org/10.5281/zenodo.13305825.

References

[1] J. Sedmidubsky, P. Elias, P. Budikova, P. Zezula, Content-based management of human motion data: Survey and challenges, IEEE Access 9 (2021) 64241–64255.

[2] M. Jánošová, P. Budikova, J. Sedmidubsky, Personalized similarity models for evaluating rehabilitation exercises from monocular videos, in: 17th International Conference on Similarity Search and Applications (SISAP 2024), Springer, Cham, 2024, pp. 73–87, http://dx.doi.org/10.1007/978-3-031-75823-2_7.

[3] A. Černek, J. Sedmidubsky, P. Budikova, REHAB24-6: Physical therapy dataset for analyzing pose estimation methods, in: 17th International Conference on Similarity Search and Applications (SISAP 2024), Springer, Cham, 2024, pp. 18–33, http://dx.doi.org/10.1007/978-3-031-75823-2_2, Best Paper Award.

[4] R.B. Neupane, K. Li, T.F. Boka, A survey on deep 3D human pose estimation, Artif. Intell. Rev. 58 (24) (2024) 1–53, http://dx.doi.org/10.1007/s10462-024-11019-3.

[5] S. Dubey, M. Dixit, A comprehensive survey on human pose estimation approaches, Multimedia Syst. (2022) 1–29, http://dx.doi.org/10.1007/s00530-022-00980-0.

[6] T. Deng, Y. Sun, Recent advances in deterministic human motion prediction: A review, Image Vis. Comput. 143 (2024) 104926, http://dx.doi.org/10.1016/j.imavis.2024.104926, URL https://www.sciencedirect.com/science/article/pii/S0262885624000295.

[7] Z. Niu, K. Lu, J. Xue, X. Qin, J. Wang, L. Shao, From methods to applications: A review of deep 3D human motion capture, IEEE Trans. Circuits Syst. Video Technol. 34 (11) (2024) 11340–11359, http://dx.doi.org/10.1109/TCSVT.2024.3423411.

[8] K. Sun, B. Xiao, D. Liu, J. Wang, Deep High-Resolution representation learning for human pose estimation, in: CVPR, 2019, pp. 5693–5703.

[9] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, C. Lu, AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time, IEEE Trans. Pattern Anal. Mach. Intell. (2022).

[10] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann, BlazePose: On-device Real-time Body Pose tracking, 2020, arXiv:2006.10204.

[11] D. Maji, S. Nagori, M. Mathew, D. Poddar, YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss, 2022, arXiv:2204.06806.

[12] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022, arXiv:2207.02696.

[13] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, Y. Wang, MotionBERT: A unified perspective on learning human motion representations, in: IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 15085–15099.

[14] W. Li, H. Liu, H. Tang, P. Wang, L. Van Gool, MHFormer: Multi-hypothesis transformer for 3D human pose estimation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 13147–13156.

[15] Z. Tang, Z. Qiu, Y. Hao, R. Hong, T. Yao, 3D human pose estimation with spatio-temporal criss-cross attention, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 4790–4799.

[16] H. Xu, E.G. Bazavan, A. Zanfir, W.T. Freeman, R. Sukthankar, C. Sminchisescu, GHUM & GHUML: Generative 3D human shape and articulated pose models, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 6184–6193.

[17] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár, Microsoft COCO: Common objects in context, 2015, arXiv:1405.0312.

[18] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 36 (7) (2014) 1325–1339.

[19] M. Fieraru, M. Zanfir, S.-C. Pirlea, V. Olaru, C. Sminchisescu, AIFit: Automatic 3D human-interpretable feedback models for fitness training, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 9919–9928.

[20] C. Wang, Y. Li, Z. Xiong, Y. Luo, Y. Cao, Lower body rehabilitation dataset and model optimization, in: 2021 IEEE International Conference on Multimedia and Expo, ICME, 2021, pp. 1–6.

[21] Y. Tang, J. Liu, A. Liu, B. Yang, W. Dai, Y. Rao, J. Lu, J. Zhou, X. Li, FLAG3D: A 3D fitness activity dataset with language instruction, 2023, arXiv:2212.04638.

[22] S. An, Y. Li, U. Ogras, MRI: Multi-modal 3D human pose estimation dataset using mmwave, RGB-D, and inertial sensors, in: 36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022, pp. 27414–27426.

[23] Z. Zhao, S. Kiciroglu, H. Vinzant, Y. Cheng, I. Katircioglu, M. Salzmann, P. Fua, 3D pose based feedback for physical exercises, 2022, arXiv:2208.03257.

[24] M. Capecci, M.G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriù, L. Romeo, F. Verdini, The KIMORE dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation, IEEE Trans. Neural Syst. Rehabil. Eng. (TNSRE) 27 (7) (2019) 1436–1448.

[25] A. Weitz, L. Colucci, S. Primas, B. Bent, InfiniteForm: A synthetic, minimal bias dataset for fitness applications, 2021, arXiv:2110.01330.

[26] R. Aguilar-Ortega, R. Berral-Soler, I. Jiménez-Velasco, F.J. Romero-Ramírez, M. García-Marín, J. Zafra-Palma, R. Muñoz-Salinas, R. Medina-Carnicer, M.J. Marín-Jiménez, UCO physical rehabilitation: New dataset and study of human pose estimation methods on physical rehabilitation exercises, Sensors 23 (21) (2023) http://dx.doi.org/10.3390/s23218862, URL https://www.mdpi.com/1424-8220/23/21/8862.

[27] S.M. Nguyen, M. Devanne, O. Remy-Neris, M. Lempereur, A. Thepaut, A medical Low-Back pain physical rehabilitation database for human body movement analysis, in: 2024 International Joint Conference on Neural Networks, IJCNN, 2024, pp. 1–8, http://dx.doi.org/10.1109/IJCNN60899.2024.10650036.

[28] A. Vakanski, H.-p. Jun, D. Paul, R. Baker, A Data Set of human body movements for physical rehabilitation exercises, Data 3 (1) (2018) http://dx.doi.org/10.3390/data3010002, URL https://www.mdpi.com/2306-5729/3/1/2.

[29] D. Leightley, M.H. Yap, J. Coulson, Y. Barnouin, J.S. McPhee, Benchmarking human motion analysis using kinect one: An open source dataset, in: 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA, 2015, pp. 1–7, http://dx.doi.org/10.1109/APSIPA.2015.7415438.

[30] A. Miron, N. Sadawi, W. Ismail, H. Hussain, C. Grosan, IntelliRehabDS (IRDS)—A dataset of physical rehabilitation movements, Data 6 (5) (2021) http://dx.doi.org/10.3390/data6050046, URL https://www.mdpi.com/2306-5729/6/5/46.

[31] M.S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, et al., The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset, IEEE Trans. Affect. Comput. 7 (4) (2015) 435–451.

[32] A. Mallya, T.-C. Wang, K. Sapra, M.-Y. Liu, World-consistent video-to-video synthesis, in: A. Vedaldi, H. Bischof, T. Brox, J.-i. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 359–378.

[33] J. Lei, X. Hu, Y. Wang, D. Liu, PyramidFlow: High-Resolution defect contrastive localization using pyramid normalizing flow, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 14143–14152.

[34] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, A. Germanidis, Structure and content-guided video synthesis with diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7346–7356.

[35] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.

[36] I. Sárándi, A. Hermans, B. Leibe, Learning 3D human pose estimation from dozens of datasets using a Geometry-Aware autoencoder to bridge between skeleton formats, in: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2023, pp. 2956–2966.

[37] H. Joo, N. Neverova, A. Vedaldi, Exemplar Fine-Tuning for 3D human model fitting towards In-the-Wild 3D human pose estimation, 2021, arXiv:2004.03686 URL https://arxiv.org/abs/2004.03686.

[38] S. Juraev, A. Ghimire, J. Alikhanov, V. Kakani, H. Kim, Exploring human pose estimation and the usage of synthetic data for elderly fall detection in Real-World surveillance, IEEE Access 10 (2022) 94249–94261, http://dx.doi.org/10.1109/ACCESS.2022.3203174.