UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMATICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PAULO DE OLIVEIRA GUEDES

Title: PersonalRAC Personalized Few-shot Exercise Counting

Recife

2025

PAULO DE OLIVEIRA GUEDES

Title: PersonalRAC Personalized Few-shot Exercise Counting

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Veronica Teichrieb

Coorientador (a):Lucas Silva Figueiredo

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Guedes, Paulo de Oliveira.
PersonalRAC Personalized Few-shot Exercise Counting / Paulo de Oliveira Guedes. - Recife, 2024.
52f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Pós-Graduação Acadêmica em Ciência da Computação, 2024.
Orientação: Veronica Teichrieb.
Coorientação: Lucas Silva Figueiredo.
Inclui referências e apêndice.

1. Aprendizado de Máquina; 2. Aprendizado Profundo; 3. Aprendizado com Poucos exemplos; 4. Telerreabilitação. I. Teichrieb, Veronica. II. Figueiredo, Lucas Silva. III. Título.

UFPE-Biblioteca Central

Paulo de Oliveira Guedes

“PersonalRAC: Personalized Few-shot Exercise Counting”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional

Aprovada em: 20/12/2024.

BANCA EXAMINADORA

_____

Prof. Dr. Silvio de Barros Melo
Centro de Informática / UFPE

_____

Prof. Dr. Thales Miranda de Almeida Vieira
Instituto de Computação / UFAL

_____

Profa. Dra. Fátima de Lourdes dos Santos Nunes Marques
Escola de Artes, Ciências e Humanidades / USP

_____

Profa. Dra. Veronica Teichrieb
Centro de Informática / UFPE
(**orientadora)**

À minha família que sempre me incentivou e me deu condições para percorrer esse caminho. Por todo amor e confiança depositados em mim, meus sinceros agradecimentos.

À professora Verônica Teichrieb, por ser minha orientadora, sua ajuda nesse trabalho e em outras situações foram imprescindíveis, assim como meus coorientadores Lucas Silva Figueiredo e Alana E F Da Gama.

Ao Voxar Labs e todos seus integrantes por me acolherem nessa jornada de conhecimento.

Aos meus antigos amigos, que sempre estiverem junto comigo.

Aos meus colegas universitários, pelo companheirismo e pela troca de experiências que me permitiram crescer não só como profissional, mas também como pessoa.

ACKNOWLEDGEMENTS

The authors would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (process 88887.715851/2022-00) for partially funding this research.

RESUMO

A tarefa de Contagem de Ações Repetitivas (Repetitive Action Counting - RAC) é uma área de crescente interesse em pesquisas, com diversas tecnologias sendo desenvolvidas no campo. No entanto, os métodos existentes de ponta, treinados em datasets genéricos disponíveis atualmente, não são adequados para reconhecer movimentos personalizados. Essa capacidade tem o potencial de beneficiar áreas de aplicação como fisioterapia e fitness, permitindo a criação de exercícios únicos e personalizados para pacientes ou clientes e seu acompanhamento com mínimo esforço. Para abordar essa questão, introduzimos o sistema Personalized Repetition Action Count (PersonalRAC), um método inovador capaz de contar ações em cenários de dados escassos (ou seja, implementando uma abordagem de aprendizado com poucos exemplos - few-shot learning).O PersonalRAC integra Aprendizado com Poucos Exemplos (Few-shot Learning), Contagem de Ações Repetitivas e Reconhecimento de Ações baseadas em Esqueletos. Nosso sistema opera com um número mínimo de exemplos de treinamento em vídeos não cortados, identificando autonomamente os pontos de início e fim das ações, o que facilita o registro de novos exercícios de maneira prática. Para alcançar isso, utilizamos o conceito de poses salientes, anotando um subconjunto do dataset Fit3D para essa funcionalidade e propondo uma divisão para few-shot desse conjunto de dados. O sistema processa vídeos de usuários realizando exercícios e extrai informações de esqueletos utilizando o MediaPipe. Essas informações são processadas para torná-las mais confiáveis para a próxima etapa. O modelo MotionBERT, especializado em detecção de ações, analisa as informações processadas, e a saída é encaminhada para um módulo de contagem de repetições. Os resultados experimentais demonstram a eficácia e robustez do sistema em contar repetições com precisão em diversos tipos de exercícios. Nosso sistema atinge um desempenho de ponta nos cenários few-shot e few-shot multi-câmera no dataset Fit3D, obtendo respectivamente um MAE de 0,33 (melhoria de 44,07%) e um OBO de 0,64, e um MAE de 0,22 (melhoria de 53,19%) e um OBO de 0,71.

Palavras-chaves: Aprendizado de Máquina, Aprendizado Profundo, Aprendizado com Poucos exemplos, Telerreabilitação .

ABSTRACT

The task of Repetitive Action Counting (RAC) is an area of increasing research interest, with numerous technologies being developed in the field. However, existing state-of-the-art methods, trained on currently available generic datasets, are not fit for recognizing personalized movements. Such capability has the potential to benefit application fields like physiotherapy and fitness by enabling the creation of unique tailored exercises for patients or clients and tracking them with minimal effort. To address this issue, we introduce the Personalized Repetition Action Count (PersonalRAC) system, a novel method capable of counting actions in low-data scenarios (i.e., implementing a few-shot learning approach). PersonalRAC integrates Few-shot Learning, Repetitive Action Counting, and Skeleton Action Recognition. Our system operates with minimal training examples on untrimmed videos by autonomously identifying start and end points, facilitating the easy registration of new exercises. To achieve this, we leverage the concept of salient poses, annotating a subset of Fit3D dataset for this capability and proposing a few-shot division of it. The system processes videos of users performing exercises and extracting skeleton information using MediaPipe. The information is processed to make it more reliable for the next stage. The MotionBERT model for action detection analyzes this processed information, and the output passes to a repetition counting module. Experimental results demonstrate the system's effectiveness and robustness in accurately counting repetitions across various exercise types. Our system achieves state-of-the-art performance in the few-shot and few-shot multi-cam settings on the Fit3D dataset, with respectively MAE of 0.33 (44.07% improvement) and an OBO of 0.64, and 0.22 MAE (53.19% improvement) and 0.71 OBO.

Keywords: Machine Learning, Deep Learning, Few-shot Learning, Telerehabilitation.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

1.1 CONTEXTUALIZATION

Personalized Action Recognition is an approach that adapts human action recognition systems to consider the individual characteristics of each user(ZUNINO; CAVAZZA; MURINO, 2016). Technologies for that are valuable in various contexts, such as video game controllers (YIN et al., 2020), sport exercises (PATALAS-MALISZEWSKA et al., 2023; SORO et al., 2019), industrial work (TABORRI et al., 2019), and rehabilitation (PRABHU; O'CONNOR; MORAN, 2020; ZHANG; SU; HE, 2020). These systems offer advantages by tailoring action execution to individual users rather than generalizing across subjects (ZUNINO; CAVAZZA; MURINO, 2017). In specific application scenarios, this level of customization is not only beneficial but also essential.

In the field of physical therapy, there is an increasing demand for systems capable of recognizing exercises accurately, assessing their correctness, providing feedback, and verifying whether the prescribed session was actually performed. Among these features, repetition counting plays a crucial role. However, counting is not merely a quantitative measure; it implicitly verifies whether the exercise was performed correctly. A system should only count a repetition if it meets the expected execution criteria, making recognition and correctness assessment inseparable tasks. This capability is fundamental for healthcare professionals, as it enables them to monitor the execution of rehabilitation sessions and ensure that patients adhere to prescribed routines (PRABHU; O'CONNOR; MORAN, 2020).

The idea behind several existing technologies is to count actions using rules such as speed and angle together (GAMA et al., 2015; FIERARU et al., 2021); in contrast, others use machine learning techniques that directly process RGB images from available cameras (LEVY; WOLF, 2015; DWIBEDI et al., 2020a; HU et al., 2022; DWIBEDI et al., 2020b); or else, extract the user's skeleton representation for the task (SABATER et al., 2021; MEMMESHEIMER et al., 2022; ZHU et al., 2023a; MEMMESHEIMER; THEISEN; PAULUS, 2020). Despite the diversity of approaches, many systems fail to generalize to real-world scenarios, where exercise start and end timestamps are not predefined and where variations in execution must be considered.

For these technologies to work, it is generally necessary to use training data. However, collecting representative data in the field of physical therapy, for instance, poses significant challenges due to privacy concerns, need for specialized equipment, and the diversity of patient conditions. Building comprehensive datasets requires extensive time and resources, as it

involves capturing various exercises performed by individuals with varying capabilities and limitations. Such scarcity of data hampers the development of robust models capable of effectively recognizing personalized exercises (VARGHESE et al., 2023; TSIOURIS et al., 2020).

Furthermore, most existing algorithms struggle to perform well with smaller amounts of training data. Methods using one-shot or few-shot learning scenarios would ease the applicability, being desirable considering that data labeling is time-consuming and training on larger datasets is computationally expensive. However, this capability is rare among the techniques capable of learning from limited data in this domain, which may be attributed to the complexity of human motion and the high variability in exercise execution among different individuals. The lack of effective few-shot learning approaches limits the ability of professionals in the field of physical health to personalize exercises for their patients without extensive data collection and model retraining. At the same time, personalization is essential for these professionals, both for providing exercises to their clients and patients at home as well as for applications in the telerehabilitation area (which is being proven to be very effective (van Egmond et al., 2018; PERETTI et al., 2017)).

Yet another significant limitation is the inability of current systems to accurately identify the start and end timestamps of activity repetitions in untrimmed, real-world videos. While in practical applications users perform exercises without predefined start and end points, many research studies focus on previously trimmed video segments as inputs, meaning that these timestamps are known in advance. This reliance on pre-segmented data means that models may not generalize well to real-world scenarios where the beginning and end of exercises are not explicitly marked. This challenge extends to commercial products (KEMTAI, 2023; POCKETFISIO, 2024; SENCY, 2024), which often inherit the same limitations.

This research introduces the Personalized Repetition Action Count (PersonalRAC) model, which integrates few-shot learning and skeleton-based action recognition to achieve Repetitive Action Counting (RAC) in low-data scenarios. PersonalRAC is designed to function with minimal examples for model training, enabling it to autonomously identify the exercise start and end timestamps, thereby easing the registration process of new personalized exercises by healthcare professionals. Additionally, by adapting action recognition techniques, the model can accurately count repetitions of personalized exercises in untrimmed videos. Furthermore, tailoring skeleton representations within this framework significantly enhances accuracy in few-shot learning scenarios. We also present an updated state-of-the-art dataset with salient pose annotations, contributing to advancing the field.

1.2 OBJECTIVES

The objective of this research is to Develop a method for accurately counting repetitions of exercises using minimal training data, addressing the challenges associated with limited data availability, the need for personalized exercise creation, and real-time feedback.

1.3 CONTRIBUTIONS

The main contributions are summarized as follows:

- **Salient Pose Annotations Strategy for RAC:** We introduce a new approach to annotate exercises by defining salient poses that effectively capture the key moments of an exercise. This method enhances the accuracy of action recognition and repetition counting, particularly in untrimmed videos.
- **Few-Shot Personalized Repetition Action Counting:** The PersonalRAC model is designed to operate effectively in few-shot learning scenarios, making recognizing and counting repetitions of personalized exercises with minimal training data possible. This approach is particularly beneficial for tailored applications in rehabilitation and fitness, where each user may perform exercises according to their specific demand.
- **Skeleton-Based Action Recognition Modularization:** The research demonstrates how skeleton-based action recognition models, such as MotionBERT, can be adapted for Repetitive Action Counting tasks.

2 RELATED WORK

As we design a model that combines Skeleton Action Recognition to achieve Few-shot Learning for Repetitive Action Counting, we must examine the core principles of RAC, its challenges, and the methodologies proposed to overcome these issues. It is also essential to discuss the few-shot learning paradigm, exploring its potential contributions to action recognition tasks and the limitations of these techniques when applied to repetitive action counting. Additionally, it is crucial to find a dataset focused on repetition counting, similar to the proposed physiotherapy scenario, containing diversity in angles and types of exercises.

2.1 AVAILABLE DATASETS

When analyzing datasets related to Repetitive Action Counting (RAC), we consider a set of five mainly used in the field, developed by DWIBEDI et al. (2020a), HU et al., (2022), ZHANG et al., (2020a), FIERARU et al., (2021), and LAFAYETTE et al., (2022), being the first three focused on in-the-wild repetition counting and the latter two geared toward exercise counting in more controlled scenarios. Table 1 summarizes the datasets focused on Repetitive Action Count, comparing their main focus, number of classes, consistency, and angle variety. The Consistency column indicates, after our analysis, whether exercises within a class are performed uniformly. For example, in the case of the bench press, variations such as inclined or flat, using dumbbells or barbells, are considered. If a dataset combines all these variations into a single class, we labeled it as 'Not Consistent.' Conversely, datasets that distinguish these variations into separate classes or have one kind of execution are labeled as 'Consistent,' ensuring greater standardization and reliability. We consider consistency as an important attribute, in particular given our focus on physiotherapy, where exercises performed differently should be treated as distinct. On the other hand, having a larger number of classes provides a more robust assessment of RAC and different angles.

DWIBEDI et al., (2020a) introduces one of the first datasets with a large number of videos and classes, but not necessarily being human-centric, with videos such as playing the violin, bird wing flapping, and some cases of physical exercises. A subset of the latter could be made for the focus of the work, as it has a wide variety of angles. However, not being consistent between repetitions creates great difficulty for the physiotherapy scenario. ZHANG

et al., (2020a) proposed a similar approach but focused on human actions. However, according to HU et al., (2022), the dataset design still brings several limitations: "*i) no interruption to actions, either from internal or external; ii) only containing uniform action frequency in an individual video; iii) the lack of long-range videos; iv) coarse-grained ground truth annotation, etc.*". Supported by these claims, HU et al., (2022) built a new dataset focused to solve these aforementioned limitations, however with a reduced number of classes and as pointed by our analysis, a lack of consistency.

LAFAYETTE et al., (2022) proposes a dataset that closely aligns with a physiotherapy setting, offering a good diversity of individuals and also consistent execution standards. However, it includes a limited number of exercises and only two angle variations, both frontal. Finally, FIERARU et al., (2021) presents a scenario more focused on fitness exercises (closely resembling clinical physiotherapy), with a good range of different exercises, being all classes are consistent, including a significant number of individuals (11), and four camera angles—two frontal and two rear-facing. In most cases, it only contains uniform action frequency in an individual video, but there are different cadences between people and between their recordings of the same exercise.

Reference	Main focus	Number of Classes	Consistency	Angle Variety
Countix (2020a)	Every kind of repetition	100	No	Many
Ufcrep (2020a)	Human actions repetition	101	No	Many
RepCount (2022)	Exercise repetition	9	No	Many
Fit3D (2021)	Exercise repetition	26	Yes	4
LAFAYETTE (2022)	Skeleton representation	12	Yes	2

Table 1 – Comparison of datasets for Repetitive Action Counting, focusing on their primary application, number of distinct action classes, consistency in execution, and camera angle diversity.

2.2 REPETITIVE ACTION COUNTING

RAC refers to counting actions within videos (HU et al., 2022; DWIBEDI et al., 2020a), often dealing with untrimmed footage. Since untrimmed videos capture continuous scenes without isolating specific actions, this presents a challenge in identifying and segmenting relevant actions within complex sequences. Consequently, techniques typically focus on detecting and analyzing actions within the entire video context, considering multiple repetitions, different activities, and varying conditions. Additionally, algorithms designed for this task, particularly those using camera inputs, can be split in three categories: Rule-based methods, Generic RAC

models, and Exercise Focused models.

Rule-based algorithms rely on methods that consider angles and velocities (GAMA et al., 2015; FIERARU et al., 2021). For example, the approach proposed by FIERARU et al., (2021) leverages estimated 3D poses as an intermediate representation, making it robust to variations in motion quality and the number of repetitions. This method uses a two-stage algorithm: it first initializes repetition intervals by assuming a fixed period, employing auto-correlation to determine the most likely period and starting point. Then, it refines these intervals through nonlinear constraint optimization to ensure accurate alignment of repetition segments. However, these traditional approaches often face challenges when adding new exercises to the current model, as well as not coping well with camera angles that are different from what they have been optimized for and often need large datasets to work, which makes limited data scenarios quite challenging.

On the other hand, Generic RAC models (DWIBEDI et al., 2020b; ZHANG et al., 2020b; LEVY; WOLF, 2015). For example, (LEVY; WOLF, 2015) employs a convolutional neural network (CNN) to segment and count repetitive motions in videos by detecting periodic patterns. Its design allows it to generalize across a wide variety of repetitive actions, leveraging entropy-based motion analysis to identify segments corresponding to repetitions. However, while this approach is effective for generic repetition counting, it lacks the specialization needed to handle the nuances of human exercise recognition and personalization (Few-shot learning). and those specifically targeting human exercises

Finally we have the Exercise Focused models (HU et al., 2022; KIM; LEE, 2021; LI; XU, 2024; LUO et al., 2024). For example,(HU et al., 2022) (state-of-the-art) utilizes a transformer-based architecture to process spatial and temporal information from video frames, specifically tailored to exercise repetition counting. Integrating self-attention mechanisms captures complex dependencies between video frames, enabling accurate identification and counting of exercise repetitions. TransRAC also benefits from robustness to varying camera angles, thanks to its training on large, diverse datasets. However, the dependence on extensive labeled data creates a barrier to scalability, as adding new exercise classes would require a significant effort dedicated on data collection and retraining. Although some of the RAC methods discussed can be trained, none of them are designed to be learned in a training set automatically, none were designed for Few-shot learning.

2.3 FEW-SHOT LEARNING

Although we could not find RAC methods explicitly designed for low-data scenarios, several approaches in the field of Few-shot Skeleton Action Recognition are noteworthy (SABATER et al., 2021; MEMMESHEIMER et al., 2022; ZHU et al., 2023a; MEMMESHEIMER; THEISEN; PAULUS, 2020). It is important to note that these methods work on trimmed videos, which makes it impossible to use them in real-world applications for counting. The focus is on performing action recognition effectively in low-data environments, for tasks such as one-shot or few-shot action classification, where only a limited number of examples are available.

The approach proposed by MEMMESHEIMER, THEISEN, and PAULUS (2020) extends deep metric learning to multi-modal inputs, further enhancing its ability to generalize across modalities in one-shot action recognition. Despite its robust design, the framework shares the same limitation as Skeleton-DML, being constrained to recognizing actions rather than performing detailed temporal analysis necessary for RAC tasks.

The method introduced by SABATER et al. (2021) targets one-shot action recognition in challenging therapy scenarios. It employs spatio-temporal feature extraction combined with a deep neural network that learns to recognize actions from a single demonstration. While this approach excels in scenarios with well-defined action boundaries, its reliance on trimmed clips with predefined start and end points makes it unsuitable for RAC in untrimmed videos.

Similarly, the Skeleton-DML (MEMMESHEIMER et al., 2022) introduces a deep metric learning approach tailored for skeleton-based one-shot action recognition. Skeleton-DML achieves remarkable generalization from minimal examples by designing an embedding space that clusters similar actions while separating dissimilar ones. However, the method is designed for action classification and does not address the temporal segmentation for counting repetitions.

MotionBERT (ZHU et al., 2023a) offers a unified perspective on learning human motion representations. It effectively integrates spatial and temporal dynamics by employing a transformer-based architecture, achieving state-of-the-art performance in action detection tasks. It combines its architecture with Skeleton-DML techniques in low-data scenarios for better performance but does not treat untrimmed videos.

In summary, while these methods are effective for action detection, they are inherently unsuitable for RAC. Their design relies on trimmed videos, where the start and end of actions are predefined, making it impossible to partition and count repetitions accurately in untrimmed, real-world scenarios. As such, they lack the necessary capabilities to address the challenges

posed by repetitive action counting tasks.

3 PRELIMINARY WORKS

We initially explored and experimented with various techniques to develop a model that integrates Few-Shot Learning and Skeleton Action Recognition to achieve Repetitive Action Counting (RAC) in low-data scenarios. Those early investigations are detailed in the following subsections.

3.1 FEW-SHOT ACTION DETECTION AND DYNAMIC TIME WARPING

The initial focus was on action detection, a well-established area in the state of the art. Several papers address the challenge of few-shot learning in the context of skeleton-based approaches, as discussed in the previous chapter. So, the search for few-shot skeleton action detection led us to the current state-of-the-art model, MotionBERT (ZHU et al., 2023b).

We began by conducting experiments on a physiotherapy dataset (LAFAYETTE et al., 2022), concentrating on three upper limb exercises: Elbow Flexion, Shoulder Abduction, and Shoulder Flexion. The dataset included both frontal and inclined views of the exercise executions. We undertook the task of segmenting the dataset into individual repetitions for the training and testing processes. The training dataset consisted of data from one of the six participants, using only the frontal view, while the test dataset comprised the remaining individuals.

The results obtained are presented in Table 2. It can be observed that even in a few-shot learning scenario, the algorithm achieved an average accuracy close to 90%.

However, the model operates only on pre-segmented videos, that is, with well-defined start and end points. In real-world applications, it is essential to automatically identify the beginning and end of an exercise to make predictions effectively, both in scenarios where different exercises are performed consecutively and for repetition counting.

Table 2 – Performance of the MotionBERT model in recognizing selected upper limb exercises under different viewing angles.

Test	Frontal view	Rotated view
Elbow Flexion	95.34%	97.30%
Shoulder Abduction	95.51%	85.02%
Shoulder Flexion	80%	72.29%
Mean	**88.35%**	**85.02%**

To address this limitation, our first attempt at counting exercise repetitions was based on the concept of Dynamic Time Warping (DTW) (MüLLER, 2007). The DTW algorithm can align two temporal series and measure the alignment distance between those. The core idea was to measure the similarity between two-time series of the mean angular acceleration of the joints. The reference series was taken from the training individual and compared with the series from the real-time individual, as illustrated in Figure 1.

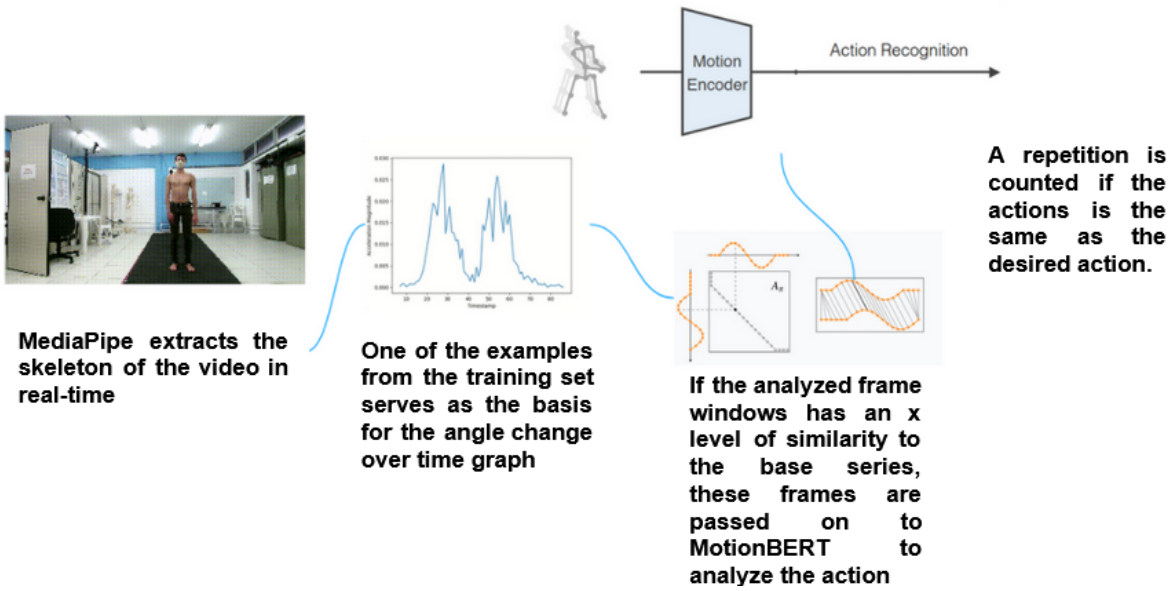

Figure 1 – Real-time repetition counting pipeline using DTW and MotionBERT.

Although, in theory, the approach could solve the defined problem, the experiments in practice revealed significant challenges in defining the windowing parameters, which rendered the system ineffective. The main technical challenge with the windowing approach lies in the dynamic and variable nature of exercise execution. Exercises performed by different individuals or even by the same individual at different times could vary significantly in speed, amplitude, and motion smoothness. Defining a fixed or adaptive window size to segment these temporal series effectively proved to be problematic. Windows that were too small failed to capture the complete motion pattern of an exercise repetition, while larger windows included irrelevant data from adjacent repetitions or transitional movements. As a result, the DTW algorithm often aligned mismatched segments, causing incorrect repetition counts or excessive computational overhead.

Moreover, the DTW algorithm introduced additional complexity when applied in this context. The alignment process focused on minimizing the overall distance between series but lacked sensitivity to the boundaries of individual repetitions. Consequently, the DTW outputs

were ambiguous, often merging consecutive repetitions into a single alignment or misaligning smaller sections within a repetition. This behavior made it difficult to extract accurate start and end points for each exercise repetition, which is crucial for real-world applications where such precision is required. These technical challenges highlighted the limitations of DTW for this specific task and underscored the need for more robust methods capable of handling untrimmed video data with varying temporal dynamics.

By addressing these challenges, our research shifted towards developing a model capable of directly recognizing the exercise start and end timestamps while counting repetitions effectively, eliminating the need for external segmentation methods like DTW.

3.2 EXPERIMENTS ON EXISTING RAC METHODS

As a consequence of exploring temporal analysis, we identified other examples related to Repetitive Action Counting (RAC) in the literature, as mentioned in the previous chapter. In particular, the PoseRAC model (YAO; CHENG; ZOU, 2023) drawn special attention, as it is designed for counting repetitions in exercises. The PoseRAC method uses the concept of salient to define key parts of exercises and other repetitive movements. Each exercise is defined as the transition between two salient poses. In general terms, once the user moves from one salient pose to the other a repetition is counted. However, the method does not inherently perform well in few-shot learning scenarios.

To address this limitation, we designed and developed a set of refinements to the technique to enhance the robustness of the skeleton representation, aiming to facilitate the learning procedure given that a limited set of training examples were provided. We tested these refinements on two subsets of the PoseRAC dataset. The first subset included three example videos for each exercise class, captured from different execution angles, referred to as the Few-Shot Multi-Cam setup. The second subset consisted of only one example video per exercise, referred to as Few-Shot. The results of these experiments can be seen in Table 3, the metrics **OBO(Off-By-One)** and **MAE (Mean Absolute Error)** are explained in Chapter 5. While analyzing the results from these experiments, we observed the potential of these modifications in the input skeleton as eligible to improve RAC results in low-data environments. Our proposed strategies for skeleton representation and metrics are carried on for further experiments and therefore are presented in detail in the next chapter will detail.

However, upon delving into the orignal available implementation of the PoseRAC technique

Method	MAE ↓	OBO ↑
Few-Shot Multi-Cam		
Default Representation (DR)	0.36	0.38
Viewpoint Invariance (VI)	**0.35**	0.39
Relational Vectors (RV)	0.40	**0.41**
Few-Shot		
Default Representation (DR)	0.58	0.26
Viewpoint Invariance (VI)	0.55	0.28
Relational Vectors (RV)	**0.51**	**0.30**

Table 3 – Performance comparison of different skeleton representation techniques in few-shot learning setups using the Poserac (YAO; CHENG; ZOU, 2023) dataset.

by the authors, we encountered inconsistencies in the metrics used to produce the results. We noticed a particularly strange behavior during the evaluation step, where the PoseRAC model considered, among a large set of exercise classes, the *best exercise class* with results closest to the ground truth, even if it was *not the actual target exercise* being performed. For example, in a squat exercise where the annotated number of repetitions was 12, consider that the model predicted 12 repetitions for elbow flexion and 5 for squats; then the model would take the value from elbow flexion (12) as the correct one to represent results, rather then using the actual squat result (5); other examples can be observed in Table 4. Such inconsistency was reviewed carefully and led to the discarding of the method.

Exercise Repetitions	Video 1: Pushup	Video 2: Pushup	Video 3: Pushup
Ground Truth	10	17	5
Pushup	0	12	5
Bench Press	8	16	0
Squat	11	2	0

Table 4 – PoseRAC evaluation inconsistency where the model selected the closest prediction to the ground truth, rather than the actual target exercise.

We considered this behavior as not an accurate evaluation method given it hinders real-world application. After modifying the code to count repetitions only for the specific target exercise, the Mean Absolute Error (MAE) and the Off-By-One (OBO) errors were approximately 0.85 and 0.1 in the Original Few-Shot Multi-Cam configuration, respectively. The obtained results showed to be unacceptably high for practical use. However, the experience with PoseRAC triggered the idea of using salient poses and implementing them with proper recognition methods, such as the MotionBERT (ZHU et al., 2023a).

4 METHODOLOGY

This chapter outlines the methodology for developing a Repetitive Action Counting (RAC) model based on skeleton action recognition and few-shot learning. By identifying key salient poses within each exercise, our approach ensures accurate tracking of repetitions with minimal data. We detail the process of annotating these poses, the techniques used to extract and process skeleton data, and the few-shot learning strategy applied to handle low-data scenarios. Additionally, we describe the integration of these skeleton representations with the action recognition module and the design of the repetition counting module. Each component is optimized for flexibility and accuracy across various exercises and camera viewpoints.

4.1 SALIENT POSES DATASET ANNOTATION

Our approach focuses on developing an action detector to count exercise repetitions by dividing each exercise into two salient poses, as shown in Figure 2 with the Fit3D dataset (FIERARU et al., 2021). This division allows for effective tracking of temporal progression. Understanding these salient poses is essential for distinguishing between different phases of an exercise and accurately counting repetitions. Salient poses can be understood as the key positions that compose the exercise. For example, in a side lateral raise, salient pose 1 is defined as when the arms are extended towards the sides of the body, almost in the anatomical position, and salient 2 is when the arms are raised above the shoulder abduction movement. We employed techniques to identify these distinct poses within each exercise.

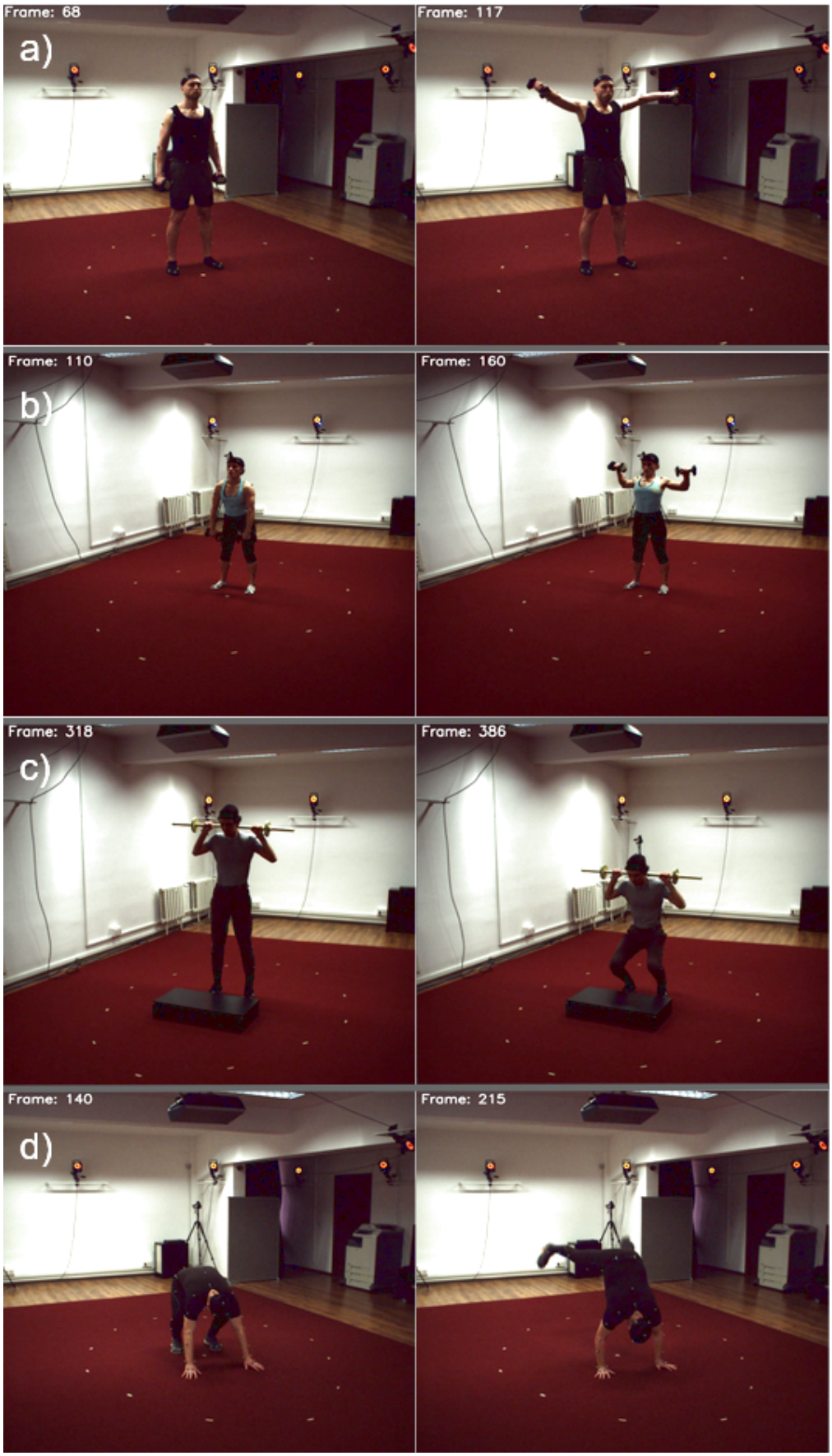

Figure 2 – Annotations of salient poses from the Fit3D dataset, in the left pose salient 1 and in the right pose salient 2: side lateral raise **(a)**, dumbbell high pulls **(b)**, squat **(c)**, and mule kick **(d)**.

We validate our system using the Fit3D (FIERARU et al., 2021) dataset. The adheres to a clearly defined protocol during filming and execution. Despite variations in execution due to individual anatomical differences and other factors, the dataset maintains a high level of consistency. Additionally, the extensive variety of exercises included in the dataset significantly contributes to a more comprehensive understanding of our algorithm's performance across different exercises.

The Fit3D dataset comprises a training set and a test set with eight subjects and three subjects performed by instructors and trainees. The videos are recorded from four different viewpoints, as shown in Figure 3, with a resolution of 900x900 and a frame rate of 50 fps. The test set includes one random camera viewpoint per sequence.

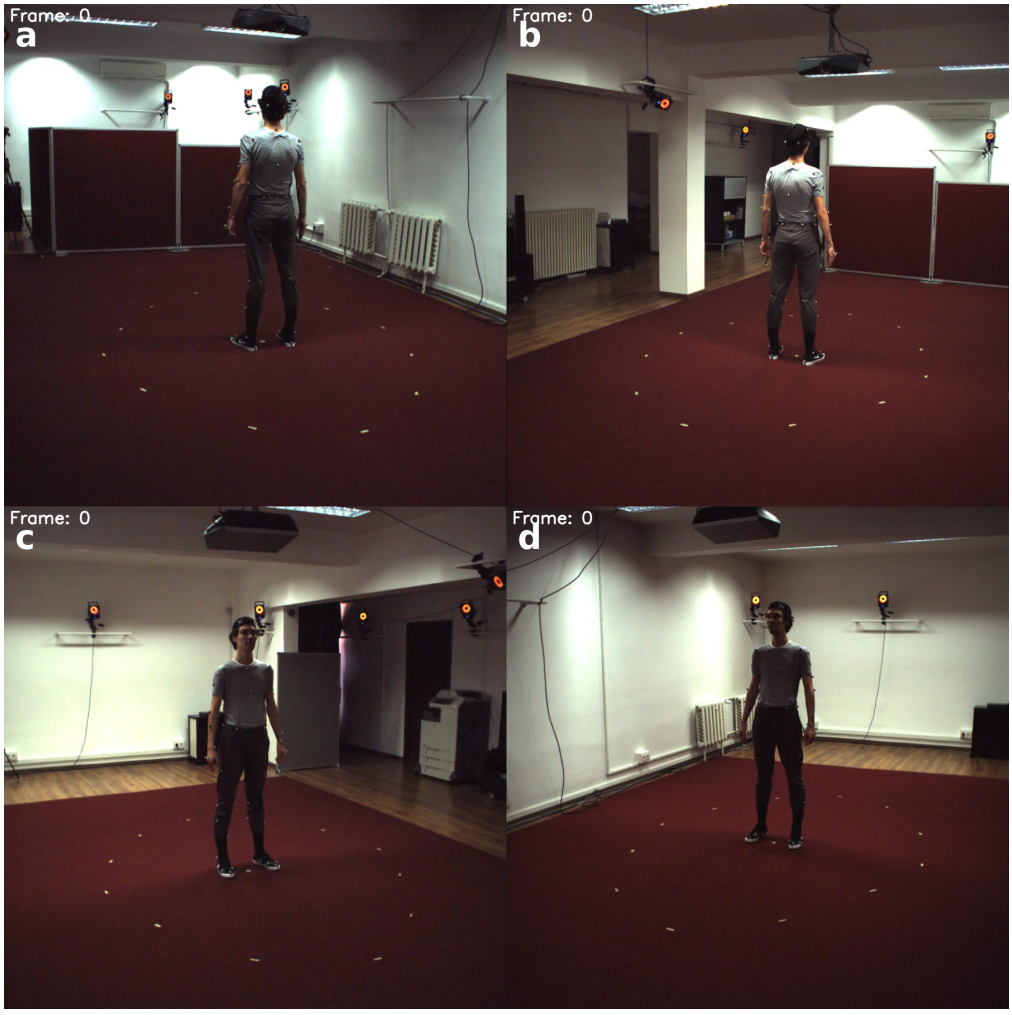

Figure 3 – Illustration of the four camera viewpoints captured in the dataset: (a), (b), (c), and (d). These views provide varied perspectives for exercise execution analysis.

The videos are categorized into Warmups, Barbell Exercises, Dumbbell Exercises, and Equipment-Free Exercises, comprising 47 exercises. For PersonalRAC, 22 exercises were selected based on the following criteria:

- Warmup exercises were skipped
- Sequences of salient poses should not present major complexity regarding the ordering of repetitions, as is the case of unilateral exercises, for example (1, 2-left side, 1, 2-right side, 1, ...)

One way of classifying the 22 exercises is to classify them as upper limb exercises, lower limb exercises, and full body:

- **Upper limb**: band pull apart, barbell row, barbell shrug, diamond pushup, dumbbell biceps curls, dumbbell hammer curls, dumbbell high pulls, dumbbell overhead shoulder press, one arm row, overhead trap raises, pushup, side lateral raise, and w raise
- **Lower limb**: dumbbell reverse lunge, and squat
- **Full body**: barbell dead row, deadlift, mule kick, overhead extension thruster, and standing ab twists

4.2 PERSONALRAC

In this dissertation we introduce the PersonalRAC, a developed technique that uses video from a single RGB camera as input to recognize repetitions of a desired personalized action. The model brings a new approach integrating skeleton action recognition using a few-shot learning method to achieve repetition counting for customized motions, gestures, and exercises. As illustrated in Figure 4, the model comprises five key steps. Initially, skeleton salient pose information is extracted using MediaPipe BlazePose (BAZAREVSKY et al., 2020). Subsequently, the extracted skeleton data undergoes processing in a dedicated module to explore other representations to enhance its quality and consistency. The processed skeletons are then analyzed by the action detection module powered by MotionBert (ZHU et al., 2023a), the current state-of-the-art in few-shot skeleton action detection. Finally, the predictions from the action detector are passed to the repetition counting module, which accurately counts the repetitions performed. These steps will be detailed below.

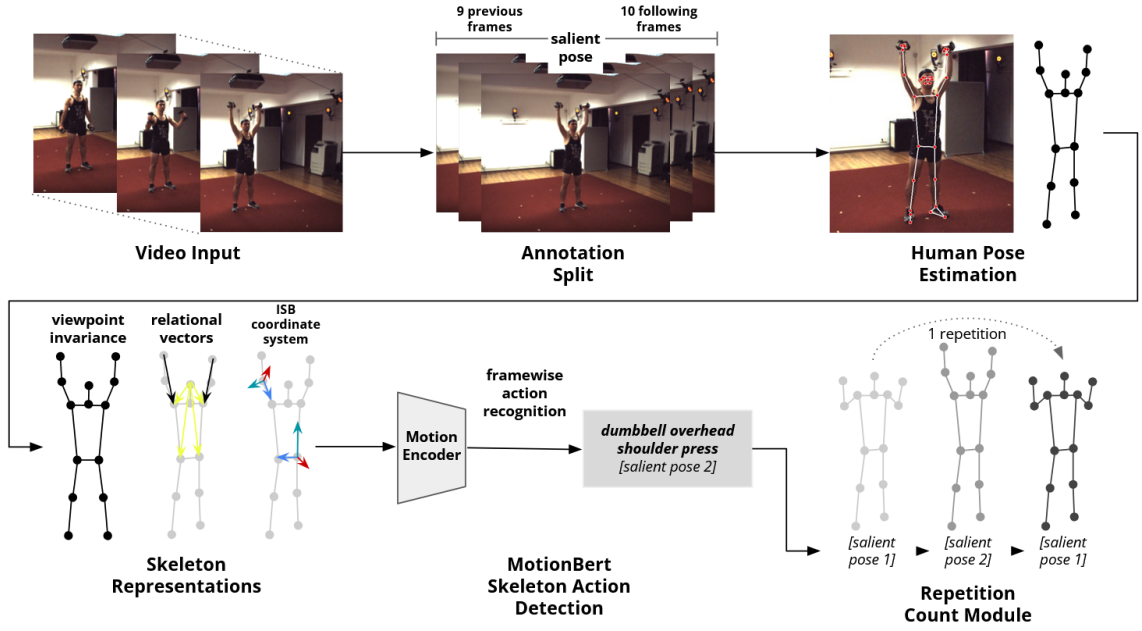

Figure 4 – Overview of the PersonalRAC model. The model begins with video input, where the exercise is performed. The video is then split into segments around salient poses, with 10 frames before and after each pose. Human pose estimation is applied to these segments to generate skeleton representations. These representations are processed through different methods (viewpoint invariance, relational vectors, and ISB coordinate model) before being passed to the MotionBERT model for skeleton action detection. Finally, the detected actions are fed into the Repetition Count Module, which tracks the number of repetitions based on the sequential triggering of the salient poses.

4.2.1 Few-shot Strategy

We implemented a few-shot division approach to evaluate our technique trained in a low-data scenario. We randomly selected one individual from the Fit3D dataset, specifically "s08." Consequently, the available data consists of only one person with limited repetitions per exercise. From this individual, we created two protocols: one called "**Few-Shot Multi-Cam**" which includes data from all four camera angles, and another simply named "**Few-Shot**", which uses only camera angle (c) from the Figure 3.

4.2.2 Skeleton Representations

After capturing the skeleton using MediaPipe, we apply processing methods to enhance the robustness of the skeleton data. All methods process skeletons in three dimensions and convert them to two dimensions to match the format required by MotionBERT. The skeleton data processed for train and test is a set of 20 frames composed by the salient pose frame, the 9 frames that precedes it and the 10 frames that follow it, as seen in Figure 4. The training

and test data consist of a window of 2 skeleton frames.

4.2.2.1 Default Representation (DR)

The extracted skeleton by the human pose estimation step (Figure 4) is represented by the joint positions within the image coordinates. Each joint position is normalized and stores the 2D coordinates x and y, as shown in Figure 5. We consider this as the default representation (DR) of the tracked user.

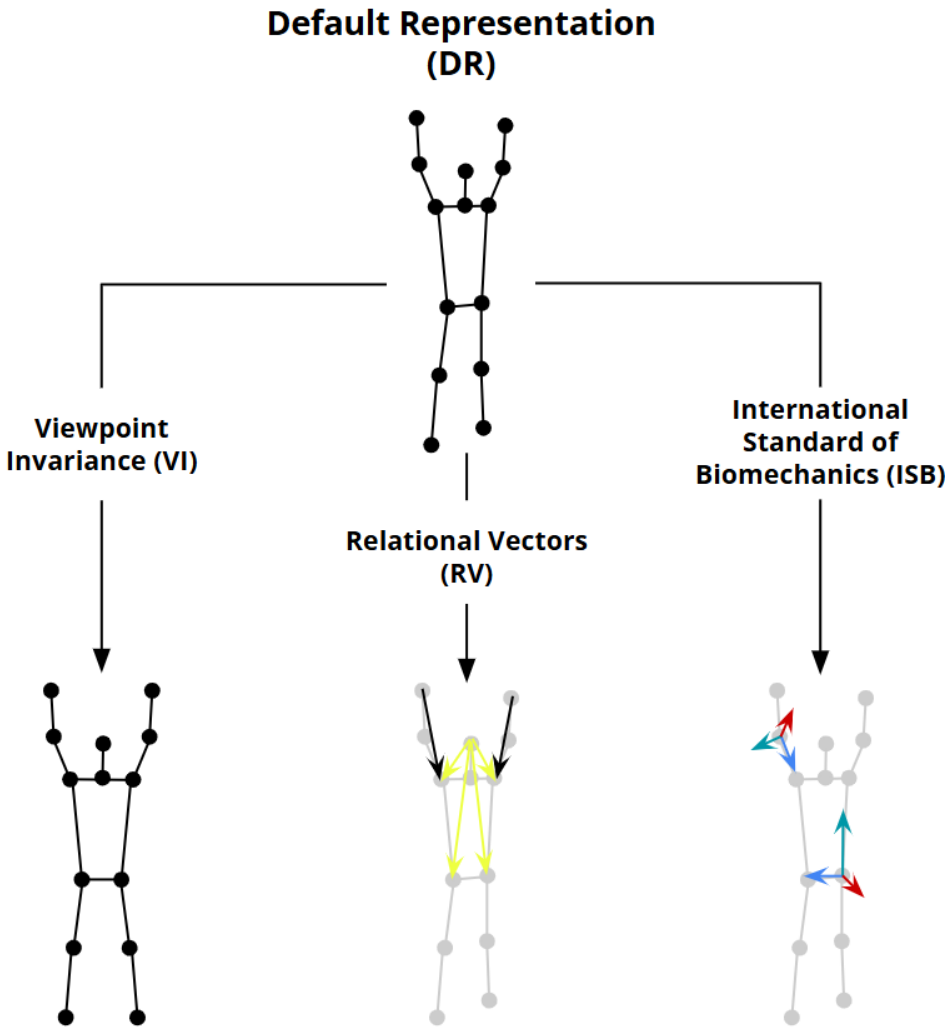

Figure 5 – Comparison of different skeleton representation strategies. The representations include Default Representation (DR), Viewpoint Invariance (VI), and Relational Vectors (RV).

To be properly used by the RAC procedure the DR representation requires a conversion of the MediaPipe skeleton format to the format used by the H36M that involves both direct mappings and calculated (indirect) keypoints to ensure compatibility with the 17-joint H36M format. Direct conversions include joints with clear correspondences in MediaPipe, such as

the nose, shoulders, elbows, wrists, hips, knees, and ankles. For example, the right and left shoulder in H36M are directly mapped to the respective shoulder keypoints in MediaPipe. Indirect conversions are required for keypoints not explicitly defined in MediaPipe, such as the mid-hip, torso, and head. The mid-hip is calculated as the midpoint between the left and right hips, the torso as the average of the shoulders and hips, and the head as the midpoint between the eyes. Then, by using the converted skeleton as a base, it is possible to create all other proposed representations (Figure 5).

4.2.2.2 Relational Vectors (RV)

In this representation, the skeleton data is extracted using image coordinates, and relational vectors are generated by connecting each keypoint to its first and second neighbors. This approach creates a richer set of vectors that capture the spatial relationships between body parts, providing more detailed information about the body's movement.

For instance, the neck is connected to its first neighbors, which include the shoulders, nose, and belly, as well as its second neighbors, the elbows and the hips, as shown in yellow in Figure 5. This method expands the number of vectors from 17 to 36, offering a more comprehensive view of the body's joint interactions. By including both close and slightly more distant relationships between keypoints, the model should better understand the overall structure and dynamics of movement.

4.2.2.3 Viewpoint Invariance (VI)

For this representation, the skeleton is extracted using real-world coordinates. The intention is to achieve viewpoint invariance by considering that users should always have their hips perfectly oriented to face the camera frontally. Hence, we calculate a transformation to rotate all joints accordingly, considering the hip as the base of the new coordinate system for the model. Using this strategy, we rotate the entire skeleton to a canonical pose to achieve viewpoint invariance, as illustrated in Figure 5 where it can be perceived that the VI joints and bones are slightly rotated if compared to the DR original skeleton.

The VI process is based on matrix rotation. We calculate a matrix that will be used to reorient the skeleton based on a set of vectors to be considered as the new coordinate system. The matrix is derived by establishing the primary axes of the body: the x-axis, defined by the

vector that connects the left and right hips, and the y-axis, defined by the vector connecting the neck to the center of the hips. The z-axis is then determined perpendicular to the x and y axes. These vectors (used as axes) are further normalized to ensure they form an orthonormal base.

Once the rotation matrix is established, it is applied to each skeleton joint by rotating each joint's coordinates relative to the center of the hips. This approach ensures the entire skeleton is consistently oriented, with the hips facing forward. Finally, the rotated skeleton is converted into 2D coordinates, which are stored for further analysis.

4.2.2.4 International Standard of Biomechanics (ISB)

In physiotherapy, it is usual to define the coordinates of the user's joints and bones in terms of the International Standard of Biomechanics (ISB). In summary, the ISB proposes to define the user skeleton as a tree graph using the hip center as the root and the following joints accordingly placed in the tree hierarchy.

Navigating the skeleton as a tree defines each new joint in the following level using its parent node as a base coordinate system, as illustrated in Figure 5. Given that the ISB defines each joint as a vector using the previous body part as the base of its coordinate system, it does not consider global relationships (e.g., if the user's hand is pointing upward or downward), but local relationships (e.g., if the hand is pointing in the same direction as the elbow).

For the ISB coordinate model representation, the skeleton is extracted using real-world coordinates. This method applies the ISB technique as described in DAGAMA2019396. The skeleton is adjusted to have the same number of points as in the H36m format.

4.2.3 Skeleton Action Detection

We trained a model for each exercise class to allow for the easy addition of new exercises as needed. We used the MotionBERT (ZHU et al., 2023b) action recognition module, with the pre-trained action recognition network (*x-sub*, *ft*) along with the linear probing followed by the fine-tuning technique proposed in (KUMAR et al., 2022). The network is trained to identify only between the two salient poses from the training set. The evaluation metrics are carried out in the following module.

4.2.4 Repetitive Counting Module

The PersonalRAC model approach to Repetitive Action Counting (RAC) leverages the adaptability of skeleton-based action detection algorithms to enhance its ability to count exercise repetitions accurately. Although MotionBERT is used in our implementation, the underlying principles and architecture of PersonalRAC can be extended to other skeleton-based action detection models.

Our enhanced method introduces an optimization process performed on the training set, after the model is trained, uses a grid search to find counting parameters specific to each exercise. This involves optimizing two separate certainty thresholds—thresholds1 for salient pose 1 and thresholds for salient pose 2—and applying offsets (offset1 and offset2) to the model's output probabilities for these poses.

The grid search process involves optimizing the model parameters for each exercise, enabling it to adapt to variations in movement patterns and camera angles, thereby improving counting accuracy. The process includes the following steps:

- **Parameter Range Definition**: We define reasonable ranges and step sizes for each parameter based on prior knowledge and preliminary experiments. Offsets may range from -0.2 to 0.2 with increments of 0.1, while thresholds may range from 0 to 0.8 with similar increments. This creates a comprehensive grid of possible parameter combinations.
- **Exhaustive Evaluation**: We evaluate every possible combination of parameters within the defined ranges for each exercise. This involves running the counting algorithm on the training set and recording the counting performance for each parameter set.
- **Parameter Selection**: The optimal parameters for each exercise are those that result in the highest OBO and, subsequently, the lowest MAE.

The videos for network evaluation are considered in a real-use context, as shown in Figure 6. After passing through a sigmoid function and applying the optimized parameters, the model outputs are analyzed every two frames. The counting algorithm proceeds as follows:

- **Initiation of Counting**: The counting process begins when the adjusted probability of salient pose 1 (after applying offset1) exceeds the threshold thresholds1.

- **Transition Detection**: The model then looks for salient pose 2, requiring that its adjusted probability (after applying offset2) exceeds the threshold thresholds2 and is higher than the adjusted probability of salient pose 1.
- **Counting a Repetition**: Subsequently, salient pose 1 must be identified again with its adjusted probability exceeding thresholds1 and higher than that of salient pose 2. When this condition is met, one repetition count is added.

By optimizing these parameters for each exercise, PersonalRAC can more accurately detect the subtle differences between poses in various exercises. This tailored approach enhances the model's robustness and accuracy in counting repetitions across different exercises and real-world scenarios.

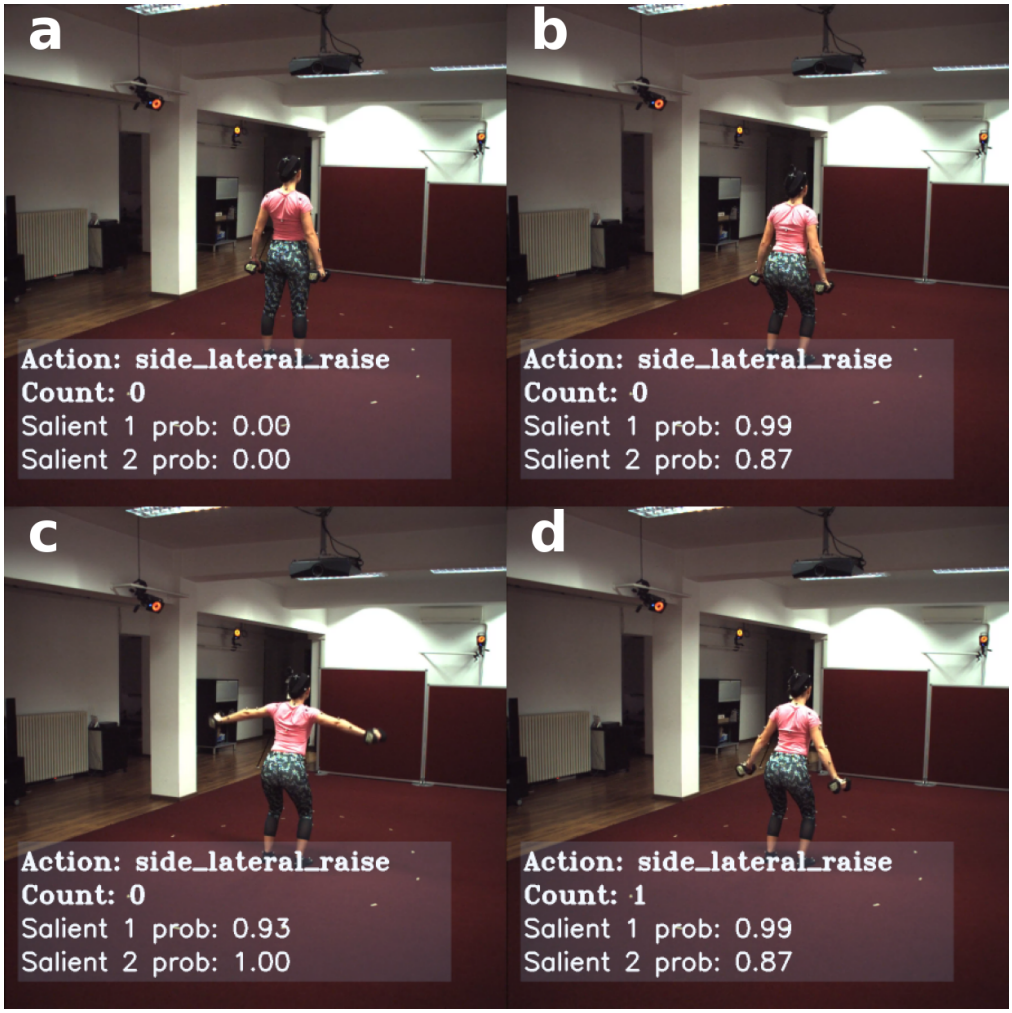

Figure 6 – Example of a real-time evaluation: The exercise begins in (a), progresses to (b) where the first salient pose (Salient 1) is detected, then moves to (c) where Salient 2 is detected, and finally returns to (d) where Salient 1 is detected again, incrementing the exercise count.

5 EXPERIMENTS AND RESULTS

To evaluate the effectiveness of our approach, we conducted three distinct experiments using the dataset, the Fit3D with our salient pose notation. The first experiment utilized the Full Dataset for training, the second was the Few-Shot Multi-Cam, and the final was the Few-Shot. Each experiment was conducted using our proposed technique and compared against the state-of-the-art method, TransRAC (HU et al., 2022).

5.1 SETUP

The training process followed the linear probing and fine-tuning strategy outlined in (KUMAR et al., 2022). Initially, the model was trained for 25 epochs in the linear probing phase, with the best result carried forward to the fine-tuning phase, which also ran for 25 epochs.

All experiments were performed on an Nvidia RTX 3080 Ti GPU. In line with MotionBERT's training guidelines, we used a learning rate of 0.0001 for both the fully connected layers and the backbone network. Additionally, a learning rate decay of 0.99 and a dropout rate of 0.5 were applied to prevent overfitting.

5.2 METRICS

The metrics used for Repetitive Action Counting are **Mean Absolute Error (MAE)** and **Off-By-One (OBO)**.

- **MAE (Mean Absolute Error)** is defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}$$

  Where $\hat{y}_i$ is the predicted value and y_i is the true value. The metric represents the normalized absolute error between y_i and $\hat{y}_i$. While it includes the term "absolute," it is subsequently normalized by dividing by the number of misrecognized repetitions.

- **OBO (Off-By-One)** is a metric that counts the number of predictions that are off by exactly one repetition, divided by the total number of predictions:

$$\text{OBO} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(|\hat{y}_i - y_i| \leq 1)$$

  Where $\mathbf{1}$ is the indicator function that returns 1 if the condition is true and 0 otherwise.

5.3 BASELINE METHOD

When discussing Repetition Action Counting (RAC), particularly in the context of exercises, TransRAC (HU et al., 2022) stands out as the state-of-the-art method. Therefore, it was selected as the baseline for testing to evaluate its performance on a new dataset and in low-data scenarios, as we could not find a RAC technique designed explicitly for few-shot learning.

To conduct the experiments, we first had to adapt our annotation method to work with TransRAC. The annotation in TransRAC involves marking the start and end frames of each repetition, noting any pauses that may occur between each one. Since there are no pauses in the Fit3D dataset, the only conversion required was to use salient pose 1 as the beginning and the next occurrence of pose 1 as the end of a repetition.

5.4 RESULTS

The results for the three scenarios (Full Dataset, Few-Shot Multi-Cam, and Few-Shot) (Chapter 4) are presented in Table 5. TransRAC emerges as the top-performing model on the Full Dataset, with the lowest MAE of 0.14 and an OBO score of 1.00, indicating high consistency and robustness across the entire dataset. The PersonalRAC variants, specifically those using Default Representation (DR) and Viewpoint Invariance (VI), show reasonable performance, although they fall short of TransRAC. The PersonalRAC methods, such as RV (Relational Vectors) and ISB (Individual Skeleton-Based), show a significant increase in MAE, pointing toward potential limitations in scaling these methods when more data is available.

In the Few-Shot Multi-Cam setting, the PersonalRAC method employing Viewpoint Invariance (VI) displays superior performance, achieving an MAE of 0.22 and an OBO of 0.71, surpassing TransRAC and all other PersonalRAC variants. We hypothesize that this performance highlights the applicability of viewpoint invariance techniques in scenarios with limited data from multiple camera perspectives, where a robust skeleton representation can substantially improve the model's accuracy.

Finally, in the Few-Shot scenario with single-camera data, the PersonalRAC model using Relational Vectors (RV) achieves the best results, with an MAE of 0.33 and an OBO score of 0.64. We would argue that this skeleton representation method can be advantageous in handling data scarcity, given that it creates additional relational vectors to encode relationships between detached joints, easing the task of learning some salient poses with fewer examples.

Method	MAE ↓	OBO ↑
Full Dataset		
AIFit	0.25	0.86
TransRAC	**0.14**	**1.00**
PersonalRAC (DR)	0.32	0.65
PersonalRAC (VI)	0.28	0.62
PersonalRAC (ISB)	0.39	0.58
PersonalRAC (RV)	0.35	0.64
Few-Shot Multi-Cam		
TransRAC	0.47	0.00
PersonalRAC (DR)	0.29	0.68
PersonalRAC (VI)	**0.22**	**0.71**
PersonalRAC (ISB)	0.30	0.65
PersonalRAC (RV)	0.32	0.62
Few-Shot		
TransRAC	0.59	0.00
PersonalRAC (DR)	0.52	0.44
PersonalRAC (VI)	0.50	0.48
PersonalRAC (ISB)	0.62	0.50
PersonalRAC (RV)	**0.33**	**0.64**

Table 5 – Performance comparison of the RAC methods for the three different dataset scenarios: **Full Dataset**, **Few-Shot Multi-Cam**, and **Few-Shot**. The metrics used are Mean Absolute Error (MAE, lower is better) and Off-By-One (OBO, higher is better).

5.4.1 Exercises Analysis

To better understand how classification occurs per exercise, we can examine in detail the chart of the best method for MAE of the Full Dataset (VI), Figure 7, for Few-Shot Multi-Cam (VI), Figure 8, and for Few-Shot (VI), Figure 9. In these Figures, it is possible to observe the OBO and the MAE for each exercise. Each scenario offers unique insights into how PersonalRAC handles different data environments, with notable performance and method effectiveness differences. Figures for each dataset scenario's skeleton representations are available in the Appendix A.

Examining the performance across all three data scenarios reveals a pattern of similarity for certain exercises, highlighting PersonalRAC's consistency in handling specific types of movements. Exercises that involve simpler, repetitive upper-body movements, such as "dumbbell scaptions," "side lateral raise," and "overhead trap raises," consistently show low MAE and

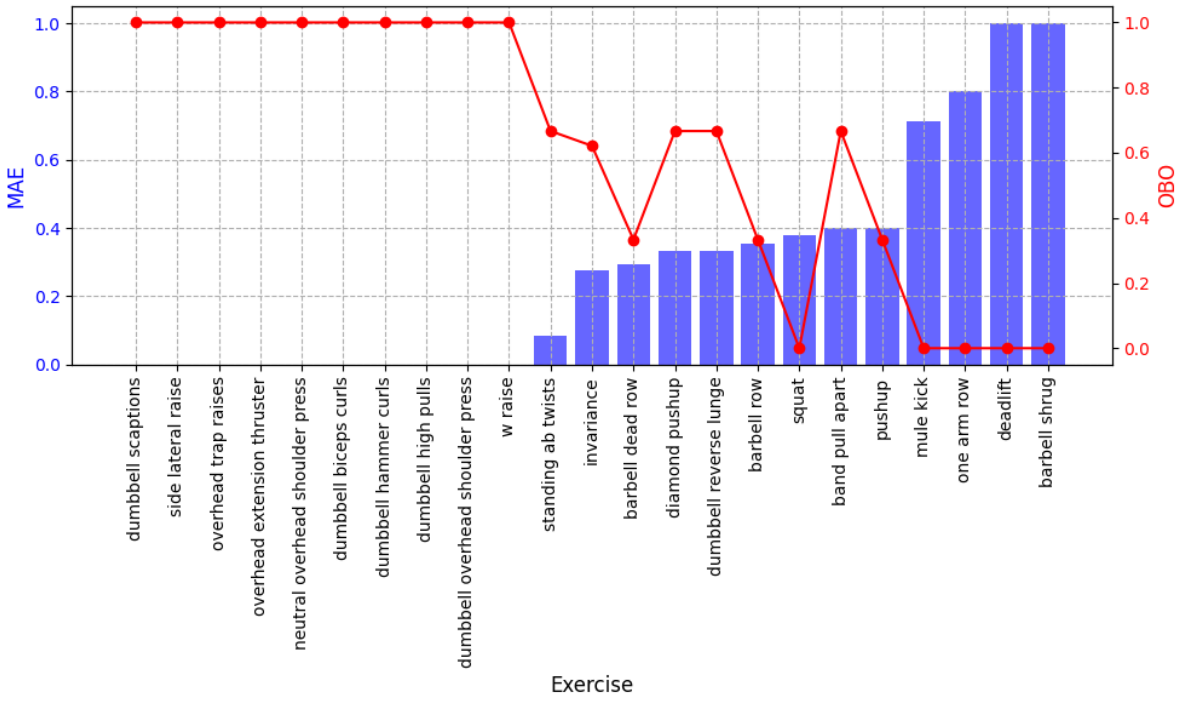

Figure 7 – Breakdown of MAE and OBO per exercise for the best-performing method (VI) on the Full Dataset split.

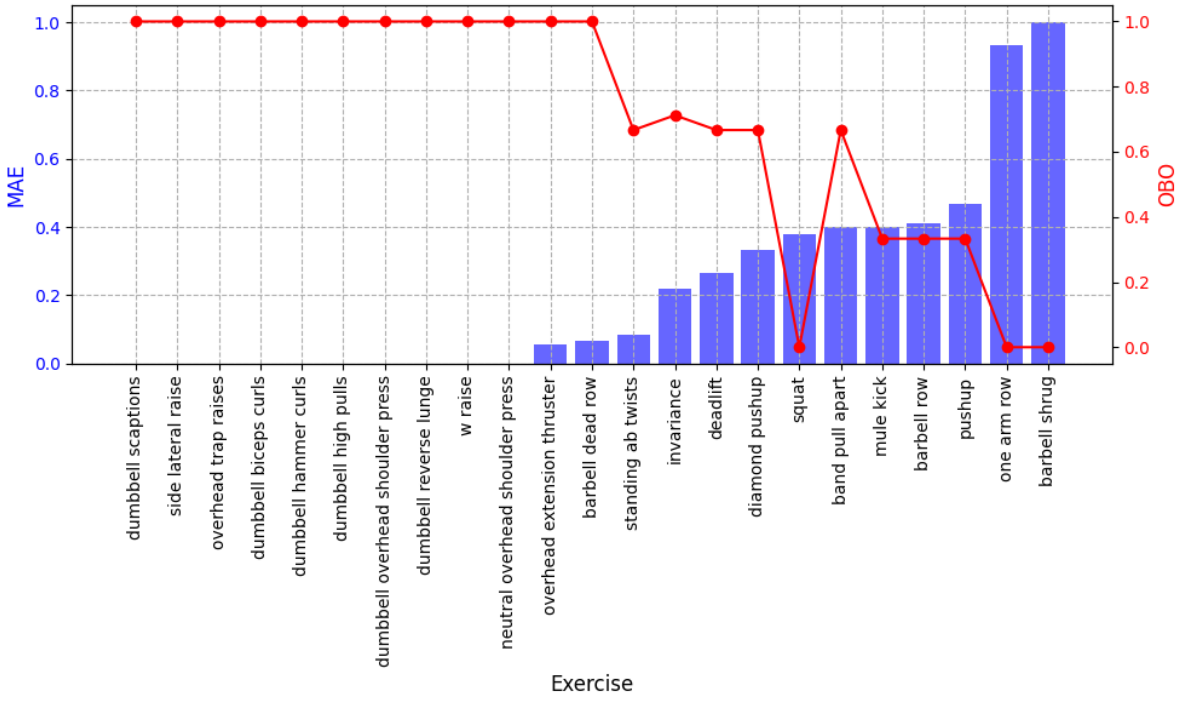

Figure 8 – Breakdown of MAE and OBO per exercise for the best-performing method (VI) on Few-shot Multi-Cam split.

high OBO values in all scenarios. This similarity suggests that PersonalRAC is particularly well-suited for exercises where joint movements are more predictable, isolated, and have a clear distinction between pose salient 1 and 2, as can be seen in Figure 10. This allows it

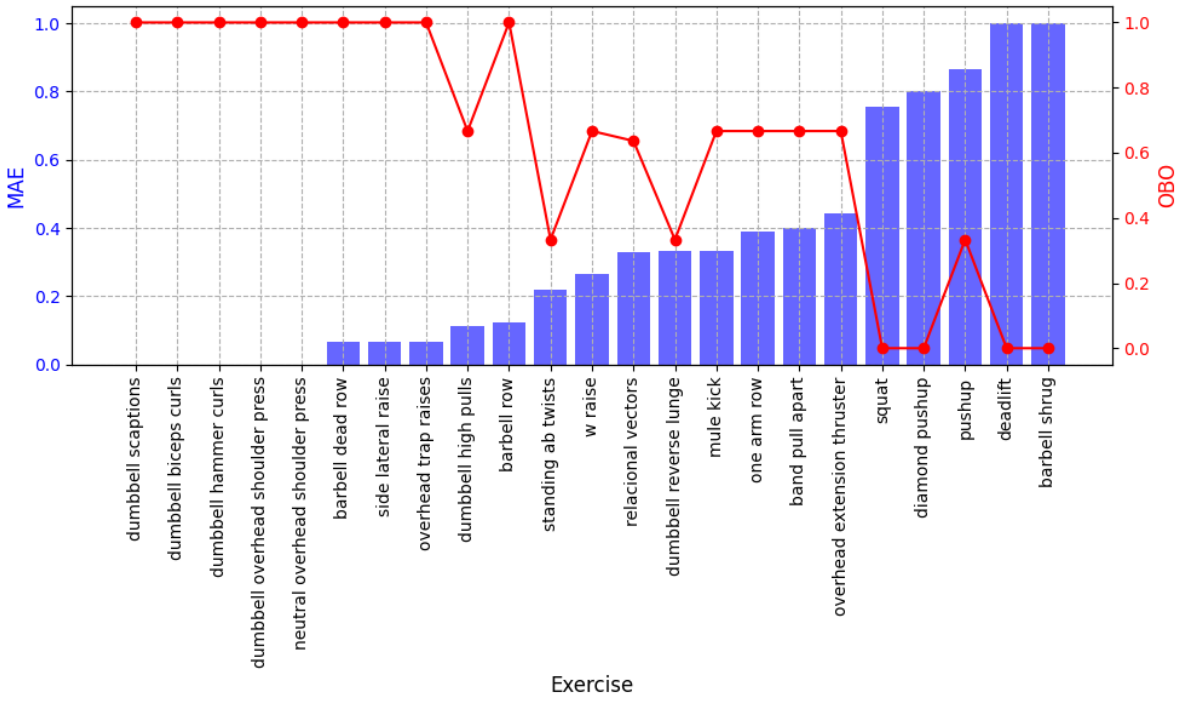

Figure 9 – Breakdown of MAE and OBO per exercise for the best-performing method (RV) on Few-Shot split.

to achieve high accuracy regardless of data availability or camera perspective. The model's ability to recognize these movements consistently across different conditions indicates that the underlying skeletal representation techniques effectively capture the essential features of these more straightforward exercises.

Conversely, some exercises, such as "barbell shrug," "deadlift," and "pushup," consistently exhibit higher MAE and lower OBO scores. In the case of the barbell shrug, this is due to the fact that the pose salients are very close to each other, making it challenging for the model to learn the skeleton's variation, as seen in Figure 10. For the deadlift, the complexity arises from it being a compound exercise, where pose salient 2 is very similar to the start of the exercise (before holding the bar) and the end of the exercise (after putting the bar down), potentially leading to extra counts, as shown in Figure 10. Additionally, changes in camera angle can lead to losses in skeleton detection, in addition to the fact that MotionBERT does not use the skeleton's depth coordinate. Consequently, predictions requiring depth perception become challenging, which is particularly noticeable with the "pushup" exercise, illustrated in Figure 10. In this row, the last frame shows low confidence in detecting a person, failing to identify the skeleton. Even when methods are applied to enhance the robustness of skeleton data, if the detector struggles to identify certain parts, it results in inconsistent training.

PersonalRAC performs well on simple, isolated, and repetitive exercises, particularly those

involving predictable upper-body motions. Its design allows it to effectively capture and analyze straightforward joint dynamics, where distinct phases in the exercise are easily recognized and consistently reproduced across different data conditions. However, PersonalRAC struggles with exercises involving complex, multi-joint interactions and compound movements, where body mechanics are more intricate, and depth perception becomes crucial. Without explicit depth information, PersonalRAC encounters challenges in accurately distinguishing phases that appear similar from certain angles, leading to inaccuracies in exercises requiring detailed three-dimensional tracking. Additionally, exercises with subtle or overlapping pose variations introduce difficulty, as the model may misinterpret these as transitions or fail to capture them consistently, highlighting potential areas for refinement in handling complex and dynamic movements.

5.4.2 Out-of-Dataset Evaluation

To further assess the robustness and generalization capability of PersonalRAC beyond the training dataset, we conducted an out-of-dataset evaluation. This analysis aimed to verify the model's ability to (i) selectively count repetitions of a specified exercise while ignoring unrelated movements and (ii) maintain accuracy across varying camera viewpoints. Two experimental video demonstrations illustrate these aspects:

- **Exercise-Specific Counting** (Video Link): This evaluation presents an individual performing multiple exercises, with the system configured to recognize and count only squats. The results confirm that PersonalRAC can isolates the target exercise, ensuring that only correctly performed squats contribute to the final count while disregarding unrelated actions.
- **Viewpoint Robustness** (Video Link): In this experiment, an individual performs a selected exercise while continuously rotating. The model consistently detects the salient pose transitions and accurately counts repetitions irrespective of camera viewpoint, demonstrating its robustness to varying perspectives and enhancing its applicability in unconstrained real-world settings.

These results reinforce the effectiveness of PersonalRAC in practical scenarios, highlighting its ability to recognize and count repetitions with high precision while remaining robust to

exercise variability and viewpoint changes.

5.4.3 Lessons Learned and Limitations

While the PersonalRAC model demonstrates advancements in personalized exercise counting by allowing the registration of new exercises through few-shot learning and skeleton action recognition, three key limitations can be pointed out.

- **Limited Exercise Complexity:** The system's effectiveness varies based on exercise complexity. It performs well on exercises with clear, isolated joint movements and distinct phases but struggles with compound movements or exercises involving complex body dynamics and multi-joint interactions. In particular, exercises with overlapping or subtle pose variations pose challenges for accurate action recognition.
- **Challenges with Depth Perception** Another limitation is the system's lack of depth perception, which affects its performance on exercises where three-dimensional tracking is essential. For instance, exercises like pushups require accurate depth information to distinguish between similar poses and prevent miscounts. The absence of depth data leads to difficulties in accurately capturing movements.
- **Robustness to Occlusion** While viewpoint invariance improves robustness across different angles, extreme perspectives or occlusions still pose significant challenges. Overhead, low-ground, or highly angled views can lead to skeleton detection errors and reduce action recognition accuracy. Similarly, occlusions (e.g., body parts blocking key joints) impact detection, especially for lower-body movements.

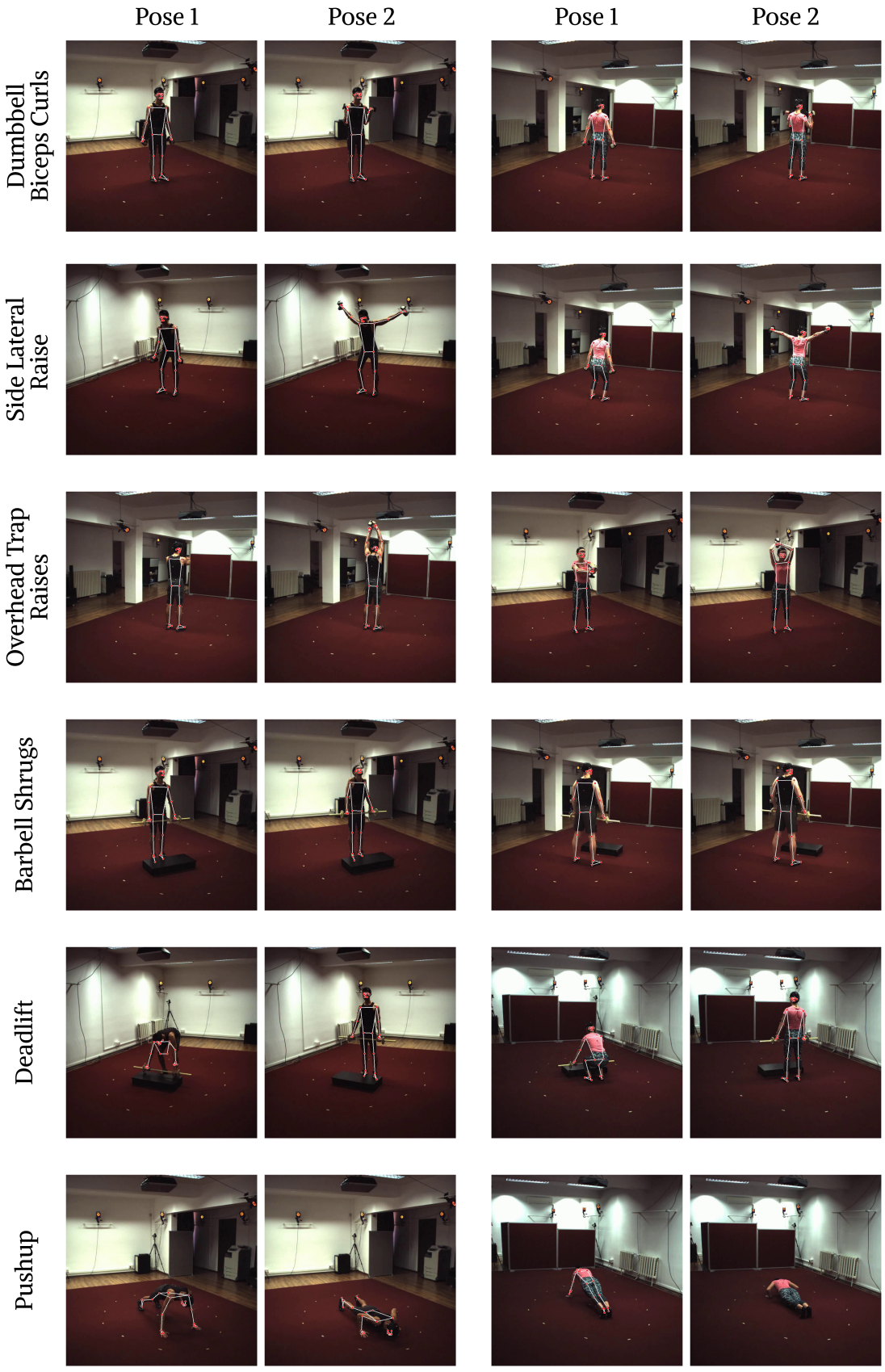

Figure 10 – Example of salient poses with the skeleton drawn from individuals in the test set, focusing on bringing variability in camera angle. Each row represents an exercise, and each column represents a different pose or individual.

6 CONCLUSION

According to our knowledge, PersonalRAC is the first solution to approach the problem of Repetition Action Counting (RAC) using a few-shot strategy. Such few-shot capability allows, for example, personal trainers, physiotherapists, and others to easily register new personalized exercises focused on specific clients and patients. This way, the potential of the PersonalRAC system aims to advance the technological support for applications dedicated to personalized exercise monitoring and rehabilitation.

The PersonalRAC model offers a novel approach to exercise repetition counting, achieving a 44.07% enhancement in MAE and a 0.64 increase in OBO in the few-shot division. Additionally, in the Few-Shot Multi-Cam setting, the model demonstrates a 53.19% improvement in MAE and a 0.71 increase in OBO, highlighting its effectiveness in challenging low-data environments. However, three key limitations persist: the model struggles with complex multi-joint exercises and subtle pose variations, lacks depth perception for exercises requiring precise three-dimensional tracking, and faces challenges with extreme camera angles and occlusions despite improvements in viewpoint invariance. By integrating Few-Shot Learning, Repetitive Action Counting, and Skeleton Action Recognition, our method surpasses the state-of-the-art performance of existing approaches that use entire datasets for training while paving the way for more personalized and accessible solutions.

Future work will involve benchmarking these various action detection models within the PersonalRAC model and evaluating their performance in RAC tasks across different exercises and environments. Using skeleton detection models that are more robust to occlusion is also a possible way to improve results and test different skeleton action detection models, especially those that can work with the depth dimension to address issues observed in exercises like "pushups."

REFERENCES

BAZAREVSKY, V.; GRISHCHENKO, I.; RAVEENDRAN, K.; ZHU, T.; ZHANG, F.; GRUNDMANN, M. *BlazePose: On-device Real-time Body Pose tracking*. 2020. Disponível em: <https://arxiv.org/abs/2006.10204>.

DWIBEDI, D.; AYTAR, Y.; TOMPSON, J.; SERMANET, P.; ZISSERMAN, A. Counting out time: Class agnostic video repetition counting in the wild. In: *CVPR*. Computer Vision Foundation / IEEE, 2020. p. 10384–10393. ISBN 978-1-7281-7168-5. Disponível em: <http://dblp.uni-trier.de/db/conf/cvpr/cvpr2020.html#DwibediATSZ20>.

DWIBEDI, D.; AYTAR, Y.; TOMPSON, J.; SERMANET, P.; ZISSERMAN, A. Counting out time: Class agnostic video repetition counting in the wild. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2020.

FIERARU, M.; ZANFIR, M.; PIRLEA, S.; OLARU, V.; SMINCHISESCU, C. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2021. p. 9914–9923.

GAMA, A. D.; FALLAVOLLITA, P.; TEICHRIEB, V.; NAVAB, N. Motor rehabilitation using kinect: A systematic review. *Games for Health Journal*, Mary Ann Liebert Inc, v. 4, n. 2, p. 123–135, abr. 2015. ISSN 2161-7856. Disponível em: <http://dx.doi.org/10.1089/g4h.2014.0047>.

HU, H.; DONG, S.; ZHAO, Y.; LIAN, D.; LI, Z.; GAO, S. Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 19013–19022.

KEMTAI. 2023. Accessed: 2024-10-25. Disponível em: <hhttps://kemtai.com/>.

KIM, J.; LEE, D. Activity recognition with combination of deeply learned visual attention and pose estimation. *Applied Sciences*, v. 11, n. 9, 2021. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/11/9/4153>.

KUMAR, A.; RAGHUNATHAN, A.; JONES, R.; MA, T.; LIANG, P. *Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution*. 2022. Disponível em: <https://arxiv.org/abs/2202.10054>.

LAFAYETTE, T.; KUNST, V.; MELO, P.; GUEDES, P.; TEIXEIRA, J.; VASCONCELOS, C.; TEICHRIEB, V.; GAMA, A. D. Validation of angle estimation based on body tracking data from rgb-d and rgb cameras for biomechanical assessment. *Sensors*, v. 23, p. 3, 12 2022.

LEVY, O.; WOLF, L. Live repetition counting. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015. p. 3020–3028.

LI, X.; XU, H. Repetitive action counting with motion feature learning. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. [S.l.: s.n.], 2024. p. 6499–6508.

LUO, Y.; YI, J.; FARHA, Y. A.; WOLTER, M.; GALL, J. *Rethinking temporal self-similarity for repetitive action counting*. 2024. Disponível em: <https://arxiv.org/abs/2407.09431>.

MEMMESHEIMER, R.; HÄRING, S.; THEISEN, N.; PAULUS, D. Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. [S.l.: s.n.], 2022. p. 3702–3710.

MEMMESHEIMER, R.; THEISEN, N.; PAULUS, D. *SL-DML: Signal Level Deep Metric Learning for Multimodal One-Shot Action Recognition*. 2020.

MüLLER, M. Dynamic time warping. *Information Retrieval for Music and Motion*, v. 2, p. 69–84, 01 2007.

PATALAS-MALISZEWSKA, J.; PAJąK, I.; KRUTZ, P.; PAJAK, G.; REHM, M.; SCHLEGEL, H.; DIX, M. Inertial sensor-based sport activity advisory system using machine learning algorithms. *Sensors (Basel, Switzerland)*, v. 23, 2023.

PERETTI, A.; AMENTA, F.; TAYEBATI, S. K.; NITTARI, G.; MAHDI, S. S. Telerehabilitation: Review of the state-of-the-art and areas of application. *JMIR Rehabil Assist Technol*, v. 4, n. 2, p. e7, Jul 2017. ISSN 2369-2529. Disponível em: <http://rehab.jmir.org/2017/2/e7/>.

POCKETFISIO. 2024. Accessed: 2024-10-25. Disponível em: <https://www.pocketfisio.com/>.

PRABHU, G.; O'CONNOR, N. E.; MORAN, K. Recognition and repetition counting for local muscular endurance exercises in exercise-based rehabilitation: A comparative study using artificial intelligence models. *Sensors (Basel, Switzerland)*, v. 20, 2020. Disponível em: <https://api.semanticscholar.org/CorpusID:221360688>.

SABATER, A.; SANTOS, L.; SANTOS-VICTOR, J.; BERNARDINO, A.; MONTESANO, L.; MURILLO, A. C. One-shot action recognition in challenging therapy scenarios. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. [S.l.: s.n.], 2021. p. 2777–2785.

SENCY. 2024. Accessed: 2024-10-25. Disponível em: <https://www.sency.ai/>.

SORO, A.; BRUNNER, G.; TANNER, S.; WATTENHOFER, R. Recognition and repetition counting for complex physical exercises with deep learning. *Sensors*, MDPI AG, v. 19, n. 3, p. 714, fev. 2019. ISSN 1424-8220. Disponível em: <http://dx.doi.org/10.3390/s19030714>.

TABORRI, J.; BORDIGNON, M.; MARCOLIN, F.; DONATI, M.; ROSSI, S. Automatic identification and counting of repetitive actions related to an industrial worker. *2019 II Workshop on Metrology for Industry 4.0 and IoT (MetroInd4.0IoT)*, p. 394–399, 2019.

TSIOURIS, K. M.; TSAKANIKAS, V. D.; GATSIOS, D.; FOTIADIS, D. I. A review of virtual coaching systems in healthcare: Closing the loop with real-time feedback. *Frontiers in Digital Health*, Frontiers Media SA, v. 2, set. 2020. ISSN 2673-253X. Disponível em: <http://dx.doi.org/10.3389/fdgth.2020.567502>.

van Egmond, M.; van der Schaaf, M.; VREDEVELD, T.; VOLLENBROEK-HUTTEN, M.; van Berge Henegouwen, M.; KLINKENBIJL, J.; ENGELBERT, R. Effectiveness of physiotherapy with telerehabilitation in surgical patients: a systematic review and meta-analysis. *Physiotherapy*, v. 104, n. 3, p. 277–298, 2018. ISSN 0031-9406. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031940618300750>.

VARGHESE, M. M.; RAMESH, S.; KADHAM, S.; DHRUTHI, V. M.; KANWAL, P. Real-time fitness activity recognition and correction using deep neural networks. *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, p. 1–6, 2023.

YAO, Z.; CHENG, X.; ZOU, Y. Poserac: Pose saliency transformer for repetitive action counting. *arXiv (Cornell University)*, Cornell University, 03 2023.

YIN, J.; WU, Y.; ZHU, C.; YIN, Z.; LIU, H.; DANG, Y.; LIU, Z.; LIU, J. Energy-based periodicity mining with deep features for action repetition counting in unconstrained videos. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 31, p. 4812–4825, 2020.

ZHANG, H.; XU, X.; HAN, G.; HE, S. Context-aware and scale-insensitive temporal repetition counting. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2020. p. 667–675.

ZHANG, H.; XU, X.; HAN, G.; HE, S. Context-aware and scale-insensitive temporal repetition counting. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2020.

ZHANG, W.; SU, C.; HE, C. Rehabilitation exercise recognition and evaluation based on smart sensors with deep learning framework. *IEEE Access*, v. 8, p. 77561–77571, 2020.

ZHU, W.; MA, X.; LIU, Z.; LIU, L.; WU, W.; WANG, Y. Motionbert: A unified perspective on learning human motion representations. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2023. p. 15085–15099.

ZHU, W.; MA, X.; LIU, Z.; LIU, L.; WU, W.; WANG, Y. Motionbert: A unified perspective on learning human motion representations. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2023.

ZUNINO, A.; CAVAZZA, J.; MURINO, V. *Revisiting Human Action Recognition: Personalization vs. Generalization*. 2016. Disponível em: <https://arxiv.org/abs/1605.00392>.

ZUNINO, A.; CAVAZZA, J.; MURINO, V. Revisiting human action recognition: Personalization vs. generalization. In: _____. *Image Analysis and Processing - ICIAP 2017*. Springer International Publishing, 2017. p. 469–480. ISBN 9783319685601. Disponível em: <http://dx.doi.org/10.1007/978-3-319-68560-1_42>.

APPENDIX A – DETAILEDS RESULTS

Here are the detailed result figures for all methods. For the Full Dataset, we have Figures 11, 12, 13, and 14; for Few-Shot Multi-Cam, Figures 15, 16, 17, and 18; and for Few-Shot, Figures 19, 20, 21, and 22 representing Default Representation (DR), Relational Vectors (RV), Viewpoint Invariance (VI), and International Standard of Biomechanics (ISB), respectively.

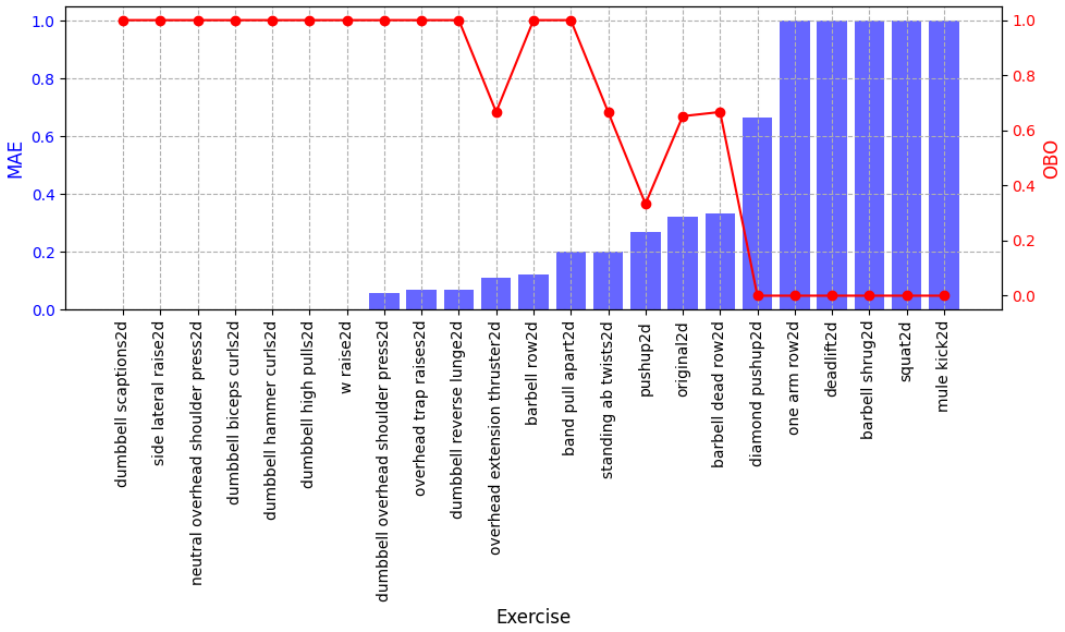

Figure 11 – Breakdown of MAE and OBO per exercise for DR in the Full dataset split.

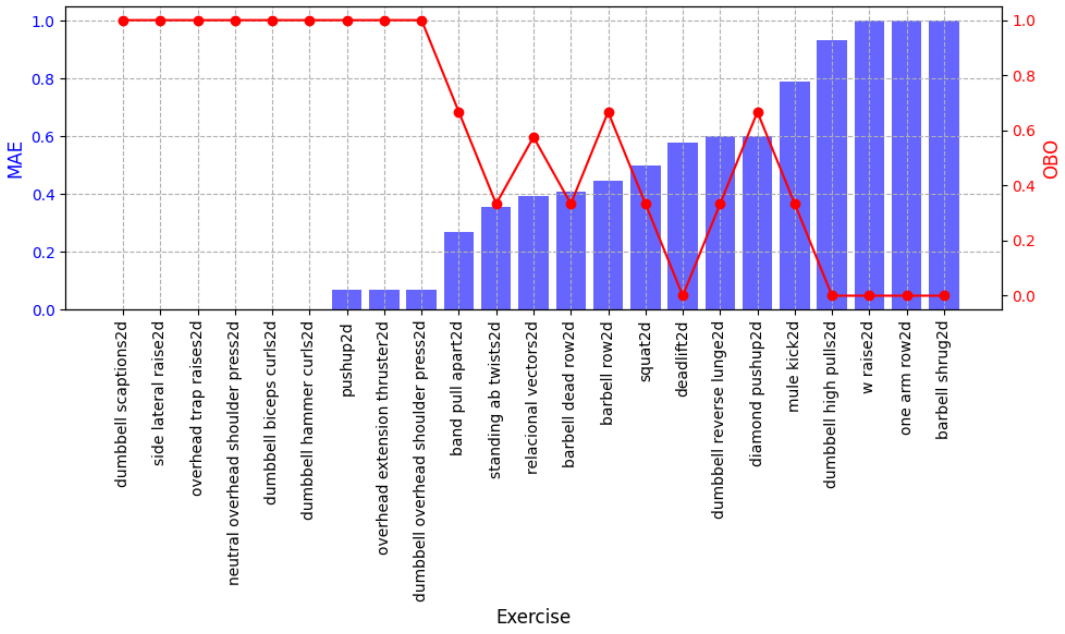

Figure 12 – Breakdown of MAE and OBO per exercise for RV in the Full dataset split.

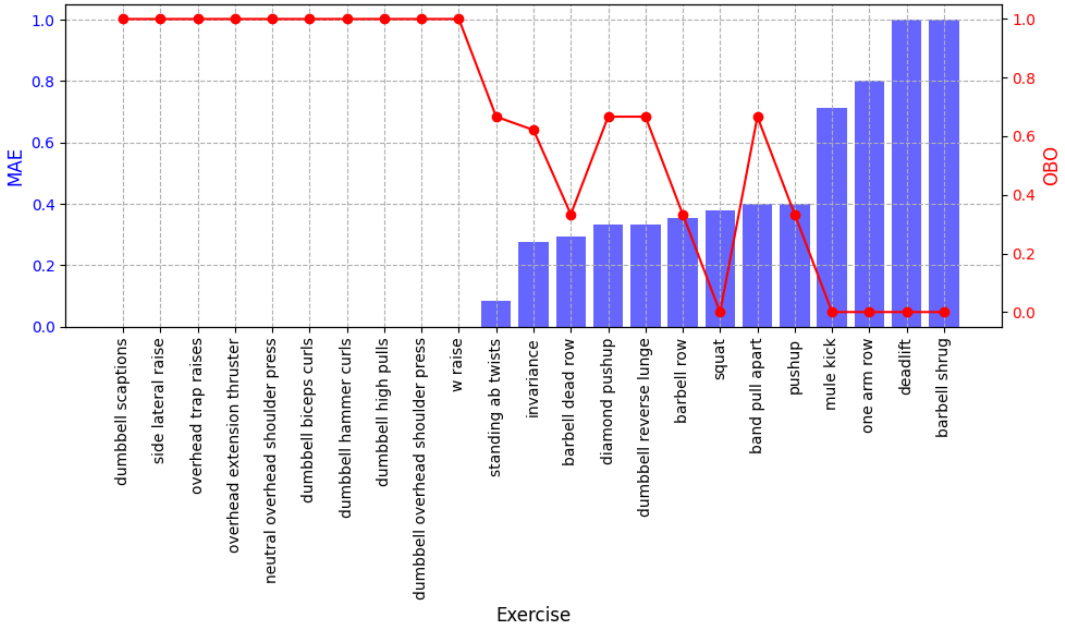

Figure 13 – Breakdown of MAE and OBO per exercise for VI in the Full dataset split.

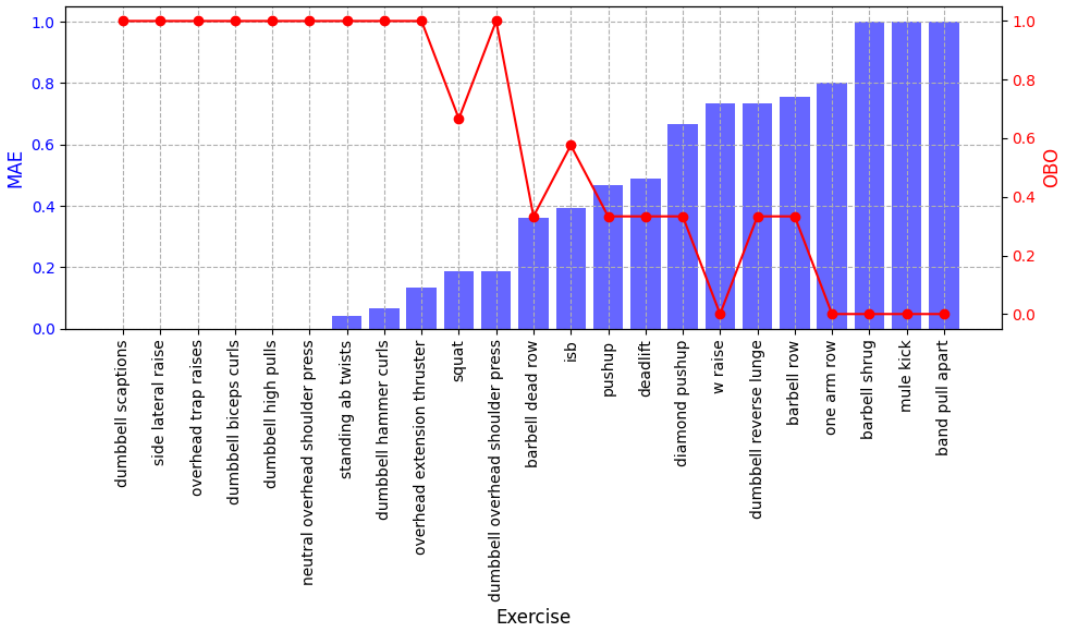

Figure 14 – Breakdown of MAE and OBO per exercise for ISB in the Full dataset split.

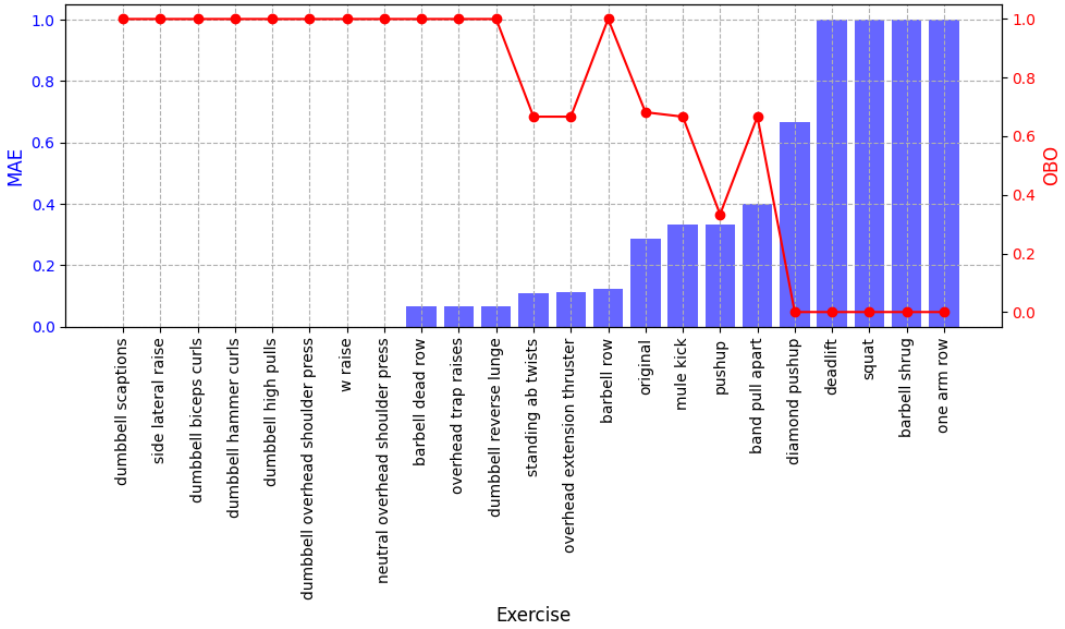

Figure 15 – Breakdown of MAE and OBO per exercise for DR in the Few-Shot Multi-Cam dataset split.

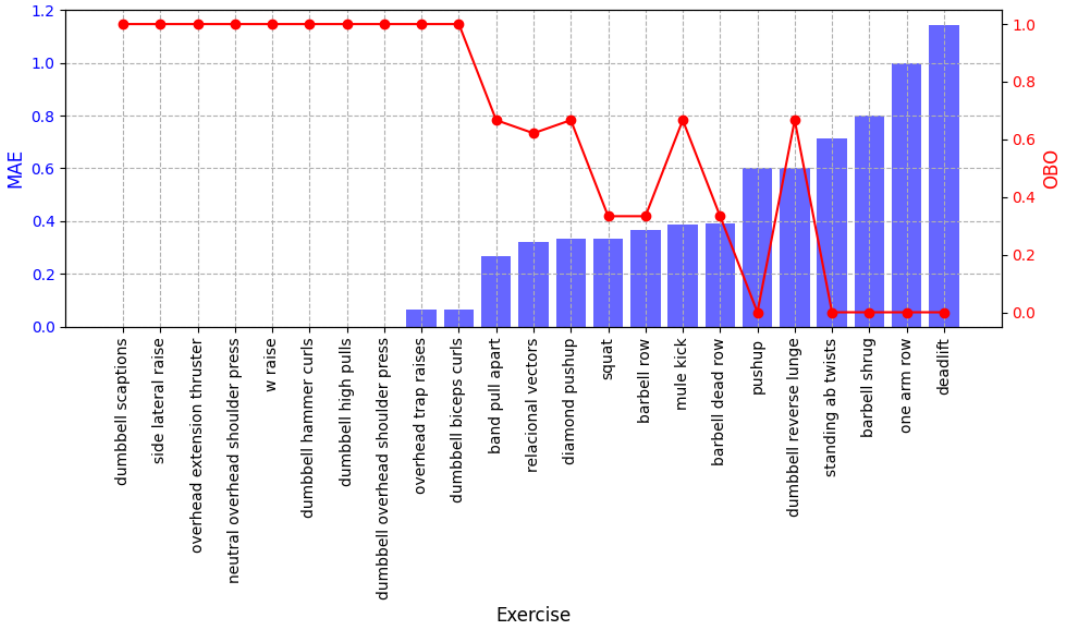

Figure 16 – Breakdown of MAE and OBO per exercise for RV in the Few-Shot Multi-Cam dataset split.

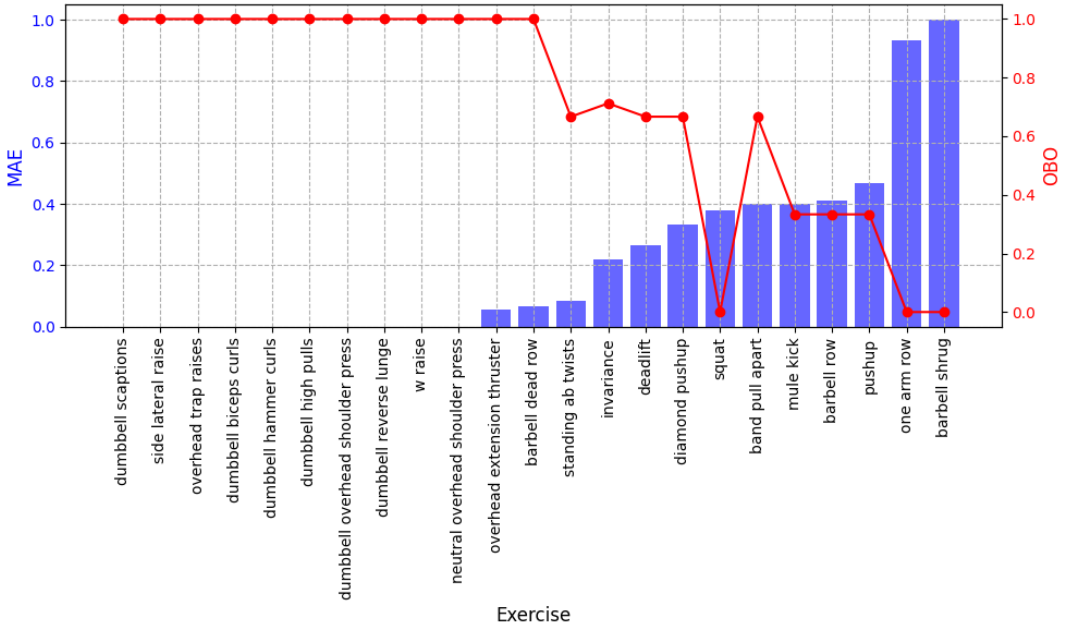

Figure 17 – Breakdown of MAE and OBO per exercise for VI in the Few-Shot Multi-Cam dataset split.

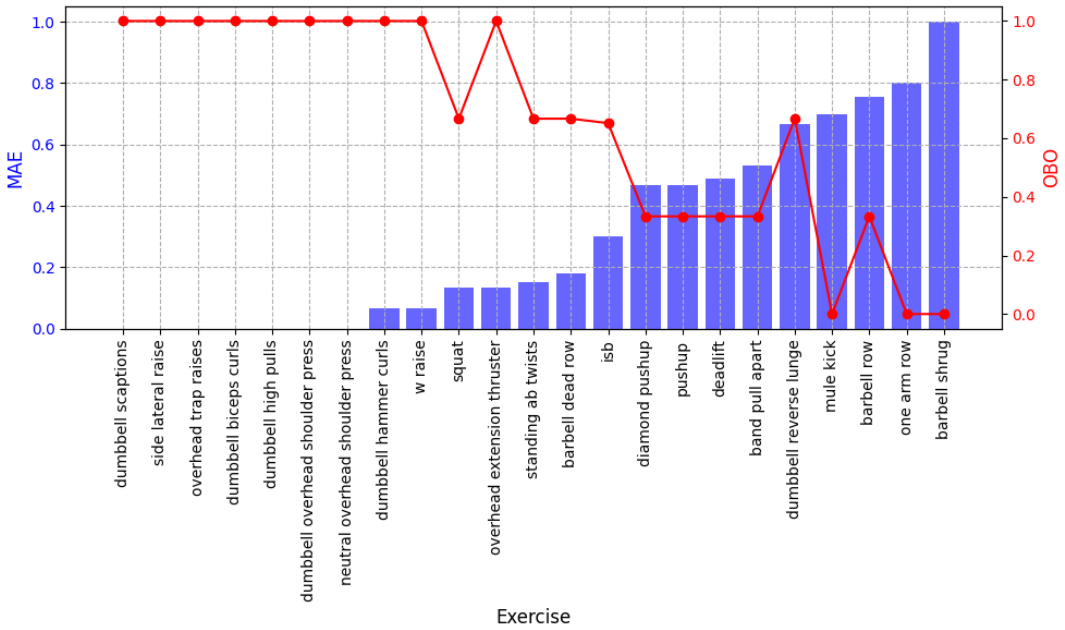

Figure 18 – Breakdown of MAE and OBO per exercise for ISB in the Few-Shot Multi-Cam dataset split.

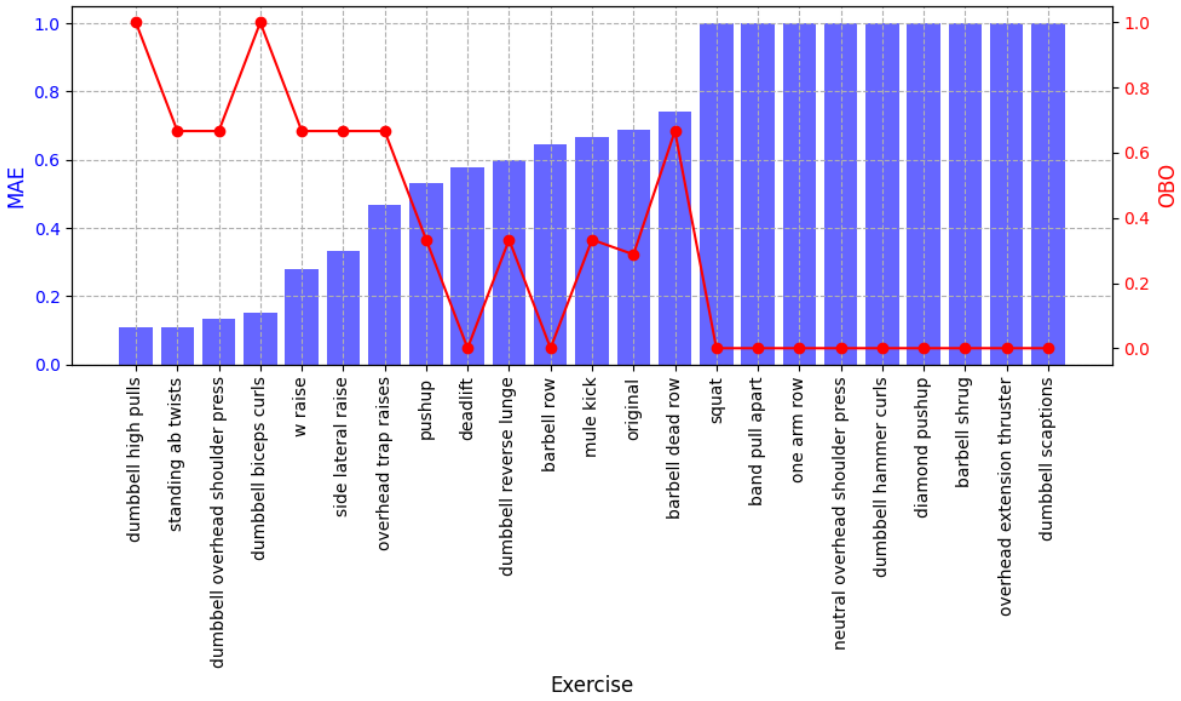

Figure 19 – Breakdown of MAE and OBO per exercise for DR in the Few-shot dataset split.

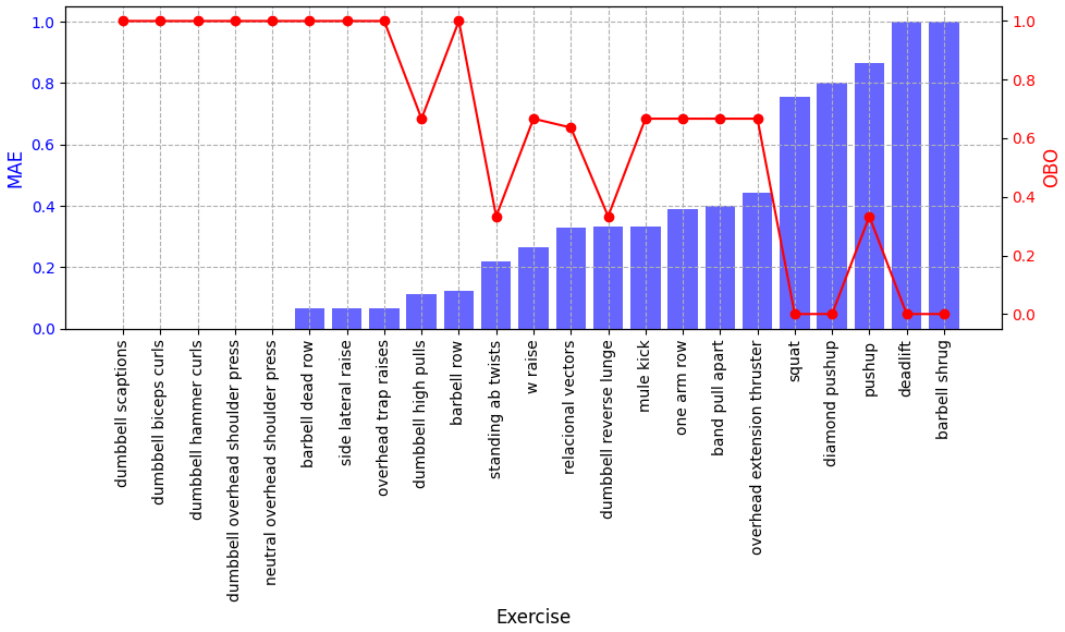

Figure 20 – Breakdown of MAE and OBO per exercise for RV in the Few-Shot dataset split.

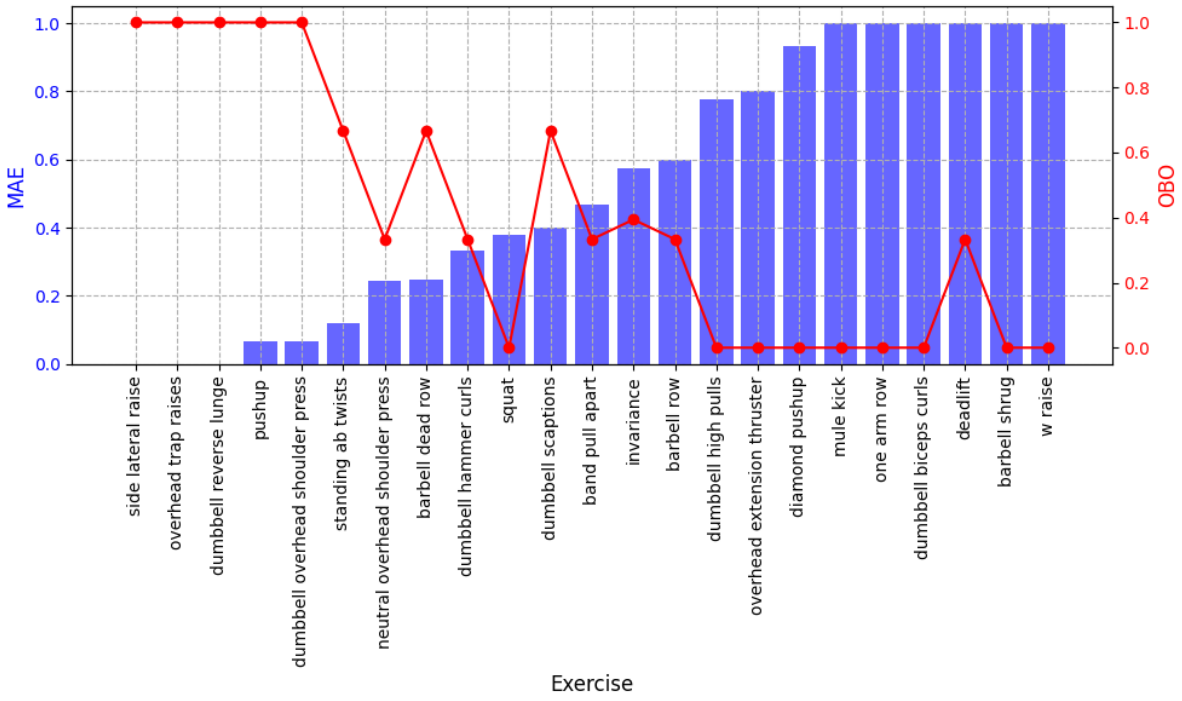

Figure 21 – Breakdown of MAE and OBO per exercise for VI in the Few-sho dataset split.

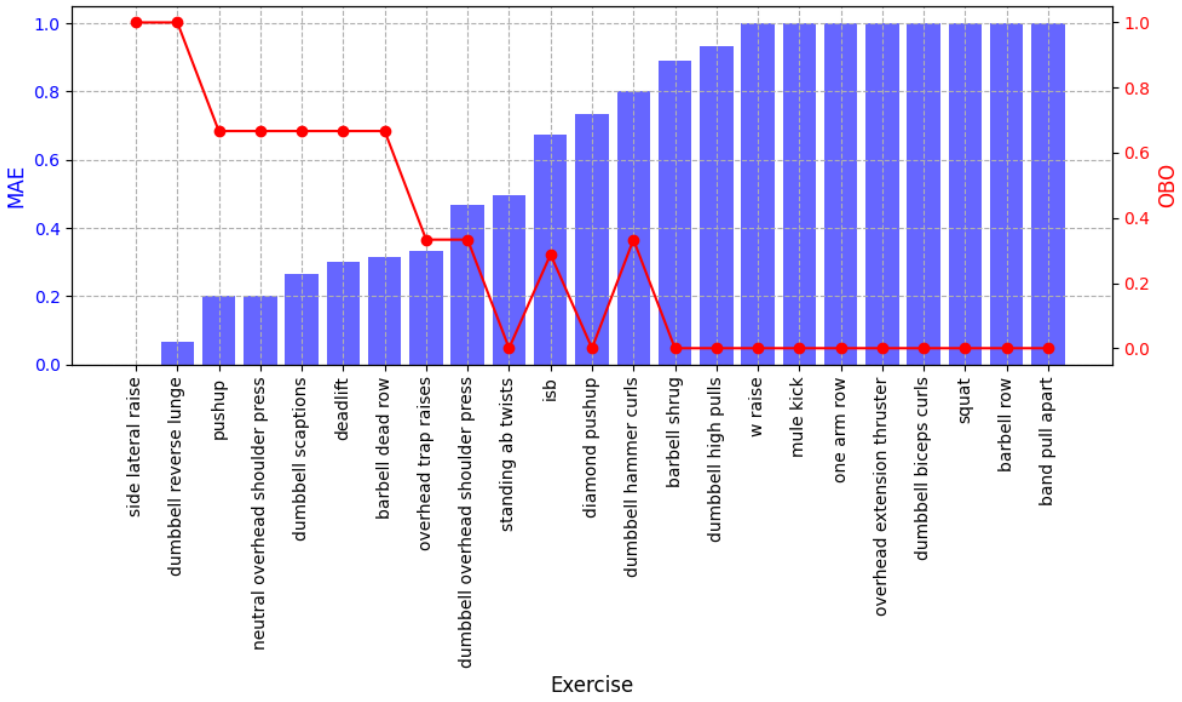

Figure 22 – Breakdown of MAE and OBO per exercise for ISB in the the Few-sho dataset split.