Evaluating the Accuracy of Cloud-based 3D Human Pose Estimation Tools: A Case Study of MOTiO by RADiCAL

Hamza Khalloufi[1]*, Mohamed Zaifri[2], Abdessamad Benlahbib[3], Fatima Zahra Kaghat[4], Ahmed Azough[5]
Laboratory of Informatics, Signals, Automatics and Cognitivism (LISAC) Faculty of Sciences Dhar El Mahraz, University Sidi Mohamed Ben Abdellah, Fez, Morocco[1, 2, 3]
Research Center, Pôle Universitaire Léonard de Vinci, Paris, France[4, 5]

***Abstract*—The use of 3D Human Pose Estimation (HPE) has become increasingly popular in the field of computer vision due to its various applications in human-computer interaction, animation, surveillance, virtual reality, video interpretation, and gesture recognition. However, traditional sensor-based motion capture systems are limited by their high cost and the need for multiple cameras and physical markers. To address these limitations, cloud-based HPE tools, such as DeepMotion and MOTiON by RADiCAL, have been developed. This study presents the first scientific evaluation of MOTiON by RADiCAL, a cloud-based 3D HPE tool based on deep learning and cloud computing. The evaluation was conducted using the CMU dataset, which was filtered and cleaned for this purpose. The results were compared to the ground truth using two metrics, the Mean per Joint Error (MPJPE) and the Percentage of Correct Keypoints (PCK). The results showed an accuracy of 98 mm MPJPE and 96% PCK for most scenarios and genders. This study suggests that cloud-based HPE tools such as MOTiON by RADiCAL can be a suitable alternative to traditional sensor-based motion capture systems for simple scenarios with slow movements and little occlusion.**

***Keywords*—3D; human pose estimation; animation; evaluation; motion tracking**

I. Introduction

Due to its crucial applications in human-computer interaction, surveillance, virtual reality [36], video interpretation, gesture recognition, and many other fields, as depicted in Fig. 1, 3D human body pose estimation (3D HPE) has attracted substantial interest in computer vision. Nevertheless, despite recent advancements, motion capture (MoCap) systems still rely on costly sensor-based systems consisting of multiple-camera setups and heavy motion capture suits with physical markers that allow position estimation. A considerable number of studies have been conducted using several approaches. However, the most significant advances in that field have been made in recent years thanks to breakthroughs in deep learning and convolutional neural networks.

Recently, cloud-based 3D HPE tools, such as MOTiON by RADiCAL and DeepMotion, have become more popular for a variety of reasons, including the needless for a powerful computer since processing is done in the cloud, the intuitive graphical user interface, and the ready-to-use outputs by almost all 3D computer graphics software. These tools are generally based on deep learning techniques and offer the ability to directly convert 2D videos to 3D coordinate files through FBX motion frames in a short time. Despite the widespread adoption of those tools, scientific evaluation of their accuracy has yet to be published to determine whether they can serve as an alternative to conventional sensor-based motion capture systems.

In this research, we are especially interested in evaluating the accuracy and suitability of 3D HPE tools based on deep learning and cloud computing. We aim to address the following research questions:

- How accurate are cloud-based 3D HPE tools in estimating human poses compared to ground truth data?
- Can cloud-based 3D HPE tools serve as a feasible alternative to traditional motion capture systems in various scenarios?

Therefore, MOTiON by RADiCAL 3D HPE tool was chosen as a case study. To achieve this goal, we now go over how the CMU dataset was cleaned and filtered for use in this study. The dataset contains multiple scenarios, each of which includes a range of actions seen in videos and the resulting 3D human poses. These videos were used to obtain 3D coordinates for each human joint. For quantitative evaluation, the results were compared to the ground truth after Procrustes alignment [1]. Several metrics, including Mean per Joint Position Error (MPJPE) and Percentage of Correct Keypoints (PCK), were used to evaluate the results of each scenario and both genders. The second sort of evaluation is qualitative, in which one frame from each situation is selected and visually evaluated.

Quantitative results revealed that the MOTiON by RADiCAL tool is adequate for most scenarios and genders. However, it has several limitations, particularly for occlusion and dynamic motions. After data analysis, it is considered that these cloud-based tools could advantageously replace the expensive traditional tools for simple scenarios with slow movements and little occlusion. As for qualitative results, nine scenarios have been accurately estimated, whereas the skeleton or some of its components were misaligned in the other scenarios.

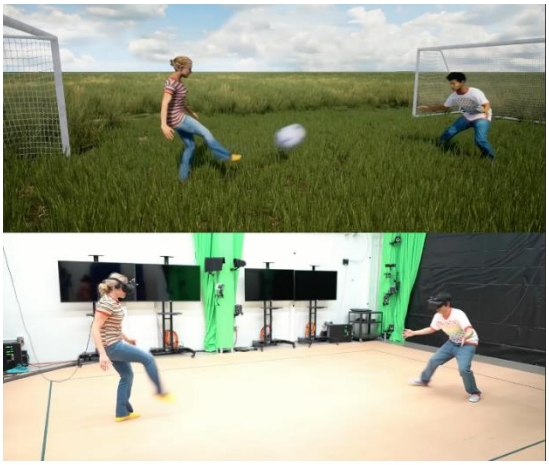

Fig. 1. Facebook's markerless body tracking for VR from a single sensor.

This study is structured to provide a comprehensive evaluation of cloud-based 3D HPE tools, with a specific focus on the MOTiON by RADiCAL system. The structure and goals of this research are aligned to achieve several key contributions:

1) Presents the first comprehensive scientific evaluation of a case study of cloud-based 3D HPE, assessing its accuracy using robust metrics.

2) It contributes to the broader understanding of the potential and limitations of cloud-based 3D HPE tools in various realistic scenarios.

3) The findings could potentially influence future developments in 3D HPE technology, enhancing the accessibility and applicability of cloud-based motion capture solutions.

The organization of the paper is as follows: the related works in Section II reviews prior studies and developments within the field, detailing advancements and challenges in 3D HPE, setting the stage for the current research. The methodology in Section III details the datasets employed in the study, the evaluation metrics used specifically MPJPE and PCK and the experimental setup designed to test the efficacy of the 3D HPE tool. Results in Section IV, presents both quantitative and qualitative analyses that compare the performance of MOTiON by RADiCAL against established ground truth data, highlighting the tool's accuracy and operational characteristics in various scenarios. The discussion in Section V interprets these results, exploring their implications for the field of 3D HPE and discussing potential limitations of the study. Finally, the conclusion in Section VI summarizes the key findings and proposes directions for future research, suggesting how improvements could enhance the utility and accuracy of cloud-based 3D HPE tools.

II. Related Works

A. 3D Human Pose Estimation

Over the past few years, there has been a growing interest in 3D HPE due to its ability to provide accurate information about the 3D structure of the human body. 3D HPE seeks to predict the location of body joints in 3D space. It can be applied to a variety of situations (e.g., 3D animation movies, extended realities, and cloud-based 3D action estimation). Even though 2D HPE has recently seen significant advancements, 3D HPE is still a challenging task to complete. Recent research in the field of computer vision has been focused on the extraction of 3D human pose estimation (HPE) from monocular images or videos. Those are a 2D representation of a 3D scene, resulting in the loss of one dimension. As a result, researchers have been working on developing algorithms and techniques to accurately estimate the 3D pose of human subjects from these 2D images. 3D human pose estimation can be a well-defined problem that is solvable using information fusion methods if there are multiple perspectives or additional sensors such as IMU and LiDAR available. However, one drawback of using deep learning models for this task is their high data dependence and sensitivity to data collection circumstances. Extensive amounts of annotated data are necessary for these models to learn accurate representations of input and output spaces, and factors such as lighting conditions, camera positions, and background can influence their performance negatively.

While obtaining accurate two-dimensional posture annotations for human datasets is relatively straightforward, obtaining accurate three-dimensional pose annotations is considerably more challenging and cannot be done manually. Furthermore, datasets are often collected in controlled indoor settings that focus on specific activities, making them biased towards these scenarios. Recent studies have shown that models trained on such biased datasets tend to perform poorly when applied to other datasets, as demonstrated by cross-dataset inference [2], [3].

1) Single-person 3D HPE: The strategies of Single-person 3D HPE can be categorized as model-free or model-based methods. The first one can be divided into two categories:

a) Direct estimate techniques: instead of first estimating the 2D pose representation, some algorithms in 3D human pose estimation employ direct estimation techniques, as seen in [4], [5], to directly infer the 3D human position from 2D images. Recent advancements include the study by H Ye et al., which enhances real-time 3D pose estimation efficiency through orthographic projection techniques, simplifying the direct estimation process from images without intermediate 2D pose estimation [33].

b) 2D to 3D lifting techniques: The process of inferring 3D poses from intermediate 2D pose pairings is inefficient because it requires multiple network inferences. Human body models are not used in the model-free approaches for recreating 3D human representations. Standard 2D HPE models are used in the first stage to estimate the 2D posture, and 2D to 3D lifting is used in the second stage to construct the 3D pose, such as [6], [7], and [5]. 2D heatmaps rather than 2D poses were used as intermediate representations to estimate 3D posture ([8] and [9]). Through distance matrix regression, Moreno-Noguer [10] deduced the 3D human position from the distances between the joints in the 2D and 3D body (EDMs). When normalization techniques are used, EDMs are invariant to scaling invariance as well as in-plane

image rotations and translations. A Paired Ranking Convolutional Neural Network (PRCNN) was created by Wang et al. [11] to predict the depth ranking of pairwise human joints. The 3D pose was then regressed from the 2D joints and the depth ranking matrix using a coarse-to-fine pose estimator. Li and Lee [12], Sharma et al. [13], and Jahangiri and Yuille [14] were the first to develop numerous, different 3D pose hypotheses. Recent work by C Han et al. introduces uncertainty learning to improve the accuracy and robustness of 3D pose estimations from single images, effectively enhancing this lifting process [34].

Parametric body models, such as kinematic and volumetric models, are utilized by model-based methods to estimate human position, as illustrated in Fig. 2.

The kinematic model represents the body as a series of joints and articulating bones, and in recent years, it has garnered increasing attention in the field of 3D human pose estimation. Pavllo et al. [15] suggested a temporal convolution network for estimating 3D posture from sequential 2D sequences using 2D keypoints. A Short-Term Long Memory (LSTM) unit and shortcut connections were employed in a recurrent neural network to leverage temporal information from human pose data [16].

The Skinned Multi-Person Linear (SMPL) model is among the most commonly utilized volumetric models in the field of 3D HPE, as evidenced by its implementation in works such as [17], [18].

2) Multi-person 3D HPE: There are two approaches for 3D multi-person human pose estimation from monocular RGB images or videos, which are classified into top-down and bottom-up categories. These approaches are illustrated in Fig. 3.

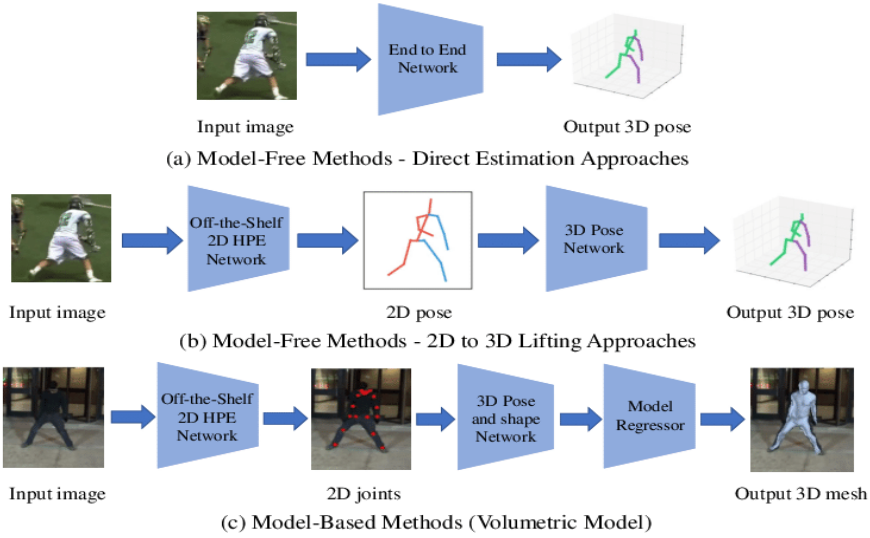

Fig. 2. Frameworks of 3D single-person pose estimation [19] (a) This method is done in one stage, i.e., directly from RGB image to 3D pose. (b) The approaches perform 3D HPE using a two-stage approach, i.e., it performs 2D HPE first and then uses the 2D keypoints to get 3D ones. (c) The 3D mesh is obtained using a regression stage on the 3D HPE outputs.

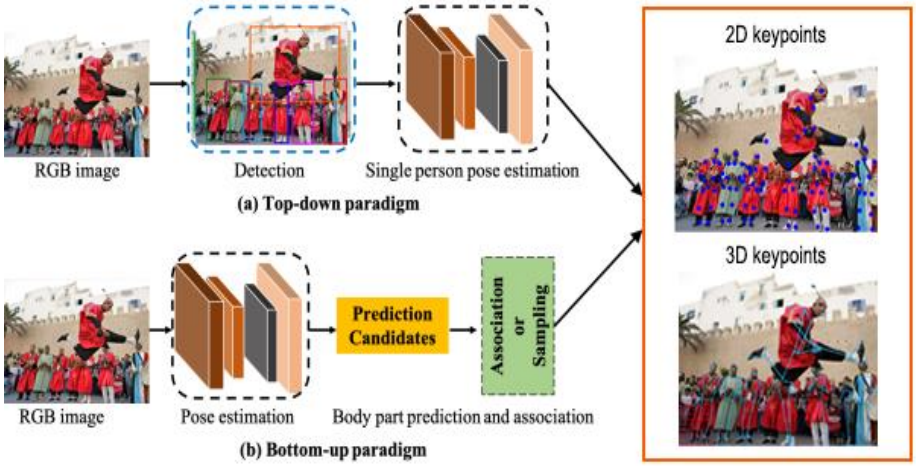

Fig. 3. Frameworks of 2D and 3D multi-person pose estimation [27] (edited). (a) The top-down approach uses person detection techniques to determine the number of persons detected in the frame, and then it applies a 2D single-person estimation framework. (b) The bottom-up approach identifies each joint in the image and then associates each one with individuals.

a) Top-down approaches: They use human detection to estimate each person's position. Each time a person is identified, 3D pose networks estimate their root (the human body's central joint) coordinate and their 3D root-relative posture. Rogez et al. [20] targeted candidate areas of each individual to produce prospective postures and then utilized a regressor to improve the pose suggestions jointly. The LCR-Net technique, which involves localization, classification, and regression, performed well on datasets collected in controlled environments, but not on images captured in natural settings. To address this limitation, LCR-Net++ was introduced, which utilizes synthetic data augmentation during training to improve performance. [21]. The 3D multi-person HPE module was enhanced with scene constraints and semantic segmentation [22]. The 3D temporal assignment problem was also tackled by the Hungarian matching approach for video-based multi-person 3D HPE, which achieved impressive results in [23], [24]. L Jin et al. introduced a single-stage method that integrates human detection and pose estimation, simplifying the process and enhancing efficiency by directly estimating 3D poses from detected individuals in a single network pass, demonstrating significant improvements over traditional multi-stage methods [35].

b) Bottom-up approaches: First, generate joint positions and depth maps for all body joints. They then assign body parts to each individual based on the root depth and relative depth of the body component [25], [26]. How to categorize human body joints is a fundamental difficulty for these techniques. Methods at a lower level exploit the common latent space between two distinct modalities.

B. Datasets for 3D HPE

Obtaining precise 3D labeling for 3D human pose estimation datasets is a difficult endeavor that necessitates the use of motion capture techniques such as MoCap and wearable IMUs. Since the 3D HPE deep learning-based needs larges datasets to train, validate, and test their models, several 3D posture datasets are created due to this need.

1) HumanEva Dataset [28]: It includes seven calibrated video sequences (4 grayscale and three colors) with ground truth 3D annotation taken by a ViconPeak commercial MoCap system. The database comprises four scenarios executing six common actions in a 3m × 2m area: walking, jogging, pointing, throwing and catching a ball, boxing, and combination.

2) Human3.6M [29]: One of the most commonly used datasets for indoor 3D human pose estimation from monocular images and videos. The dataset features 11 professional actors (six males and five females) performing 17 actions (such as smoking, taking photos, and talking on the phone) in a laboratory environment captured from four different perspectives.

3) The CMU Graphics Lab Motion Capture Database (CMU) [30]: CMU is one of the most publicly large databases of motion capture data. Numerous researchers within the scientific world have utilized it to develop previous models of human motion. However, the dataset is poorly synced and contains some films unsuitable for HPE due to multiple actors in each scene. The database comprises more than 100 scenarios executing several actions in a 3m × 8m area.

III. Methodology

After extracting videos from CMU and their associated BVH pose files, a preprocessing stage comprising: cleaning (i.e., avoid corrupted sequences or those that do not verify the necessary conditions), reorganization (i.e., reclassifying all sequences into 12 scenarios), and synchronization (because the BVH poses files are not synchronized with the associated videos) was performed. The sequences in video format were then processed in the cloud with RADiCAL. Both BVH poses of CMU and RADiCAL were rendered using a virtual camera to get the 3D coordinates of each joint. Then two evaluation types were performed; the first one was quantitative, which compared both poses of RADiCAL (as predicted results) and those of CMU (as a ground-truth one). The other evaluation type is a qualitative one based on visual analysis of the predicted 3D human pose scenario according to the ground truth. The workflow was summarized in Fig. 4.

A. Data Preprocessing

The CMU Graphics Lab Motion Capture Database (CMU) was obtained using 12 Vicon MX-40 infrared cameras, each capable of collecting 4-megapixel pictures at 120 Hz. The cameras are positioned around a 3m x 8m rectangle area in the center of the room. The actor wears a black jumpsuit with 41 markers affixed to it while infrared Vicon cameras detect the markings. The pictures captured by the numerous cameras are triangulated to provide three-dimensional data.

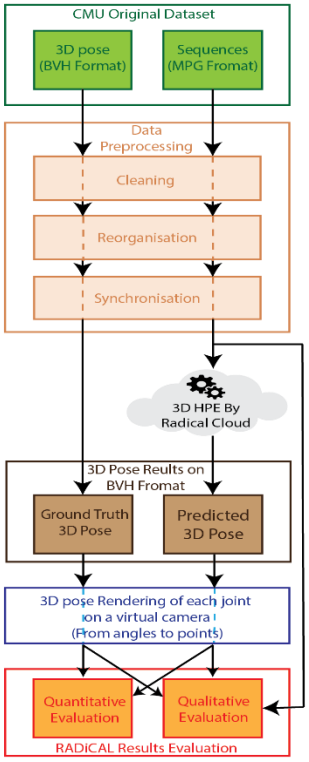

Fig. 4. Overview of the suggested approach.

Despite the dataset covering many scenarios (more than 100), several sequences are without their corresponding videos. Many videos are corrupted, contain more than one actor, or do not contain all body parts. In addition, all files, including BVH and Videos, need to be synced. Therefore the dataset was edited following the three steps below:

1) Cleaning: some corrupted videos were eliminated, and the rest were repaired by hiding the second actor, if that is possible.

2) Reorganization: the sequences were classified into 12 essential scenarios, as shown in Table I.

3) Synchronization: since all BVH frames are not synced with videos, a manual process was manually done using Blender. Also, the sequences captured with 120 or 60 FPS were decreased to 30 FPS since the RADiCAL support only motion capturing with 30 FPS.

B. MOTiON by RADiCAL

MOTiON by RADiCAL is a model-based 3D HPE AI-driven and cloud-based software that converts 2D movies into complete 3D animation with 6 degrees of freedom. The animation data is stored with 30 FPS into FBX (Filmbox), a format that allows the exchange of geometric and animation data between 3D animation software, such as Blender.

For the study, the sequences in MPG format were imported to the RADiCAL cloud then the HPE was processed using the RADiCAL model. After a few moments, the FBX files were done. In order to compare those results to CMU's ground truth, the FBX output files were converted to BVH format using Blender. The output skeleton and the joints are shown in Fig. 5.

C. BVH Projecting to 3D Coordinates

The motion capture of videos from the RADiCAL and CMU datasets is stored in BVH format, including the root transaction coordinates and Euler angles for each joint. As illustrated in Fig. 6, all the coordinates, including those in the BVH files, were projected to a 3D virtual camera to obtain the 3D coordinates of each joint.

Algorithm 1 computes the joint coordinates in camera space from a BVH file containing joint hierarchy and motion data. The algorithm starts by defining the camera intrinsic matrix K, which represents the camera's internal parameters such as focal length and principal point. The camera extrinsic matrix C is also defined, which represents the camera's external parameters such as position and orientation in global space.

$$K = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \quad (1)$$

Algorithm 1: Extract 3D Joint Coordinates from BVH File.

***Algorithm**: Compute Joint Coordinates in Camera Space from BVH File*

Input: BVH file with joint hierarchy and motion data
Output: 3D joint coordinates in camera space for each frame of motion data

1) Load BVH file and extract joint hierarchy and motion data.
2) Define camera intrinsic matrix K with focal lengths f_x and f_y and principal point coordinates c_x and c_y.
3) Define camera extrinsic matrix C with rotation matrix R and position vector P.
4) For each frame of motion data, traverse the joint hierarchy in forward kinematics to compute global joint positions.
5) Transform global joint positions to camera coordinates using K and C.
6) Output the 3D joint coordinates in camera space for each frame of motion data.

End algorithm.

TABLE I. CMU DATASET COMPONENT AFTER CLEANING, FILTERING, AND CLASSIFICATION

Number of scenarios	Number of sequences	Number of views	Frequency	Scenarios	Number of frames
12	279	1	30 FPS	Animal behaviors	62 454
				Climbing	981
				Daily activities	6 306
				Dancing	1 122
				Home activities	56 359
				Jumping	1 553
				Reactions	12 852
				Running	332
				Sitting	4 548
				Sport	9 714
				Walking	16 493
				Working	11 952
				Female	71 759
				Male	112 907
				All	184 667

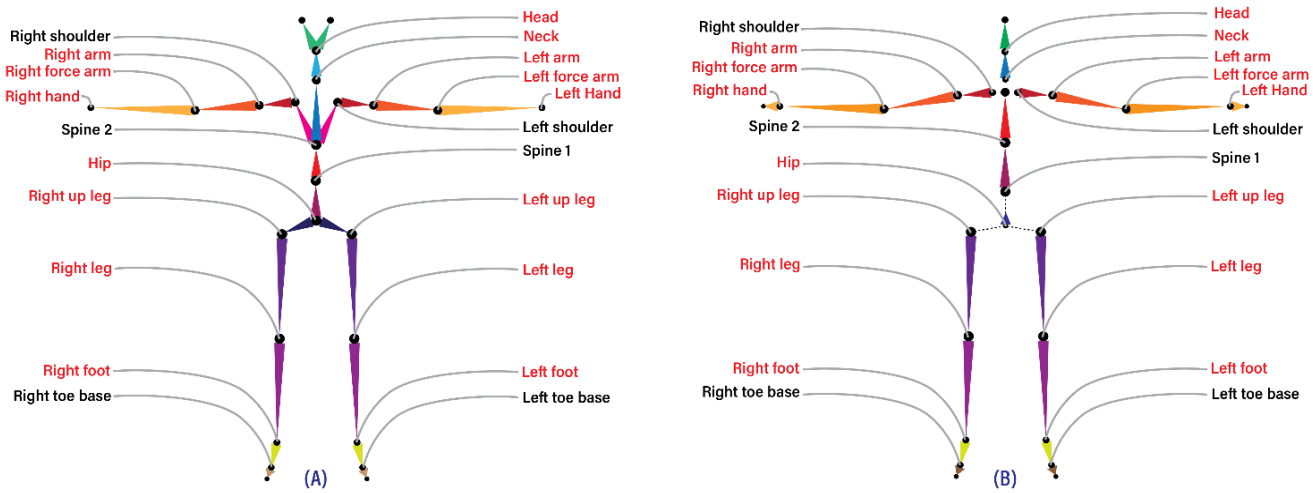

Fig. 5. Skeleton model hierarchy of CMU. (B): Skeleton model hierarchy of RADiCAL. The red ones are the chosen joints to perform the quantitative evaluation.

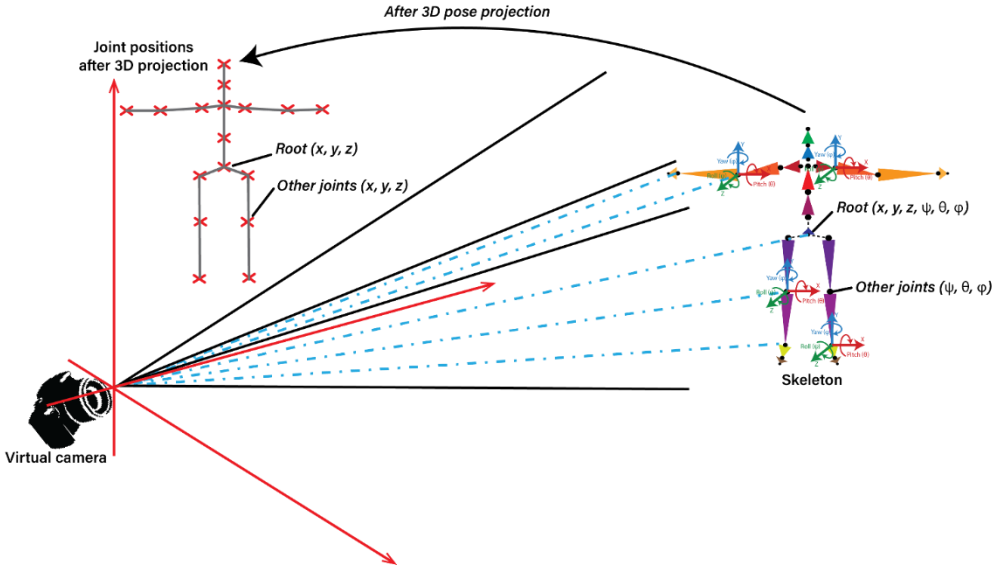

Fig. 6. Rendering process of CMU and RADiCAL skeletons. The purpose is to obtain the 3D pose coordinate of joints from angles.

where, f_x and f_y are the focal lengths of the camera in x and y directions, and c_x and c_y are the coordinates of the principal point of the camera.

$$C = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{pmatrix} \quad (2)$$

where, r_{ij} is the rotation matrix that describe the camera's orientation in global space, and t_i is translation offset. The joint positions and orientations for each frame in the motion data are then computed using forward kinematics, with the root joint's global position and orientation serving as the initial values. The global positions and orientations of child joints are then computed by traversing the joint hierarchy, and the resulting global joint positions are transformed to camera coordinates using the intrinsic and extrinsic matrices. This transformation can be represented mathematically as:

$$\begin{pmatrix} X_{position\_camera} \\ Y_{position\_camera} \\ Z_{position\_camera} \end{pmatrix} = K \times C \times \begin{pmatrix} X_{position\_global} \\ Y_{position\_global} \\ Z_{position\_global} \\ 1 \end{pmatrix} \quad (3)$$

where, $(X_{position\_global}, Y_{position\_global}, Z_{position\_global})$ the global joint position in 3D space, and the resulting is $(X_{position\_camera}, Y_{position\_camera}, Z_{position\_camera})$ is the joint position in camera coordinates. This algorithm provides the way to extract joint positions in camera space from a BVH file.

D. Skeleton Scaling and Evaluation Metrics

1) Skeleton scaling: Since the skeletons of CMU and Radical are not similar, the Procrustes analysis was used to determine the scale [1], rotation, and translation. Given correspondences of points $A_j \in R^3$ and $B_j \in R^3$ of the joint j find scaling, rotation, and translation transformation, called similitude transformation that satisfies:

$$A_j = sRB_j + T \quad (4)$$

For $R \in SO(3), T \in R, and\ s\ \in R^+$

2) Evaluation metrics: Our experiments use two metrics. The first is the mean per-joint position error (MPJPE [29]) between the ground-truth 3D pose and the predicted 3D pose, which is calculated using the Eq. (5): Then, we calculate the mean error across all poses and actions in the dataset.

$$MPJPE = \frac{1}{m}\sum_{i=1}^{m}\left\|p_i^3 + \overline{p_i^3}\right\|_2 \quad (5)$$

For a given skeleton comprising m joints, p_i^3 denotes the actual 3D pose of joint i, whereas $\overline{p_i^3}$ signifies the predicted 3D pose of the same joint.

The second metric is the Percentage of Correct Keypoints for 3D Pose Estimation (PCK3D) [31], a 3D version of the PCK utilized for 2D pose estimation [32]. If the estimated joint location is within a reasonable distance of the ground-truth joint, it is considered to be accurately estimated. Then, the proportion of accurately calculated joints is computed. As in earlier research, the neighborhood threshold is chosen at 150mm [31], corresponding to about half the head size. This statistic is more expressive and robust than MPJPE, highlighting joint mispredictions more clearly. A 15 keypoints were examined, which are indicated in red in Fig. 5.

IV. Results

As stated previously, qualitative and quantitative evaluations were performed. With the restructured CMU dataset, the initial step was to obtain the MPJPE and the PCK by scenario and gender. The second sort of evaluation consisted of picking 3D postures of various scenarios and visually analyzing the results' accuracy.

A. Quantitative Evaluation

The results obtained using MOTiON by RADiCAL cloud-based were compared with the ground-truth 3D poses from the reconstructed CMU dataset using two metrics measurements (MPJPE and PCK). The 3D poses were classified by gender and scenario to assess the accuracy of each one. Then the accuracy of each joint was discussed.

1) Comparing by joints: In this evaluation, 15 crucial joints were analyzed, as depicted in Fig. 5 where the red joints are highlighted. The results are presented in Table II and Fig. 7, displaying the highest mean error values of the Middle Hip, Left Wrist, and Right Wrist joints. While, the lowest mean error values were obtained for the Shoulders, Knees, Nose, and Nick.

2) Comparing by scenarios: Fig. 8 and Table III show that the MPJPE varied from 90,7 mm to 119,1 mm, depending on the scenario. The walking scenario was the most accurate, with an MPJPE of 90.7 mm, whereas the running scenario was the least accurate.

Each scenario's MPJPE (walking, jumping, dancing, reaction, and animal behavior) was under 100 mm while they were near one another, except for home activities, who's MPJPE was just under 100 mm. The MPJPE is more than 100 mm for the remaining scenarios (daily activities, working, climbing, sitting, sports, and running). Expect "Running," "Sitting," and "Sport"; all the scenarios were accurate with higher than 90% of correct joints according to the PCK values of each one. The scenarios: "Home activities," "Jumping," "Reactions," and "Walking" had a PCK near 100%. Expect running and sports scenarios with a standard deviation of around 70 mm. Every other scenario was within 50 mm.

TABLE II. Mean Error by Joints

Joints	Mean (mm)	Standard deviation (mm)
Nose	85.28	36.11
Neck	87.15	27.30
Right Shoulder	77.96	33.85
Right Elbow	99	41.29
Right Wrist	116.96	77.72
Left Shoulder	70.52	30.86
Left Elbow	103.61	49.12
Left Wrist	122.73	85.58
Middle Hip	170.49	20.33
Right Hip	94.78	24.28
Right Knee	80.91	37.07
Right Ankle	92.17	59.06
Left Hip	91.13	25.73
Left Knee	89.61	42.77
Left Ankle	99.07	50.24

TABLE III. MPJPE by Scenarios

Scenarios	MPJPE (mm)	Standard deviation (mm)	PCK (%)
Animal behaviours	92.14	52.33	96.5
Climbing	112.65	54.48	95.62
Daily activities	101.92	53.56	95.8
Dancing	92.14	63.09	94.55
Home activities	99.44	47.66	97.3
Jumping	91.30	40.58	99.03
Reactions	92.97	47.67	98.63
Running	119.10	72.22	83.68
Sitting	116.75	59.32	87.28
Sport	118.31	68.82	83.75
Walking	90.71	45.39	98.13
Working	104.79	56.47	91.57
All	98.76	52.06	95.8

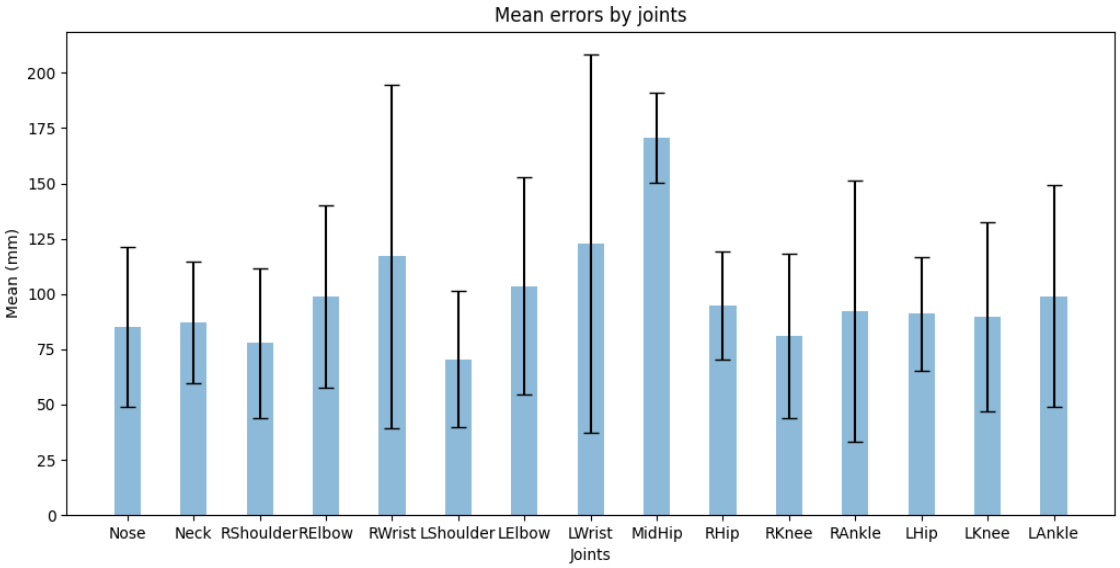

Fig. 7. The results of MPJPE by joints.

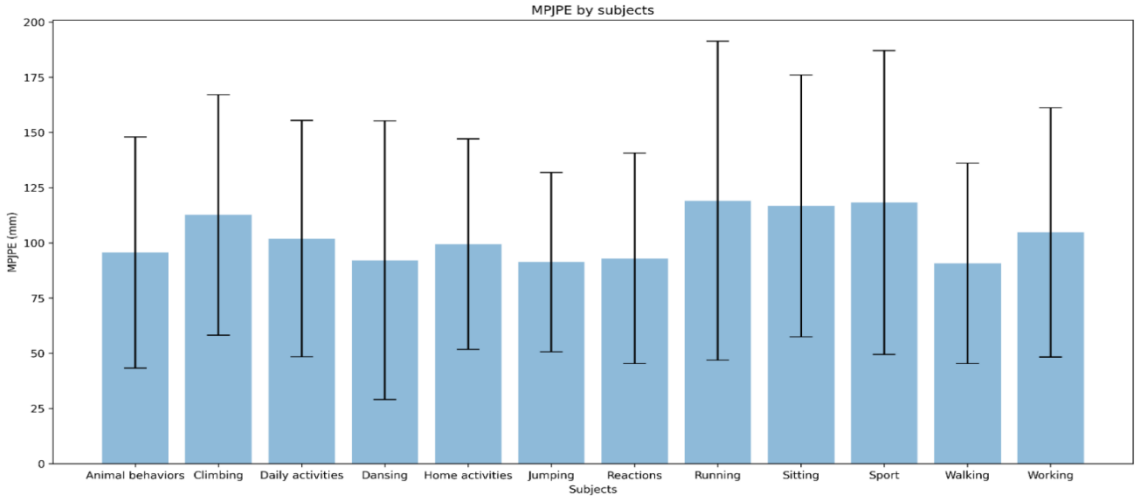

Fig. 8. The results of MPJPE by scenarios.

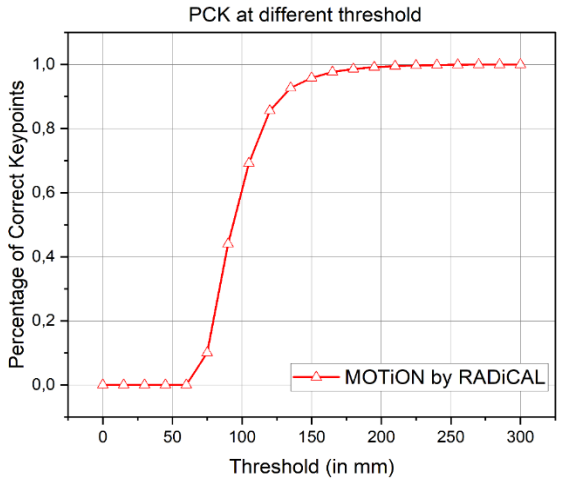

Fig. 9. The thresholds and the corresponding PCK.

Fig. 9 shows a clear progression of the Percentage of Correct Keypoints (PCK) in response to the changing Mean Per Joint Error (MPJPE) threshold. As the MPJPE threshold increases, the PCK also follows suit. Notably, at a relatively stringent threshold of 75 mm, the PCK already reaches about 44%, almost half of the keypoints estimated correctly within this error range. This trend continues and the PCK grows to about 96% when the MPJPE threshold is relaxed to 150 mm, indicating a substantial portion of the estimated keypoints are accurately detected within this margin of error. As we further expand the MPJPE threshold beyond 150 mm, the PCK continues to increase, albeit at a slower rate. The curve eventually approaches a saturation point near 100%, indicating that practically all keypoints are accurately estimated within these larger margins of error.

3) Comparing by gender: The findings of gender-based evaluations are depicted in Table IV and Fig. 10. Male and female MPJPEs were comparable, with the female MPJPE (95

mm) being 5 mm better than the male MPJPE (100 mm). The same holds for the standard deviation, which was almost identical. The PCK of both genders was almost the same, with 95.7%. As shown in Fig. 11, comparing all joints by gender reveals a lower mean error for eight male joints.

TABLE IV. MPJPE BY GENDER

	Female	Male
MPJPE (mm)	95.7	100.71
Standard deviation (mm)	53.06	51.31
PCK (%)	95.9	95.7

B. Qualitative Evaluation

One challenging frame from each scenario was selected, and RADiCAL output was visually compared to the ground-truth frame. As demonstrated in Fig. 12, the scenarios such as "Animal Behaviors," "Daily Activities," "Reactions," "Dancing," "Jumping," "Home Activities," and "Walking" imitate the ground-truth quite accurately. In addition, all skeletal parts are in their proper locations. In the remaining instances, RADiCAL correctly estimated all skeleton parts. However, its orientation was incorrect. The "Working" scenario was estimated correctly, except for the head in several frames. Some parts of the "Sports" scenario, such as the hand, were not precisely estimated.

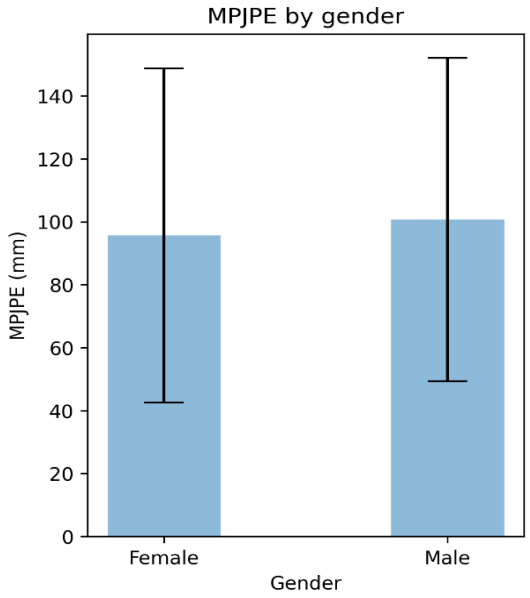

Fig. 10. The results of MPJPE by gender.

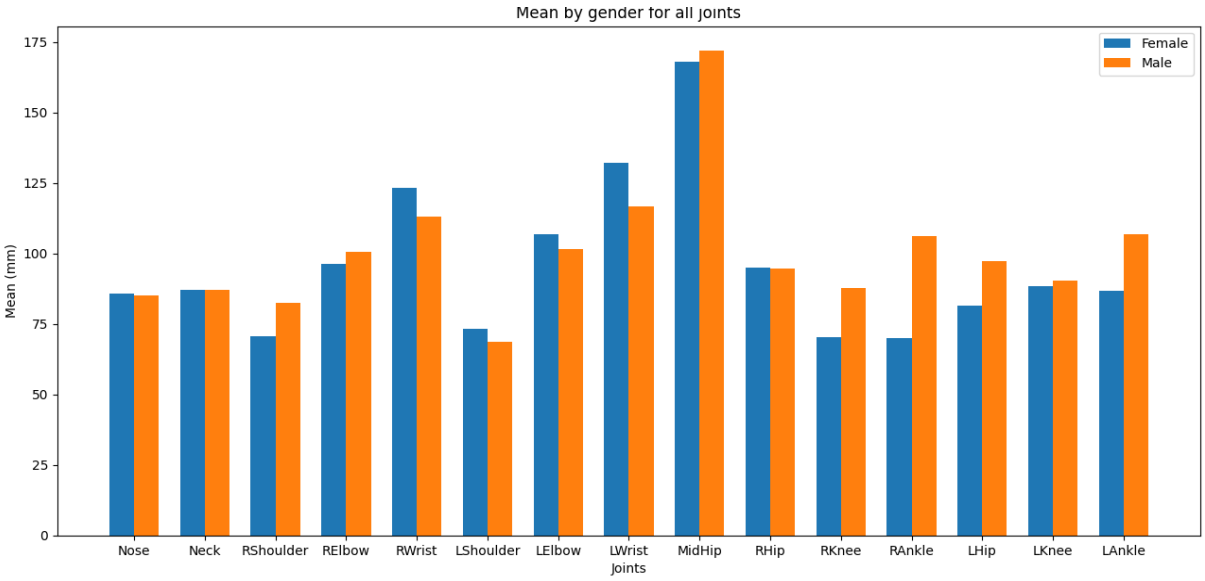

Fig. 11. The MPJPE results of comparison of joints by gender.

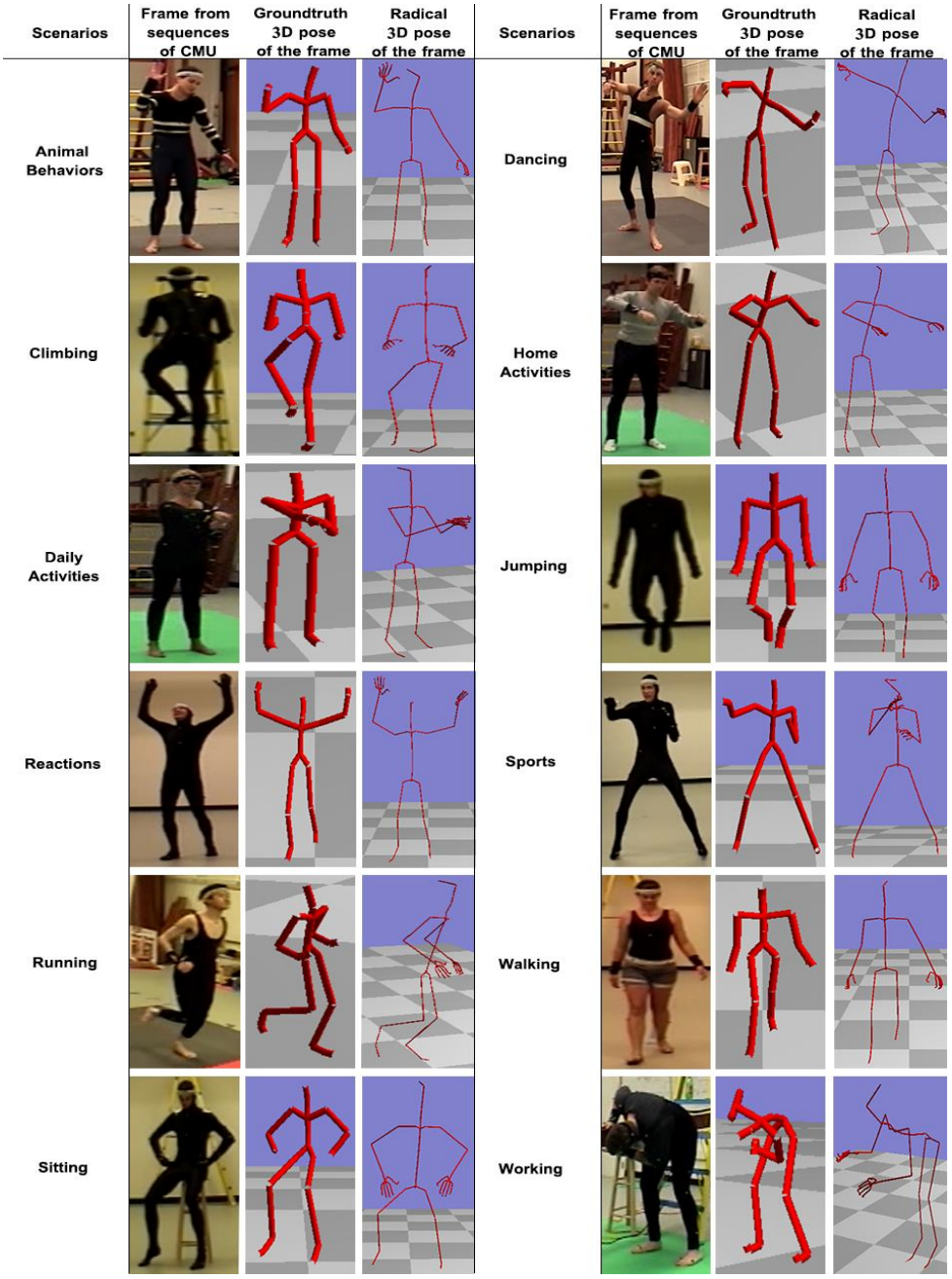

Fig. 12. The qualitative results of each scenario.

V. Discussion and Limitations

This study provides the first comprehensive evaluation of MOTiON by RADiCAL, a cloud-based 3D human motion estimation tool, revealing both its potential advantages and inherent limitations. In scenarios involving less complex actions or slower movements, such as walking or light exercise routines, our evaluation demonstrated a relatively low Mean Per Joint Position Error (MPJPE) and a high Percentage of Correct Keypoints (PCK), signaling promising performance. However, the tool's performance diminished in complex, dynamic scenarios, including sports movements, actions involving occlusion, or tasks requiring significant height variation like climbing.

The distinction in performance directly influences the range of applications suitable for MOTiON by RADiCAL. In digital content creation fields such as simple animation for games or films, or casual fitness tracking where millimeter-level precision may not be paramount, the tool's cost-effectiveness and accessibility offer substantial benefits.

However, for applications demanding high-precision motion capture, such as advanced biomechanical analysis, sports performance analysis, or precise virtual reality interaction, the current version of MOTiON by RADiCAL may not provide the necessary accuracy. The MPJPE of 98mm found in our study, while acceptable in some contexts, could

lead to significant errors in these precision-demanding applications.

VI. Conclusion

MOTiON by RADiCAL, as evaluated in this study, shows promise as a cost-effective, user-friendly alternative to traditional sensor-based motion capture systems. However, the tool's current performance suggests its best fit for applications where absolute precision is not a critical requirement.

In realms like basic animation for gaming, motion-guided user interface design, or casual fitness tracking, the tool's slight inaccuracies are unlikely to substantially impact the end result, making it a beneficial tool. Its cost and usability advantages are particularly beneficial for independent creators, small studios, or hobbyists in these fields.

However, in precision-critical applications, such as advanced biomechanical research, sports performance analysis, or high-end virtual reality systems that require nuanced interaction, the existing error levels in MOTiON by RADiCAL may be prohibitive. For these applications, traditional sensor-based systems, despite their higher cost and complexity, may remain the gold standard.

In summary, MOTiON by RADiCAL represents a significant step forward in democratizing access to 3D human motion estimation. However, its current performance limitations suggest that it is not a one-size-fits-all solution. Future research should explore ways to improve the precision of such tools to extend their applicability to a broader range of scenarios.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank MOTiON by RADiCAL for providing a discount to use their cloud to get the results.

References

[1] J. C. Gower, “Generalized procrustes analysis,” Psychometrika, vol.40, no. 1, pp.33–51, Mar. 1975.

[2] B. Wandt and B. Rosenhahn, “RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation,” arXiv, 12-Mar-2019.

[3] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, “3D Human Pose Estimation in the Wild by Adversarial Learning,” arXiv, 16-Apr-2018.

[4] G. Moon and K. M. Lee, “I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image,” in Computer Vision – ECCV 2020, vol.12352, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. Springer International Publishing, Cham, 2020, pp.752–768.

[5] G. Pavlakos, X. Zhou, and K. Daniilidis, “Ordinal Depth Supervision for 3D Human Pose Estimation,” arXiv, 10-May-2018.

[6] C.-H. Chen and D. Ramanan, “3D Human Pose Estimation = 2D Pose Estimation + Matching,” arXiv, 11-Apr-2017.

[7] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A Simple Yet Effective Baseline for 3d Human Pose Estimation,” 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp.2659–2668, Oct. 2017.

[8] B. Tekin, P. Marquez-Neila, M. Salzmann, and P. Fua, “Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation,” 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp.3961–3970, Oct. 2017.

[9] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, “HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp.2344–2353, Oct. 2019.

[10] F. Moreno-Noguer, “3D Human Pose Estimation from a Single Image via Distance Matrix Regression,” arXiv, 28-Nov-2016.

[11] M. Wang, X. Chen, W. Liu, C. Qian, L. Lin, and L. Ma, “DRPose3D: Depth Ranking in 3D Human Pose Estimation,” Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, pp.978–984, Jul. 2018.

[12] C. Li and G. H. Lee, “Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network,” arXiv, 11-Apr-2019.

[13] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, “Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp.2325–2334, Oct. 2019.

[14] E. Jahangiri and A. L. Yuille, “Generating Multiple Diverse Hypotheses for Human 3D Pose Consistent with 2D Joint Detections,” 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, pp.805–814, Oct. 2017.

[15] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, “3D human pose estimation in video with temporal convolutions and semi-supervised training,” arXiv, 29-Mar-2019.

[16] M. R. I. Hossain and J. J. Little, “Exploiting Temporal Information for 3D Human Pose Estimation,” in Computer Vision – ECCV 2018, vol.11214, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. Springer International Publishing, Cham, 2018, pp.69–86.

[17] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. De la Torre, “3D Human Shape and Pose from a Single Low-Resolution Image with Self-Supervised Learning,” arXiv, 09-Aug-2020.

[18] T. Zhang, B. Huang, and Y. Wang, “Object-Occluded Human Shape and Pose Estimation From a Single Color Image,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp.7374–7383, Jun. 2020.

[19] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, “Deep Learning-Based Human Pose Estimation: A Survey,” arXiv, 23-Jan-2022.

[20] G. Rogez, P. Weinzaepfel, and C. Schmid, “LCR-Net: Localization-Classification-Regression for Human Pose,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp.1216–1224, Jul. 2017.

[21] G. Rogez, P. Weinzaepfel, and C. Schmid, “LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images,” IEEE Trans. Pattern Anal. Mach. Intell., pp.1–1, 2019.

[22] A. Zanfir, E. Marinoiu, and C. Sminchisescu, “Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp.2148–2157, Jun. 2018.

[23] A. Benzine, F. Chabot, B. Luvison, Q. C. Pham, and C. Achard, “PandaNet: Anchor-Based Single-Shot Multi-Person 3D Pose Estimation,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp.6855–6864, Jun. 2020.

[24] C. Wang, J. Li, W. Liu, C. Qian, and C. Lu, “HMOR: Hierarchical Multi-person Ordinal Relations for Monocular Multi-person 3D Pose Estimation,” in Computer Vision – ECCV 2020, vol.12348, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. Springer International Publishing, Cham, 2020, pp.242–259.

[25] M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, and R. Cucchiara, “Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation,” arXiv, 01-Apr-2020.

[26] Q. Nie, Z. Liu, and Y. Liu, “Unsupervised 3D Human Pose Representation with Viewpoint and Pose Disentanglement,” in Computer Vision – ECCV 2020, vol.12364, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds. Springer International Publishing, Cham, 2020, pp.102–118.

[27] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent Advances in Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective," arXiv, 23-Apr-2021.

[28] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," Int. J. Comput. Vis., vol.87, no. 1–2, pp.4–27, Mar. 2010.

[29] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no. 7, pp.1325–1339, Jul. 2014.

[30] "Carnegie Mellon University - CMU Graphics Lab - motion capture library," http://mocap.cs.cmu.edu/, accessed Sep. 29. 2022. .

[31] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision," arXiv, 04-Oct-2017.

[32] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation," arXiv, 17-Sep-2014.

[33] H. Ye, W. Zhu, C. Wang, R. Wu, and Y. Wang, 'Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection', in Computer Vision – ECCV 2022, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, pp. 142–159. doi: 10.1007/978-3-031-20068-7_9, 2022.

[34] C. Han, X. Yu, C. Gao, N. Sang, and Y. Yang, 'Single image based 3D human pose estimation via uncertainty learning', Pattern Recognition, vol. 132, p. 108934, Dec. 2022.

[35] L. Jin et al., 'Single-Stage Is Enough: Multi-Person Absolute 3D Pose Estimation', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13086–13095, 2022.

[36] H. Khalloufi, M. Zaifri, M. Kadri, A. Benlahbib, F. Z. Kaghat, and A. Azough, 'El-FnaVR: An Immersive Virtual Reality Representation of Jemaa El-Fna in Marrakech for Intangible Cultural Heritage Experiences', IEEE Access, vol. 12, pp. 9331–9349, 2024.