2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Multi-View 2D to 3D Lifting Video-Based Optimization: A Robust Approach for Human Pose Estimation with Occluded Joint Prediction*

Daniela Rato[1], Miguel Oliveira[1], Vítor Santos[1], Angel Sappa[2] and Bogdan Raducanu[3]

Abstract— **In the context of robotics, accurate 3D human pose estimation is essential for enhancing human-robot collaboration and interaction. This manuscript introduces a multi-view 2D to 3D lifting optimization-based method designed for video-based 3D human pose estimation, incorporating temporal information. Our technique addresses key challenges, namely robustness to 2D joint detection error, occlusions, and varying camera perspectives. We evaluate the performance of the algorithm through extensive experiments on the MPI-INF-3DHP dataset. Our method demonstrates very good robustness up to 25 pixels of 2D joint error and shows resilience in scenarios involving several occluded joints. Comparative analyses against existing 2D to 3D lifting and multi-view methods showcase good performance of our approach.**

I. INTRODUCTION

3D human pose estimation is essential for human-robot collaboration, as it equips robots with the ability to comprehend and respond to human movements in a three-dimensional space. This technology facilitates seamless and intuitive interactions by enabling robots to interpret gestures, body language, and spatial relationships. Accurate pose estimation ensures precise coordination between humans and robots, enhancing safety and efficiency in shared workspaces [1], [2]. It allows robots to adapt their actions based on human poses, allowing smoother collaboration in diverse applications such as manufacturing, healthcare, and assistive robotics [1], [3]. Thus, it forms a fundamental bridge for effective communication and cooperation between humans and robots, promoting a productive collaborative environment.

However, despite the advances in 3D human pose estimation, current methods often struggle to handle occluded joints effectively. Occlusions occur when certain body parts are temporarily hidden from view, challenging the ability of the algorithm to predict the complete pose accurately. Occlusions can be caused by either an object in front of the human or by the human itself (self-occlusions) when part of the body occludes certain joints. In the context of human-robot collaboration, predicting occlusions becomes crucial. When robots cannot accurately perceive occluded joints, it may lead to misinterpretations of human actions, potentially resulting in errors or accidents. For instance, if a robot fails to recognise that an arm of the person is temporarily hidden behind an object, it might misunderstand the intended action, impacting the collaborative task. Therefore, developing robust algorithms that can predict and account for occluded joints is paramount for enhancing the reliability and safety of human-robot collaboration scenarios. It ensures that the robot can adapt appropriately even when parts of the human body are temporarily occluded, contributing to a more effective and secure collaborative environment.

To address the limitations of existing methods, this manuscript introduces a novel approach to 3D human pose estimation represented in Fig. 1. We employ a 2D to 3D lifting optimization technique, leveraging information gathered from multiple video frames. Unlike traditional methods that rely solely on individual frames, our algorithm considers temporal information to predict occluded joints more robustly. By analysing a sequence of frames, the algorithm is able to handle situations where certain joints are temporarily occluded. Additionally, we also use information specific to each human skeleton to improve the accuracy of the 3D estimation and predict the pose of occluded joints. We evaluate our framework on a representative 3D human pose estimation dataset, the MPI-INF-3DHP Dataset [4], and present comparative results with other state-of-the-art methods. The contributions of this paper can be summarised as follows:

- to propose a multi-camera video-based 3D human pose estimation algorithm;
- to predict accurately the position of occluded 3D joints;
- to compare with other 3D human pose estimation state-of-the-art approaches.

II. RELATED WORK

Human pose estimation refers to the process of determining the configuration of a human body within a given scene. This can be achieved through two primary methods: skeleton-based [5], [6], [7], [1] and mesh-based approaches [8], [9], [10], [11], [12], [13]. A more simplified representation of the human body is provided by the skeleton-based method, which views it as a collection of joints connected by links. Alternatively, the mesh-based approach approximates the human body as a triangular mesh, providing a more detailed albeit more complex representation.

*This work was supported by the Foundation for Science and Technology (FCT) under the grant 2021.04792.BD and by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020.

[1]Daniela Rato, Miguel Oliveira and Vítor Santos are with the Institute of Electronics and Informatics Engineering of Aveiro and the Department of Mechanical Engineering, Campus Universitário de Santiago, University of Aveiro, 3810-193 Aveiro, Portugal, and with the LASI – Intelligent System Associate Laboratory, Portugal danielarato@ua.pt, mriem@ua.pt, vitor@ua.pt

[2]Angel Sappa is with the Computer Vision Center, Edificio O, Campus UAB, Barcelona, 08193 Bellaterra, Spain, and with the ESPOL Polytechnic University, Guayaquil, Ecuador asappa@cvc.uab.cat asappa@espol.edu.ec

[3]Bogdan Raducanu is with the Computer Vision Center, Edificio O, Campus UAB, Barcelona, 08193 Bellaterra, Spain bogdan@cvc.uab.cat

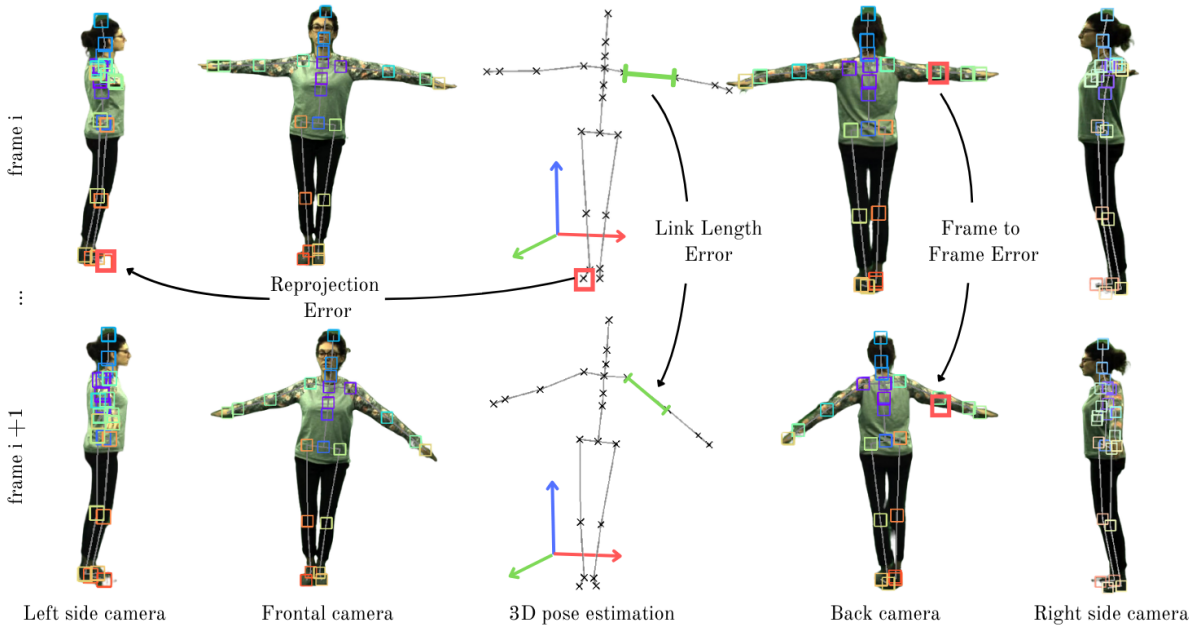

Fig. 1. Schematic representation of the proposed approach. The main framework is divided into three key components: the reprojection component, the link length component, and the frame-to-frame component. The reprojection component aims to minimize the distance between the projection of the 3D joints and their 2D detections. The link length component aims to uniformize the tridimensional link length in all the frames. And the frame-to-frame component helps predict the position of occluded joints using the position of the same joint in adjacent frames.

In the context of 3D human pose estimation, the goal is to infer the spatial pose of the human body in three-dimensional space, typically making use of image data. Skeleton-based 3D human pose estimation techniques are commonly categorized into two main methodologies: direct estimation [1], [14], [15] and 2D to 3D lifting [7], [5], [16], [17]. Direct estimation techniques aim to predict the complete 3D pose without leveraging pre-existing 2D pose information, with the capacity of models to learn the spatial representation of the human directly from visual cues [18]. On the contrary, 2D to 3D lifting approaches operate under the assumption of prior knowledge of 2D human poses captured from one or more viewpoints, focusing exclusively on the transition from 2D to 3D space [18].

Choi et al. generate multiple 3D pose candidates for each identical 2D keypoint and use a diffusion-based framework to effectively sample diverse 3D poses from an off-the-shelf 2D detector (Choi, 2023). The algorithm outputs 3D joint errors of around $50\,\mathrm{mm}$ on the Human3.6M dataset [19]. Martinez et al. use a simple deep feedforward network to lift 2D keypoints to 3D joint positions [16]. Using the 2D ground-truth joint values, the algorithm outperforms state-of-the-art methods by $30\,\%$ on Human3.6M [19], and using off-the-shelf 2D detectors, the algorithm obtains around $60\,\mathrm{mm}$ of 3D joint error. Pavvlo et al. estimate 3D poses in videos using a fully convolutional model based on dilated temporal convolutions over 2D keypoints [17], and the authors also propose a semi-supervised training method that back-projects the estimated 3D poses back into 2D keypoints. 3D joint errors round $50\,\mathrm{mm}$ when testing on the Human3.6M dataset [19].

Although learning-based methods are predominant in the human pose estimation field [7], [16], [20], [21], [22], [23], [24], [25], [26], they can sometimes fall short due to domain gaps or varying intrinsic parameters [5]. Optimization methods usually use a frame-to-frame independent analysis, which does not allow for understanding the context of the entire problem and usually cannot achieve the performance of learning-based methods, but its case-by-case analysis can be an advantage in understanding scenarios where learning approaches struggle. Some optimization methods use A Skinned Multi-Person Linear Model (SMPL) [10] mesh models as a baseline for optimization [8], [9], [11]. Bogo et al. fit the SMPL mesh model to the pre-obtained 2D joints by optimizing the distance between the project 3D model joints and the previously detected 2D joints [8]. Results presented on the HumanEva dataset [27] average $80\,\mathrm{mm}$ of 3D joint error. Choutas et al. use learned optimization and propose a method based on the Levenberg-Marquardt algorithm [9]. Results on the 3DPW dataset [28] are around $50\,\mathrm{mm}$ of 3D joint error. Müller et al. propose an optimization method focused on the detection of human self-contact that includes contact constraints [11]. 3D joint errors are around $100\,\mathrm{mm}$ on MPI-INF-3DHP [4] and $85\,\mathrm{mm}$ on 3DPW [28]. Regarding skeleton-based human pose estimation optimization techniques, Jiang et al. propose a zero-shot diffusion-based optimization method focused on cross-domain and in-the-wild 3D human pose estimation also from 2D keypoint values [5]. 3D joint errors on 3DPW [28] are around $70\,\mathrm{mm}$, on Human3.6M [19] are around $50\,\mathrm{mm}$ and on MPI-INF-3DHP [4] are around $70\,\mathrm{mm}$.

In conclusion, 3D human pose estimation, as explored

through various methodologies such as skeleton-based and mesh-based approaches, faces inherent challenges with errors typically ranging between 50 and 70 millimetres. Despite significant progress, there is still room for improvement in achieving higher accuracy and robustness. The challenge of occluded joints remains an issue in human posture estimation. Existing solutions, including learning-based methods and optimization techniques, exhibit limitations in handling diverse scenarios.

Unlike the previously mentioned optimization methods, our approach stands out by strategically incorporating temporal information and relying on unique characteristics specific to each individual human skeleton. By using temporal patterns and personalised features, our algorithm not only enhances the precision of 3D pose estimations but is also successful in predicting the poses of occluded joints. This is a distinctive feature compared to traditional optimization methods, which often lack the capacity to adapt to dynamic temporal changes and handle occlusions effectively.

III. METHOD

The proposed video-based optimization approach uses the least-squares method to determine the 3D position of each joint in a predefined skeleton. This approach uses 2D to 3D lifting, meaning that it assumes that the 2D poses are known in a certain image (in pixels) and also requires the extrinsic parameters of the cameras in the system.

Besides the information from the 2D keypoints, our approach uses temporal information that helps detect occluded joints while trying to predict the movement of the occluded joint by extrapolating from frames where that joint was previously seen. It integrates knowledge from the anatomical configuration of the human skeleton by aiming to homogenize the three-dimensional length of each skeletal link across all frames. More precisely, the objective is to ensure uniformity in the three-dimensional length of each link across the entire sequence of frames.

The optimization problem is solved using a nonlinear least squares method. This algorithm aims to find the parameter vector θ that minimises the sum of squared residuals $Q(\theta) = \sum_{i=1}^{n} r_i^2$, where $r_i = y_i - f(x_i, \theta)$ represents the difference between the observed data y_i and the model prediction $f(x_i, \theta)$, where x_i represents the input features or predictors used in the model to make predictions. The Jacobian matrix $\mathbf{J}$ is a key component, containing partial derivatives of the residuals with respect to the parameters: $J_{ij} = \frac{\partial r_i}{\partial \theta_j}$. The NLS method iteratively updates the parameter estimates using the linearized system of equations represented by eq. 1

$$\theta_{k+1} = \theta_k - (\mathbf{J}^\intercal \mathbf{J})^{-1} \mathbf{J}^\intercal r \ , \tag{1}$$

where $\mathbf{J}^\intercal$ is the transpose of the Jacobian matrix, and r is the vector of residuals. This update rule adjusts the parameter estimates in the direction that reduces the sum of squared residuals, and the process is repeated iteratively until convergence is achieved. The final θ represents the optimal parameter that provides the best fit of the nonlinear model to the observed data. The optimization is handled as a sparse problem because the parameters do not influence all of the residuals, and it is solved with the trust region reflective algorithm [29], which is a suitable method for large bounded sparse problems.

A. Objective Function

The objective function is used as a baseline for our optimization process. This function aims to determine the 3D coordinates, X, Y, Z, for each joint j and each frame f, where the error is minimum. There are three main error components in the objective function: the reprojection error, e_{rp}; the link length error, e_{ll}; and the frame-to-frame error, e_{ff}. It is defined as:

$$f_{obj} = \underset{(X,Y,Z)_{j,f}}{\arg\min} \sum_j e_{ll} + \sum_j \sum_f (e_{rp} + e_{ff}) \tag{2}$$

The error components are detailed as follows.

1) Reprojection Residuals: The reprojection error is computed between the calculated joint 3D coordinates. It is defined as:

$$e_{rp} = \left\| \text{proj}\Big((X,Y,Z)_{j,f}, \lambda_i\Big) - d_{j,f,i} \right\| \cdot c_{j,f,i} \ , \tag{3}$$

where $(X,Y,Z)_{j,f}$, for each joint j, and each frame f, projected to the frame in question for each camera image i, with the intrinsic and extrinsic parameters λ_i, and the coordinates 2D detection of that joint, $d_{j,f,i}$ in pixel. The confidence value for each joint in each frame and camera $c_{j,f,i}$ is also used as a multiplying factor for the reprojection residuals, highlighting joints that have high confidence detection values. The confidence value is provided by the 2D detector.

2) Link Length Residuals: Eq. 4 expresses the error to be optimized for the link length portion of the objective function. It is defined as:

$$e_{ll} = \sqrt{\frac{\sum (l_{j,f} - \bar{l})^2}{F}} \tag{4}$$

The link length residuals calculate the tridimensional length of the skeleton links for each frame. For each link l, the average link length $\bar{l}$ for all the frames needs to be calculated. The residual for a link equals the standard deviation of that link, meaning the root square of the sum of the square of the difference between the link length for each frame and the average link length for that link divided by the total number of frames F. This forces all the frames to have the same link length for a determined link. This should be true because the links of the human body are rigid and, for the same human, will not change its tridimensional length.

3) Frame to Frame Residuals: Eq. 5 expresses the frame-to-frame error. It is defined as:

$$e_{ff} = \begin{cases} \|(X,Y,Z)_{j,f} - (X,Y,Z)_{j,f-1}\| & \text{if } j \text{ occluded} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where the error for a joint j is the Euclidean distance between the coordinates of that joint in frame f and the coordinates of that joint in the previous frame (for all frames but the

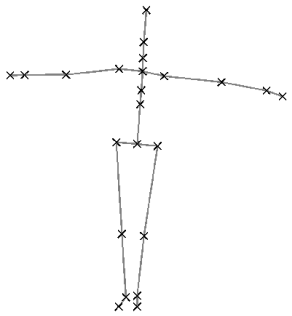

Fig. 2. Representation of the skeleton used in the experiments.

first). This residual is only activated for occluded joints or when the current distance between joints is bigger than a given threshold; in this case, we use $150\,\mathrm{mm}$ since it allows the baseline of the algorithm to be mainly the reprojection function, only activating the frame-to-frame residuals when the prediction of the joint position is very different from the previous frame. Since the standard value for the 3DPCK metrics explained in Section IV-A and proposed in [4] also considers a correctly identified keypoint as a keypoint within the $150\,\mathrm{mm}$ threshold, this value also aligns with that. This residual assumes that in consecutive frames, and with relatively high frame rates, the position of a given joint cannot be significantly different.

Adding the frame-to-frame error residual, $e_f f$, to the optimization improves 3D detection quality by ensuring temporal coherence and spatial continuity. It encourages smooth transitions of joint positions between frames, helping maintain the realism of motion and predict occluded joints based on their last known positions. By combining temporal and spatial information, the algorithm becomes more robust to noise and occlusions, leading to more accurate and reliable 3D pose estimations.

IV. EXPERIMENTS AND RESULTS

This section presents the results and experiments developed to prove the accuracy of our method. We present comparative results and three separate experiments where we evalute the impact of the initial 2D detection noise, the number of occluded joints and the number of cameras.

A. Dataset and Metrics

For all the tests and evaluations presented in this manuscript, we use a 3D human pose estimation dataset, the MPI-INF-3DHP dataset [4], which provides several scenarios, 8 cameras with different points of view, and 2D and 3D ground-truth labels. For the context of the following experiments, we use only 4 out of the 8 available cameras. As the skeleton model, we use the MPI-INF-3DHP skeleton with 23 joints represented by Fig. 2. To assess the performance of the algorithm, we use established evaluation metrics commonly utilized in the field of 3D human pose estimation: MPJPE (Mean Per Joint Position Error) and 3DPCK (3D Percentage of Correct Keypoints). The MPJPE is represented by eq. (6),

$$\text{MPJPE} = \frac{\sum_{f=0}^{F}\sum_{j=1}^{J}\|P_{j,f} - P_{gt}\|}{F \cdot J}\ , \tag{6}$$

where $P = (X, Y, Z)$. The MPJPE quantifies the average error per joint position, determined by the Euclidean distance between the ground truth (X_{gt}, Y_{gt}, Z_{gt}) and the estimated joint positions $(X_{j,f}, Y_{j,f}, Z_{j,f})$ for each joint j and frame f, divided by the total number of frames F and the total number of joints, J.

The 3DPCK is represented by eq. (7),

$$\text{3DPCK} = \frac{J_{correct}}{J} \times 100\ , \tag{7}$$

where $J_{correct}$ is the number of correct joints and J is the total number of joints. The 3DPCK measures the percentage of correctly identified 3D keypoints. A detection is a true positive if the Euclidean distance between the estimated joint position and its corresponding ground truth falls within a specified threshold. In alignment with standard practices from other state-of-the-art approaches, we applied a threshold value of $150\,\mathrm{mm}$ like suggested in [4]. These metrics collectively provide a comprehensive evaluation of the accuracy of the algorithm in estimating 3D human poses.

B. Comparative Analysis

Table I presents a comparative analysis of the proposed methodology with other state-of-the-art algorithms on the MPI-INF-3DHP dataset [4]. The evaluated algorithms include Bouazizi et al. [22], Bouazizi et al. [23], Kocabas et al. [24], Jiang et al. [5], Yu et al. [25], and Pavvlo et al. [17] as benchmark references, which have been discussed in the section II.

We limit our comparison to 2D to 3D lifting state-of-the-art methodologies, as these are the most similar to our approach. For our methodology, we use two variants of the approach with two different errors (10 and 20 pixels) in the 2D detected keypoints, where we vary the magnitude of the noise added to the ground-truth 2D detections. This is done to emulate the fact that 2D detectors are not perfect. In any case, it is important to note that state-of-the-art 2D detectors consistently perform better than 10 pixels.

Our algorithm outperforms all other methodologies in MPJPE values. The MPJPE stands at 18.06 for a 10-pixel error scenario and increases to 36.40 for a more challenging 20-pixel error scenario. These results prove the robustness and accuracy of the proposed approach, demonstrating its efficacy in achieving accurate 3D human pose estimation under conditions with varying degrees of 2D pixel errors.

Our approach exhibits state-of-the-art performance for several key reasons, setting it apart from existing methodologies. One factor is the used video-based approach. Our algorithm improves the simple reprojection function by taking into account how human poses change between frames by using the frame-to-frame approach. Utilizing information from multiple frames enables our algorithm to gather context and

TABLE I

COMPARATIVE ANALYSIS WITH OTHER 2D TO 3D LIFTING STATE-OF-THE-ART METHODOLOGIES ON THE MPI-INF-3DHP [4] DATASET.

Methodology	Optimization	Multi-view	Video	MPJPE ↓
Kocabas et al. [24]		✓		109.0
Bouazizi et al. [22]		✓	✓	93.0
Pavvlo et al. [17]			✓	86.6
Bouazizi et al. [23]		✓		65.9
Jiang et al. [5]	✓			55.2
Zhao et al. [26]			✓	27.8
Yu et al. [25]			✓	27.8
Ours (20px)	✓	✓	✓	36.4
Ours (10px)				18.1

refine estimations by considering the consistency of pose configurations across frames. This not only improves the robustness but also enhances the ability to handle dynamic and occluded complex movements.

Furthermore, our approach relies on the unique characteristics of each human skeleton. The link length component of the objective function (see section III-A.2) will estimate different link lengths for each human. By tailoring the optimization process to the individual characteristics of skeletons, our algorithm achieves a higher degree of precision. This customized optimization contributes significantly to mitigating errors and enhances the overall accuracy of 3D pose estimations by guaranteeing that the link length does not change in different frames.

The proposed approach, using 2D detection with a 10 pixel error, gives the best results overall. It is followed by Yu et al. [25] and Zhao et al. [26]. These good results may be related with the fact that all the mentioned approaches are video-based, which allows to improve 3D poses by leveraging information from all the frames.

C. Impact of 2D Joint Detection Error

The following experiment evaluates the impact of the 2D joint detection pixel error on the 3D joint results. The test set consisted of 500 frames (roughly 20 s) from the MPI-INF-3DHP [4]. To evaluate the impact of the quality of 2D keypoints detection on the outcome, we used the 2D ground-truth keypoints values as input. We added a systematic absolute pixel error to all the keypoints in a random direction. The added error varied from 0 to 100 pixels. The test dataset did not include any occluded joints, and the optimization used 4 of the 8 available cameras.

Fig. 3 shows a plot of the evolution of the indicators MJPJE and 3DPCK indicators with the increase of the 2D joint detection error. It presents an analytical view of the correlation between 2D joint detection errors and the subsequent impact on the detection of human 3D poses. The MJPJE, illustrated by the ascending orange curve, shows a gradual increase in millimeters as 2D joint errors in pixels rise. This positive correlation underscores the sensitivity of 3D pose predictions to inaccuracies in 2D joint localization. The trend suggests that as the precision of 2D joint detection diminishes, the accuracy of predicting the spatial positions of joints in the 3D space becomes compromised. The 3DPCK, represented by the descending blue curve, reflects the percentage of accurately estimated 3D keypoints in relation to increasing 2D joint errors. The decline in 3DPCK underscores a more pronounced sensitivity to higher 2D detection errors.

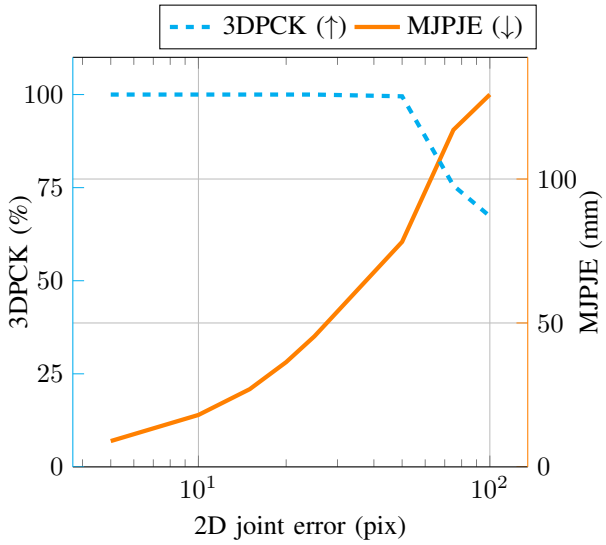

Fig. 3. Impact of 2D joint detection error in detection of human 3D poses. A detailed explanation of the indicators can be found in section IV-A.

Nevertheless, the algorithm demonstrates robust performance up to a 2D joint detection error of 25 pixels. Within this range, both the MJPJE and 3DPCK show favorable behaviors. The MJPJE remains relatively low, up to 50 mm, indicating an accurate prediction of 3D joint positions, while the 3DPCK remains consistently high, demonstrating a high percentage of correctly estimated keypoints.

Up to the 25-pixel threshold, the algorithm effectively compensates for minor inaccuracies in 2D joint detection, showcasing resilience to moderate 2D detection errors. Beyond 25 pixels, however, the performance trends diverge, with both MJPJE and 3DPCK responding more sensitively to increasing 2D joint errors.

D. Impact of Occlusions

This subsection aims to determine the robustness of the algorithm to occlusions. For this, we designed two different experiments: one where we randomly occluded an increasing number of 2D joints that served as input for optimization; in the second experiment, we occluded the same joint in all cameras for a period of time and evaluated how precisely the position of that joint was being predicted.

1) Random Occlusions: This experiment aims to assess the influence of 2D joint occlusions on the prediction of 3D joint values. The test set contains 500 frames from the MPI-INF-3DHP dataset [4]. To simulate occlusions, we systematically remove keypoints from the ground-truth 2D keypoints. The quantity of deleted keypoints per frame and point-of-view ranged from 0 to 15. Additionally, an

absolute error of 10 pixels was added to each ground-truth keypoint value. This controlled variation in occlusion and error scenarios allows for a comprehensive evaluation of the robustness of the algorithm under realistic conditions.

Fig. 4 shows the results obtained from the experience mentioned earlier. Comparative analyses evaluate the efficacy of a simple reprojection function, optimized through the least squares method (plots with the tag $_{reproj}$), against our proposal.

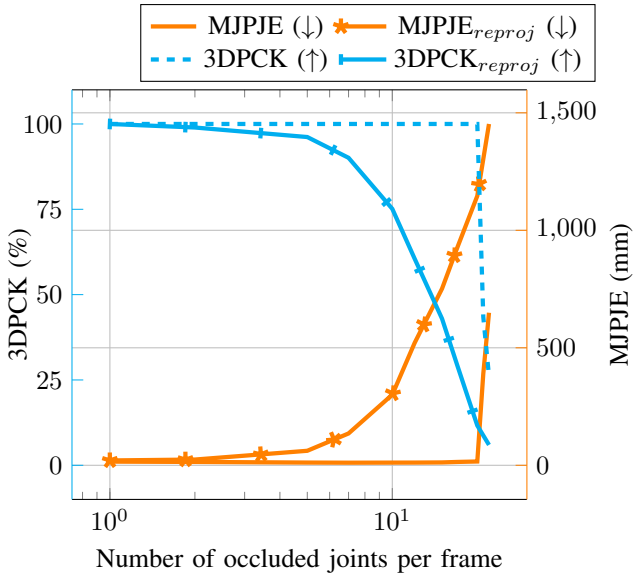

Fig. 4. Impact of occluded 2D joints in the detection of human 3D poses, where simple lines represent the performance of our proposal and marked lines represent the performance of an optimization using only the reprojection of 3D coordinates to 2D images as the objective function.

In the MJPJE plot, our algorithm (denoted as "MJPJE") consistently outperforms the reprojection function ("MJPJE$_{reproj}$"), particularly with higher occurrences of occluded joints. The stability in MJPJE values for our methodology signifies very good precision in joint position estimation, even in highly occluded scenarios, with the performance of the algorithm maintaining resilience.

In the 3DPCK plot, our algorithm (denoted as "3DPCK") shows higher correctness percentages, even with a substantial number of occluded joints. In contrast, the reprojection approach ("3DPCK$_{reproj}$") evidences a decline in correctness with escalating occluded joints, highlighting the efficacy of our algorithm in sustaining keypoint accuracy under challenging occlusive scenarios and significantly improving the performance of the reprojection function.

In conclusion, the figure successfully proves that our proposal greatly improves the reprojection function, particularly in scenarios where joints are occluded. The observed stability in performance highlights the robustness and potential of our solution, enhancing its utility for precise 3D human pose estimation within intricate real-world scenarios.

2) Consistent Occlusions: This experiment evaluates the impact of occluded joints in all points of view for a period of time in 3D joint poses. For this, the dataset used, for each

TABLE II

IMPACT OF THE NUMBER OF CAMERAS IN THE MPJPE (mm) AND 3DPCK (%) IN THE MPI-INF-3DHP DATASET.

# cameras	MPJPE ↓	3DPCK ↑
2	47.4	96.2
3	13.8	100
4	11.6	100

10 normal frames, had 5 frames where the left elbow was occluded in all cameras. The dataset also had 10 pixels of 2D joint error in every joint. The obtained MJPJE was $18.08\,\mathrm{mm}$ and $100\,\%$ 3DPCK. Regarding the left elbow, the 3D joint error was $20.30\,\mathrm{mm}$, which is slightly above average but proves that the position of that joint was well predicted by the proposed approach.

E. Impact of Number of Cameras

This experiment intends to determine the impact of the number of cameras used to optimize the quality of the 3D human poses. The test set contains 500 frames from MPI-INF-3DHP [4]. We chose a test set with frames with 5 randomly occluded joints and 10 pixels of error.

Table II shows the results obtained when calibrating the same dataset with a varying number of cameras. We can conclude that even with the constraint of optimizing only 2 cameras, the performance remains robust. The MPJPE is $47.4\,\mathrm{mm}$, indicating good precision in estimating joint positions. The 3DPCK is registered at $96.2\,\%$, a very good accuracy considering the limited number of cameras. As the number of cameras increases to 3 and 4, the precision further improves, as is evident in the reduced MPJPE values (13.8 and 11.6, respectively) and the $100\,\%$ 3DPCK accuracy. This analysis emphasizes the resilience of the algorithm, demonstrating good performance even in scenarios where only two cameras are utilized.

From a practical standpoint, these results have significant implications for real-world applications. The ability to achieve accurate 3D human pose estimation with only two cameras makes the system more feasible and cost-effective for deployment in real world scenarios, where the number of available cameras might be limited. The method's robustness in low-camera scenarios enhances its versatility and potential for broader adoption across different fields and use cases. Additionally, the point of view of the cameras also influences the quality of detection, as optimal camera placement can further enhance the accuracy and reliability of the system.

V. CONCLUSION AND FUTURE WORK

In conclusion, our comprehensive experimentation and analysis have proved the robustness and efficacy of our proposed algorithm for 3D human pose estimation. The conducted experiments, including the impact of 2D joint detection error, occlusions, and varying numbers of cameras, have proved the resilience of the algorithm under diverse conditions. Additionally, the algorithm showed excellent performance even in scenarios with significant 2D joint

error and exhibited a high degree of robustness in handling several occluded joints. In our future works, we aim to evaluate the robustness and generalisation capabilities of our algorithm by testing it on in-the-wild 3D human pose datasets, such as the 3DPW dataset [28]. This will provide insight into the performance of the algorithm under more diverse and dynamic real-world conditions. Adapting to in-the-wild datasets poses challenges such as dealing with varied lighting, backgrounds, and increased occlusions, necessitating enhanced robustness and generalization strategies, including advanced data augmentation techniques and more sophisticated occlusion handling mechanisms. Additionally, we plan to improve the practical applicability of this method by working towards real-time implementation. Our proposed strategy involves optimizing the last n frames, with n yet to be defined (for instance, 5 frames), instead of optimizing the entire set simultaneously. The optimization process will be recursive, utilizing the last optimized position as the initial guess for the subsequent frame. This will allow the application of the algorithm in our robotics laboratory, where real-time is essential. The adapted algorithm will be crucial in constraining the robot's movement around humans, facilitating safer interactions, and enabling future human-robot collaboration tasks within our laboratory environment.

Extending the approach to mesh-based representations could enhance the realism and detail of 3D human pose estimation by capturing intricate body shapes and surface deformations. This could improve pose precision in complex scenarios and offer more accurate reconstructions. However, mesh-based models are computationally intensive and require sophisticated algorithms, making them challenging to integrate with existing skeleton-based datasets and evaluation metrics. Despite these challenges, the potential benefits make this a promising area for future research and development.

REFERENCES

[1] H. M. Clever, A. Kapusta, D. Park, Z. Erickson, Y. Chitalia, and C. C. Kemp, "3D human pose estimation on a configurable bed from a pressure image," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2018, pp. 54–61.

[2] S. Hu, C. Zheng, Z. Zhou, C. Chen, and G. Sukthankar, "Lamp: Leveraging language prompts for multi-person pose estimation," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2023, pp. 3759–3766.

[3] A. Casalino, S. Guzman, A. Maria Zanchettin, and P. Rocco, "Human pose estimation in presence of occlusion using depth camera sensors, in human-robot coexistence scenarios," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2018, pp. 1–7.

[4] 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," in *Fifth Int. Conf. on 3D Vision*. IEEE, 2017.

[5] Z. Jiang, Z. Zhou, L. Li, W. Chai, C.-Y. Yang, and J.-N. Hwang, "Back to optimization: Diffusion-based zero-shot 3D human pose estimation," in *IEEE/CVF Winter Conf. on Applications of Computer Vision*, 2024.

[6] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, "Cascaded deep monocular 3D human pose estimation with evolutionary training data," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020.

[7] J. Choi, D. Shim, and H. J. Kim, "Diffupose: Monocular 3D human pose estimation via denoising diffusion probabilistic model," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2023, pp. 3773–3780.

[8] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *European Conf. on Computer Vision*, 2016.

[9] V. Choutas, F. Bogo, J. Shen, and J. Valentin, "Learning to fit morphable models," in *European Conf. on Computer Vision*, 2022.

[10] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, pp. 248:1–248:16, 2015.

[11] L. Müller, A. A. A. Osman, S. Tang, C.-H. P. Huang, and M. J. Black, "On self-contact and human pose," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021.

[12] D. F. Henning, C. Choi, S. Schaefer, and S. Leutenegger, "BodySLAM++: Fast and tightly-coupled visual-inertial camera and human motion tracking," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2023, pp. 3781–3788.

[13] S. Schaefer, D. F. Henning, and S. Leutenegger, "Glopro: Globally-consistent uncertainty-aware 3D human pose estimation and tracking in the wild," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2023, pp. 3803–3810.

[14] R. A. Güler and I. Kokkinos, "Holopose: Holistic 3D human reconstruction in-the-wild," in *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 10 876–10 886.

[15] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3D people," in *IEEE/CVF Int. Conf. on Computer Vision*, 2021, pp. 11 159–11 168.

[16] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *2017 IEEE Int. Conf. on Computer Vision*, 2017, pp. 2659–2668.

[17] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Conf. on Computer Vision and Pattern Recognition*, 2019.

[18] 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," *ACM Comput. Surv.*, vol. 56, 2023.

[19] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1325–1339, 2014.

[20] B. T. Soroush Mehraban, Vida Adeli, "Motionagformer: Enhancing 3D human pose estimation with a transformer-gcnformer network," in *IEEE/CVF Winter Conf. on Applications of Computer Vision*, 2024.

[21] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *IEEE/CVF Int. Conf. on Computer Vision*, 2023.

[22] A. Bouazizi, U. Kressel, and V. Belagiannis, "Learning temporal 3d human pose estimation with pseudo-labels," in *17th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, 2021, pp. 1–8.

[23] A. Bouazizi, J. Wiederer, U. Kressel, and V. Belagiannis, "Self-supervised 3d human pose estimation with multiple-view geometry," in *2021 16th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2021, pp. 1–8.

[24] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3D human pose using multi-view geometry," pp. 1077–1086, 2019.

[25] B. X. Yu, Z. Zhang, Y. Liu, S.-h. Zhong, Y. Liu, and C. W. Chen, "Glagcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video," in *IEEE/CVF Int. Conf. on Computer Vision*, 2023, pp. 8818–8829.

[26] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "PoseFormerV2: Exploring frequency domain for efficient and robust 3d human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8877–8886.

[27] 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, pp. 4–27, 2010.

[28] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using imus and a moving camera," in *European Conf. on Computer Vision*, 2018.

[29] M. A. Branch, T. F. Coleman, and Y. Li, "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems," *SIAM Journal on Scientific Computing*, vol. 21, pp. 1–23, 1999.