Enhancing 3D Human Pose Estimation: A Novel Post-Processing Method

Elham Iravani[1,2][a], Frederik Hasecke[2][b], Lukas Hahn[2][c] and Tobias Meisen[1][d]
[1]*University of Wuppertal, Gaußstraße 20, Wuppertal, Germany*
[2]*APTIV, Am Technologiepark 1, Wuppertal, Germany*
{*elham.iravani, meisen*}*@uni-wuppertal.de,* {*frederik.hasecke, lukas.hahn*}*@aptiv.com*

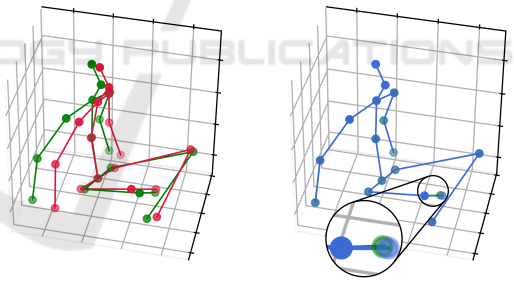
Keywords: Human Pose Estimation, Absolute Pose Estimation, Pose Refinement.

Abstract: Human Pose Estimation (HPE) is a critical task in computer vision, involving the prediction of human body joint coordinates from images or videos. Traditional 3D HPE methods often predict joint positions relative to a central body part, such as the hip. Transformer-based models like PoseFormer (Zheng et al., 2021), MHFormer (Li et al., 2022b), and PoseFormerV2 (Zhao et al., 2023) have advanced the field by capturing spatial and temporal relationships to improve prediction accuracy. However, these models primarily output relative joint positions, requiring additional steps for absolute pose estimation. In this work, we present a novel post-processing technique that refines the output of other HPE methods from monocular images. By leveraging projection and spatial constraints, our method enhances the accuracy of relative joint predictions and seamlessly transitions them to absolute poses. Validated on the Human3.6M dataset (Ionescu et al., 2013), our approach demonstrates significant improvements over existing methods, achieving state-of-the-art performance in both relative and absolute 3D human pose estimation. Our method achieves a notable error reduction, with a 33.9% improvement compared to PoseFormer and a 27.2% improvement compared to MHFormer estimations.

1 INTRODUCTION

The accurate estimation of human poses from images and videos is a foundational task across diverse application domains, including human-computer interaction, augmented and virtual reality, and healthcare. The principal objective of HPE is to accurately detect and represent human joint positions. However, the majority of HPE algorithms estimate joint positions relative to a joint, which can lead to inconsistencies in some applications. Moreover, these methods may result in skeleton structures with non-symmetrical limb lengths and different body dimensions for the same subject in different frames of a video. Such discrepancies can significantly affect the accuracy and reliability of subsequent applications that rely on precise human pose data.

To address these limitations, we propose a novel post-processing algorithm for refining estimated 3D human joint poses from monocular images. The objective is to enhance the precision of a given joint pose estimation and to ensure a consistent skeletal structure across the entire sequence by maintaining uniform limb proportions. Our method leverages geometrical concepts and 2D joint poses. Inputs to our algorithm include estimated 3D human poses from an existing HPE model, 2D pose data, and optional body dimen-



a) PoseFormer vs. ground truth b) Ours vs. ground truth

Figure 1: Example of refinement on a frame of the Human3.6M dataset, test subject *S9*, action *SittingDown 1*. **a)** (Red): PoseFormer; (Green): ground truth. **b)** (Blue): Ours; (Green): ground truth. (It should be noted that the ground truth and the predicted pose are overlapping for the most part after the refinement, hence why it's hard to see a difference. Please note the right foot for the two colors, ground truth (green) and prediction (blue) which shows a slight difference.)

[a] https://orcid.org/0000-0003-1961-2130
[b] https://orcid.org/0000-0002-6724-5649
[c] https://orcid.org/0000-0003-0290-0371
[d] https://orcid.org/0000-0002-1969-559X

Iravani, E., Hasecke, F., Hahn, L. and Meisen, T.
Enhancing 3D Human Pose Estimation: A Novel Post-Processing Method.
DOI: 10.5220/0013316600003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 209-220
ISBN: 978-989-758-728-3; ISSN: 2184-4321

sions specifications, collectively supporting a refined, robust output suitable for precision-critical applications.

Our algorithm refines estimated 3D joint poses in camera coordinates, utilizing either prior body dimensions or body dimensions estimated using 2D poses. Summarizing our contributions:

- Body dimensions refinement: We introduce a method that utilizes 2D joint poses and an estimated ground plane to derive the subject body dimensions.
- Estimating absolute 3D poses from root relative poses: Our proposed approach estimates absolute joint poses by moving the skeleton along the root joint's projection ray, minimizing the 2D projection error for the entire skeleton. This enables accurate prediction of 3D poses.
- Main contribution - refining 3D HPE: With the skeleton model modified to include refined body dimensions, enhancing the accuracy of 3D pose estimations by adjusting along the projection rays to match spatial constraints.

Recent advancements in human pose estimation, notably through models like PoseFormer and MHFormer, have driven substantial progress. Our approach achieves considerable improvements over these models. Using prior body dimensions, we achieve a reduction in the average Mean Per Joint Position Error (MPJPE) across all actions of the Human3.6M dataset by 33.9% compared to PoseFormer and 27.2% compared to MHFormer estimations. Furthermore, by refining 3D joint poses using estimated body dimensions, our results show a 7.9% reduction in the average MPJPE compared to PoseFormer and a 6.3% reduction compared to MHFormer. These results underscore the effectiveness of our approach, marking a significant advance in the precision of human pose estimation. Figure 1 also shows an example of how our method improves the 3D HPE.

The paper is structured as follows: In Section 2, we review related work in human pose estimation, focusing on recent advancements in transformer-based models and geometry-based methods. Section 3 details our proposed post-processing algorithm, including body dimensions refinement, absolute pose estimation, and the overall 3D joint pose refinement methodology. In Section 4, we present our experimental setup and evaluation metrics, followed by a comprehensive analysis of the results, highlighting the performance improvements over existing models. Section 5 discusses the implications of our findings and examines the limitations of our approach. Finally, Section 6 concludes the paper and suggests potential directions for future research in refining 3D human pose estimation.

2 RELATED WORKS

Human Pose Estimation (HPE) has achieved significant progress in recent years, largely driven by advances in deep learning. Traditional motion capture systems can generate 3D pose annotations in controlled laboratory settings, but their effectiveness diminishes in natural, uncontrolled environments. One major challenge in the context of monocular RGB images and videos is resolving depth ambiguities. This challenge arises because the conversion of 3D information into 2D projections inherently loses one dimension, making it an ill-posed inverse problem. As noted by (Zheng et al., 2023) in their comprehensive survey, the majority of research in 3D HPE from monocular images or videos has struggled with these depth ambiguities.

In contrast, multi-view approaches encounter challenges in accurately associating multiple viewpoints. Some studies have sought to overcome these limitations by incorporating additional sensors, such as depth sensors, inertial measurement units (IMUs), and radio frequency devices. (Yu et al., 2018; Kadkhodamohammadi et al., 2017; Zhi et al., 2020) However, these approaches often constrained by cost and the requirement for specialized hardware.

Additionally, deep learning models in this field tend to rely heavily on large, diverse datasets and are sensitive to the conditions of the data collection environment. These constraints highlight the ongoing need for more robust, adaptable methods in 3D HPE research (Zheng et al., 2023). The advent of deep learning revolutionized HPE, especially through the application of convolutional neural networks (CNNs).

In 2D HPE, a seminal work by Toshev and Szegedy, introducing DeepPose (Toshev and Szegedy, 2014) utilizes a cascade of CNNs to predict human poses. This approach marked a significant leap in accuracy by leveraging the hierarchical structure of CNNs to learn feature representations at multiple scales. Subsequently, several state-of-the-art methods have been developed, establishing new benchmarks in monocular 2D HPE. OpenPose (Cao et al., 2017) introduced Part Affinity Fields (PAFs), which encode the location and orientation of limbs, thereby enabling real-time multi-person pose estimation. The High-Resolution Network (HRNet) (Sun et al., 2019) maintains high-resolution representations throughout the network, significantly enhancing pose estimation accuracy by integrating high-resolution feature maps

with multi-scale information. The Cascaded Pyramid Network (CPN) (Chen et al., 2018), addresses challenges such as occluded and invisible keypoints by employing a two-stage process: GlobalNet for coarse prediction and RefineNet for refining hard keypoints.

Similarly, the architecture of CNNs also facilitate 3D Human Pose Estimation (HPE). PoseNet (Martinez et al., 2017) employed a straightforward yet effective approach using a fully connected network on detected poses, demonstrating competitive performance and underscoring the importance of efficient architectural design. In addition to traditional convolutional networks, transformer architectures (Hassanin et al., 2022; Li et al., 2022a; Zhang et al., 2022; Zheng et al., 2021; Li et al., 2022b; Zhao et al., 2023) have been employed to model spatio-temporal correlations in 3D human pose estimation. PoseFormer (Zheng et al., 2021), introduced by Zheng et al., employs a spatial-temporal transformer to model human joint relationships within frames and temporal correlations across frames. Building upon this, (Li et al., 2022b) proposed MHFormer, a Multi-Hypothesis Transformer that generates multiple plausible pose hypotheses to address depth ambiguity and self-occlusion.

(Li et al., 2023) introduces the Pose-Oriented Transformer (POT), which incorporates a pose-oriented self-attention mechanism and distance-related position embeddings to explicitly model the interactions between body joints based on their hierarchical structure. Furthermore, they present an Uncertainty-Guided Refinement Network (UGRN), which refines initial 3D pose predictions by considering the estimated uncertainty of each joint, employing an uncertainty-guided sampling strategy and self-attention mechanism.

Geometry-based methods have also been instrumental in advancing HPE from monocular cameras, as they apply geometric constraints and leverage 3D information to improve pose estimation accuracy. Integral Pose Regression (Sun et al., 2018) introduced a method that directly regresses joint coordinates in 3D space by leveraging geometric constraints within a deep learning framework, unifying 2D heatmaps and 3D joint locations into a single representation. Geometry-Aware Methods (Kocabas et al., 2020) developed a model that leverages 3D human body models and optimizing pose estimation by aligning 2D poses with 3D model projections, thereby preserving the geometric consistency of the predicted poses. Additionally, Weakly-Supervised Learning (Rhodin et al., 2018) introduced a weakly-supervised approach that applies geometric constraints from synchronized multi-view images, enhancing 3D pose prediction even with limited labeled data.

While deep learning methods have demonstrated remarkable capabilities in various domains, we observe a significant gap in incorporating physical constraints to ensure realistic and consistent outcomes. This limitation has also been noted by (Zheng et al., 2021) and (Li et al., 2022b), particularly in the context of addressing challenges like complex poses and occlusions. Therefore, in this work, we aim not to further enhance the already well-established strengths of deep learning Transformer methods but rather to tackle their apparent weaknesses in handling these challenges. Geometric methods, which leverage spatial relationships between 2D joint locations to infer 3D poses, offer a complementary perspective that can help address these limitations. By integrating such approaches, we aim to achieve a more robust framework capable of addressing the inherent complexities of real-world scenarios.

3 METHOD

The primary objective of our method is to improve the accuracy of estimated 3D human joint poses through a post-processing approach that leverages geometric principles and 2D pose estimation. As shown in Figure 2, this process incorporates estimated 3D HPE (from existing algorithms), 2D HPE, and optional body dimensions. For effective refinement of 3D pose estimation, the algorithm requires 3D joint positions in camera coordinates. Additionally, it can incorporate either prior body dimensions, derived from subject-specific knowledge, or estimated body dimensions, approximated using 2D HPE. Our approach comprises three main phases: body dimensions refinement, absolute pose estimation, and 3D joint pose refinement, each of which is discussed in detail in the following subsections.

3.1 Body Dimensions Refinement

For improved absolute pose estimation using 2D joint poses and enhanced joint pose refinement, it is advantageous to obtain accurate body dimensions, enabling more precise adjustments. The input requires estimated 3D poses that provide initial body dimension estimates. However, as noted earlier, these body dimensions - when generated by current HPE methods - are frequently inaccurate and can vary across frames for a same subject. Additionally, they often result in asymmetrical limb lengths within the skeletal structure.

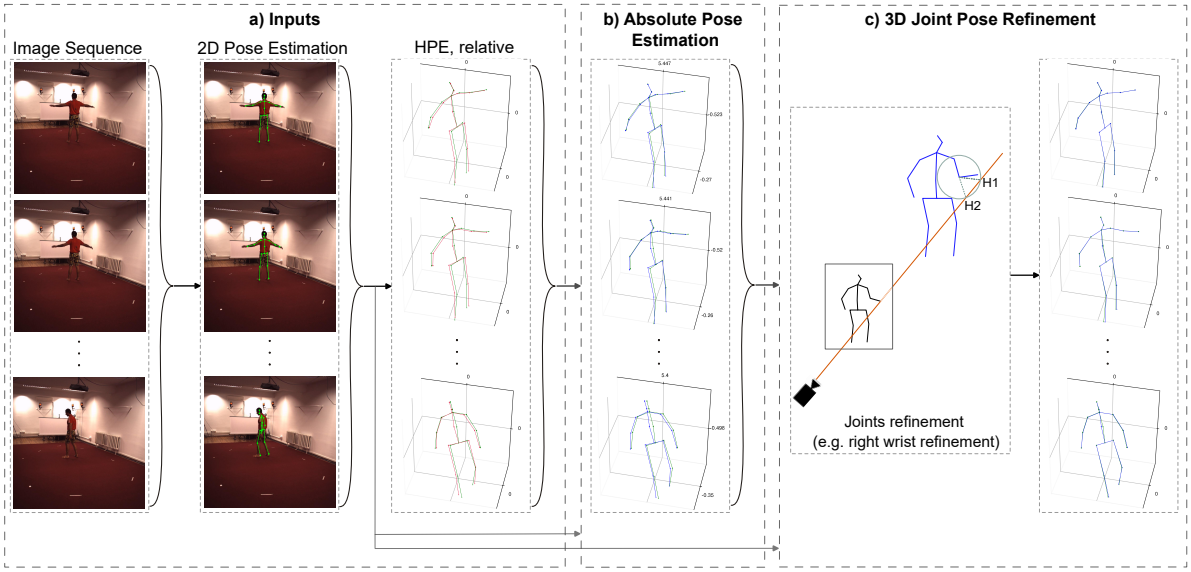


Figure 2: a) Images and 2D joint poses which are used by a 3D HPE method to provide 3D hip relative joint poses for our method. b) In post-processing phase, it estimates joints poses in camera coordinates by utilizing 2D joint poses. c) By having the joint poses in camera coordinate and 2D joint poses, our method will refine the joint poses. (Green): Ground truth; (Red): PoseFormer HPE; (Blue): Our method results. (Where the hip is in (0, 0, 0), it means the result shows in hip-relative).

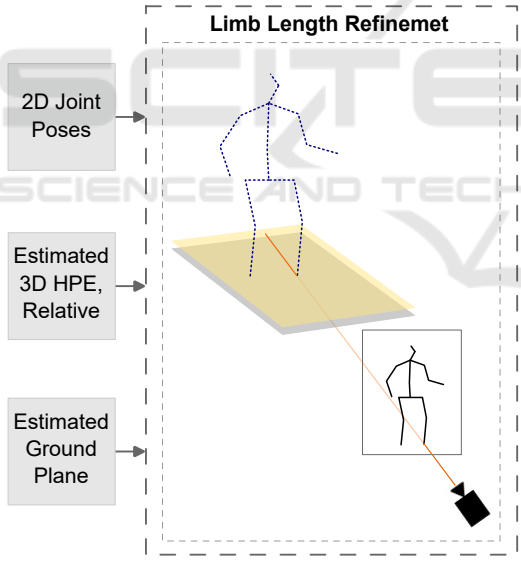


Figure 3: Body dimensions refinement uses 2D joint poses and estimated 3D HPE and ground plane. Camera ray, in this image for the right foot, intersected with a plane (yellow), which is passing through the right foot 3D position and parallel to the ground plane (gray), starts the body dimensions refinement procedure.

In 3D HPE input, the initial 3D poses are provided in hip-relative coordinates. In the world coordinate system, the ground plane is defined by $z = 0$. This plane is then transformed into the camera coordinate system, denoted as $plane_{ground\_camera}$. For each joint i (specifically the *Foot* joints in this context), the undistorted 2D position on the image plane is represented by $p_i^{2D} = (u_i, v_i)$. The corresponding ray in the camera coordinate system can be expressed as:

$$r_i : \begin{cases} z_i > 0, \\ x_i = (u_i - c_x) z_i / f_x, \\ y_i = (v_i - c_y) z_i / f_y \end{cases} \quad (1)$$

where c_x, c_y, f_x, f_y are the camera parameters. In order to determine 3D position of the *Foot* joint, as illustrated in Figure 3, the ray r_i is intersected with a plane $plane_{Foot}$, which is parallel to the ground plane in the camera coordinate system. Assuming $plane_{Foot}$ (the yellow plane shown in Figure 3) is positioned at a height $z = h$ in the transformed camera coordinates, where h represents an approximation of the *Foot* joint distance from the ground plane (the gray plane shown in Figure 3), the intersection point $p_{Foot} = (x_{Foot}, y_{Foot}, z_{Foot})$ can be determined by solving the following equation:

$$plane_{Foot} : Ax_i + By_i + Cz_i = 0; \quad \text{where} \begin{cases} x_i = (u_i - c_x) z_i / f_x, \\ y_i = (v_i - c_y) z_i / f_y \end{cases} \quad (2)$$

Here, the coefficients of A, B, C are defined from the $plane_{Foot}$. Subsequently, the entire set of body joint poses from the initial frame is transformed into the camera coordinate system, resulting in new positions $P_{camera} = \{p_i^{camera} \mid i = 1, \ldots, N\}$. Each joint p_i^{camera}

is then further refined by aligning it along its projection ray (r_i) and following the vector direction from its initial parent joint. The line for each joint is defined by the vector vec_i, which extends from the initial parent joint $p^{camera}_{parent(i)}$ to the initial child joint $p^{camera}_{child(i)}$. The adjusted joint position is found by knowing that the adjusted child joint $p^{adjusted}_{child(i)}$ lies along vector vec_i direction from the adjusted parent joint $p^{adjusted}_{parent(i)}$. The intersection of this line with r_i ensures the joint position aligns accurately with 2D projection constraints:

$$L_i(t) = p^{adjusted}_{parent(i)} + t vec_i \tag{3}$$

The intersection point $p^{adjusted}_{child(i)}$ is then calculated:

$$\text{where } \begin{cases} x_i = (u_i - c_x) z_i / f_x, \\ y_i = (v_i - c_y) z_i / f_y \end{cases} \tag{4}$$

where t is determined by solving for the intersection of the line $L_i(t)$ with the projected ray r_i.

3.2 Absolute Pose Estimation

Estimating 3D HPE in camera coordinates is beneficial for accurately capturing spatial relationships and scale, thereby providing a reliable foundation for refining finer details through 2D HPE projections. To improve the accuracy of absolute pose estimation using 2D poses, our method leverages body dimensions. Their accuracy directly impacts the quality of the absolute pose estimation. This approach involves two preliminary steps: refining the body dimensions and adjusting the skeleton in the camera coordinate system. First, refined body dimensions are calculated following the methodology described in Section 3.1. These refined lengths are then used to adjust the dimensions of the skeleton in the relative pose, modifying each limb segment and thereby updating the 3D joint positions accordingly. The absolute pose is estimated by translating the entire skeleton along the hip joint's projection ray to minimize the 2D projection error. The objective is thus to minimize discrepancies between the projected 2D joint poses and the observed 2D joint positions.

Utilizing the transformation matrix from the previous frame can help improve the accuracy of predictions in camera coordinates. Additionally, inaccurate estimations from the input data may sometimes lead to premature convergence, resulting in a less precise pose estimation. To address this, when processing video input for 3D HPE refinement, the previous frame is also used for absolute pose estimation to enhance continuity and accuracy.

3.3 3D Joint Pose Refinement

As it illustrates in part c of Figure 2, our approach uses undistorted 2D poses to refine the joint poses by aligning them more closely with 2D projection rays in camera coordinates. Most current models that predict 3D poses estimate them in relative joint poses, which can, as stated, result in asymmetrical skeletons and inconsistent body dimensions across frames. Thus, our initial step is to estimate body dimensions or use predefined body dimensions as input. This enables a more accurate absolute pose estimation, setting a strong foundation for the subsequent refinement phase.

Following the adjustment of 3D poses using new body dimensions, absolute poses are estimated as described in Section 3.2. This results in a set of 3D joint poses in camera coordinates, which are further refined by applying the updated body dimensions. To continue the refinement, 2D poses are retrieved from undistorted images. The refinement process begins at the *hip* joint, which serves as the base point. Let P_{hip} represent the 3D coordinates of the *hip* joint, and assume w.l.o.g that J_i is the next joint (e.g., spine) to be refined. Given the link length $L_{\text{hip-to-spine}}$ between the hip and spine, we can establish a 3D ray, r_{spine}, from the camera through the undistorted 2D position of the spine in the image. This ray can be parameterized the same way as in Equation 1.

The 3D ray r_{spine} represents an infinite line in 3D space, where all points on this ray project to the same 2D position as the spine's 2D pose in the image. The potential 3D positions for the spine joint are the points on the aforementioned ray that are precisely $L_{\text{hip-to-spine}}$ units away from the hip joint, yielding at most two possible positions, $(J^{(1)}_{spine}, J^{(2)}_{spine})$. This requires finding the intersection between the sphere centered at P_{hip} with radius of $L_{\text{hip-to-spine}}$ and the ray r_{spine}. In part c of Figure 2, an example is shown illustrating two possible hypotheses for the right wrist refinement. In order to determine the correct position, we select the one closer to the initial estimate of J_{spine} by comparing Euclidean distances:

$$J^{\text{refined}}_{\text{spine}} = \arg \min_{J^{(1)}_{\text{spine}}, J^{(2)}_{\text{spine}}} \| J^{(i)}_{\text{spine}} - J_{\text{spine}} \| \tag{5}$$

This refinement process is applied iteratively to each joint, beginning at the hip and proceeding sequentially through the remaining joints (e.g., spine, shoulders, knees). For each joint J_i, given a body dimension $L_{\text{parent-to-child}}$ and a projection ray r_i, the position is refined using the same method as described above. By applying this method, even a single frame can be refined effectively. However, utilizing frame se-

Table 1: Comparison of the estimated body dimensions error and PoseFormer average body dimensions error (mm) on the Human3.6M dataset, test subject *S9*.(Average limb length error of all four recorded viewpoints of the same sequence used independently) (**Bold**: Improved).

Link		GT to avg PoseFormer	GT to refined
Thorax	Neck	0.75	2.42
Neck	Head	2.90	**2.41**
Thorax	R Shoulder	16.78	**4.66**
R Shoulder	R Elbow	8.87	**5.11**
R Elbow	R Wrist	2.49	2.99
Thorax	L Shoulder	17.50	**4.66**
L Shoulder	L Elbow	8.76	**5.11**
L Elbow	L Wrist	2.64	2.99
Hip	R Hip	13.38	**5.75**
R Hip	R Knee	18.59	**1.55**
R Knee	R Foot	22.75	**7.54**
Hip	L Hip	13.42	**5.75**
L Hip	L Knee	18.18	**1.55**
L Knee	L Foot	23.96	**7.54**
Hip	Spine	4.15	**2.64**
Spine	Thorax	1.68	2.14
Average		11.05	**4.05**

quences we noticed an enhancement in temporal consistency and reduce potential ambiguities.

4 RESULTS

For the evaluation, we used PoseFormer (Zheng et al., 2021) and MHFormer (Li et al., 2022b) as 3D pose estimation model on the Human3.6M (Ionescu et al., 2013) dataset to provide the input. The Human3.6M dataset, designed for 3D HPE and activity recognition, includes 3.6 million annotated 3D human poses captured from videos of 11 actors performing 17 activities from multiple camera angles. Each frame is precisely annotated with a marker-based motion capture system, providing accurate 3D joint coordinates. Following prior works (Zheng et al., 2021; Li et al., 2022b), we evaluated our method using both test sets of *S9* and *S11*. It is important to note that 3D poses estimated by PoseFormer and MHFormer are provided in hip-relative coordinates. In order to utilize the projection rays and refine the estimation, it is necessary to estimate the absolute poses in camera coordinates. The PoseFormer estimates were generated using a model trained on ground truth 2D poses with 81 frames, whereas MHFormer estimates were produced using a model trained on predicted 2D poses with 351 frames.

4.1 Body Dimensions Refinement

As discussed in Section 3, it is beneficial to refine the body dimensions from 3D joint poses input. The estimated 3D joint poses from PoseFormer shows that the body dimensions have some issues. Some of these issues could be resolved with relatively straightforward modifications. For instance, the mirrored limbs exhibit disparate sizes which could be modified easily. Moreover, sizes fluctuate over time for a single subject. Table 1 shows how body dimensions were refined by utilizing 2D joint poses for test set *S9* using the method described in Section 3.1. The second column shows the error between the Ground Truth (GT) body dimensions and the average body dimensions over frames estimated by PoseFormer. The third column shows the error between the ground truth body dimensions and the estimated body dimensions. Notably, the lower body, which showed the largest inaccuracies in PoseFormer estimates, benefited most from the refinement process. Overall, the body dimensions error was reduced by **63.3%**.

4.2 3D Human Pose Refinement

4.2.1 Quantitative Results

To evaluate the refinement results, we report the Mean Per Joint Position Error (MPJPE) evaluation metric in millimeter for joint poses in hip relative. MPJPE quantifies the average distance between corresponding joints in predicted and ground truth poses, providing a measure of pose estimation accuracy. This metric is used to assess the performance of human pose estimation methods.

Table 2 presents the results of all 15 action sequences in the test sets *S9* and *S11*. Both the estimation and refinement processes utilize 2D ground truth poses. To facilitate the adjustment of poses, the body dimensions were provided in two different ways: We show both the results using the prior body dimensions of the test subjects, referred to as *Ours2*, and using estimated body dimensions, referred to as *Ours1*. When no prior knowledge about the subject is given, our method estimates the body dimensions using the approach described in Section 3.1. The results show that the enhancement of body dimensions leads to more precise refinements, resulting in improved accuracy.

As previously stated in Section 3, absolute pose estimation is also calculated by using 2D joint poses. This process involves moving the estimated 3D body skeleton on the hip projection ray in order to minimize overall 2D projection errors.

Table 2 demonstrates the effectiveness of this method in improving pose estimation accuracy. If the absolute poses become more precise, they can further enhance the final refinement process. The last column presents the average of MPJPE across all actions.

Table 2: Quantitative comparison of 3D human pose estimation on the Human3.6M dataset; MPJPE (mm) values of each action using ground truth 2D poses. ("PoseFormer + Ours": Our method used PoseFormer estimation as input to do the post-processing; "MHFormer + Ours": Our method used MHFormer estimation as input to do the post-processing.) - (1: Post-Processing using estimated body dimensions; 2: Post-Processing using prior body dimensions) - (**Bold**: The first best; Underline: The second best).

	Dir.	Dis.	Eat.	Greet.	Phone	Photo	Pose	Purch.	Sit.	Sit.D.	Smoke	Wait.	Walk.D.	Walk.	Walk.T.	Avg
(Li et al., 2023)	32.9	38.3	28.3	33.8	34.9	38.7	37.2	30.7	34.5	39.7	33.9	34.7	34.3	26.1	28.9	33.8
MHFormer (f = 351)	34.8	39.8	34.4	37.0	38.2	44.3	38.7	36.2	45.2	48.3	38.6	38.6	38.3	27.4	27.5	38.3
MHFormer + Ours[1]	31.3	37.0	33.9	35.1	35.9	40.4	37.8	32.3	41.3	41.3	35.8	37.6	35.0	29.0	27.8	35.9
MHFormer + Ours[2]	23.9	29.7	27.7	26.7	26.6	32.8	29.5	24.4	34.6	34.9	28.3	29.5	27.5	18.0	17.4	27.9
PoseFormer (f = 81)	29.9	33.5	29.9	31.0	30.2	33.2	34.7	31.3	37.8	38.6	31.6	31.5	28.9	23.3	23.1	31.6
PoseFormer + Ours[1]	26.4	29.8	28.8	29.6	28.5	29.5	34.0	26.2	33.8	33.3	29.4	29.5	25.5	23.9	23.1	29.1
PoseFormer + Ours[2]	**17.7**	**21.1**	**22.2**	**21.0**	**20.3**	**22.5**	**24.1**	**17.2**	**27.2**	**27.5**	**21.5**	**21.4**	**18.0**	**13.4**	**12.6**	**20.9**

Table 3: Quantitative comparison of 3D human pose estimation on the Human3.6M dataset; MPJPE (mm) values of each joint using ground truth 2D pose. ("PoseFormer + Ours": Our method used PoseFormer estimation as input to do the post-processing) - (1: Post-Processing using estimated body dimensions; 2: Post-Processing using prior body dimensions) - (**Bold**: The first best; Underline: The second best).

	Head	Neck	R Shoulder	R Elbow	R Wrist	L Shoulder	L Elbow	L Wrist	R Hip	R Knee	R Foot	L Hip	L Knee	L Foot	Spine	Thorax
PoseFormer (f = 81)	34.2	29.3	32.9	42.9	50.5	29.7	37.4	47.7	17.0	24.9	48.4	17.0	29.4	52.5	18.9	23.9
PoseFormer + Ours[1]	32.0	28.4	30.4	42.7	48.7	28.2	37.0	45.9	13.1	19.8	43.5	13.3	19.9	44.7	19.4	27.3
PoseFormer + Ours[2]	**25.8**	**19.9**	**21.7**	**32.5**	**40.0**	**21.4**	**29.7**	**38.5**	**3.2**	**13.7**	**30.0**	**3.2**	**14.0**	**29.6**	**13.9**	**18.9**

Ours2, which show the estimations using prior body dimensions, reveal a reduction in the average MPJPE across all actions by **33.9%** compared to PoseFormer and by **27.2%** compared to MHFormer, with even larger improvements for certain individual actions. The other columns confirm improvements across all actions.

Ours1, refinement of 3D joint poses using estimated body dimensions, achieving a **7.9%** reduction in the average MPJPE relative to PoseFormer and a **6.3%** reduction compared to MHFormer. The other columns indicate improvements for all actions, except for the *"Walking"* and *"WalkingTogether"* actions. Furthermore, our post-processing method demonstrates a markedly superior outcome compared to (Li et al., 2023), in which a refinement method for 3D HPE was also employed. Moreover, Figures 6 and 7 present a detailed frame-wise MPJPE comparison, highlighting the effectiveness of *Our1* and *Our2* across 400 frames.

In Table 3, a joint-by-joint quantitative comparison is presented. *Our2* shows the reduction in MPJPE achieved by our method for all joints, with a particularly notable improvement observed for the lower body, particularly for *"R Hip"* and *"L Hip"*, when using prior body dimensions. In *Our1*, a reduction in MPJPE is achieved for all joints except the *Spine* and *Thorax* when our method employed the estimated body dimensions as described in Section 3.1. This may occur because this is a non-rigid link, and their positions are influenced by the overall body posture rather than being directly related to link lengths. Consequently, variations in these regions are less sensitive to adjustments in estimated body dimensions, leading to a smaller impact on MPJPE. For additional joint-based comparisons across selected actions, please refer to Figures 8 to 12 in the appendix. The presented figures illustrate the extent to which each joint has been individually refined through post-processing.

4.2.2 Qualitative Results

To highlight the effectiveness of our approach, we present a visual comparison between the predicted 3D poses and the ground truth. As illustrated in Figure 4, our method demonstrates improved accuracy compared to PoseFormer, particularly in challenging scenarios. We evaluate our method with both using estimated body dimensions and prior body dimensions on the Human3.6M, test subject *S9*, actions *"WalkDog", "Discussion", "Directions", "Smoking"* and *"Sitting"*. The results clearly show that our predictions align more closely with the ground truth, showcasing the robustness and precision of our approach.

5 DISCUSSION

The proposed post-processing method effectively refines joint poses, ensuring accurate limb proportions in human pose estimation. Improved skeleton scaling directly contributes to more precise joint pose refinements.

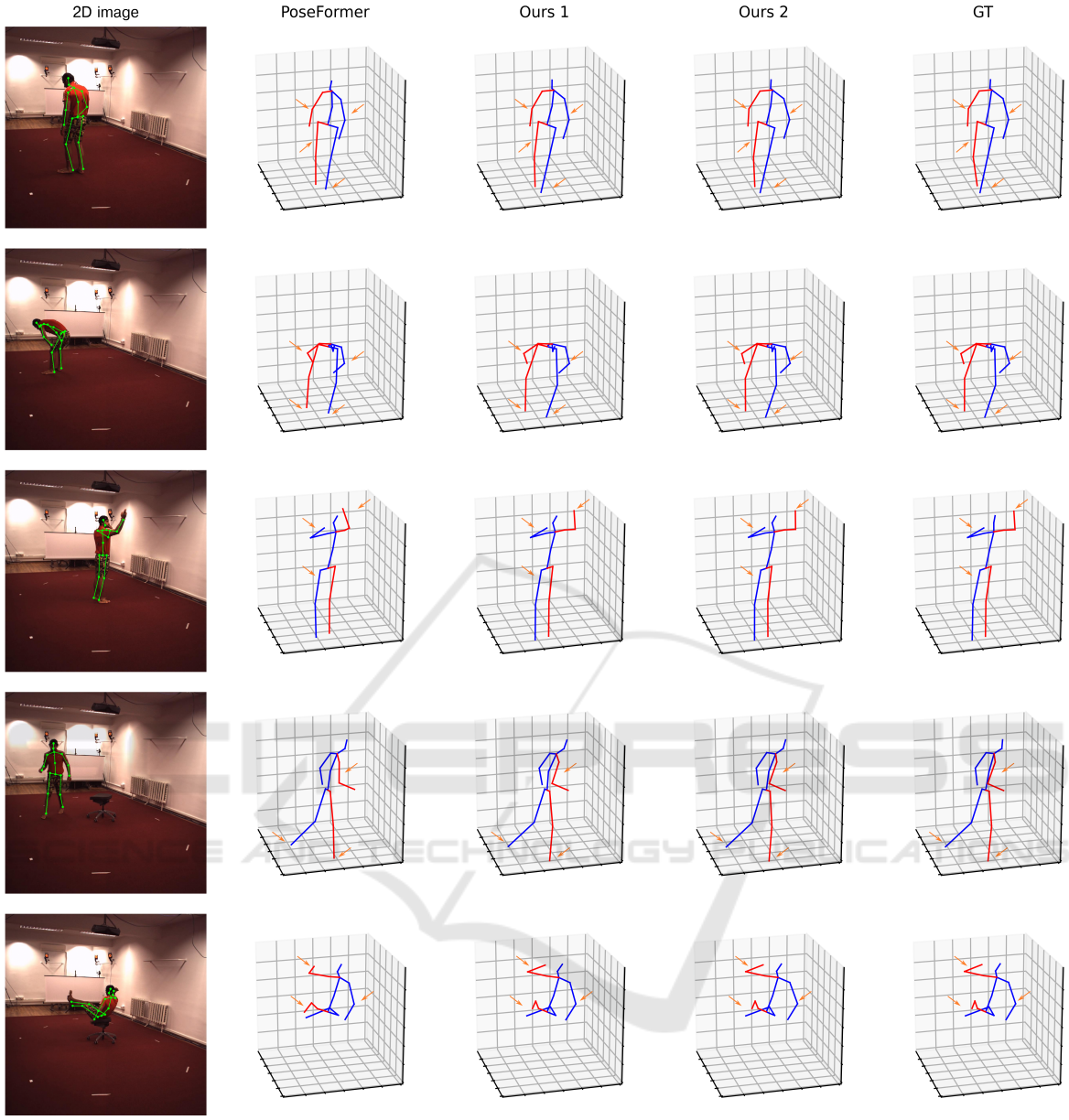
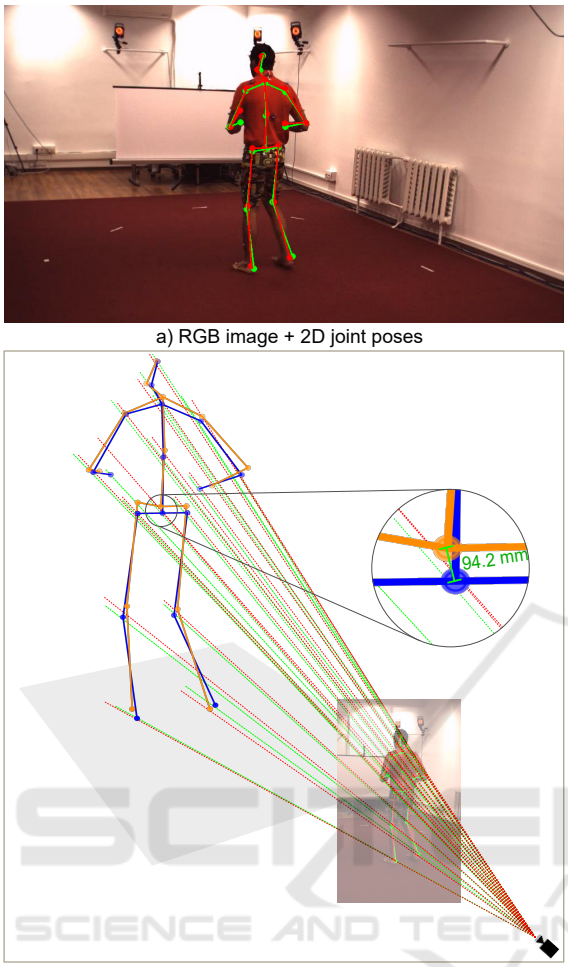


Figure 4: Qualitative comparison of 3D human pose estimation results between our proposed method and the PoseFormer baseline on the Human3.6M dataset, test subject *S9*, action *WalkDog*, *Discussion*, *Directions*, *Smoking* and *Sitting*.; ("Ours 1": Our method used PoseFormer estimation as input to do the post-processing using estimated body dimensions; "Ours 2": Our method used PoseFormer estimation as input to do the post-processing using prior body dimensions.); Locations with huge improvements are highlighted with an orange arrow. (Blue): Left; (Red): Right.

As presented in Section 4, our refinement method delivers strong results using both prior body dimensions and estimated body dimensions based on 2D poses. Improvements were consistently observed, with some actions showing significant refinements. Models capable of predicting 2D joint poses, such as CPN (Chen et al., 2018), often fail to produce smooth, accurate 2D predictions over sequential frames, which can lead to large deviations from the actual joint positions. To examine the impact of 2D prediction accuracy on 3D HPE refinement, we compared results using CPN-derived 2D joint poses with those obtained using ground truth 2D poses. We observed that the average MPJPE across all actions and test sets for our method was 49.47 mm when using CPN poses, compared to 44.79 mm with PoseFormer, illustrating the critical role of accurate 2D predictions in 3D pose estimation. Figure 5 illustrates how 2D



a) RGB image + 2D joint poses

b) Projection rays from camera through virtual image and 3D HPE

Figure 5: The effect of 2D pose estimation on absolute 3D HPE. a) One frame of test set *S9*, action *Directions* of the Human3.6M dataset. (Green): Ground truth 2D joint poses; (Red): CPN 2D joint pose estimation. b) Green rays from the camera are related to each joints form ground truth and the red rays are related to CPN result. (Blue): The result of our method when it used ground truth 2D joint poses; (Orange): The result of our method when it used CPN 2D joint poses.

pose estimation accuracy affects absolute 3D HPE. In this example, the absolute hip joint estimated using CPN 2D joint pose estimation differs 94.2 mm from the one estimated using ground truth 2D poses. The accuracy of 2D HPE also affects the correct prediction of occluded joints. Therefore, if an occluded joint is misestimated in 2D HPE, the probability that our method identifies it as occluded in 3D HPE would be low.

A review of both absolute and relative poses refinement reveals numerous cases where MPJPE for both absolute and relative poses are approximately 1 mm, when our post-processing method is applied, demonstrating the method's ability to perform highly precise refinements, especially when the absolute pose estimation is accurate. Figure 1 shows an example frame of *SittingDown* action, test set *S9*, on Human3.6M dataset, where the MPJPE was reduced from 41.31 mm to 1.15 mm in camera coordinates, and the relative pose error was reduced to 0.90 mm. This example underscores the significant enhancements achievable with our post-processing approach.

6 CONCLUSION

In this paper, we propose a novel post-processing technique that focuses primarily on refining joint poses output of human pose estimation models from a monocular camera to address critical limitations found in existing methods. By estimating absolute poses and ensuring consistent limb proportions, our approach improves the accuracy and reliability of skeleton representations. This improvement is essential for applications that rely on precise human poses in camera coordinates. Our method demonstrates competitive performance, outperforming state-of-the-art techniques on the Human3.6M dataset, underscoring its effectiveness in advancing 3D human pose estimation.

REFERENCES

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., and Sun, J. (2018). Cascaded pyramid network for multi-person pose estimation.

Hassanin, M., Khamiss, A., Bennamoun, M., Boussaid, F., and Radwan, I. (2022). Crossformer: Cross spatio-temporal transformer for 3d human pose estimation. *arXiv preprint arXiv:2203.13387*.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2013). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339.

Kadkhodamohammadi, A., Gangi, A., de Mathelin, M., and Padoy, N. (2017). A multi-view rgb-d approach for human pose estimation in operating rooms. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 363–372. IEEE.

Kocabas, M., Athanasiou, N., and Black, M. J. (2020). Vibe: Video inference for human body pose and shape estimation.

Li, H., Shi, B., Dai, W., Zheng, H., Wang, B., Sun, Y., Guo, M., Li, C., Zou, J., and Xiong, H. (2023). Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1296–1304.

Li, W., Liu, H., Ding, R., Liu, M., Wang, P., and Yang, W. (2022a). Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293.

Li, W., Liu, H., Tang, H., Wang, P., and Gool, L. V. (2022b). Mhformer: Multi-hypothesis transformer for 3d human pose estimation.

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649.

Rhodin, H., Salzmann, M., and Fua, P. (2018). Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 750–767.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703.

Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y. (2018). Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545.

Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660.

Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., and Liu, Y. (2018). Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296.

Zhang, J., Tu, Z., Yang, J., Chen, Y., and Yuan, J. (2022). Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242.

Zhao, Q., Zheng, C., Liu, M., Wang, P., and Chen, C. (2023). Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., and Shah, M. (2023). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37.

Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., and Ding, Z. (2021). 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhi, T., Lassner, C., Tung, T., Stoll, C., Narasimhan, S. G., and Vo, M. (2020). Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 492–509. Springer.

APPENDIX

To provide further illustration, the figures included in this appendix offer additional context and detailed insights in comparison to PoseFormer estimations.

Three actions of *SittingDown*, *Directions* and *Photo* from test set *S9* are presented to provide further illustration. Figure 8 to Figure 10, compare the MPJPE across all joints and all frames of the aforementioned actions to demonstrate the extent of the improvement in each joint when the method utilized prior body dimensions. Figure 11 to Figure 13 present the same comparison when the method employed estimated body dimensions.

Figure 6 and Figure 7 illustrate the MPJPE over 400 frames of the action *SittingDown* and *Directions* for test set *S9* of the Human3.6M dataset. The figures demonstrate the extent to which the refinement process reduced the MPJPE, utilizing both prior body dimensions and estimated body dimensions.

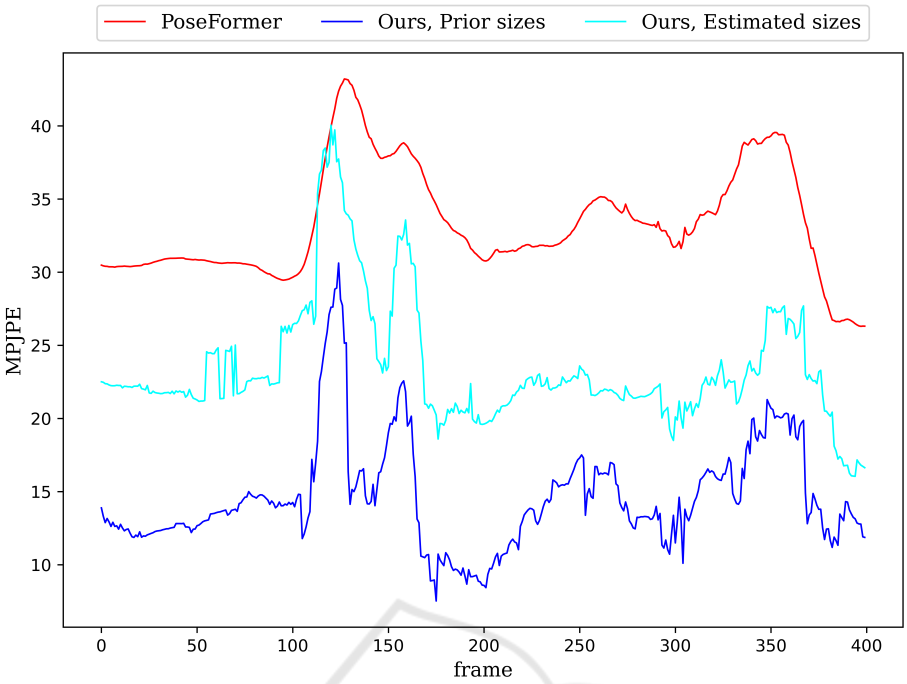


Figure 6: Comparison of the MPJPE values across 400 frames of test set *S9*, action *SittingDown* of the Human3.6M dataset. (ground truth 2D pose and prior body dimensions vs. ground truth 2D pose and estimated body dimensions vs. PoseFormer).

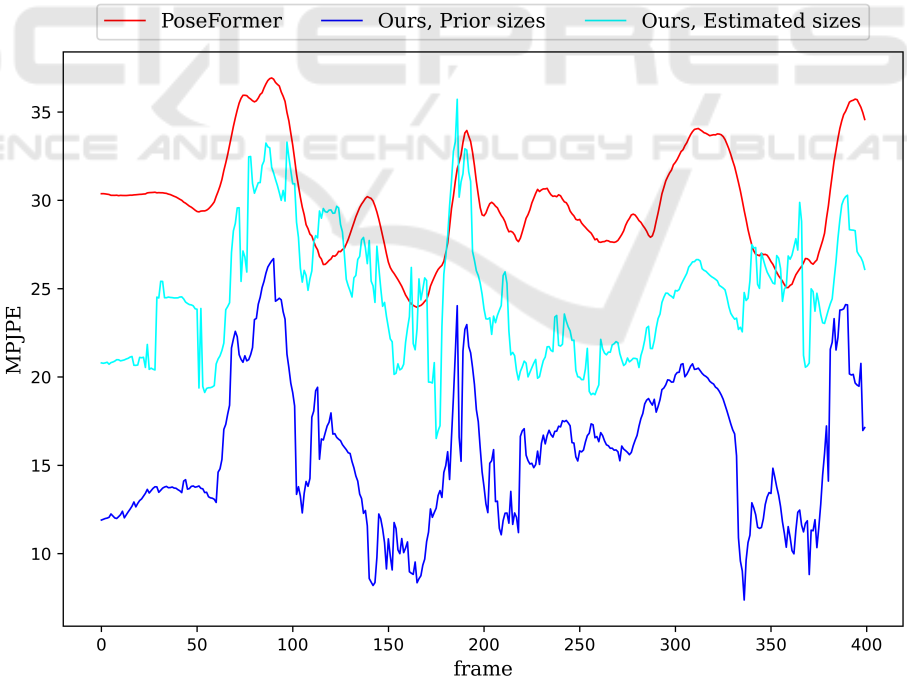


Figure 7: Comparison of the MPJPE values across 400 frames of test subject *S9*, action *Directions* of the Human3.6M dataset. (ground truth 2D pose and prior body dimensions vs. ground truth 2D pose and estimated body dimensions vs. PoseFormer).

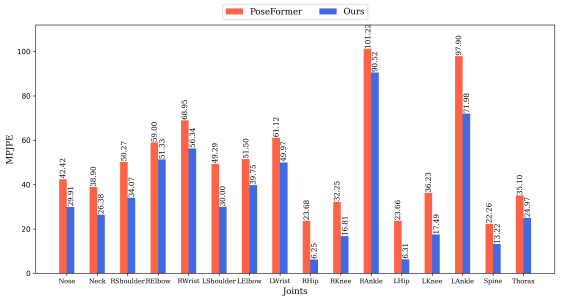


Figure 8: Comparison of the MPJPE values of all joints across all frames of test set *S9*, action *SittingDown* of the Human3.6M dataset. (using ground truth 2D pose and prior body dimensions).

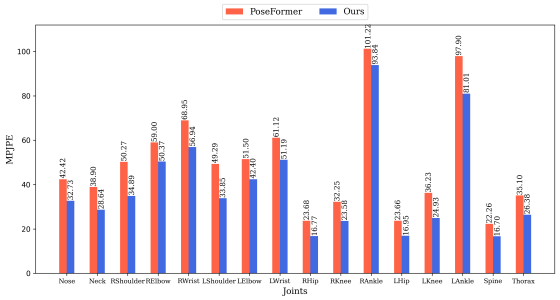


Figure 11: Comparison of the MPJPE values of all joints across all frames of test set *S9*, action *SittingDown* of the Human3.6M dataset. (using ground truth 2D pose and estimated body dimensions).

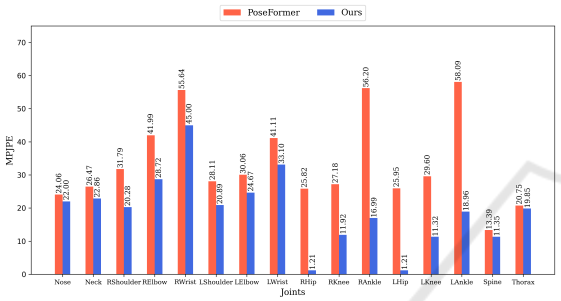


Figure 9: Comparison of the MPJPE values of all joints across all frames of test set *S9*, action *Directions* of the Human3.6M dataset. (using ground truth 2D pose and prior body dimensions).

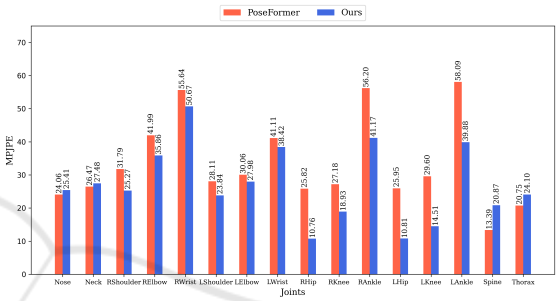


Figure 12: Comparison of the MPJPE values of all joints across all frames of test set *S9*, action *Directions* of the Human3.6M dataset. (using ground truth 2D pose and estimated body dimensions).

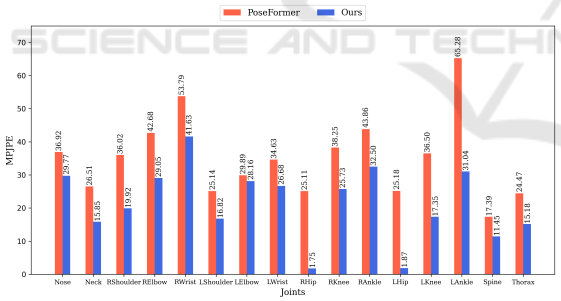


Figure 10: Comparison of the MPJPE values of all joints across all frames of test set *S9*, action *Photo 1* of the Human3.6M dataset. (using ground truth 2D pose and prior body dimensions).

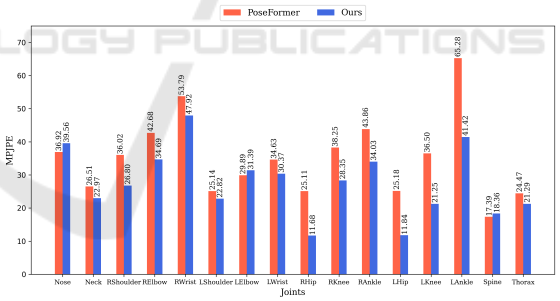


Figure 13: Comparison of the MPJPE values of all joints across all frames of test set *S9*, action *Photo 1* of the Human3.6M dataset. (using ground truth 2D pose and estimated body dimensions).